



MATA TRIPURA SUNDARI OPEN UNIVERSITY

Tripura

ECM-6214

SCHOOL OF HUMANITIES AND SOCIAL SCIENCES

Statistical Methods

**MASTER OF ARTS
(ECONOMICS)**

Course Code: ECM-6214
Course Name: Statistical Methods



MASTER OF ARTS (ECONOMICS)

PROGRAMME DESIGN COMMITTEE

Prof. Masood Parveez
Vice Chancellor – Chairman
MTSOU, Tripura

Prof. Mohd. Nafees Ahmad Ansari
Director of Distance Education
Aligarh Muslim University, Aligarh

Prof. Kaustuva Barik
Professor of Economics
IGNOU, New Delhi

Prof. Shiva Kumar GN
Dean , Academics
MTSOU, Tripura

Prof. Saka Bhanu Baba Saheb
Professor of Economics
MTSOU, Tripura

Prof. Jayanti Lal Jain
Professor of Economics
Mangalayatan University, Aligarh

Prof. C.R.K. Murty
Professor of Distance Education
IGNOU, New Delhi

Prof. Narayan Prasad
Professor of Economics
MTSOU, Tripura

COURSE WRITERS

Prof. Saka Bhanu Baba Saheb
Professor of Economics
MTSOU, Tripura
ECM-6211 Microeconomic II

Ms. Pem Eton
Assistant Professor of Economics
MTSOU, Tripura
ECM-6213 Fundamentals of Rural Development

Dr. Poonam Gupta
Assistant Professor of Economics
Mangalayatan University, Aligarh
ECM-6215 Agricultural Economics

Mr. Sumanta Debbarma
Assistant Professor of Economics
MTSOU, Tripura
ECM-6212 Macroeconomics-II

Dr. Bhism Bharti Rai
Assistant Professor of Economics
MTSOU, Tripura
ECM-6214 Statistical Methods

COURSE EDITORS

Prof. Saka Bhanu Baba Saheb
Professor of Economics
MTSOU, Tripura

Dr. Vijeta Banwari
Associate Professor
IGNOU, New Delhi

Dr. Indu Varshney
Associate Professor of Economics
Dharm Samaj College, Aligarh

Prof. Jayanti Lal Jain
Professor of Economics
Mangalayatan University, Aligarh

Dr. Saugato Sen Associate
Professor of Economics
IGNOU, New Delhi

FORMAT EDITORS

Dr. Nitendra Singh
Associate Professor of English
MTSOU, Tripura

Dr. Dabinjoy Tripura
Assistant Professor of English
MTSOU, Tripura

Dr. Vikrant Rajput
Assistant Professor of English
MTSOU, Tripura

MATERIAL PRODUCTION

1. Mr. Panthor Debbarma
Assistant Registrar (Material Production)
MTSOU, Tripura

2. Ms. Rupa Jamatia
3. Mr. Pratik Kanti Roy
4. Ms. Bwkhwi Jamatia

5. Mr. Soubhik Paul
6. Mr. Abhijit Roy Bardhan

Contents

Page No.

Block I: Fundamentals of Statistics and Data Representation 5-73

Unit 1: Introduction to Statistics: Meaning and Importance – Definitions, functions, limitations, importance, statistical tools in economics

Unit 2: Statistical Processes and Data Collection – Statistical processes, primary and secondary data, sampling, accuracy, processing tools

Unit 3: Data Classification and Tabulation – Classification types, frequency distribution, tabulation methods, tables

Unit 4: Graphic Representation of Data – Diagrams, histograms, frequency polygon, ogive, graphical methods

Block II: Central Tendency and Dispersion Measures in Statistics 74-171

Unit 5: Measures of Central Tendency – Mean, Median and Mode – Arithmetic mean, median, mode, applications

Unit 6: Harmonic and Geometric Mean – Concepts, formulas, uses, comparisons

Unit 7: Measures of Dispersion – Mean deviation, standard deviation, combined standard deviation

Unit 8: Skewness and Kurtosis – Measures of asymmetry, moments, peaked ness of distribution

Block III: Elementary Probability, Random Variables and Distribution 172-285

Unit 9: Elementary Probability – Probability concepts, conditional probability, Bayes' Theorem

Unit 10: Random Variables and Mathematical Expectation – Random variables, expectation, distribution functions, density functions

Unit 11: Binomial Distributions – Binomial model, Bernoulli distribution, solved problems

Unit 12: Poisson and Normal Distributions – Poisson distribution, normal curve, applications

Block IV: Correlation And Regression Analysis, Curve Fitting 286-345

Unit 13: Correlation: Karl Pearson's Coefficient of Correlation – Correlation measurement, Pearson coefficient

Unit 14: Rank Correlation and Scatter Diagram – Rank methods, scatter plots, relationship analysis

Unit 15: Regression Lines and Coefficients – Regression equations, coefficients, standard error of estimate

Unit 16: Curve Fitting: Method of Least Squares – Curve fitting techniques, least squares method, examples

Block V: Hypothesis Testing and Analysis of Variance 346-446

Unit 17: Sampling and Testing of Hypothesis – Sampling, standard error, tests for mean, proportion, deviations

Unit 18: Z-Test, T-Test and F-Test – Small sample tests, Student's t-test, F-test, Z-test

Unit 19: Chi-Square Test and Goodness of Fit – Chi-square distribution, independence test, goodness of fit

Unit 20: Introduction to Analysis of Variance – One-way ANOVA, two-way ANOVA, Fisher's z transformation



BLOCK I: FUNDAMENTALS OF STATISTICS AND DATA REPRESENTATION

UNIT 1 INTRODUCTION TO STATISTICS: MEANING AND IMPORTANCE

STRUCTURE

- 1.1 Introduction
- 1.2 Objectives
- 1.3 Definition of Statistics
- 1.4 Popular definitions of Statistics
- 1.5 Function of statistics
- 1.6 Limitations of statistics
- 1.7 Importance of statistics
- 1.8 Statistical tools used in economic analysis
- 1.9 Summary
- 1.10 EXERCISES
- 1.11 References

1.1 INTRODUCTION

Statistics is a powerful tool that plays a critical role in shaping modern society. As the science of data, statistics involves collecting, organizing, analysing, interpreting, and presenting information. Its applications are vast and span across multiple domains, from economics and business to healthcare, education, and social sciences. In a world driven by data, the ability to understand and apply statistical methods has become essential for making informed decisions and solving complex problems.

The significance of statistics lies in its ability to simplify complex information, transforming raw data into actionable insights. Whether analysing market trends, conducting scientific research, or formulating

government policies, statistical methods help uncover patterns and relationships that may not be immediately apparent. This makes statistics an indispensable part of planning, forecasting, and problem-solving. For example, businesses rely on statistical analysis to understand consumer behaviour, optimize production, and forecast sales. Governments use statistical data to assess economic performance, guide policy decisions, and allocate resources effectively. Researchers apply statistical techniques to test hypotheses, validate findings, and ensure the accuracy of their studies.

One of the unique features of statistics is its applicability to virtually every field. In healthcare, statistics helps track disease outbreaks, evaluate treatment effectiveness, and improve patient care. In education, it is used to analyse student performance, design curriculum improvements, and enhance teaching methods. In environmental science, statistical models predict climate change patterns and assess the impact of human activities on ecosystems. The versatility of statistics makes it a universal language for understanding the world and driving progress.

Statistics also plays a crucial role in addressing uncertainty and variability. In many situations, decisions must be made based on incomplete or imperfect data. Statistical tools provide methods for estimating probabilities, measuring risks, and making informed predictions despite these uncertainties. For instance, economists use time series analysis to forecast economic growth, while meteorologists apply statistical models to predict weather patterns. By quantifying uncertainty, statistics enables better decision-making and minimizes the risks associated with unpredictable events.

Furthermore, statistics bridges the gap between theory and practice by providing empirical evidence to support conclusions. Unlike pure mathematics, which often deals with abstract concepts, statistics is rooted in real-world data. This practical aspect makes statistics

particularly valuable for policymakers, business leaders, and researchers who need to base their decisions on solid evidence. The ability to analyze data accurately and interpret results correctly can lead to more effective solutions and better outcomes.

Despite its many strengths, the area of statistics is not without challenges. The primary concern is the potential for data misinterpretation or misuse. When statistical methods are applied incorrectly, or data is presented out of context, the results can be misleading. This highlights the importance of statistical literacy and the need for professionals who are trained in data analysis. Additionally, statistical methods often rely on assumptions and approximations, which may not always align perfectly with real-world conditions. Recognizing these limitations is essential for using statistics responsibly and avoiding erroneous conclusions.

In conclusion, statistics is more than just numbers and calculations—it is a fundamental tool for understanding the complexities of the world. Its ability to simplify data, identify trends, and make predictions makes it indispensable in various fields. As technology advances and the volume of data continues to grow, the role of statistics will only become more critical. By mastering statistical methods, individuals and organizations can unlock valuable insights, drive innovation, and make more informed decisions. This unit will delve deeper into the meaning, functions, and importance of statistics, providing a comprehensive overview of its applications and limitations.

1.2 OBJECTIVES

After completing this unit, students will be able to

- Understand the meaning and scope of statistics.

- Differentiate between the plural and singular aspects of statistics.
- Recognize the various definitions and functions of statistics.
- Identify the limitations of statistical methods.
- Appreciate the significance of statistics in economic analysis and other fields.

1.3 DEFINITION OF STATISTICS

Statistics can be viewed from two perspectives – as a plural and singular noun. This duality reflects both the data-driven aspect and the methodological aspect of the discipline.

1.3.1 Statistics as a Plural Noun

As a plural noun, statistics refers to numerical data collected for analysis. Examples include national income figures, employment rates, or exam scores.

In 1935, Dr. W.F. Willcox listed over 100 definitions of statistics but Prof. Hoarse Sacrist gave the most exhaustive and accurate definition of statistics as given below.

"By statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other".

According to this definition, statistics should possess the following peculiarities and characteristics:

1. Statistics in an aggregate of facts: A single figure or an isolated fact does not constitute statistics. A single accident in a road at a place X, however interesting, is not statistics, but the number of road accidents

which take place at X during a year is statistics. Thus, the study deals with mass phenomena.

2. Statistics is affected to a marked extent by multiplicity of causes:

Statistics or data are affected by a large number of forces, internal or external. For example, the phenomena of falling sales in a business firm may be caused by general recession in the business. The agricultural production is affected by cultivation practices, crop pattern and rainfall etc.

3. Statistics are numerically expressed: Qualitative descriptions are not statistics unless they are expressed in numerical form. For example, tall and short, blind, deaf, average and dull are statements of qualities and are, therefore, not statistics. In order to become so they must be amenable to quantitative expressions. Qualitative knowledge has to be converted into quantitative units, 6' 3" and 4' 3" tall and short) - before it becomes statistics.

4. Statistics are enumerated or estimated according to reasonable standards of accuracy: Chances of errors in the collection of data should be minimised if correct result are to be obtained. What is reasonable standard of accuracy will depend upon the purpose of enquiry. For example, if statistics of wheat yield are required, they might be collected in quintals and not in grams while purchase and sale of gold/silver should be expressed in grams and not in quintals.

5. Statistics are collected in a systematic manner for a predetermined purpose: The object of collection of data must be clear, definite and known in advance. For example, while collecting statistics of food production, we must decide whether the country is self-sufficient in food or not.

6. Statistics are placed in relation to each other: Statistics are used to show comparison between two or more phenomena. This is only possible when a common basis is adopted in collection, for example,

supply of a particular commodity can be compared with its demand or price and not with the birth and death rates in the country. Heterogeneous data are incomparable.

1.3.2 Statistics as a Singular Noun

As a singular noun, statistics represents the scientific method of data collection, organization, presentation, analysis, and interpretation. It involves applying mathematical theories and models to draw meaningful conclusions.

According to Croxton and Cowden, *"Statistics may be defined as science of collection, presentation, analysis and interpretation of numerical data."*

Thus, the above definition of statistics broadly covers the following statistical methods:

1. **Collection of data:** This is the primary step in a statistical study and therefore, data should be collected with care by the investigator. If data is not collected carefully, the results based on such data may not be upto the mark.
2. **Organisation of data:** Data collected by an investigator are organised by means of editing, tabulating and classifying.
3. **Presentation of data:** Collected and organised data are presented in some systematic way so that statistical analysis become easier. The organised data can be presented with the help of charts, graph, diagrams, etc.
4. **Analysis of data:** The next stage in the analysis of presented data. This can be done by using various statistical methods depending upon the situation. For example, averages, standard deviation, correlation, regression etc.
5. **Interpretation of data:** The conclusions are derived from the data analysis. This stage of the statistical research is crucial and

extremely delicate. for the reason that conclusions are used to make some decisions.

1.4 POPULAR DEFINITIONS OF STATISTICS

- **Bowley**
 - *“Statistics is the science of measurement of social organism regarded as a whole, in all its manifestations”.*
 - *“Statistics may be called the science of counting”.*
 - *“Statistics may rightly be called the science of averages”.*
- **Boddington**
 - *“Statistics is a science of estimates and probabilities”.*
- **King**
 - *“The science of statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimates”.*
- **Harlows**
 - *“Statistics is the science and the art of handling aggregates of facts-observing, enumerating, recording, classifying and otherwise systematically treating them” purpose.*

- **H. Secrist**
 - *“Statistics is the study of numerical data, the interpretation of facts, and the drawing of sound conclusions.”*
- **Croxton and Cowden**
 - *“Statistics is the science of uncertainty and variability.”*

1.5 FUNCTIONS OF STATISTICS

1. **Data Collection and Presentation** – Organizing raw data into tables, charts, and graphs.

Statistics organizes vast amounts of raw data into structured formats such as tables, graphs, and charts. This makes information more accessible and understandable, facilitating quick insights and efficient communication of findings.

2. **Simplification of Complex Data** – Summarizing large datasets for easier understanding.

Large, complex datasets are distilled into meaningful summaries using statistical measures like averages and percentages. This reduces the complexity of data, allowing for easier interpretation and comparison across different scenarios.

3. **Comparison and Analysis** – Identifying relationships and patterns.

By applying statistical techniques, data can be compared across different groups, regions, or time periods. This aids in identifying relationships, disparities, and trends, which are crucial for making informed decisions.

4. **Forecasting and Prediction** – Using past data to predict future trends.

Historical data is analysed to predict future outcomes, enabling businesses and governments to plan for growth, allocate resources, and anticipate market demands or economic shifts.

5. **Decision-Making** – Guiding policy and business strategies through empirical evidence.

Statistics provides evidence-based insights that support critical decisions in fields like business, healthcare, and policy-making.

This minimizes uncertainty and enhances the likelihood of achieving desired results.

1.6 LIMITATIONS OF STATISTICS

1. **Limited to Quantitative Data** – Statistics focuses primarily on numerical data, often neglecting qualitative factors such as personal experiences, emotions, or subjective opinions, which may be essential for comprehensive analysis.
2. **Misleading Interpretations** – Statistical results can be manipulated or misrepresented to favour certain outcomes. If data is presented without context or selectively, it may lead to biased conclusions.
3. **Requires Skill and Expertise** – Effective use of statistical tools demands specialized knowledge. Without proper understanding, data can be misinterpreted, resulting in incorrect decisions and flawed policies.
4. **Assumptions and Approximations** – Many statistical models rely on assumptions that simplify real-world complexities. If these assumptions are inaccurate, the reliability of the results may diminish significantly.
5. **Incomplete or Biased Data** – The accuracy of statistical findings depends on the quality of data collected. Errors, incomplete datasets, or sampling bias can distort results, limiting the validity of conclusions.

1.7 IMPORTANCE OF STATISTICS

1. **Economic Planning** – Governments use statistical data to evaluate economic indicators like GDP, inflation, and employment rates. This informs national policies and long-term economic strategies, ensuring sustainable development.
2. **Business and Industry** – Companies apply statistics to market research, sales forecasting, and production management. Statistical analysis helps businesses identify consumer trends, optimize resources, and improve overall efficiency.
3. **Social Sciences** – In areas like sociology, psychology, and public health, statistics is used to study demographic changes, assess social programs, and measure public opinion, guiding policy and social interventions.
4. **Scientific Research** – Statistics plays a vital role in designing experiments, testing hypotheses, and validating results. It ensures the reliability of scientific discoveries by quantifying uncertainty and measuring significance.
5. **Public Policy** – Policymakers rely on statistical analysis to shape effective public policies in areas such as healthcare, education, and infrastructure. By understanding population needs through data, policies become more targeted and impactful.

1.8 STATISTICAL TOOLS USED IN ECONOMIC ANALYSIS

1. **Measures of Central Tendency** – These tools, including mean, median, and mode, provide a single value representing the center of a dataset. They summarize data effectively, highlighting the typical or average case in economic indicators.

2. **Measures of Dispersion** – Range, variance, and standard deviation illustrate the spread of data, showing how much individual data points deviate from the average. This helps assess risk and variability in economic performance.
3. **Correlation and Regression** – Correlation measures the strength of relationships between two variables, while regression models predict one variable based on another. These tools are essential for economic forecasting and understanding market dependencies.
4. **Index Numbers** – Index numbers track changes in economic variables such as price levels and industrial production over time. They are key for measuring inflation, cost of living, and overall economic growth.
5. **Time Series Analysis** – Time series methods analyze data points collected at regular intervals, revealing trends, seasonal patterns, and cyclical movements. This aids in long-term economic planning and short-term forecasting.

CHECK YOUR PROGRESS

1. Define Statistics and explain its meaning and scope.

.....

.....

.....

2. Explain the importance of Statistics in economics and social sciences.

.....

.....

.....

3. Discuss the main functions and limitations of Statistics.

1.9 SUMMARY

This unit provided an overview of statistics, its definitions, and its importance across various domains. The functions, limitations, and practical applications were explored to highlight how statistics influences decision-making and economic analysis.

1.10 EXERCISES

1. Define statistics and explain its twofold meaning.
2. List and explain three key functions of statistics.
3. Discuss the limitations of using statistical methods.
4. How does statistics contribute to economic analysis? Provide examples.
5. Objective Type Questions
 1. **Which of the following best defines statistics as a singular noun?**
 - a) A collection of facts and figures
 - b) The process of collecting, organizing, and interpreting data
 - c) A set of numerical data
 - d) A list of economic indicators
 2. **Which of the following is not a measure of central tendency?**
 - a) Mean
 - b) Mode
 - c) Standard Deviation
 - d) Median

3. **Statistics primarily deals with:**
- a) Qualitative data only
 - b) Quantitative data only
 - c) Both qualitative and quantitative data
 - d) Neither qualitative nor quantitative data
4. **Which statistical tool is used to measure the dispersion of data?**
- a) Correlation
 - b) Index Numbers
 - c) Standard Deviation
 - d) Regression
5. **What does time series analysis help in?**
- a) Analyzing the relationship between two variables
 - b) Comparing data from different sources
 - c) Tracking data over a period to identify trends
 - d) Calculating averages in a dataset
6. **In economic planning, statistics is mainly used to:**
- a) Write policies
 - b) Measure inflation, GDP, and unemployment rates
 - c) Study social behavior
 - d) Record historical events
7. **Which of the following is a limitation of statistics?**
- a) It can analyze qualitative data directly
 - b) It is free from errors in all cases
 - c) It relies heavily on assumptions
 - d) It provides perfect predictions without variability
8. **Correlation measures the:**
- a) Spread of data
 - b) Relationship between two variables
 - c) Average of data points
 - d) Frequency of occurrences

9. **Which of the following is an example of using statistics in public policy?**
- a) Forecasting market trends for businesses
 - b) Designing experiments in a laboratory
 - c) Analyzing population growth to plan infrastructure
 - d) Predicting the weather patterns
10. **What is the primary objective of using index numbers in economic analysis?**
- a) To calculate the sum of all data points
 - b) To measure changes in economic variables over time
 - c) To analyze relationships between unrelated data
 - d) To find the average of a data set

Answers:

- 1. b) The process of collecting, organizing, and interpreting data
- 2. c) Standard Deviation
- 3. b) Quantitative data only
- 4. c) Standard Deviation
- 5. c) Tracking data over a period to identify trends
- 6. b) Measure inflation, GDP, and unemployment rates
- 7. c) It relies heavily on assumptions
- 8. b) Relationship between two variables
- 9. c) Analysing population growth to plan infrastructure
- 10. b) To measure changes in economic variables over time



1.11 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 2 STATISTICAL PROCESSES AND DATA COLLECTION

STRUCTURE

- 2.1 Introduction
- 2.2 Objectives
- 2.3 Definition of Statistical Processes
- 2.4 Components of Statistical Processes
- 2.5 Methods of Data Collection
 - a. Primary Data Collection
 - b. Secondary Data Collection
- 2.6 Sampling Techniques
- 2.7 Importance of Data Accuracy
- 2.8 Statistical Tools for Data Processing
- 2.9 Application of Data Collection in Economic Analysis
- 2.10 Summary
- 2.11 Exercises
- 2.12 References

2.1 INTRODUCTION

Statistics forms the backbone of decision-making processes across various fields, from economics and business to healthcare and engineering. At the heart of statistical analysis lies the process of data collection, which serves as the foundation for generating meaningful insights. The accuracy and reliability of statistical outcomes heavily depend on the quality of data collected. Therefore, understanding the methods and processes involved in gathering, organizing, and analyzing data is crucial for producing valid results.

Statistical processes encompass the systematic approach to collecting, summarizing, and interpreting data to extract valuable information. This

unit delves into the core aspects of data collection, the various methods employed, and the importance of ensuring data accuracy. By mastering these concepts, individuals can enhance their ability to conduct research, forecast trends, and make informed decisions in professional and academic contexts.

2.2 OBJECTIVES

After completing this unit, students will be able to

- Understand the concept of statistical processes and their significance.
- Identify various methods of data collection and distinguish between primary and secondary data.
- Analyse and apply sampling techniques in different scenarios.
- Comprehend the role of statistical tools in processing and interpreting collected data.
- Evaluate the relevance of data collection in economic analysis and business planning.

2.3 DEFINITION OF STATISTICAL PROCESSES

Statistical processes refer to the structured activities involved in gathering, analysing, and interpreting data. These processes enable researchers and analysts to identify patterns, test hypotheses, and draw conclusions that inform decisions. A well-defined statistical process typically follows a cycle involving problem definition, data collection, data analysis, and reporting.

2.4 COMPONENTS OF STATISTICAL PROCESSES

1. **Problem Definition** – Clearly identifying the question or problem that requires investigation. This measure makes ensuring that the focus stays in line with the study's goals.
2. **Data Collection** – Gathering relevant data using appropriate methods such as surveys, experiments, or observations. The overall dependability of the analysis is directly impacted by how well this stage is done.
3. **Data Organization** – Structuring and arranging data into usable formats such as tables, spreadsheets, or databases. This facilitates easier manipulation and analysis.
4. **Data Analysis** – Applying statistical techniques to interpret the data and derive insights. This may involve descriptive statistics, inferential methods, or advanced modelling.
5. **Reporting and Interpretation** – Presenting findings in a clear, concise manner through charts, graphs, or reports. This step bridges the gap between data and actionable knowledge.

2.5 METHODS OF DATA COLLECTION

Statistics organizes vast amounts of raw data into structured formats such as tables, graphs, and charts. This makes information more accessible and understandable, facilitating quick insights and efficient communication of findings.

Large, complex datasets are distilled into meaningful summaries using statistical measures like averages and percentages. This reduces the complexity of data, allowing for easier interpretation and comparison across different scenarios.

By applying statistical techniques, data can be compared across different groups, regions, or time periods. This aids in identifying relationships, disparities, and trends, which are crucial for making informed decisions. Historical data is analysed to predict future outcomes, enabling businesses and governments to plan for growth, allocate resources, and anticipate market demands or economic shifts.

Statistics provides evidence-based insights that support critical decisions in fields like business, healthcare, and policy-making. This minimizes uncertainty and enhances the likelihood of achieving desired results.

The kind of study, the kind of data needed, and the resources available all influence the data gathering techniques. Primary and secondary data are two main groups into which data collecting may be divided.

2.5.1 Primary Data Collection

Primary data refers to information gathered directly from original sources. This method is tailored to meet the specific needs of a research project, ensuring relevance and accuracy.

Methods of Collecting Primary Data:

1. **Surveys and Questionnaires** – Surveys are one of the most popular ways to get information from people by asking them structured questions. They can be conducted in person, via phone, or online.
2. **Interviews** – Direct interaction between the researcher and the respondent. Interviews can be structured, semi-structured, or unstructured, depending on the level of flexibility required.
3. **Observations** – Recording behaviours or events as they occur naturally. Observational data is often used in social sciences, psychology, and market research.
4. **Experiments** – Conducting controlled tests to observe the effects of varying conditions. Experimental data is crucial in scientific research and product development.

5. **Focus Groups** – Gathering a small group of individuals to discuss specific topics, providing qualitative insights into consumer behaviour and preferences.

Advantages of Primary Data:

- High relevance to the research issues.
- Greater control over data accuracy and reliability.
- Flexibility in data collection methods.

Disadvantages of Primary Data:

- Time-consuming and costly.
- Requires extensive planning and resources.

2.5.2 Secondary Data Collection

Secondary data involves using information previously gathered by other researchers, institutions, or organizations. This data is often readily available and can be accessed through publications, reports, and online databases.

Sources of Secondary Data:

1. **Government Reports** – Data on population, economic indicators, and social trends.
2. **Research Articles and Journals** – Peer-reviewed studies that offer valuable insights.
3. **Company Records** – Financial statements, sales reports, and operational data.
4. **Online Databases** – Aggregated data from various sources, providing extensive datasets for analysis.

Advantages of Secondary Data:

- Cost-effective and time-saving.
- Easy to access large volumes of data.
- Useful for exploratory research and trend analysis.

Disadvantages of Secondary Data:

- May not be specific to the current research problem.
- Risk of outdated or inaccurate information.
- Lack of control over data collection methods.

2.6 SAMPLING TECHNIQUES

The practice of choosing a subset of people or data points from a broader population is known as sampling. In order to make generalizations and predictions, proper sampling guarantees that the chosen group accurately represents the total population.

Types of Sampling Techniques:

1. **Random Sampling** – Each person in the population has an equal probability of being chosen.
2. **Stratified Sampling** – Samples are taken from each stratum once the population is split up into subgroups (called strata) according to particular traits.
3. **Cluster Sampling** – After dividing the population into clusters, entire clusters are chosen at random.
4. **Systematic Sampling** – Selecting every n th individual from a list or population.
5. **Convenience Sampling** – Selecting individuals who are easily accessible, often leading to bias.

2.7 IMPORTANCE OF STATISTICS

Data accuracy is a cornerstone of effective statistical analysis, ensuring that the insights derived are valid and actionable. Inaccurate data can mislead decision-makers, resulting in costly mistakes, missed opportunities, and compromised outcomes. The significance of accurate data spans across industries, influencing everything from financial planning and market research to healthcare diagnostics and policy-making.

- **Informed Decision-Making** – Accurate data provides a solid foundation for strategic decisions, reducing uncertainty and guiding resource allocation.
- **Risk Reduction** – By minimizing errors, accurate data helps organizations avoid costly miscalculations and operational inefficiencies.
- **Enhanced Forecasting** – Reliable data improves the accuracy of predictive models, allowing businesses and governments to anticipate future trends more effectively.
- **Operational Efficiency** – Accurate data ensures smooth operations by reducing redundancies, optimizing supply chains, and streamlining workflows.
- **Customer Trust and Satisfaction** – Businesses that use accurate data for personalized services and product development are more likely to have a higher chance to develop lasting relationships with customers.
- **Scientific Validity** – In research, accurate data ensures the reproducibility of results, reinforcing the credibility of scientific studies.
- **Economic Planning** – Governments rely on accurate economic data to draft policies, allocate resources, and respond to socio-economic challenges.
- **Performance Monitoring** – Accurate data allows organizations to track progress, measure performance, and adjust strategies as needed.
- **Accountability and Transparency** – Reliable data fosters trust among stakeholders by demonstrating clear, evidence-based results.
- **Competitive Advantage** – Companies that prioritize data accuracy gain a competitive edge by responding to market changes with precision and agility.

Ultimately, the importance of data accuracy lies in its ability to drive innovation, improve efficiency, and ensure that decision-making is grounded in reality rather than assumptions.

2.8 STATISTICAL TOOLS DATA PROCESSING

Statistical tools are essential to transforming raw data into valuable insights. Once data is collected, it needs to be organized, analysed, and interpreted effectively to draw meaningful conclusions. Various statistical tools are used to process and analyse data, ranging from simple descriptive methods to complex inferential techniques.

1. **Descriptive Statistics** – This includes measurements like mean, median, mode, and standard deviation, which summarize the central tendency, spread, and variability of the data. Descriptive statistics help in understanding the basic characteristics of a dataset, provide a summary of the main trends and patterns.
2. **Inferential Statistics** – These resources enable analysts to make predictions or generalizations about a population based on a sample. Techniques like hypothesis testing, confidence intervals, and p-values are used to infer the likelihood of an event or outcome occurring.
3. **Regression Analysis** – Regression models help to recognize the connection between two or more variables. By identifying patterns and trends, regression analysis enables predictions and assists in identifying cause-and-effect relationships.
4. **Time Series Analysis** – This tool helps to analyze data collected over time, helping to determine trends, cycles, and seasonal variations. Time series analysis is crucial in forecasting future values based on historical data.
5. **Correlation Analysis** – This method aids in understanding how two variables move in respect to one another by measuring the direction and strength of the link between them.

These statistical tools allow analysts to make data-driven decisions, support hypothesis testing, and uncover underlying patterns in large datasets.

2.9 APPLICATION OF DATA COLLECTION IN ECONOMIC ANALYSIS

Data collection is central to economic analysis, providing insights that guide policy decisions, business strategies, and financial forecasting. By collecting data on key economic indicators, analysts can assess the current state of the economy, predict future trends, and make informed decisions.

- **Monetary Policy Decision-Making** – Central banks rely on data such as inflation rates, interest rates, and money supply to set monetary policies that stabilize the economy.
- **Labor Market Analysis** – Employment and unemployment data help policymakers design labour policies aimed at reducing joblessness and improving workforce participation.
- **Economic Growth and Development** – Data on GDP, industrial production, and investment help track national economic performance and design strategies to foster growth.
- **Trade and International Economics** – Data on imports, exports, and trade balances allow economists to evaluate the effects of trade policies, tariffs, and global economic relationships.
- **Income Distribution and Poverty Analysis** – Data on household income, poverty levels, and wealth distribution guide the development of social welfare programs and economic equality measures.

- **Consumer Behaviour** – Data collected on consumer spending, preferences, and demand patterns enables businesses to adapt their products and marketing strategies.
- **Public Policy Formulation** – Data on social, healthcare, and education indicators inform policy decisions aimed at improving public welfare and allocating resources efficiently.
- **Market Forecasting** – Economic data helps businesses predict market trends, plan for future demand, and optimize pricing strategies.
- **Investment Decisions** – Accurate economic data supports investors in making informed decisions about capital allocation, stock investments, and financial risk management.
- **Evaluating Economic Programs** – Data collection helps assess the efficiency of government programs and initiatives, allowing for adjustments and enhancements in the provision of public services.

These applications underscore the critical role of data gathering in shaping sound economic policies, guiding business operations, and ensuring economic stability.

CHECK YOUR PROGRESS

1. What is meant by statistical investigation? Explain its main stages.

.....

.....

.....

2. Distinguish between primary data and secondary data.

.....

.....

-
3. Explain the different methods of collecting primary data.
-
-
-

2.10 SUMMARY

This unit explored the processes and methods involved in data collection, highlighting its importance in statistical analysis. By understanding sources of primary and secondary data, sampling methods, also significance of data accuracy, learners gain essential skills for conducting effective research and making data-driven decisions.

2.11 EXERCISES

1. Define and explain the components of statistical processes.
2. Differentiate between primary and secondary data collection methods with examples.
3. Discuss the advantages and limitations of using surveys for data collection.
4. Explain the concept of sampling and describe any three sampling techniques in detail.
5. How does data accuracy affect the outcomes of statistical analysis? Suggest ways to ensure accuracy.
6. Describe the role of statistical tools in economic analysis and provide real-world examples.
7. Compare and contrast random sampling and stratified sampling. Which method is more suitable for population studies and why?

8. List and explain the benefits and drawbacks of using secondary data in research.
9. What are the different methods of collecting primary data? Illustrate with examples from business or healthcare.
10. Design a small survey (5-10 questions) to collect data on customer satisfaction for a retail store. Explain how you would ensure the data collected is reliable and valid.

Objective Type Questions

11. **Which of the following is a source of primary data?**
 - a) Government census reports
 - b) Research articles
 - c) Personal interviews
 - d) Online databases
12. **Which sampling method divides a population into subgroups before sampling?**
 - a) Cluster Sampling
 - b) Stratified Sampling
 - c) Systematic Sampling
 - d) Random Sampling
13. **What is the main advantage of secondary data?**
 - a) It is specific to the current research question.
 - b) It saves time and costs.
 - c) It is always free from errors.
 - d) It guarantees data accuracy.
14. **Which of the following is NOT a primary data collection method?**
 - a) Surveys
 - b) Observations
 - c) Focus Groups
 - d) Company financial records

15. **Systematic sampling selects data by:**
- a) Choosing clusters randomly
 - b) Selecting every n th item from a population
 - c) Dividing the population into strata
 - d) Selecting individuals based on convenience
16. **The purpose of descriptive statistics is to:**
- a) Test hypotheses
 - b) Draw conclusions about a population
 - c) Summarize and present data
 - d) Establish causality between variables
17. **Which term refers to the spread of data points around the mean?**
- a) Median
 - b) Correlation
 - c) Standard Deviation
 - d) Regression
18. **What is the basic reason for conducting pilot surveys?**
- a) To analyze the whole population
 - b) To check the feasibility and refine the main survey
 - c) To collect final data for analysis
 - d) To measure economic growth
19. **Which of the following is a limitation of using primary data?**
- a) It is outdated.
 - b) It is difficult to access.
 - c) It is time-consuming and expensive to collect.
 - d) It may not be specific to the research objective.
20. **Which sampling method selects entire groups rather than individuals?**
- a) Stratified Sampling
 - b) Random Sampling

- c) Cluster Sampling
- d) Convenience Sampling

Answers (Objective Type Questions):

- c) Personal interviews
- b) Stratified Sampling
- b) It saves time and costs
- d) Company financial records
- b) Selecting every n th item from a population
- c) Summarize and present data
- c) Standard Deviation
- b) To test the feasibility and refine the main survey
- c) It is time-consuming and expensive to collect
- c) Cluster Sampling

2.12 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 3 DATA CLASSIFICATION AND TABULATION

STRUCTURE

- 3.1 Introduction
- 3.2 Objectives
- 3.3 Definition of Data Classification
- 3.4 Types of Data Classification
- 3.5 Principles of Data Classification
- 3.6 Frequency Distribution
- 3.7 Tabulation of Data
- 3.8 Types of Tables
- 3.9 Steps in Data Tabulation
- 3.10 Importance of Data Classification and Tabulation
- 3.11 Summary
- 3.12 EXERCISES
- 3.13 References

3.1 INTRODUCTION

In statistics, data categorization and tabulation are essential methods for arranging and displaying data in a comprehensible manner. Data, in its raw form, can be overwhelming and difficult to interpret. Therefore, the process of classification helps to categorize data into distinct groups or classes, while tabulation involves presenting data in tables, making it easier to understand, analyze, and draw conclusions. These techniques are essential in turning raw information into usable insights, facilitating the analysis of large datasets across various fields, such as economics, business, healthcare, and social sciences.

Data categorization and tabulation provide structure, enabling researchers to identify patterns, trends, and relationships within the data.

By organizing data in a systematic manner, researchers can apply statistical methods effectively and ensure that the analysis is based on well-organized, structured, and interpretable data. This unit will explore the concepts of data classification, types of classification, methods of tabulation, and their importance in statistical analysis.

3.2 OBJECTIVES

After completing this unit, students will be able to

- Recognize the significance of data tabulation and classification.
- Determine the various data kinds and classification techniques.
- Apply the procedure of classification to organize raw data.
- Learn how to present data using different kind of tables.
- Analyze data presented in tabular form.
- Understand the importance of accurate data categorization in statistical analysis.

3.3 DEFINITION OF STATISTICAL PROCESSES

Data classification is the process of organizing data into distinct categories or groups based on shared characteristics or attributes. This is a fundamental step in statistical analysis, as it simplifies large datasets and makes it easier to analyze, interpret, and extract meaningful insights. Data classification helps in identifying patterns, comparing variables, and facilitating decision-making by grouping similar data points together.

There are two main types of data classification: **qualitative** (categorical) and **quantitative** (numerical). Qualitative classification involves grouping data based on non-numeric characteristics, such as gender, color, or occupation. For example, classifying people into categories like "male" or "female" is a qualitative classification. On the other hand,

quantitative classification involves grouping data based on numerical values, such as age, income, or temperature. It is possible to further categorize these data points into continuous and discrete groups.

Data classification must adhere to certain principles, such as **mutual exclusiveness**, where every data point needs to fall into a single category., and **exhaustiveness**, ensuring all data points are included in the classification system. Other important principles include **homogeneity**, where similar items are grouped together, and **simplicity**, which ensures the classification system is easy to understand and apply.

Ultimately, data classification is an essential tool for organizing and simplifying complex datasets, making it easier to conduct statistical analyses and derive actionable insights.

3.4 TYPES OF DATA CLASSIFICATION

Data classification involves categorizing data into meaningful groups or classes based on certain attributes or characteristics. The nature of the data and the purpose of the analysis determine the classification method used. Broadly, data can be classified into **qualitative** (categorical) and **quantitative** (numerical) types. Below are the primary types of data classification, along with relevant examples:

1. Qualitative Classification (Categorical Classification)

Qualitative classification involves grouping data based on non-numeric characteristics that describe categories or qualities. These characteristics cannot be measured or quantified but are essential for distinguishing between different groups.

Examples of Qualitative Classification:

- **Gender:** Data can be classified into categories like "Male," "Female," and "non-binary." These categories represent distinct groups based on gender identity.
- **Colors:** The data can be classified into color groups such as "Red," "Blue," "Green," and "Yellow." These are descriptive categories based on color attributes.
- **Occupation:** People can be classified based on their profession into categories like "Teacher," "Engineer," "Doctor," and "Artist." Each occupation represents a distinct qualitative category.
- **Type of Cuisine:** A restaurant review dataset can classify data into categories like "Italian," "Chinese," "Mexican," and "Indian," based on the type of cuisine served.

2. Quantitative Classification (Numerical Classification)

Quantitative classification involves grouping data based on numerical values that can be measured or counted. This type of classification is useful for numerical data that represents quantities or amounts. Quantitative data can be further divided into **discrete** and **continuous** types.

Examples of Quantitative Classification:

- **Income Levels:** Data on household income can be grouped into categories like "Low," "Medium," and "High." For example, income ranges can be classified as:
 - Low: Less than \$30,000
 - Medium: \$30,000 to \$60,000

- High: More than \$60,000
- **Age Groups:** Age data can be classified into intervals such as "0-18 years," "19-35 years," "36-50 years," and "51+ years." These groups represent age ranges.
- **Product Prices:** Price data for products can be classified into categories such as "Under \$10," "\$10-\$50," "\$51-\$100," and "Above \$100."
- **Temperature:** Temperature data can be classified into intervals such as "0-10°C," "11-20°C," and "21-30°C" to represent different temperature ranges.

3. Discrete Classification

Discrete classification involves grouping data into distinct, separate categories. These categories are finite and can often be counted. Discrete data typically represents whole numbers or counts.

Examples of Discrete Classification:

- **Number of Children:** The number of children in a household can be classified into categories like "0," "1," "2," "3," or more. This is an example of discrete data, as the number of children is countable and finite.
- **Number of Employees:** Companies can classify organizations based on the number of employees, such as "1-50," "51-200," and "201+" employees.
- **Cars in a Household:** A household can be classified based on the number of cars owned, such as "0 cars," "1 car," "2 cars," and so on.

4. Continuous Classification

Continuous classification involves grouping data that can take any value within a certain range or interval. Continuous data typically represents measurements that can be expressed in decimal form and are not restricted to whole numbers.

Examples of Continuous Classification:

- **Height:** Height data can be classified into intervals like "150-160 cm," "161-170 cm," "171-180 cm," and "181+ cm." Since height is continuous, it can take any value within a range.
- **Weight:** Weight data can be classified into ranges such as "50-60 kg," "61-70 kg," and "71-80 kg." Like height, weight is continuous and can vary within intervals.
- **Time:** Time taken to complete a task can be classified into intervals like "0-10 minutes," "11-20 minutes," and so on, with the possibility of any value within each range.

5. Ordinal Classification

Ordinal classification involves grouping data based on a ranked order or scale. The categories are ordered but not necessarily evenly spaced, meaning there is a meaningful ranking of the categories but no exact numerical value or distance between them.

Examples of Ordinal Classification:

- **Education Level:** Data on education level can be classified as "High School," "Associate's Degree," "Bachelor's Degree," and "Master's Degree." These categories have an inherent order but do not have a fixed numerical distance between them.

- **Customer Satisfaction:** Ratings like "Very Poor," "Poor," "Fair," "Good," and "Excellent" are ordinal classifications, as they represent a ranking of satisfaction levels without precise numerical values between them.
- **Hotel Star Rating:** Hotels are often classified by their star ratings, such as "1-star," "2-star," "3-star," and "5-star." While there is an order, the difference in the experience provided by each level can vary.

6. Nominal Classification

When there is no inherent order, nominal categorization entails grouping data into mutually exclusive groups. These categories are just names or labels.

Examples of Nominal Classification:

- **Blood Group:** Blood types such as "A," "B," "AB," and "O" are nominal classifications, as there is no specific order to these categories.
- **Animal Species:** Animals can be classified into different species, such as "Dog," "Cat," "Elephant," and "Tiger." Each species is distinct, but there is no ordering or ranking involved.
- **Color:** The colors "Red," "Green," "Blue," and "Yellow" are nominal categories. While they are distinct, they have no inherent order.

Data classification is a powerful tool that helps to organize and structure data in a way that makes it easier to analyze. By classify data into meaningful groups, analysts can identify patterns, draw comparisons, and come to well-informed conclusions. Whether data is qualitative or quantitative, discrete or continuous, classification provides clarity and

facilitates deeper statistical analysis, making it an essential step in the process of data analysis.

3.5 PRINCIPLES OF DATA CLASSIFICATION

The process of classification should follow a set of principles to ensure that the data is grouped logically and consistently. Some important principles include:

- **Mutual Exclusiveness** – Each data item should belong to a single, exclusive category or class, without overlap.
- **Exhaustiveness** – All Data ought to be assigned to a category. No data points should be left out.
- **Homogeneity** – Items within each class should share similar characteristics or features.
- **Simplicity** – The classification should be easy to understand and not overly complex.

3.6 FREQUENCY DISTRIBUTION

Frequency may be described as the quantity of individuals having the same measurement or lying in the identical measurement group. Frequency distribution is the distribution of frequencies over different measurements. Frequency distributions may be two types (i) Discrete frequency distributions (ii) Continuous frequency distributions. Let us discuss the method of construction of both distributions.

3.6.1 Method of Construction of Discrete Frequency Distribution

The following steps are taken:

- (i) A table with columns is prepared. In first column all the values of the variable are written in ascending order starting from the lowest to highest.

- (ii) We go through the values of the given data and insert a tally mark against each value of the variable. If this value again occurs in the data, tally mark is again inserted. For the sake of convenience, we use four and cross method that is, after four tallies fifth tally is crossly drawn.
- (iii) We count number of tallies with respect to each value of the variable and place it in the third column made for frequency.

Example 1: The number of members in each family were recorded. The figures for 25 such families are given below:

0	4	2	3	1
1	2	0	5	3
4	1	2	0	2
2	1	0	4	3
3	2	1	0	4

Express the above data by discrete frequency distribution.

Number of members	Tally marks	Frequency
0	NVI	5
1	NV	5
2	NVI	6
3	IIII	4
4	IIII	4
5	I	1
	Total	25

3.6.2 Method of Construction of Continuous Frequency Distribution

For a continuous frequency distribution, it is necessary to understand the following technical terms:

Class Interval: This name has been given to each group in which the range of the variable is divided.

Class Limits: These are the values of the class intervals. The minimum and the maximum values are called the lower- and upper-class limits.

Magnitude: The difference between the upper- and lower-class limits is called magnitude.

Mid-Value: The mid-value is that value which lies half way between lower- and upper-class limits. For example, in a class of 40 – 50, the mid value is:

$$\frac{\text{upper class limit} + \text{lower class limit}}{2} \text{ or } \frac{40 + 50}{2} = 45$$

Class intervals can be formed according to two methods:

- (a) Inclusive Method
 - (b) Exclusive Method
- (a) Inclusive Method: Under this method the upper limit of a class is included in that class. For example, if pocket money of engineering students is grouped as ₹40 – 59, 60 – 79, 80 – 99 etc., then in the group 40 – 59, we include students whose pocket money is between ₹40 and 59. If the pocket money is ₹60, he is included in the next class 60-79 and so on. If there are no whole numbers, the classes can be made 40 - 59.9, 60-79.9 and so on.
- (b) Exclusive Method: Under this method, the upper limit of one class interval is the lower limit of the other. For example, 40 –

60,60 – 80,80 – 100 and so on. We include in first group, the students whose pocket money is ₹ 40 or more but less than ₹60. If the pocket money of a student is ₹60, he is not included in the first group but in second i.e., 60-80.

3.7 TABULATION OF DATA

Tabulation refers to the process of organizing data into tables for easier interpretation and analysis. Tables summarize large datasets by presenting them in an organized, concise, and systematic format. Tabulation helps to compare different variables, detect trends, and analyze relationships between data points.

3.8 TYPES OF TABLES

There are several types of tables used in statistical analysis, each serving a specific purpose:

- **Simple Table** – Displays a single set of data points, often in a row and column format. For example, a table showing the number of students in different grades.
- **Frequency Table** – Used to display the number of occurrences (or frequency) of different values or categories in a dataset. For example, a table showing how many people fall into different age groups.
- **Two-Way Table** – A table that shows the relationship between two variables. It is commonly used to display data in rows and columns, making it easy to compare two variables side by side.
- **Grouped Table** – Organizes data into classes or intervals and shows the frequency of occurrences within each class. For example, grouping income data into different income brackets.

3.9 STEPS IN DATA TABULATION

Several crucial phases are included in the tabulation process to guarantee that the data is effectively arranged:

1. **Determine the objective** – Define the purpose of the table and the specific data to be included.
2. **Select variables** – Identify the variables that will be tabulated and decide whether they are categorical or numerical.
3. **Organize data** – Sort the data into appropriate categories or intervals based on the type of data.
4. **Design the table** – Create the table with clear rows and columns, labelling them appropriately for easy understanding.
5. **Fill in the table** – Input the data into the table and ensure that it is accurate and consistent.
6. **Interpret the table** – Analyze the table to identify trends, patterns, or relationships in the data.

3.10 IMPORTANCE OF DATA

CLASSIFICATION AND TABULATION

The importance of classification and tabulation lies in their ability to organize and present data in a way that makes it more accessible and understandable. They are essential tools for:

- **Simplifying complex data** – By grouping data into categories, classification and tabulation reduce the complexity of large datasets, making them easier to analyze.
- **Enhancing comparability** – These techniques help to compare different data points or variables in a structured format, facilitating deeper analysis and understanding.

- **Detecting patterns and trends** – Well-organized data helps identify trends, patterns, and relationships, providing valuable insights that inform decision-making.
- **Supporting decision-making** – Classification and tabulation provide a clear overview of the data, making it easier to make informed decisions based on accurate and organized information.
- **Improving efficiency** – Data classification and tabulation streamline the method of data analysis by providing structure, saving time, and reducing errors.

CHECK YOUR PROGRESS

1. What is meant by classification of data? Explain its different types.

.....

.....

.....

2. Define tabulation and explain the essential parts of a statistical table.

.....

.....

.....

3. Distinguish between simple, double, and multiple tables.

.....

.....

.....

3.11 SUMMARY

In this unit, we explored the concepts of data classification and tabulation, which are fundamental techniques in statistical analysis. Data classification helps in organizing data into meaningful categories, while tabulation presents the data in tables for easier interpretation. Both techniques are essential for simplifying large datasets, identifying patterns, and making informed decisions based on accurate, structured data. Understanding how to classify and tabulate data effectively is crucial for anyone involved in data analysis, from researchers to business professionals.

3.12 EXERCISES

1. Define data classification and explain its importance in statistical analysis.
2. Differentiate between qualitative and quantitative data classification.
3. Describe the principles of data classification.
4. List the types of tables used in statistical analysis and explain their purpose.
5. How does tabulation help in presenting data for analysis?
6. Explain the difference between a simple table and a frequency table.
7. Create a frequency table for a set of data on student grades.
8. Discuss the significance of mutual exclusiveness and exhaustiveness in classification.
9. What is a two-way table, and how is it used in statistical analysis?
10. Design a grouped table for data on monthly income in a community.

Objective Type Questions

1. **Identify an example of qualitative data classification.**

- a) Age groups
- b) Income levels
- c) Gender
- d) Height

Answer: c) Gender

2. **What is the primary purpose of data classification in statistics?**

- a) To calculate the mean
- b) To organize and group data into categories
- c) To perform hypothesis testing
- d) To measure standard deviation

Answer: b) To organize and group data into categories

3. **Which of the following data classifications is related to non-numeric attributes?**

- a) Continuous data
- b) Discrete data
- c) Qualitative data
- d) Quantitative data

Answer: c) Qualitative data

4. **Which classification method involves grouping data based on numerical measurements that can take any value within a range?**

- a) Ordinal classification
- b) Discrete classification
- c) Continuous classification
- d) Nominal classification

Answer: c) Continuous classification

5. **Identify the example of ordinal classification.**

- a) Blood group categories
- b) Hotel star ratings

- c) Types of fruits
- d) Species of animals

Answer: b) Hotel star ratings

6. **In discrete classification, data is classified into:**

- a) Categories with continuous ranges
- b) Finite, countable groups
- c) Ranked order based on quality
- d) Non-numeric groups

Answer: b) Finite, countable groups

7. **Which of the following is an example of nominal classification?**

- a) Education level
- b) Customer satisfaction ratings
- c) Animal species
- d) Height categories

Answer: c) Animal species

8. **Which principle of data classification ensures that no data point is left out of the classification system?**

- a) Simplicity
- b) Exhaustiveness
- c) Mutual exclusiveness
- d) Homogeneity

Answer: b) Exhaustiveness

9. **What type of data classification is used when grouping students based on their grades as "A," "B," "C," and "D"?**

- a) Ordinal classification
- b) Nominal classification
- c) Continuous classification
- d) Quantitative classification

Answer: a) Ordinal classification

10. **Identify an example of continuous data.**

- a) Number of children in a family

b) Age of individuals

c) Blood group

d) Favourite color

Answer: b) Age of individuals

3.13 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.



UNIT 4 GRAPHIC REPRESENTATION OF DATA

STRUCTURE

- 4.1. Introduction
- 4.2. Objectives
- 4.3. Diagrammatic Representation of Data
 - 4.3.1. Rules for Constructing diagrams
 - 4.3.2. Types of diagrams
 - 4.3.3. Advantages of diagram representation
 - 4.3.4. Limitations of diagram representation
- 4.4. Graphical Representation of Data
 - 4.4.1. The Histogram
 - 4.4.2. The Frequency Polygon
 - 4.4.3. The Frequency Curve
- 4.5. Cumulative Frequency Curve or Ogive
- 4.6. Summary
- 4.7. Exercises
- 4.8. References

4.1 INTRODUCTION

In the study of statistics, the effective presentation of data is crucial for understanding and analysis. Among various methods, graphical representation plays a pivotal role in making data more comprehensible, engaging, and interpretable. Although the main focus of this unit is on **graphical representation**, it is essential first to understand **diagrammatic representation**, which lays the groundwork for graphical techniques.

The unit begins with a section on *diagrammatic representation of data*, introducing basic rules for constructing diagrams, various types of diagrams, and discussing their advantages and limitations. This section provides a visual and intuitive way to present categorical and numerical data in a simplified manner.

Following this, the unit delves into the core area of *graphical representation*, focusing on techniques suited for frequency data. You will learn about **histograms**, **frequency polygons**, **frequency curves**, and **cumulative frequency curves (ogives)**, all of which are vital tools in statistical analysis and interpretation.

Through this unit, learners will not only gain the skills to construct these graphical tools but also develop the ability to choose the most appropriate representation based on the type and purpose of the data.

4.2 OBJECTIVES

After studying this unit, learners will be able to:

- Understand the distinction between diagrammatic and graphical representation of data.
- Apply the general rules for constructing effective diagrams.
- Identify and construct various types of diagrams such as bar diagrams, pie charts, and pictograms.
- Recognize the advantages and limitations of diagrammatic data presentation.
- Construct and interpret different types of graphs, including:
 - Histograms
 - Frequency polygons
 - Frequency curves
 - Cumulative frequency curves (ogives)

- Select appropriate graphical techniques for presenting frequency data.
- Analyse and draw conclusions from graphical data representations.

4.3 DIAGRAMMATIC REPRESENTATION OF DATA

4.3.1. Rules for Constructing diagrams

The following are some of the important points to be kept in mind while constructing a Diagram:

1. The Diagram should be simple, impressive and attractive.
2. A proper scale should be used so that the Diagram should not be unnecessarily large or small.
3. A proper title should be given to the Diagram.
4. An index must be given for identification of different shades or colours used.
5. There are a number of Diagrams for statistical data. Extreme care should be taken in the selection of a suitable Diagram.
6. Footnotes should also be given at the bottom of the Diagrams.

4.3.2 Types of diagrams

The choice of the types of Diagrams by which the data are to be presented is a difficult job. There is a large variety of Diagrams and extreme care must be taken while choosing a Diagram for a particular type of data. The following are the common Diagrams:

1. One Dimensional Diagrams
2. Two Dimensional Diagrams

3. Pictograms or Pictorial Diagrams

4. Cartograms or maps

1. One Dimensional Diagrams

In this type of Diagrams, we consider only one dimension and these are in the shapes of lines or bars. Accordingly, we call them line Diagrams or bar Diagrams. A bar is a thick line whose width is shown merely for attraction, and is not taken into consideration. A bar can be drawn either to a common horizontal base or vertical base. Vertical bars on horizontal base are more popular and proper comparisons can be made by them. Bar Diagrams are of the following types:

- (i) Simple Bar Diagram
- (ii) Multiple Bar Diagram
- (iii) Subdivided Bar Diagram
- (iv) Percentage Bar Diagram

(i) Simple Bar Diagram: In a simple bar Diagram, only one variable is presented. It is generally used to show changes in the magnitudes of a phenomenon over time (Chronological) or place (Geographical). Bars can be drawn vertically or horizontally side by side at equal distance from one another. Bars of equal width are formed with their length proportional to various values of the variable. A scale is placed at the side of the Diagram.

Example 1: Draw a bar Diagram to the following data relating to the marks (outs of 50) of 10 students of chemical engineering.

Roll No.:	1	2	3	4	5	6	7	8	9	10
Marks:	15	20	25	30	35	37	20	15	40	15

Solution: Simple bar Diagram showing marks of 10 students (Vertical bars)

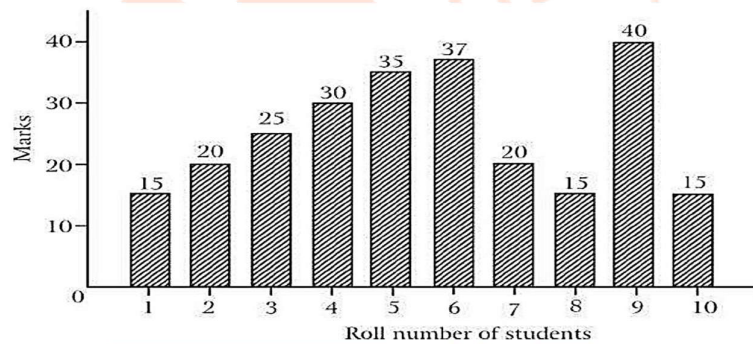


Figure 4.1: Example of simple bar diagram

The simple bar Diagram or horizontal bars can be drawn by taking marks of students on y -axis and roll number of students on x -axis.

(ii) Multiple Bar Diagram: These types of bars are very useful for direct comparison between two or more independent quantities. Two or more bars in different shades for different series but for the same period are drawn together on a common scale.

Example 2: Draw a multiple bar Diagram from the following data:

Year	(Profit in thousands)	
	Company A	Company B
1998 – 1999	120	90
1999 – 2000	135	95
2000 – 2001	140	108
2001 – 2002	145	120
2002 – 2003	150	130

Solution: Diagram showing profit of company A and company B.

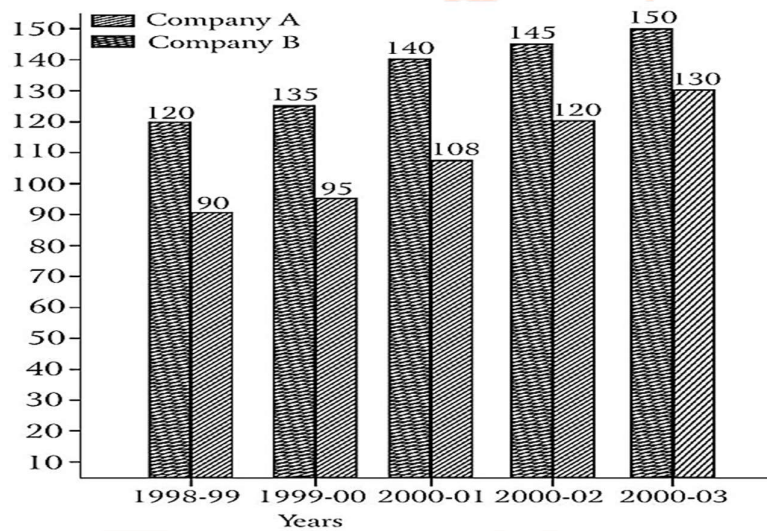


Figure 4.2: Example of multiple bar diagram

(iii) Sub-divided Bar Diagram: When a given quantity can be broken in parts, all the parts with the total can be shown by means of sub-divided bars. This is done by dividing a single bar into several parts representing various sub-divisions of the given aggregate.

Example 3: Represent the following data by sub-divided bar Diagram.

Examination	No. of Students appeared		
	2002	2003	2004
M.Sc.	200	250	400
M.Com.	250	250	300
M.A	150	200	300

Solution:

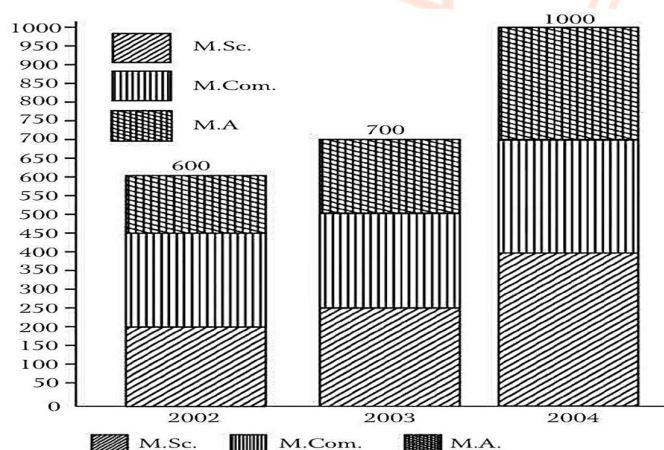


Figure 4.3: Example of sub-divided bar diagram

(iv) Percentage Bar Diagram: Sub-divided bar Diagram may also be constructed on the basis of percentage. These types of Diagrams are very useful for depicting and emphasising relative changes. In this case all the bars are taken of equal length i.e., 100.

Example 4: Represent the data of example 3 by percentage bar Diagram.

Solution: For representing the data in percentage bar Diagram, percentage of students for each faculty is to be obtained.

Table showing % of students, faculty wise for each year.

Year	Percentage of Students			
	M.Sc.	M.Com	M.A	Total
2002	33.3	41.7	25.0	100.0
2003	35.7	35.7	28.6	100.0
2004	40.0	30.0	30.0	100.0

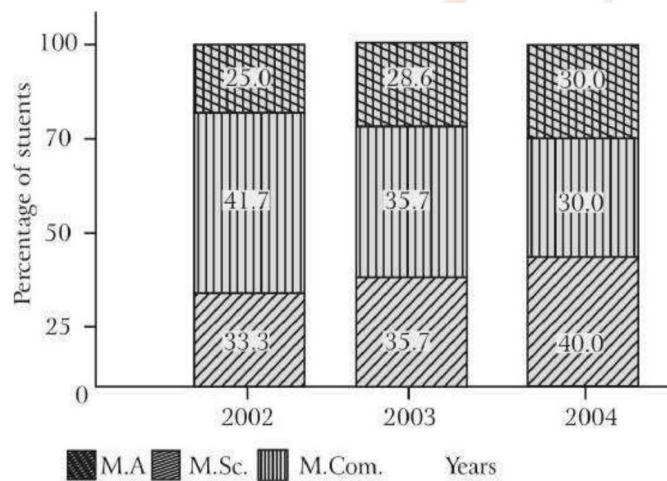


Figure 4.4: Example percentage bar diagram

2. Two-dimensional diagrams or area diagrams

In bar Diagrams, the length of the bar was considered and width was taken only for attraction. In two-dimensional Diagrams we consider length as well as width of the geometrical figures. These Diagrams are also called area Diagrams because the area of the geometrical figure is taken into consideration. The Diagrams which are commonly used are as given below:

- (i) Rectangle Diagrams
- (ii) Square and Circle Diagrams
- (iii) Pie-Diagrams

(i) Rectangle Diagrams: Rectangles are two-dimensional Diagrams in which length and width are considered. The length and width of each rectangle may vary and therefore, their area will also vary. The area of rectangle represents the proportion of the values of the variable. Percentage rectangular Diagram may also be drawn to compare the family budgets.

Example 5: Represent the following data by a suitable Diagram showing family budget of two families:

Items of Expenditure	Family A (₹)	Family B (₹)
Food	1800	2200
Clothing	600	1000
House Rent	1000	2000
Fuel and Electricity	400	500
Miscellaneous	200	300
Total	4000	6000

Solution: Here the total expenditure of the family A is ₹ 4000 and that of family B is ₹6000. For the rectangular Diagram, we take the width of the rectangles in proportion to the total income i.e., 2:3 (4000:6000), and the height shows the subdivision of expenditures.

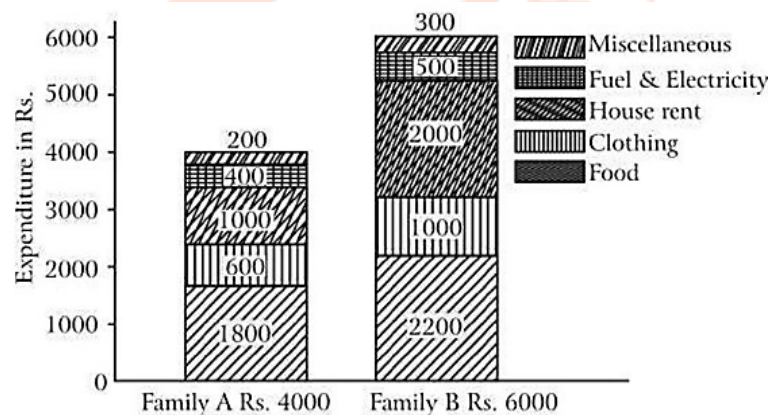


Figure 4.5 Example of rectangular diagram

(ii) Square and Circle Diagrams: When the difference between the proportion of two quantities is very large, rectangle Diagrams are not very convenient. For example, if proportion of expenditure in two families is 1:14 or so, the width of the rectangles will be in the ratio 1:14, which is not a convenient ratio to draw rectangle Diagram. In such situations we use square or circle Diagrams. In a square Diagram the

areas of the squares are taken proportional to the given quantities, while in circles the areas of circles are taken proportional to the given quantities. Thus, for square Diagram the sides of the square are proportional to the square root of the given quantities and for circle Diagram the radius of the circle is taken proportional to the square roots of the given quantities.

The bases of the square are taken on the same horizontal line and circles are so constructed, that their centres lie on the same horizontal line. The method is illustrated below.

Example 6: Represent the following data by square and circle Diagrams.

Area Under Rice

Country	India	Iraq	Sri Lanka	Trinidad
Area (000 hectares)	9403	2642	721	68

Solution:

Country	Area (000) hectares	Square root	Sides of the square (cms)
India	9403	96.9	4.85
Iraq	2642	51.4	2.57
Sri Lanka	721	26.8	1.34
Trinidad	68	8.2	0.41

In the above table square roots are calculated and each figure has been reduced after division by a suitable number say 20. The resultant figures may be taken as the sides of the squares in inches/cms.

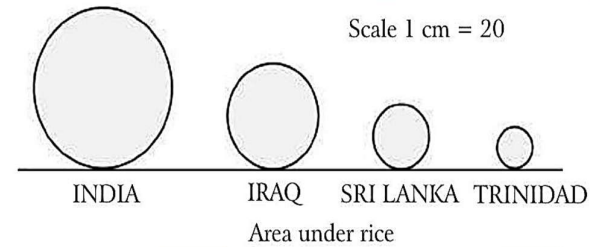
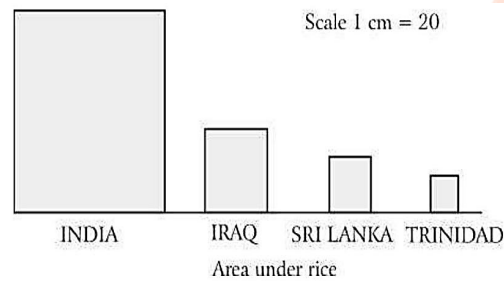


Figure 4.6 Square diagram

Figure 4.7 Circle

(iii) Pie Diagrams: Pie Diagrams are more attractive to look at and are simple to draw. They are constructed by dividing a circle into sectors in order to represent the component parts. The sum of the angles at the centre of a circle equal 360° . The various component parts should be reduced in proportion to this and then the sectors in the circle marked off.

Example 7: Represent the following data by a pie Diagram.

Infant Mortality in different Cities

City	Mumbai	Delhi	Kanpur	London	Paris
Infant Mortality	174	168	220	93	66

Solution: We will convert these values into corresponding degrees in the circle taking the total infant mortality 721 as equal to 360° . The calculation of degrees is shown below:

Cities	Infant Mortality	Angles (in degree)
Mumbai	174	$\frac{174}{721} \times 360 = 86.88^\circ \approx 87^\circ$
Delhi	168	$\frac{168}{721} \times 360 = 83.88 \approx 84^\circ$
Kanpur	220	$\frac{220}{721} \times 360 = 109.84 \approx 110^\circ$
London	93	$\frac{93}{721} \times 360 = 46.43 \approx 46^\circ$
Paris	66	$\frac{66}{721} \times 360 = 32.95 \approx 33^\circ$
Total	721	360°

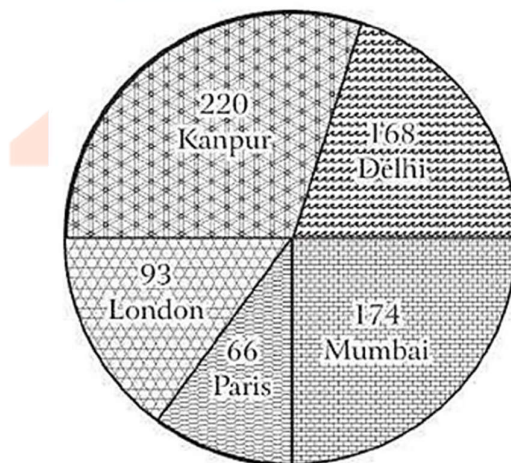


Figure 4.8 Example of pie diagram

4.3.3 Advantages of diagram representation

Diagram representations of data offer several advantages in statistical analysis:

- **Simplifies complex data:** Diagrams condense large amounts of data into visual summaries that are easier to understand.

- **Identifies trends and patterns:** By representing data visually, diagrams help in recognizing trends, patterns, and relationships that might be difficult to discern from raw data.
- **Facilitates comparison:** Diagrams allow easy comparison of different data sets or variables, making it easier to highlight differences and similarities.
- **Aids in decision-making:** Well-constructed diagrams provide valuable insights that can aid in making informed decisions based on data trends.
- **Engages audiences:** Diagrams and charts are more engaging for audiences compared to text-based data, making them effective tools in presentations and reports.

4.4.4 Limitations of diagram representation

Despite their advantages, diagram representations also have some limitations:

- **Simplification can lead to loss of details:** Diagrams often simplify data, which may lead to the omission of important nuances or details.
- **Misleading visuals:** Poorly designed diagrams (e.g., inappropriate scaling, selective data inclusion) can mislead viewers and distort the actual interpretation of the data.
- **Limited for complex data:** Some data, especially very large or multidimensional data sets, may not be adequately represented by simple diagrams.

4.4 GRAPHIC REPRESENTATION OF DATA

Diagrams or geometrical constructions, like bars, rectangles cubes etc., discussed in the previous section give a visual effect. Students of life

sciences and others use it for visual presentation and analysis of the data. In many cases, a graphic presentation of a frequency table gives more concise and clear information about a frequency distribution. Graphic methods involve continuous curves. They are powerful aids for quick visualisation of complex situations. Let us discuss graphs of frequency distribution.

Graphs of Frequency Distribution Frequency distribution can be represented graphically. Such graphs give a better picture than the data when arranged in tabular form. Mainly four types of graphs are used in presenting a frequency distribution.

1. The Histogram
2. The Frequency Polygon
3. The Frequency Curve
4. Cumulative Frequency Curve or Ogive

4.4.1. The Histogram: It consists of rectangles erected with bases equal to class interval of frequency distribution and the heights of rectangles are proportional to the frequencies of the respective classes in such a way that the areas of rectangles are directly proportional to the corresponding frequencies. Each rectangle is joined with the other and thus it looks like a continuous figure. The sum of the frequencies is represented by the total area of the histogram.

Construction of Histogram: Construction of histogram can be divided into two parts.

- (i) When class intervals are equal.
- (ii) When class intervals are not equal.

When Class Intervals are Equal: The class intervals are shown on the x-axis and the frequencies on y-axis. Let us clarify it by the following example

Example 1. Represent the given data by a histogram.

Class	30-35	35-40	40-45	45-50	50-55	55-60	60-65
Frequency	1	5	22	14	11	6	1

Solution.

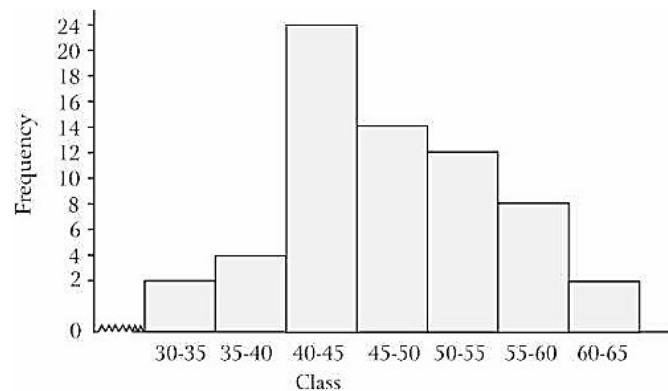


Figure 4.9 Example of histogram I

When Class Intervals are not Equal: In case of unequal class intervals, frequencies are adjusted, otherwise the histogram would give a misleading picture. Thus, before constructing a histogram, all larger classes must be expressed as multiples of smaller classes, and then the corresponding class frequencies are divided by these multiples. The method is illustrated below.

Example 2. Represent the following data by a histogram:

Weekly wages (Rupees)	Number of workers
110-115	6
115-120	18
120-125	25
125-130	15
130-140	12
140-160	12
160-180	8

Solution. The class intervals are unequal and therefore, frequencies are to be adjusted by finding an adjustment factor. Since the lowest class interval is 5, the frequencies of class interval 130-140 shall be divided by 2 because the class interval is double. Similarly, the frequencies of class interval 140-160 and 160-180 shall be divided by 4 each. In other words, the adjustment factor for class interval 110-115, 115-120, 120-125 and 125-130 is 1 each, while for 130-140, it is 2 and for 140-160 and 160-180, it is 4 each. This has been summarised in the following table.

Weekly wages (Rupees)	Number of workers	Adjustment factor $5=1$	Frequency
110-115	6	1	6
115-120	18	1	18
120-125	25	1	25
125-130	15	1	15
130-140	12	2	6
140-160	12	4	3
160-180	8	4	2



Figure 4.10 Example of histogram II

4.4.2. The Frequency Polygon

The histogram can be shown as a frequency polygon. To do it we link the mid points of the class intervals. We can say that a frequency polygon is a line graph of class frequency plotted against mid points (class marks). The end points are extended to join the x-axis at the midpoint of the empty classes on both sides which contain zero frequency. Note also that the area under the frequency polygon is equal to the area of the histogram. Frequency polygon may also be constructed by taking the mid points of the various class intervals and plotting the frequency corresponding to each mid-point and then joining them with straight lines. The figure obtained in this manner would be same as that of obtained by earlier method. But in this case, histograms are not constructed.

Example 3. Draw a frequency polygon from the following data:

Classes	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	6	12	16	18	8	4

Solution. We find the mid value of each class and record them as below:

Classes	Mid-value	Frequency
---------	-----------	-----------

0-10	5	6
10-20	15	12
20-30	25	16
30-40	35	18
40-50	45	8
50-60	55	4

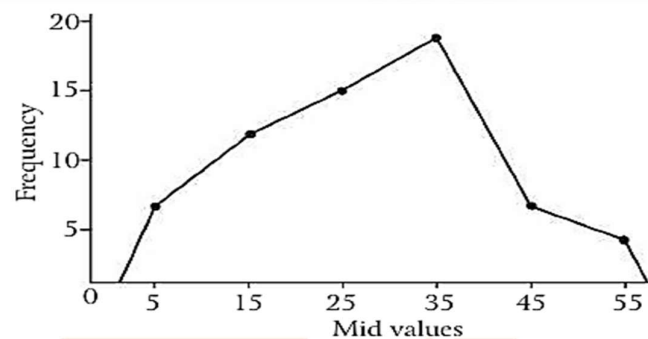


Figure 4.11 Example of frequency polygon

4.4.3. The Frequency Curve

A smooth frequency curve is drawn through the various points of the frequency polygon. This curve is drawn in such a manner that the angularity of the polygon is removed and the area under the curve is approximately equal to the area of the frequency polygon. By smoothing the curve all accidental variations present in the data are eliminated and it gives a regular and continuous curve without distorting the facts. For drawing a correct frequency curve, it is necessary to construct frequency polygon and frequency polygon can be constructed by histogram or without histogram as stated earlier.

Example 4. Draw a histogram, frequency polygon and frequency curve from the following data:

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	4	6	10	20	12	2

Solution.

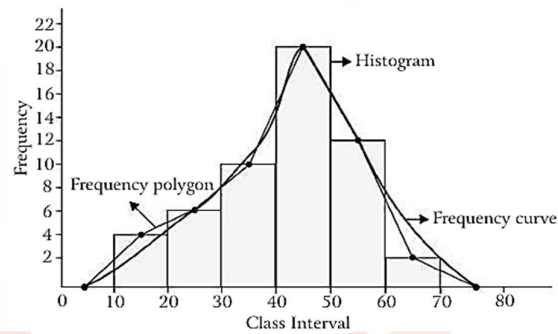


Figure 4.12 Example of frequency curve

4.4.4. Cumulative Frequency Curve or Ogive

A cumulative frequency curve or ogive is a graphical representation of a cumulative frequency distribution. Let us have the following illustration:

Example 5. Draw histogram, frequency polygon and cumulative frequency curve from the following data:

Class	10-15	15-20	20-25	25-30	30-35
Frequency	7	19	27	15	12

Solution.

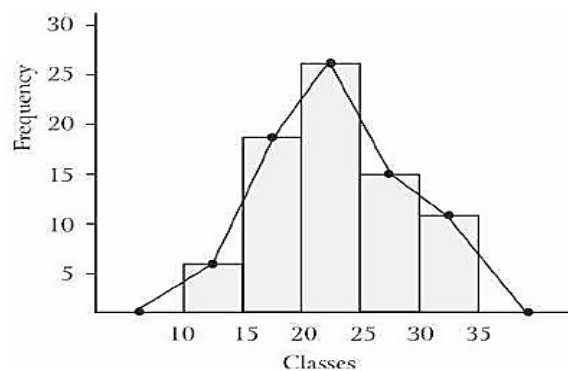


Figure 4.13: Histogram and Frequency Polygon

For cumulative frequency curve we will find the cumulative frequencies as given below:

Class (less than)	Frequency	Cumulative frequency
15	7	7
20	19	26
25	27	53
30	15	68
35	12	80

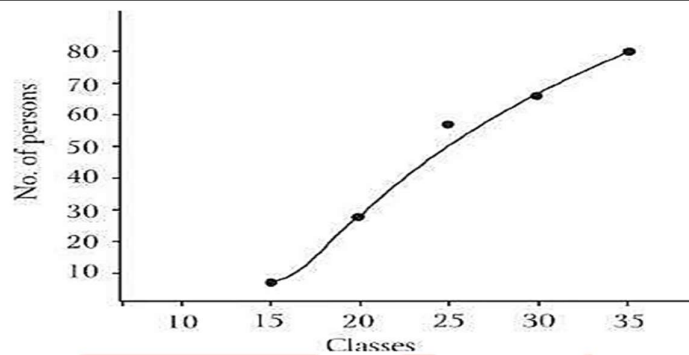


Figure 4.14: Cumulative Frequency Curve (Ogive)

CHECK YOUR PROGRESS

1. Explain the meaning and importance of graphical representation of data.

.....

.....

.....

2. Describe the different types of diagrams used for presenting statistical data.

.....

.....

.....

3. Distinguish between bar diagrams, histograms, frequency polygons, and ogives.

4.5 SUMMARY

- **Diagrammatic representation** serves as a preliminary step towards graphical representation and includes simple visual formats like bar diagrams, pie charts, and pictograms. These are useful for presenting categorical data in an accessible and appealing form.
- While diagrams are easy to construct and interpret, they may lack precision when dealing with large datasets or complex distributions.
- **Graphical representation** is especially valuable for summarizing and analysing frequency distributions. This unit introduced four main graphical tools:
 - **Histograms**, which represent grouped frequency data using adjoining bars.
 - **Frequency polygons**, which connect midpoints of histogram bars with line segments.
 - **Frequency curves**, which provide a smooth outline of the distribution.
 - **Ogives**, or cumulative frequency curves, which help in locating medians, quartiles, and percentiles.

Mastery of these techniques equips learners to visualize statistical data meaningfully and make informed interpretations.

4.6 EXERCISES

1. What do you mean by diagram representation of data? How they are useful?
2. Give the importance and utility of diagram. What are their limitations?
3. Write a note on diagram representation of data. Give important points to be kept in mind while constructing a diagram.
4. Discuss the merits and demerits of diagram representation of statistical data. Explain various bar Diagrams used in statistics.
5. Define the utility of diagrams used for presentation of the data.
6. What do you mean by area diagram? Name the diagram which are commonly used under this head.
7. Discuss the various types of two-dimensional diagrams. Explain with examples, how these are prepared?
8. Describe the square and circle Diagrams and state the method of constructing these Diagrams.
9. Depict the following data by a bar Diagram.

Roll No.:	1	2	3	4	5	6	7	8	9	10
Marks:	15	20	35	17	19	27	23	10	40	30

10. Represent the following data by a sub-divided bar Diagram.

Year	1999	2000	2001	2002	2003
Export (in crores)	110	120	125	135	132
Import (in crores)	95	99	106	111	117

11. Describe graphical representation of data.

12. What is the difference between 'Diagram' and 'Graph'? What is histogram? How will you draw it? Also write whether it is a graph or a diagram?
13. Describe the various modes of graphical representation of a frequency distribution.
14. What do you mean by cumulative frequency? How a cumulative frequency table is formed?
15. What do you mean by frequency curve? Give different type of frequency curves known to you.
16. Explain briefly the various methods that are used for graphical representation of frequency distribution.

4.7 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

BLOCK II: CENTRAL TENDENCY AND DISPERSION MEASURES IN STATISTICS

UNIT 5 MEASURES OF CENTRAL TENDENCY- MEAN MEDIAN AND MODE

STRUCTURE

- 5.1 Introduction
- 5.2 Objectives
- 5.3 Various measures of central tendency
- 5.4 Arithmetic mean
- 5.5 Median
- 5.6 Mode
- 5.7 Summary
- 5.8 EXERCISES
- 5.9 References and Suggested Reading

5.1 INTRODUCTION

The concept of measures of central tendency is foundational in statistics. These measures provide a way to summarize a data set by identifying a single value that represents the "center" or most typical value of the data. The three primary measures of central tendency are **mean**, **median**, and **mode**. Each has unique properties and applications depending on the nature of the data.

5.2 OBJECTIVES

After completing this unit, students will be able to

- Understand the concept of central tendency and why it is important in summarizing data.

- Differentiate between the mean, median, and mode, and understand when to use each measure.
- Calculate the mean, median, and mode for different types of data sets.
- Analyze real-world data and interpret the meaning of these measures within context.

5.3 VARIOUS MEASURES OF CENTRAL TENDENCY

We know that a measure of central tendency is a quantity around which the data tend to cluster. Under different types of situations, data tend to cluster around different quantities. Thus, a single average is not suitable under all types of situations. Consequently, there are five measures of central tendency which are in common use. These are:

- Arithmetic Mean or Simply Mean
- Geometric Mean
- Harmonic Mean
- Median
- Mode

Before discussing these variety of averages, let us, be familiar with basic notations and terminology used. In this unit, we will discuss only Arithmetic Mean together with median and mode. There is a separate unit (Unit 6) for Geometric and Harmonic mean.

Notations and Terminology

Let the symbol ' x_i ' denote the N values $x_1, x_2, \dots, x_N$ assumed by a variable x . The letter ' i ' in x_i is called a 'subscript or index'. The symbol $\sum_{i=1}^N x_i$ is used to denote the sum of all the x_i 's from $i = 1, 2, \dots, N$. i.e., by definition

$$\sum_{i=1}^N x_i = x_1 + x_2 + \cdots + x_N$$

Without any confusion we often denote this sum simple by

$$\sum x, \sum x_i, \sum_i x_i$$

The symbol $\sum$ is the Greek capital letter sigma.

Similarly,

$$1 \quad \sum_{i=1}^N x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N$$

$$2 \quad \sum_{i=1}^N a x_i = a x_1 + a x_2 + \cdots + a x_N = a(x_1 + x_2 + \cdots + x_N) = a \sum_{i=1}^N x_i$$

where 'a' is some constant. More simply,

$$\sum a x = a \sum x$$

3 If a, b and c are any constants

$$\sum (a x + b y + c z) = a \sum x + b \sum y + c \sum z$$

$$4 \quad \sum_{i=1}^N x_i^2 = x_1^2 + x_2^2 + \cdots + x_N^2 = \sum x^2$$

$$5 \quad \sum_{i=1}^N (x_i - a) = \sum_{i=1}^N x_i - \sum_{i=1}^N a = \sum_{i=1}^N x_i - N a$$

Example 1: A variable x assumes the values $x_1 = 2, x_2 = -5, x_3 = 4$ and $x_4 = -8$ find:

- (a) $\sum x$
- (b) $\sum x^2$
- (c) $\sum (x + 6)$
- (d) $\sum (x - 6)^2$.

Solution: Given that $x_1 = 2, x_2 = -5, x_3 = 4$ and $x_4 = -8$

(a)

$$\sum x = x_1 + x_2 + x_3 + x_4 = 2 - 5 + 4 - 8 = -7$$

(b)

$$\sum x^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 4 + 25 + 16 + 64 = 109$$

(c)

$$\sum (x + 6) = \sum x + 4 \times 6 = -7 + 24 = 17$$

(d)

$$\begin{aligned} \sum (x - 6)^2 &= (x_1 - 6)^2 + (x_2 - 6)^2 + (x_3 - 6)^2 + (x_4 - 6)^2 \\ &= (2 - 6)^2 + (-5 - 6)^2 + (4 - 6)^2 + (-8 - 6)^2 \\ &= 16 + 121 + 4 + 196 = 337 \end{aligned}$$

Example 2: If $\sum_{i=1}^5 x_i = -4$ and $\sum_{i=1}^5 x_i^2 = 10$

find (a) $\sum_{i=1}^5 (2x_i + 5)$

(b) $\sum_{i=1}^5 x_i(x_i - 1)$

Solution: (a) $\sum_{i=1}^5 (2x_i + 5) = \sum_{i=1}^5 2x_i + \sum_{i=1}^5 5$

$$\begin{aligned} &= 2 \sum x_i + 5 \times 5 \\ &= 2(-4) + 25 = 17 \end{aligned}$$

$$\begin{aligned} \text{(b) } \sum_{i=1}^5 x_i(x_i - 1) &= \sum_{i=1}^5 x_i^2 - \sum_{i=1}^5 x_i \\ &= 10 - (-4) = 14 \end{aligned}$$

5.4 ARITHMETIC MEAN

It is the most popular and important measure of central tendency. In daily life it is called average while in statistics it is called arithmetic mean or

simply mean. It is defined as the sum of the items divided by the number of items. Thus, for the case of N values $x_1, x_2, \dots, x_N$ of the variate x , the arithmetic means usually denoted by $\bar{x}$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

If the variate x takes the values $x_1, x_2, \dots, x_N$, with frequencies $f_1, f_2, \dots, f_N$ respectively. That is x takes the value x_i with frequency f_i ($i = 1, 2, \dots, N$) then the arithmetic mean, $\bar{x}$ is given by

$$\begin{aligned} \bar{x} &= \frac{f_1 x_1 + f_2 x_2 + \dots + f_N x_N}{f_1 + f_2 + \dots + f_N} = \frac{\sum_{i=1}^N f_i x_i}{\sum_{i=1}^N f_i} \\ &= \frac{1}{N} \sum_{i=1}^N f_i x_i, \text{ where } N = \sum_{i=1}^N f_i \end{aligned}$$

Example 3: Find arithmetic mean from the following discrete series:

$x:$	9	8	7	6	5	4	3	2	1	0
$f:$	2	3	4	6	11	10	6	5	2	1

Solution: The given data is in the form of frequency distribution.

Hence,

$$\begin{aligned} \bar{x} &= \frac{x_1 f_1 + x_2 f_2 + \dots + x_{10} f_{10}}{f_1 + f_2 + \dots + f_{10}} \\ &= \frac{(9 \times 2) + (8 \times 3) + (7 \times 4) + (6 \times 6) + (5 \times 11) + (4 \times 10)}{2 + 3 + 4 + 6 + 11 + 10 + 6 + 5 + 2 + 1} \\ &= \frac{18 + 24 + 28 + 36 + 55 + 40 + 18 + 10 + 2}{50} = \frac{231}{50} = 4.62 \end{aligned}$$

Example 4: Find the arithmetic mean of the first N natural numbers.

Solution: The first N natural numbers are $1, 2, 3, \dots, N$. Then arithmetic mean of these numbers is

$$\begin{aligned}\bar{x} &= \frac{1 + 2 + 3 + \dots + N}{N} = \frac{N(N + 1)}{2 \times N} \\ &\left(\text{Because sum of } N \text{ natural number is } \frac{N(N + 1)}{2} \right) \\ &= \frac{(N + 1)}{2}\end{aligned}$$

Short-cut Method for Finding the Arithmetic Mean

The previous method for finding mean is usually known as Direct Method which is lengthy and time consuming, if data are large. In such a situation, short-cut method is used in which the origin is shifted to some point A [$(x - A)$ is also called change of origin]. Let $x_1, x_2, \dots, x_N$ be the values of the variate with frequencies $f_1, f_2, \dots, f_N$ respectively. Let us denote ' d ' as deviation of x_i from a point A so that

$$d_i = x_i - A \text{ or } x_i = d_i + A$$

Now by definition

Mean

$$\begin{aligned}(\bar{x}) &= \frac{\sum_{i=1}^N f_i x_i}{\sum_{i=1}^N f_i} = \frac{\sum_{i=1}^N f_i (d_i + A)}{\sum_{i=1}^N f_i} \\ &= \frac{\sum_{i=1}^N f_i d_i + A \sum_{i=1}^N f_i}{\sum_{i=1}^N f_i} = \frac{\sum_{i=1}^N f_i d_i}{\sum_{i=1}^N f_i} + A \\ &= A + \frac{\sum_{i=1}^N f_i d_i}{N} \\ &= \text{Assumed value} + \text{Mean of the new variate ' } d\end{aligned}$$

Sometimes in case of data with equal class interval, the calculation can be further reduced by dividing these deviations by a common quantity '

h ' called class interval $\left(\frac{x-A}{h} \text{ or } \frac{x}{h} \text{ is called change of scale}\right)$, then we define

$$u_i = \frac{x_i - A}{h} = \frac{d_i}{h} \text{ i.e., } d_i = hu_i$$

Substituting the value of d_i in (1), we have

$$\begin{aligned}\bar{x} &= A + \frac{\sum_{i=1}^N f_i hu_i}{N} \\ &= \text{Assumed value} + h \text{ times mean of new variate 'u' }\end{aligned}$$

(1) and (2) are called the short-cut formulae for the determination of mean.

Example 5: Calculate the mean from the following data:

Marks below	No. of students
10	15
20	35
30	60
40	84
50	96
60	127
70	198
80	250

Solution: As data are given in the form of cumulative frequency distribution, we find the (ordinary) frequency distribution as below:

Marks	Frequency
0 – 10	15
10 – 20	$35 - 15 = 20$

20 – 30	60 – 35 = 25
30 – 40	84 – 60 = 24
40 – 50	96 – 84 = 12
50 – 60	127 – 96 = 31
60 – 70	198 – 127 = 71
70 – 80	250 – 198 = 52

Now the arithmetic mean is calculated as below:

Marks	Mid values (x)	Frequency (f)	u $= \frac{x - 35}{10}$	fu
0 – 10	5	15	-3	-45
10 – 20	15	20	-2	-40
20 – 30	25	25	-1	-25
30 – 40	35	24	0	0
40 – 50	45	12	+1	+12
50 – 60	55	31	+2	+62
60 – 70	65	71	+3	+213
70 – 80	75	52	+4	+208
		$N = 250$		$\sum fu$ $= 385$

$$\begin{aligned}\bar{x} &= A + \frac{\sum fu}{N} \times h = 35 + \frac{385}{250} \times 10 \\ &= 35 + 15.4 = 50.4\end{aligned}$$

Example 6: Find the A.M from the following data:

Class	100	110	120	130	140	150	160	170
Interval:	– 110	– 120	– 130	– 140	– 150	– 160	– 170	– 180
Frequen cy:	4	6	20	32	33	17	8	2

Solution:

Class Interval	Mid values (x)	Frequency (f)	u $= \frac{x - 145}{10}$	f_u
100 – 110	105	4	-4	-16
110 – 120	115	6	-3	-18
120 – 130	125	20	-2	-40
130 – 140	135	32	-1	-32
140 – 150	145	33	0	0
150 – 160	155	17	+1	+17
160 – 170	165	8	+2	+16
170 – 180	175	2	+3	+6
		$N = 112$		$\sum f_u$ $= -67$

$$\begin{aligned}
 x &= A + \frac{\sum fu}{N} \times h = 145 + \frac{(-67) \times 10}{112} \\
 &= 145 - \frac{670}{112} = 145 - 5.98 \\
 &= 139.02
 \end{aligned}$$

Properties of Arithmetic Mean

The arithmetic mean has the following important properties.

Property 1: For any distribution, the sum of the deviations from their arithmetic mean is always zero.

Proof: Let the variate x takes the values $x_1, x_2, \dots, x_i, \dots, x_N, (i = 1, 2, \dots, N)$

Let $\bar{x}$ denotes the arithmetic mean of the variate x .

Then $d_1 = x_1 - \bar{x}, d_2 = x_2 - \bar{x}, \dots, d_N = x_N - \bar{x}$ be the deviations of $x_1, x_2, \dots, x_N$ from their mean $\bar{x}$. Then,
sum of deviations

$$\begin{aligned} &= \sum_{i=1}^N d_i = \sum_{i=1}^N (x_i - \bar{x}) \\ &= \sum_{i=1}^N x_i - N\bar{x} \\ &= \sum_{i=1}^N x_i - N \left(\frac{\sum_{i=1}^N x_i}{N} \right) \\ &= \sum_{i=1}^N x_i - \sum_{i=1}^N x_i \\ &= 0 \end{aligned}$$

Property 2: If there be k distributions with frequencies $n_1, n_2, \dots, n_k$ and the respective arithmetic means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$, then the arithmetic mean $\bar{\bar{x}}$, of the distribution, obtained by combining the k distribution is given by,

$$\begin{aligned} \bar{\bar{x}} &= \frac{1}{N} \sum_{i=1}^k n_i \bar{x}_i, N = n_1 + n_2 + \dots + n_k \\ &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k} \end{aligned}$$

Proof: The first distribution has n_1 items with mean $\bar{x}_1$. Thus, the sum of items in the first distribution is $n_1\bar{x}_1$. Similarly, the sum of items of the second distribution is $n_2\bar{x}_2$..., the sum of items of k -th distribution is $n_k\bar{x}_k$.

Also, the total frequency of the combined distribution is

$$\begin{aligned} N &= n_1 + n_2 + \cdots n_k \\ \bar{\bar{x}} &= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{N} \\ &= \frac{1}{N} \sum_{i=1}^k n_i\bar{x}_i \end{aligned}$$

Property 3: If each value of the variate x is increased or decreased by a constant value then the arithmetic mean of the variate so obtained is also increased or decreased by the same constant value.

Proof: Let x is the variate under study. Let us form a new variate y as follows:

$y = x \pm a$, where ' a ' is some constant.

Applying Σ on both sides,

$$\begin{aligned} \sum_{i=1}^N y &= \sum_{i=1}^N (x \pm a) \\ &= \sum_{i=1}^N x \pm \sum_{i=1}^N a \\ &= \sum_{i=1}^N x \pm Na \end{aligned}$$

Dividing both sides by N ,

$$\begin{aligned} \frac{\sum_{j=1}^N y}{N} &= \frac{\sum_{i=1}^N x}{N} \pm a \\ \bar{y} &= \bar{x} \pm a \end{aligned}$$

Property 4: If a constant value is used to multiply or divide the variate's values, the resulting arithmetic mean is equal to the original arithmetic mean multiplied or divided by the constant value.

Proof: Let $y = kx$, where k is some constant.

Applying Σ on both sides,

$$\sum_{i=1}^N y = \sum_{i=1}^N kx$$

Dividing both sides by N ,

$$\frac{\sum_{i=1}^N y}{N} = \frac{k \sum_{i=1}^N x}{N}$$

$$\bar{y} = k\bar{x}$$

Property 5: When deviations are measured from the mean, the total of the squared deviations of a set of values is minimized.

Proof: Let us consider the frequency distribution

$$x_i, (i = 1, 2, \dots, N) \text{ and } f_i, (i = 1, 2, \dots, N)$$

Let

$$Z = \sum_{i=1}^N f_i(x_i - A)^2$$

be the sum of squares of the deviations of given values from any arbitrary point 'A'. We are to prove that Z is minimum when $A = \bar{x}$.

Z will be minimum for variation is A if

$$\begin{aligned} \text{Now, } \frac{\partial Z}{\partial A} &= -2 \\ \Rightarrow \sum_{i=1}^N f_i(x_i - A) &= 0 \end{aligned}$$

$$\frac{\partial Z}{\partial A} = 0 \text{ and } \frac{\partial^2 Z}{\partial A^2} > 0$$

$$\Rightarrow \sum_{i=1}^N f_i x_i - A \sum_{i=1}^N f_i = 0$$

or

$$A = \frac{\sum_{i=1}^N f_i x_i}{N} = \bar{x} \left(\text{because } \sum_{i=1}^N f_i = N \right)$$

$$\text{Again } \frac{\partial^2 Z}{\partial A^2} = -2 \sum_{i=1}^N f_i (-1)$$

Hence, Z is minimum at the point $A = \bar{x}$. Hence the result.

Example 7:(a) Three teachers of statistics reported mean marks of their classes consisting of 69, 64 and 71 students as 30, 26 and 18. Determine the mean marks for all three classes.

(b) Find the combined mean of the distributions in which: $N_1 = 40, N_2 = 50, N_3 = 60, \bar{x}_1 = 30, \bar{x}_2 = 40, \bar{x}_3 = 30$.

Solution: (a) We are given that

$$\begin{aligned} N_1 &= 69, & \bar{x}_1 &= 30 \\ N_2 &= 64, & \bar{x}_2 &= 26 \\ N_3 &= 71, & \bar{x}_3 &= 18 \end{aligned}$$

Combined mean is given by

$$\begin{aligned} \bar{\bar{x}} &= \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3}{N_1 + N_2 + N_3} = \frac{69 \times 30 + 64 \times 26 + 71 \times 18}{69 + 64 + 71} \\ &= \frac{2070 + 1664 + 1278}{204} = \frac{5012}{204} = 24.57 \end{aligned}$$

(b) Combined mean for three distributions is

$$\begin{aligned} \bar{\bar{x}} &= \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3}{N_1 + N_2 + N_3} = \frac{40 \times 30 + 50 \times 40 + 60 \times 30}{40 + 50 + 60} \\ &= \frac{1200 + 2000 + 1500}{150} = \frac{5000}{150} = 33.3 \end{aligned}$$

Example 8: The mean marks of 80 students was found to be 50. Later on, it was detected that a score of 55 is misread as 75. Find the correct mean of 80 students.

Solution: We are given that $N = 80, \bar{x} = 50$

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} \text{ or } \sum_{i=1}^N x_i = N\bar{x} = 80 \times 50 = 4000$$

But $\sum_{i=1}^{80} x_i = 4000$ is not correct as it includes a figure 75 which is actually 55. So we will calculate corrected $\sum_{i=1}^{80} x_i$ i.e.

$$\begin{aligned} \text{Corrected } \sum_{i=1}^{80} x_i &= \text{Incorrect } \sum_{i=1}^{80} x_i - 75 + 55 \\ &= 4000 - 75 + 55 \\ &= 3980 \\ \text{Mean } \bar{x} &= \frac{3980}{80} = 49.75 \end{aligned}$$

$\therefore$ Correct Mean

Example 9: The following frequency distribution has an arithmetic mean of 67.45. Determine the frequency that is unknown.

Classes:	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
Frequency:	15	54	?	81	27

Solution: Let the unknown frequency is 'A'.

Classes	Mid values (x)	Frequency (f)	fx
60 – 62	61	15	915
63 – 65	64	54	3456
66 – 68	67	A	67 A
69 – 71	70	81	5670
72 – 74	73	27	1971
		$\sum f = 177 + A$	$\sum fx = 12012 + 67 A$

Mean,

$$\bar{x} = \frac{\sum fx}{N} = \frac{12012 + 67A}{177 + A}$$

It is given that mean of the distribution is 67.45. Thus,

$$\begin{aligned} 67.45 &= \frac{12012 + 67A}{177 + A} \\ (177 + A)67.45 &= 12012 + 67A \\ 67.45A - 67A &= 12012 - 11938.65 \\ 0.45A &= 73.55 \\ A &= \frac{73.55}{0.45} = 163 \end{aligned}$$

Hence, unknown frequency corresponding to the class 66-68 is 163.

Example 10: Find the A.M. of the numbers 1,2,4,8, 16....

Solution: The given series is a G. P. with first term $a = 1$ and common ratio $r = 2$. Thus, the sum of $(n + 1)$ terms is

$$\frac{a(r^{n+1} - 1)}{r - 1} = \frac{1(2^{n+1} - 1)}{2 - 1} = (2^{n+1} - 1)$$

Hence, the mean of the series = $\frac{\text{Sum of } (n+1) \text{ terms}}{(n+1)}$

$$= \frac{(2^{n+1} - 1)}{(n + 1)}.$$

Example 11: Show that the A.M. of n numbers of an A.P. is equal to the mean of its initial and end term.

Solution: Consider the ' n ' terms of an A.P. are

$$a, a + d, a + 2d \dots a + (n - 1)d$$

where ' a ' is the first term and ' d ' is the common difference.

$$\begin{aligned}
 \text{Sum of } n \text{ term of an A.P.} &= \frac{n}{2} [2a + (n-1)d] \\
 \text{Mean of } n \text{ terms of an A.P.} &= \frac{\text{Sum of } n' \text{ terms of an A.P.}}{n} \\
 &= \frac{n}{2 \times n} [2a + (n-1)d] \\
 &= \frac{1}{2} [2a + (n-1)d] \\
 &= \frac{1}{2} [a + \{a + (n-1)d\}] \\
 &= \frac{[\text{First term} + \text{Last term}]}{2} \\
 &= \text{Mean of first and last term.}
 \end{aligned}$$

Weighted Arithmetic Mean

The main limitations of the A.M. is that it gives equal importance (weights) to all the objects in the series. But there are cases where the relative importance of all the items is not equal. In such a situation, weighted arithmetic mean is used. Let the value x_i of the variate x have the weight $w_i (i = 1, 2, \dots, N)$. Then

$$\text{Weighted arithmetic mean } \bar{x}_w = \frac{1}{N} \sum_{i=1}^N w_i x_i$$

where

$$N = w_1 + w_2 + \dots + w_N$$

Example 12: Show that the weighted arithmetic means of first n natural numbers whose weights are equal to the corresponding numbers is equal to $\frac{1}{3}(2n+1)$.

Solution: The first ' n ' natural numbers are

$$1, 2, 3, \dots, n$$

and their corresponding weights are

$$\therefore \text{Weighted mean } \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} = \frac{1 \times 1 + 2, 3, \dots, n}{1 + 2 + 3 + \dots + n}$$

$$\begin{aligned}
 &= \frac{1^2 + 2^2 + 3^2 + \dots + n^2}{1 + 2 + 3 + \dots + n} \\
 &= \frac{\frac{1}{6}n(n+1)(2n+1)}{\frac{1}{2}n(n+1)} \\
 &= \frac{1}{3}(2n+1)
 \end{aligned}$$

Advantages of the Arithmetic Mean

1. It is straightforward and easy to compute.
2. It is widely recognized and frequently used in practice.
3. It takes every data point into account, making it comprehensive.
4. It is suitable for further mathematical and algebraic analysis.
5. It serves as a reliable basis for comparing different data sets.

Limitations of the Arithmetic Mean

1. It is sensitive to outliers and extreme values, which can distort the result.
2. It cannot be identified just by looking at the data.
3. It may not correspond to an actual value in the data set.
4. It cannot be determined if any data point is missing.

5.5 MEDIAN

The **arithmetic mean**, as previously discussed, is calculated through mathematical methods and applies to variables that can be directly measured—such as height, weight, and similar quantities.

In contrast, the **median** represents the central or middle value in a dataset, especially useful when dealing with characteristics that cannot

be directly measured—like honesty, blindness, or intelligence. In such situations, the central tendency is identified by locating the midpoint of the data, making the median a type of **positional average**.

Partition values and **mode** also fall under this category of positional averages.

As defined by Bowley: *“If the members of a group are arranged in order according to the characteristic being measured, the value corresponding to the middle position is the median.”* In other words, the **median** is the value that divides the dataset into two equal halves—one half of the values lie below it, and the other half lie above—when the data is sorted in ascending order.

Calculation of Median

Depending upon the nature of the series median is calculated by different methods.

- 1 Median for individual series: For individual series the values are arranged in increasing or decreasing order of magnitude. If the frequency (the no. of units in the series) is odd and equal to N , say, then $\left(\frac{N+1}{2}\right)$ -th item gives the median. If frequency is even, say, $2N$, then the median is the A.M. of the N -th and $(N + 1)$ -th items.
- 2 Median in case of grouped data: For a grouped frequency distribution, the median is calculated by the following formula:

$$\text{Median} = L_1 + \frac{i}{f} \left(\frac{N}{2} - C \right)$$

where

L_1 = Lower limit of the median class (in which median lies)

i = width of the median class

f = frequency of the median class

C = cumulative frequency of the class preceding the median class

N = Total frequency.

Example 13: (i) Find out the median of the marks obtained by a batch of 9 students in a class test.

Nos.:	1	2	3	4	5	6	7	8	9
Marks:	33	32	55	47	21	50	27	12	48

(ii) If an additional student scored marks 67, then find the median.

Solution: (i) We arrange the marks of 9 students in ascending order of magnitude as given below:

21, 27, 32, 33, 47, 48, 55

Now, Median = $\left(\frac{N+1}{2}\right)$ -th item = $\left(\frac{9+1}{2}\right)$ -th item = 5-th item.

Thus, median = 5 – th item i.e., 33.

(ii) The additional student i.e., 10 -th student scored 67 marks then median will be $\left(\frac{N+1}{2}\right)$ -th item = $\left(\frac{10+1}{2}\right)$ -th item = 5.5-th item.

Thus, median is $\frac{(5\text{-th}+6\text{-th})\text{item}}{2} = \frac{33+47}{2} = 40$

Example 14: Find out the median of wages received by the staff of a certain company.

Daily wages (₹):	30	50	70	90	110	130	150	170	190
Frequency:	2	10	12	15	20	13	12	10	4

Solution: Given series is a discrete series and in this case C.M. are also calculated as follows.

Daily wages (₹)	Frequency	Cumulative frequency
30	2	2
50	10	12
70	12	24

90	15	39
110	20	59
130	13	72
150	12	84
170	10	94
190	4	98

Thus, $N = 98$

Median = $\left(\frac{N+1}{2}\right)$ -th item = $\left(\frac{98+1}{2}\right)$ -th item = 49.5-th item

Now we observe that median number 49.5 lies in the cumulative frequency 59 which belongs to the value of $x = 110$.

Thus, Median = ₹ 110

Example 15: Obtain the median from the following data of weights of children in a particular locality.

Weights:	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24
No. of children:	3	9	18	20	16	7

Solution: We form the following table.

Weights	No. of children	Cumulative frequency
0 – 4	3	3
4 – 8	9	12
8 – 12	18	30
12 – 16	20	50
16 – 20	16	66
20 – 24	7	73

$$N = 73$$

Median number = $\left(\frac{N}{2}\right)$ -th item = $\left(\frac{73}{2}\right)$ -th = 36.5-th item. Here 36.5-th item lies in the cumulative frequency 50, which is corresponding to the class 12-16. Thus, 12-16 is the median class in which

$$L_1 = 12, i = 4, f = 20, c = 30$$

Using the formula

$$\begin{aligned} \text{Median} &= L_1 + \frac{i}{f} \left(\frac{N}{2} - c \right) \\ &= 12 + \frac{4}{20} (36.5 - 30) \\ &= 12 + \frac{1}{5} \times 6.5 = 12 + 1.3 = 13.3 \end{aligned}$$

Example 16: Compute the median from the following data:

Mid-values:	115	125	135	145	155	165	175	185	195
Frequency:	6	25	48	72	116	60	38	22	3

Solution: In this question, mid-values are given in which difference between two consecutive mid-values is 10. Thus, we reconstruct the following table with class intervals.

Classes	Frequency	$c.f$	Classes	Frequency	$c.f$
110 – 120	6	6	150 – 160	116	267
120 – 130	25	31	160 – 170	60	327
130 – 140	48	79	170 – 180	38	365
140 – 150	72	151	180 – 190	22	387
			190 – 200	3	390

$$N = 390$$

$$\text{Median No.} = \left(\frac{390}{2}\right) = 195$$

Thus, median class is 150-160 in which

$$L_1 = 150, i = 10, f = 116, c = 151$$

$$\text{Medina} = L_1 + \frac{i}{f} \left(\frac{N}{2} - c \right)$$

$$= 150 + \frac{10}{116} (195 - 151)$$

$$= 150 + \frac{10}{116} \times 44 = 150 + \frac{440}{116}$$

$$= 150 + 3.79 = 153.79$$

Properties of Median

Property 1: It is a positional average.

Property 2: The sum of the deviations of the items from median ignoring $\pm$ signs is least. That is,

$$\Sigma |x - \text{median}| < \Sigma |x - \text{any value}|$$

Merits of the Median

- 1 It is easy to understand.
- 2 It is well defined and can be easily found out.
- 3 Extreme values are eliminated.
- 4 The position of the median is based on all observations.
- 5 It can be located graphically.
- 6 It can be useful in open end distributions.

Demerits of the Median

- 1 It is not a calculated figure obtained by mathematical formula.
- 2 The data has to be arranged in order of magnitude.
- 3 When the number of items is small it is erratic.

- 4 It is not capable of further algebraic treatment.
- 5 In continuous series, interpolation formula is used to calculate median.

5.6 MODE

The **mode** is the value or category that occurs most frequently in a data set. Unlike other measures of central tendency, such as the mean and median, the mode can be used for both numerical and categorical data.

The most important point regarding the mode is that it is not affected by extreme items and open-ended classes. It can only be calculated from a grouped data in which there is only one point of concentration. A distribution may have two modes. If it is so, it is called bimodal distribution.

Calculation of Mode

Mode is determined by different methods, depending upon the nature of the series.

- 1 Mode in case of individual series: In individual series we guess mode by inspection. We observe that term in the series which occurs maximum number of times. This term is called mode. For the sake of convenience, we arrange the terms.

Example 17: (i) Find out mode from the following data of sizes of shoes sold at a shop in one day.

5,10,7,8,7,4,7,5,3,7,2,6

(ii) Find out the mode from the following data:

Wages (₹):	200	250	275	350	325	225
Frequency:	4	6	10	18	9	3

Solution: (i) For the sake of convenience, we arrange the data in a array and observe which term occurs most frequently

2, 3, 4, 5, 5, 6, 7, 7, 7, 7, 8, 10

It is clear that term 7 occurs maximum number of times i.e., 4 times. Thus, mode is 7.

(ii) In this question maximum frequency is 18 and it belongs to the value 350. Thus, ₹350 is the modal wage.

2. Mode in grouped data: For a grouped frequency distribution, the mode is calculated by the following formula:

$$\text{Mode} = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where,

L_1 = Lower limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

i = width of the modal class.

Example 18: Find the mode from the following data:

Class:	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
Frequency:	29	195	241	117	52	10	6	3	2

Solution: In the above example, the highest frequency lies in the class 10-15. Hence, 10 – 15 is the modal class. For the determination of mode we use the formula

$$\text{Mode} = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where, $L_1 = 10, f_1 = 241, f_2 = 117, f_0 = 195$ and $i = 5$

$$\begin{aligned} \therefore \text{Mode} &= 10 + \frac{241 - 195}{2 \times 241 - 195 - 117} \times 5 \\ &= 10 + \frac{46 \times 5}{482 - 312} = 10 + \frac{230}{170} \\ &= 10 + 1.35 = 11.35 \end{aligned}$$

Example 19: Find out the mode from the following data:

Classes:	0	10	20	40	50	70
	– 10	– 20	– 40	– 50	– 70	– 80
Frequency:	5	15	40	32	28	5

Solution: In the above distribution, class intervals are unequal and they are to be converted into equal class intervals. If we add some classes then only two classes namely 0 – 40 and 40 – 80 will be formed which is not justified. Hence, in this case we will break the classes each of size 10. Frequencies are accordingly divided.

Classes	Frequency
0 – 10	5
10 – 20	15
20 – 30	20
30 – 40	20
40 – 50	32
50 – 60	14
60 – 70	14

70 – 80	5
---------	---

Here maximum frequency is 32 which belongs to 40 – 50 class.

Hence, 40 – 50 is the modal class in which

$$L_1 = 40, f_1 = 32, f_0 = 20, f_2 = 14, i = 10$$

Using the formula,

$$\begin{aligned} \text{Mode} &= L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \\ &= 40 + \frac{32 - 20}{2 \times 32 - 20 - 14} \times 10 = 40 + \frac{120}{30} \\ &= 40 + 4 = 44 \end{aligned}$$

Example 20: The following table shows the distribution of 100 families according to their expenditure per week. Number of families corresponding to expenditure group 10-20 and 30-40 are missing from the table. The median and mode are given to be 25 and 24 respectively. Calculate the missing frequencies and then arithmetic mean of the data.

Expenditure	No. of Families
0 – 10	14
10 – 20	?
20 – 30	27
30 – 40	?
40 – 50	15

Solution: We are given:

$$\text{Median} = 25, \text{Mode} = 24 \text{ and } N = 100$$

Let the missing frequencies are x and y respectively for classes 10-20 and 30 – 40. We know that the values of median and mode always lie in their median class and modal class. Thus, if median is 25 it means median class is 20-30.

Let us form the table for cumulative frequencies

Expenditure	No. of families	<i>c.f.</i>
0 – 10	14	14
10 – 20	x	$14 + x$
20 – 30	27	$41 + x$
30 – 40	y	$41 + x + y$
40 – 50	15	$56 + x + y$
	$N = 100$	

Applying the formula for median

$$\begin{aligned}\text{Median} &= 4 + \frac{i}{f} \left(\frac{N}{2} - c \right) \\ 25 &= 20 + \frac{10}{27} \left[\frac{100}{2} - (14 + x) \right]\end{aligned}$$

$$\begin{aligned}5 &= \frac{10}{27} (50 - 14 - x) \\ 135 &= 500 - 140 - 10x \\ 10x &= 500 - 275 \\ 10x &= 225 \\ \therefore x &= \frac{225}{10} = 22.5 = 23\end{aligned}$$

Similarly, mode is given to be 24, which must lie in the class 20-30.

Applying the formula of mode

$$\begin{aligned}
 \text{Mode} &= 4 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \\
 24 &= 20 + \frac{27 - x}{2 \times 27 - x - y} \times 10 \\
 4 &= \frac{(27 - x)}{(54 - x - y)} \times 10 \text{ (as } x = 23 \text{)} \\
 4 &= \frac{(27 - 23)}{(54 - 23 - y)} \times 10 \\
 4 &= \frac{40}{(31 - y)} \\
 4 \times (31 - y) &= 40 \\
 124 - 4y &= 40 \\
 4y &= 124 - 40 \\
 y &= \frac{84}{4} = 21
 \end{aligned}$$

Thus, the distribution becomes like this

Expenditure:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of families:	14	23	27	21	15

Let us find the arithmetic mean of the above distribution.

Expenditure	No. of families (<i>f</i>)	M.V (<i>x</i>)	<i>u</i> $= \frac{x - 25}{10}$	<i>fu</i>
0 – 10	14	5	-2	-28
10 – 20	23	15	-1	-23
20 – 30	27	25	0	0
30 – 40	21	35	+1	+21
40 – 50	15	45	+2	+30
	N = 100			Σfu = 0

$$\begin{aligned}
 A \cdot M &= A + \frac{\sum fu}{N} \times i \\
 &= 25 + 0 \\
 &= 25
 \end{aligned}$$

3 Mode by graphic method or by histogram: Like median, mode can also be located graphically for which we use histogram.

The :

- (i) The histogram of the given data are prepared and the rectangle with maximum height is considered as modal class.
- (ii) Draw a line joining the top right point of the modal class rectangle to the top right of the preceding rectangle.
- (iii) Draw another line joining the top left point of the modal class rectangle to the top left point of the succeeding rectangle.
- (iv) The point of intersection of these lines are marked and a perpendicular is drawn from this point to x -axis. The point of meeting of this perpendicular at x -axis gives the mode.

Example 21: Locate mode graphically from the data given below and verify the result.

Class:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequency:	4	8	16	12	2

Determination of mode from mean and median: In the chapter of graphic representation of data we have studied various types of frequency distributions. The main distributions are symmetrical and asymmetrical distributions. In symmetrical distribution, mean, median and mode coincide while in moderately asymmetrical distribution they will not coincide but will pull apart. The same is given below by the figures.

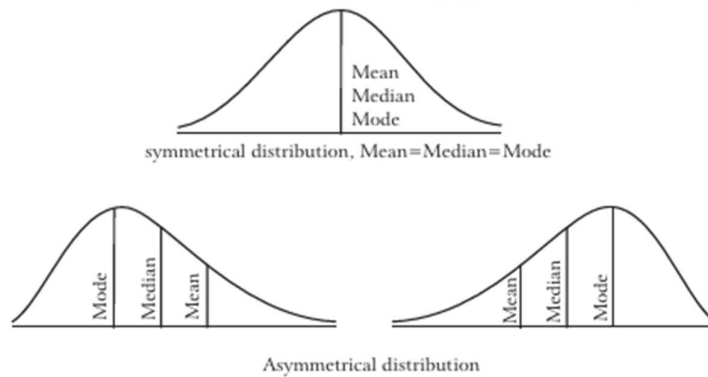


Figure 5.1 Symmetric and asymmetric distribution

In a **negatively skewed distribution**, the mean generally has the lowest value, followed by the median, while the mode tends to have the highest value. Conversely, in a **positively skewed distribution**, the mean usually has the highest value, followed by the median, whereas the mode has the lowest value.

For a moderately skewed distribution, an empirical relationship exists among the mean, median and mode. This relationship can be expressed as:

$$\text{Mean} - \text{Mode} = 3 (\text{Mean} - \text{Median})$$

From this empirical formula, the following alternative expressions can be derived:

$$\begin{aligned} \text{Mode} &= 3(\text{Median}) - 2(\text{Mean}) \\ \text{Median} &= \frac{2(\text{Mean}) + \text{Mode}}{3} \\ \text{Mean} &= \frac{3(\text{Median}) - \text{Mode}}{2} \end{aligned}$$

Example 22

- (a) Estimate the mode when the mean is 12 and the median is 16.5.
- (b) If the mean is 26 and the mode is 28.5, estimate the median.

Solution:

(a) Using

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

$$\begin{aligned}
 &= 3(16.5) - 2(12) \\
 &= 49.5 - 24 = 25.5
 \end{aligned}$$

Therefore, the estimated **mode** = **25.5**.

(b) Using

$$\begin{aligned}
 \text{Median} &= \frac{2(\text{Mean}) + \text{Mode}}{3} \\
 &= \frac{2(26) + 28.5}{3} \\
 &= \frac{52 + 28.5}{3} = \frac{80.5}{3} = 26.83
 \end{aligned}$$

Therefore, the estimated **median** = **26.83**.

Merits of Mode

1. **Simple to understand:** Mode is relatively easy to identify and interpret.
2. **Less influenced by extreme observations:** Unlike the arithmetic mean, mode is generally unaffected by exceptionally high or low values.
3. **Can be determined graphically:** In suitable frequency distributions, mode can be estimated with the help of a graph.
4. **Represents the most frequent observation:** Since it identifies the value occurring most frequently, it can provide a useful measure of the typical value.
5. **Useful for qualitative data:** Mode is particularly appropriate when the variable represents categories or attributes, such as colour, size, design or preferences.

Limitations of Mode

1. **May not always be uniquely identifiable:** Some distributions do not have a clearly defined mode. In bimodal or multimodal distributions, more than one value may have the highest frequency.
2. **Does not consider every observation:** The calculation of mode depends only on the frequency pattern and does not incorporate all observations in the dataset.
3. **Not suitable for algebraic manipulation:** Unlike the mean, mode cannot generally be subjected to further mathematical or algebraic operations.

4. **Requires appropriate grouping:** In grouped frequency distributions, the identification of the modal class may depend on suitable class intervals, which can sometimes affect the result.

CHECK YOUR PROGRESS

1. What is meant by measures of central tendency? Explain their significance.

.....
.....
.....

2. Describe the procedures for calculating Arithmetic Mean, Median and Mode.

.....
.....
.....

3. Distinguish between Mean, Median and Mode. Discuss the major advantages and limitations of each measure.

.....
.....
.....

5.7 SUMMARY

This unit introduced the principal measures used to represent the central or typical value of a dataset. The major concepts covered in the unit include:

- The meaning and significance of **measures of central tendency**.
- The concept and methods of calculating the **Arithmetic Mean**.
- The concept and procedures for determining the **Median**.
- The definition and methods of obtaining the **Mode**.
- The comparative characteristics, advantages and limitations of Mean, Median and Mode.
- The empirical relationship among Mean, Median and Mode for moderately skewed distributions.

5.8 EXERCISES

1. Which measure of central tendency is most sensitive to extreme observations or outliers?

- a) Mean
- b) Median
- c) Mode
- d) Range

2. What is the mode of the following dataset: 3, 5, 7, 5, 8, 5, 9?

- a) 3
- b) 5
- c) 7
- d) 9

3. When a dataset contains an even number of observations, how is the median determined?

- a) By selecting the middle observation
- b) By calculating the average of the two middle observations
- c) By selecting the most frequently occurring observation
- d) By selecting the largest observation

4. Which measure of central tendency is particularly suitable for qualitative or categorical data?

- a) Mean
- b) Median
- c) Mode
- d) Variance

5. What is the arithmetic mean of the following values: 10, 20, 30, 40 and 50?

- a) 25
- b) 30
- c) 35
- d) 40

6. Determine the median of the following observations:

12, 18, 22, 25, 30, 35, 40

7. Find the mode of the following dataset:

4, 8, 4, 6, 7, 8, 8, 10

8. Calculate the arithmetic mean of:

15, 20, 25, 35, 40, 45

9. The arithmetic mean of five numbers is 28. Four of the numbers are 22, 26, 32 and 30. Determine the fifth number.

10. Calculate the mean, median and mode of the following dataset:

3, 7, 7, 9, 10, 12, 7

11. A shopkeeper recorded the number of customers visiting his shop over seven consecutive days as follows:

25, 30, 22, 28, 35, 40, 25

Calculate:

- the mean number of customers;
- the median number of customers; and
- the mode of the observations.

12. The weights, in kilograms, of eight children are:

28, 35, 30, 32, 40, 28, 33, 37

Calculate:

- the mean weight; and
- the median weight.

13. The following frequency distribution represents the quiz scores obtained by students in a science class:

Score Frequency

10 2

15 4

20 5

25 3

Determine:

- the mean quiz score; and
- the mode.

14. The ages, in years, of a group of friends are:

12, 14, 16, 14, 17, 15, 14, 16, 15, 18

5.9 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., McKeane J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.



UNIT 6 HARMONIC AND GEOMETRIC MEAN

STRUCTURE

- 6.1 Introduction
- 6.2 Objectives
- 6.3 Geometric mean
- 6.4 Harmonic mean
- 6.5 Summary
- 6.6 EXERCISES
- 6.7 References

6.1 INTRODUCTION

While the arithmetic mean is widely known, the **harmonic mean** and **geometric mean** are essential in situations involving rates, ratios, and multiplicative data. The **geometric mean** is ideal for averaging values like growth rates or returns, while the **harmonic mean** is useful for rates such as speed or density.

This unit introduces both means, their formulas, and applications. Through comparisons, solved examples, and practical use cases, learners will understand when and how to apply each mean effectively.

6.2 OBJECTIVES

After going through this unit, students will be able to

- Gain an understanding of the geometric mean and harmonic.
- Differentiate between the Harmonic and Geometric Mean
- Calculate the Harmonic and Geometric Mean for different types of data sets.
- Analyze real-world data and interpret the meaning of these measures within context.

Compare the sensitivity of each measure to outliers and skewed data

6.3 GEOMETRIC MEAN

The geometric mean for N positive items is defined as the positive N -th root of the product of N items. It is usually denoted by G . For the values $x_1, x_2, \dots, x_N$ of the variate x , the geometric mean is given by,

$$G = N\sqrt{x_1 \cdot x_2 \cdot x_3 \dots x_N}$$

Taking log, we get

$$\begin{aligned}\log G &= \frac{1}{N} (\log x_1 + \log x_2 + \dots \log x_N) \\ &= \frac{1}{N} \sum_{i=1}^N \log x_i \\ G &= \text{antilog} \left[\frac{1}{N} \sum_{i=1}^N \log x_i \right]\end{aligned}$$

If the values of the variate x i.e., $x_1, x_2, \dots, x_N$ have the frequencies $f_1, f_2, \dots, f_N$ respectively, then

$$G = (x_1^{f_1} \cdot x_2^{f_2} \dots x_N^{f_N})^{1/N}$$

where

$$\sum_{i=1}^N f_i = N$$

Taking log, we get

$$\begin{aligned}
 \log G &= \frac{1}{N} [f_1 \log x_1 + f_2 \log x_2 + \dots + f_N \log x_N] \\
 &= \frac{1}{N} \sum_{i=1}^N f_i \log x_i \\
 G &= \text{Antilog} \left[\frac{1}{N} \sum_{i=1}^N f_i \log x_i \right]
 \end{aligned}$$

Note: It is clear that geometric mean is zero if any one of the items is zero. It is generally applied when all the items are positive.

Example 1: Find the geometric mean for the following data:

Classes:	100 – 200	200 – 300	300 – 400	400 – 500	500 – 600
Frequency:	15	18	30	20	17

Solution: We find the following table

Classes	Mid value (x)	Frequency (f)	$\log_{10} x$	$f \log_{10} x$
100 – 200	150	15	2.1761	32.6415
200 – 300	250	18	2.3979	43.1622
300 – 400	350	30	2.5441	76.3230
400 – 500	450	20	2.6532	53.0640
500 – 600	550	17	2.7404	46.5868
		$N = 100$		$\sum f \log_{10} x$ $= 251.7775$

Using the formula for geometric mean

$$\begin{aligned}
 \log_{10} G &= \frac{1}{N} \sum f \log_{10} x \\
 &= \frac{1}{100} (251.7775) = 2.5178 \\
 G &= \text{Antilog} (2.5178) = 329.5
 \end{aligned}$$

Example 2: If n_1 and n_2 are the sizes, G_1 and G_2 the geometric means of two series respectively. then the geometric mean G of the combined series is given by

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Solution: Let $x_{1i} (i = 1, 2, \dots, n_1)$ and $x_{2j} (j = 1, 2, \dots, n_2)$ be n_1 and n_2 items of the two series, respectively, then

$$G_1 = (x_{11}x_{12} \dots x_{1n_1})^{1/n_1}$$

$$\log G_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \log x_{1i}$$

$$G_2 = (x_{21}x_{22} \dots x_{2n_2})^{1/n_2}$$

$$\log G_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \log x_{2j}$$

Geometric mean G of the combined series is given by

$$G = (x_{11}x_{12} \dots x_{1n_1} \cdot x_{21}x_{22} \dots x_{2n_2})^{n/(n_1+n_2)}$$

$$\log G = \frac{1}{n_1 + n_2} \log(x_{11}x_{12} \dots x_{1n_1} \cdot x_{21}x_{22} \dots x_{2n_2})$$

$$= \frac{1}{n_1 + n_2} [\log(x_{11}x_{12} \dots x_{1n_1}) + \log(x_{21}x_{22} \dots x_{2n_2})]$$

$$= \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \log x_{1i} + \sum_{j=1}^{n_2} \log x_{2j} \right]$$

$$= \frac{1}{n_1 + n_2} [n_1 \log G_1 + n_2 \log G_2]$$

Properties of Geometric Mean

Property 1: The geometric mean of any series is always less than or equal to the arithmetic mean i.e., $A.M. \geq G.M.$

Proof: We shall prove the results for two variates only. The result can easily be generalised for n variates. Let x_1 and x_2 be the given variates of the series, then

$$A \cdot M = \frac{x_1 + x_2}{2} \quad \text{and} \quad G \cdot M = (x_1 \cdot x_2)^{1/2}$$

Consider the difference

$$\begin{aligned}
 \text{A.M.} - \text{G.M.} &= \frac{1}{2}(x_1 + x_2) - (x_1 \cdot x_2)^{1/2} \\
 &= \frac{1}{2}(x_1 + x_2 - 2\sqrt{x_1 x_2}) \\
 &= \frac{1}{2}(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0
 \end{aligned}$$

being a squared quantity. The equality sign will occur only when $x_1 = x_2$. Thus,

$$\text{A.M.} \geq \text{G.M.}$$

Property 2: The logarithm of the geometric mean of the geometric means of n frequency distributions is arithmetic mean of the logarithms of their respective geometric means i.e., if $G_1, G_2, \dots, G_n$ are the geometric means of n distributions with $N_1, N_2, \dots, N_n$ as the total frequency, then

$$\log G = \frac{\sum_{i=1}^n N_i \log G_i}{\sum_{i=1}^n N_i}$$

Proof: By the definition of geometric mean

$$\begin{aligned}
 G &= [(G_1)^{N_1} (G_2)^{N_2} \dots (G_n)^{N_n}]^{\frac{1}{N_1 + N_2 + \dots + N_n}} \\
 \log G &= \frac{1}{N_1 + N_2 + \dots + N_n} [N_1 \log G_1 + N_2 \log G_2 + \dots + N_n \log G_n] \\
 &= \frac{\sum_{i=1}^n N_i \log G_i}{\sum_{i=1}^n N_i}
 \end{aligned}$$

Property 3: The G.M of the ratios of corresponding observations in the two series is equal to the ratio of the geometric means of the respective series i.e., if G_1 and G_2 are the geometric means of the two series then G the geometric mean of the ratio of their observations is given by

$$G = \frac{G_1}{G_2}$$

Proof: Let $x_1, x_2, \dots, x_n$ be n observations corresponding to variate x and $y_1, y_2, \dots, y_n$ be n observations corresponding to variate y . Then by definition

$$\begin{aligned} G_1 &= (x_1 \cdot x_2 \dots x_n)^{1/n} \\ G_2 &= (y_1 \cdot y_2 \dots y_n)^{1/n} \end{aligned}$$

Now the geometric mean of their ratio is given by

$$\begin{aligned} G &= \left(\frac{x_1}{y_1} \cdot \frac{x_2}{y_2} \dots \frac{x_n}{y_n} \right)^{1/n} \\ &= \left(\frac{x_1 \cdot x_2 \dots x_n}{y_1 y_2 \dots y_n} \right)^{1/n} = \frac{(x_1 \cdot x_2 \dots x_n)^{1/n}}{(y_1 \cdot y_2 \dots y_n)^{1/n}} \\ &= \frac{G_1}{G_2} \end{aligned}$$

Merits of Geometric Mean

- 1 It is rigidly defined.
- 2 It is based upon all the observations.
- 3 It is not much effected by the extreme values.
- 4 It is most appropriate when one deals with ratio's e.g., in the case of index numbers.
- 5 It is capable of further algebraic treatment.
- 6 It gives less weight to larger items and more weight to small items.

Demerits of Geometric Mean

- 1 It is not easy to understand and involves complex numerical calculations.
- 2 It cannot be computed if any one of the items in the series is zero or negative.
- 3 It may be a value which is not present in the actual data.
- 4 It is most useful in the study of social and economic phenomenon where smaller items are to be given importance.
- 5 If some items of series are very large in magnitude and others are small then geometric mean is the most representative average.
- 6 The most important use of geometric mean is in the case of index numbers.

6.4 HARMONIC MEAN

Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus, the harmonic mean H of N observations $x_1, x_2, \dots, x_N$ is

$$H = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N}}$$

If $x_1, x_2, \dots, x_N$ have the frequencies $f_1, f_2, \dots, f_N$ respectively, then harmonic mean is given by

$$H = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{f_i}{x_i}} = \frac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_N}{x_N}}$$

In the case of class intervals, x is taken to be the mid value of the class interval.

Example 1: Calculate the harmonic mean for the following data.

Marks:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students:	4	6	10	7	3

Solution: The harmonic mean is calculated below:

Marks	Mid values (x)	Frequency (f)	$1/x$	$F \times (1/x)$
0 – 10	5	4	$1/5$	$4/5$
10 – 20	15	6	$1/15$	$6/15$
20 – 30	25	10	$1/25$	$10/25$
30 – 40	35	7	$1/35$	$7/35$
40 – 50	45	3	$1/45$	$3/45$

$$\begin{aligned}
 N = 30, \sum \frac{f_i}{x_i} &= \frac{4}{5} + \frac{6}{15} + \frac{10}{25} + \frac{7}{35} + \frac{3}{45} \\
 &= 0.8 + 0.4 + 0.4 + 0.2 + 0.06 \\
 &= 1.86 \\
 \therefore H &= \frac{N}{\sum \frac{f_i}{x_i}} = \frac{30}{1.86} = 16.12
 \end{aligned}$$

Weighted Harmonic Mean

In some cases, it becomes necessary to calculate weighted harmonic mean. Suppose an automobile covers different distances with different speeds, the average speed can be obtained by using weighted harmonic mean. The formula for calculating weighted harmonic mean is

$$\text{Weighted H.M.} = \text{Reciprocal} \frac{\Sigma(W/x)}{\Sigma W} = \frac{\Sigma W}{\Sigma \left(\frac{W}{x}\right)}$$

where $W_1, W_2, \dots, W_N$ stand for the respective weights of $x_1, x_2, \dots, x_N$.

Example 2: Calculate the average speed of a car running at the rate of 15 kms per hour during the first 30kms, 20kms per hour during the second 30 kms and at 2.5 kms per hour during the third 30 kms .

Solution: Speed Rates (x): $x_1 = 15, x_2 = 20, x_3 = 2.5$

Distance (W): $W_1 = 30, W_2 = 30, W_3 = 30$

$$\begin{aligned} \text{Average speed} &= \frac{\Sigma W}{\frac{1}{x_1} \times W_1 + \frac{1}{x_2} \times W_2 + \frac{1}{x_3} \times W_3} = \frac{30 + 30 + 30}{\frac{1}{15} \times 30 + \frac{1}{2} \times 30 + \frac{1}{2.5} \times 30} \\ &= \frac{90}{2 + 1.5 + 12} = \frac{90}{15.5} = 5.806 \text{ km/hour} \end{aligned}$$

Example 3: An aeroplane flies round a square the sides of which measure 100 miles each. The aeroplane covers at a speed of 100 miles per hour the first side, at 200 m.p.h the second side, at 300 m.p.h the third side and 400 m. p.h the fourth side. What is the average speed of the acroplane around the squarre?

Solution: Here we are to find average speed and the distance is kept fixed, therefore, we shall find harmonic mean to get the correct average speed.

$$\begin{aligned} \text{Thus, average speed (Harmonic mean) } &= \frac{400}{\frac{100}{100} + \frac{100}{200} + \frac{100}{300} + \frac{100}{400}} \\ &= \frac{4800}{25} = 192 \text{ m} \cdot \text{p} \cdot \text{h} \end{aligned}$$

Example 4: A variate takes the values $a, ar, ar^2 \dots ar^{n-1}$ each with frequency unity. If A, G and H are respectively the arithnetic mean, geonetric mean and harmonic mean, show that

$$A = \frac{a(1 - r^n)}{n(1 - r)}, G = ar^{\frac{(n-1)}{2}}, H = \frac{an(1 - r)r^{n-1}}{1 - r^n}$$

Prove that $G^2 = AH$. Also prove that $A \geq G \geq H$ unless $n = 1$ when all the three means coincide.

Solution: The values of the variate are given as $a, ar, ar^2 \dots ar^{n-1}$

Thus,

$$\begin{aligned} \text{A.M.} &= \frac{a + ar + ar^2 + \dots + ar^{n-1}}{n} = \frac{a}{n} [1 + r + r^2 + \dots + r^{n-1}] \\ &= \frac{a(1 - r^n)}{n(1 - r)} \text{ if } r < 1 \text{ (as the series is a G.P.)} \\ A &= \frac{a(1 - r^n)}{n(1 - r)} \\ \text{G.M. a.} &= (a \cdot ar \cdot ar^2 \dots ar^{n-1})^{1/n} \\ &= [a^n (r^{1+2+\dots+(n-1)})]^{1/n} \\ &= a \left(\frac{(n-1)n}{2} \right)^{\frac{1}{n}} = ar^{\frac{(n-1)}{2}} \\ G &= ar \cdot \frac{(n-1)}{2} \\ H.M &= \frac{1}{\frac{1}{a} + \frac{1}{ar} + \dots + \frac{1}{ar^{n-1}}} = \frac{1}{a} \left[1 + \frac{1}{r} + \frac{1}{r^2} + \dots + \frac{1}{r^{n-1}} \right] \\ &= \frac{an}{\left(\frac{1}{r^n} - 1 \right) / \left(\frac{1}{r} - 1 \right)} = \frac{an \left(\frac{1}{r} - 1 \right)}{\left(\frac{1}{r^n} - 1 \right)} = \frac{an(1 - r)}{r(1 - r^n)} \times r^n \\ H &= \frac{an(1 - r)r^{n-1}}{(1 - r^n)} \end{aligned}$$

Now for proving $A.H = G^2$, Let us consider

$$AH = \frac{a(1 - r^n)}{n(1 - r)} \times \frac{an(1 - r)r^{n-1}}{(1 - r^n)} = a^2 r^{n-1} = \left(ar^{\frac{n-1}{2}} \right)^2 = G^2$$

Thus,

$$AH = G^2$$

Let us prove $A > G > H$, let us consider

$$\begin{aligned}
A - G &= \frac{a}{n}(1 + r + r^2 + \dots + r^{n-1}) - a \cdot r^{\frac{(n-1)}{2}} \\
&= \frac{a}{n} \left[1 + r + r^2 + \dots + r^{n-1} - nr^{\frac{(n-1)}{2}} \right] \\
&= \frac{a}{n} \left[\left(1 - r^{\frac{(n-1)}{2}}\right) + \left(r - r^{\frac{(n-1)}{2}}\right) + \left(r^2 - r^{\frac{(n-1)}{2}}\right) \dots + \left(r^{n-1} - r^{\frac{(n-1)}{2}}\right) \right] \\
&= \frac{a}{n} \left[\left(1 - r^{\frac{(n-1)}{2}}\right) + r \left(1 - r^{\frac{(n-3)}{2}}\right) + r^2 \left(1 - r^{\frac{(n-5)}{2}}\right) + \dots + r^{n-1} \left(1 - r^{\frac{(1-n)}{2}}\right) \right] \\
&= \text{a positive quantity if } r < 1
\end{aligned}$$

$$\therefore A - G > 0 \text{ or } A > G$$

But if $n = 1$, $A - G = 0$, i.e., $A = G$

Also consider,

$$\begin{aligned}
G - H &= ar^{\frac{(n-1)}{2}} - \frac{an(1-r)r^{n-1}}{1-r^n} \\
&= ar^{\frac{(n-1)}{2}} - \frac{anr^{n-1}}{(1-r^n)/(1-r)} \\
&= ar^{\frac{(n-1)}{2}} - \frac{anr^{n-1}}{1+r+r^2+\dots+r^{n-1}} \left(\text{as } \frac{1-r^n}{1-r} = 1+r+r^2+\dots+r^{n-1} \right) \\
&= \frac{ar^{\frac{(n-1)}{2}}}{1+r+r^2+\dots+r^{n-1}} \left[1+r+r^2+\dots+r^{n-1} - nr^{\frac{(n-1)}{2}} \right] \\
&= \frac{ar^{\frac{(n-1)}{2}}}{1+r+r^2+\dots+r^{n-1}} \left[\left(1 - r^{\frac{(n-1)}{2}}\right) + \left(r - r^{\frac{(n-1)}{2}}\right) + \dots + \left(r^{n-1} - r^{\frac{(n-1)}{2}}\right) \right] \\
&= \frac{ar^{\frac{(n-1)}{2}}}{1+r+r^2+\dots+r^{n-1}} \left[\left(1 - r^{\frac{(n-1)}{2}}\right) + r \left(1 - r^{\frac{(n-3)}{2}}\right) + \dots + r^{n-1} \left(1 - r^{\frac{(1-n)}{2}}\right) \right] \\
&= \text{a positive quantity if } r < 1
\end{aligned}$$

$$\therefore G - H > 0 \text{ or } G > H$$

But if $n = 1$, $G - H = 0$, i.e., $G = H$. Hence, $A > G > H$ unless $n = 1$, in which case $A = G = H$, i.e., when all the three means coincide.

Example 5: A variate takes values 2, 4, 8, 16, ..., 2^n . Find A.M, G.M and H.M.

Solution: The A.M is

$$\begin{aligned}
 A \cdot M \cdot &= \frac{(2 + 4 + 8 + 16 + \dots + 2^n)}{n} \\
 &= \frac{(2^1 + 2^2 + 2^3 + \dots + 2^n)}{n} \\
 &= \frac{2(1 + 2 + 2^2 + \dots + 2^{n-1})}{n} \\
 &= \frac{2(2^n - 1)}{n(2 - 1)} = \frac{2(2^n - 1)}{n}
 \end{aligned}$$

$$A.M. = \frac{2^n - 1}{n}$$

$$\begin{aligned}
 G.M. &= (2 \cdot 4 \cdot 8 \dots 2^n)^{1/n} \\
 &= (2^1 \cdot 2^2 \cdot 2^3 \dots 2^n)^{1/n} \\
 &= (2^{1+2+3+\dots+n})^{\frac{1}{n}} = \left[2^{\frac{n(n+1)}{2}} \right]^{\frac{1}{n}} = 2^{\frac{n+1}{2}}
 \end{aligned}$$

$$\begin{aligned}
 H.M. &= \frac{1}{\frac{1}{n} \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \dots + \frac{1}{2^n} \right)} = \frac{1}{\frac{1}{2} \left(1 + \frac{1}{2} + \frac{1}{2^2} + \dots + \frac{1}{2^{n-1}} \right)} \\
 &= \frac{2n}{\left(1 - \frac{1}{2^n} \right) / \left(1 - \frac{1}{2} \right)} = \frac{2n \left(1 - \frac{1}{2} \right)}{\left(1 - \frac{1}{2^n} \right)} = \frac{2n \times \frac{1}{2} \times 2^n}{(2^n - 1)} \\
 &= \frac{n2^n}{(2^n - 1)}.
 \end{aligned}$$

Merits of Harmonic Mean

- 1 It is based on all the observations.
- 2 It gives largest weight to smaller units.
- 3 It is specially used in finding averages of rates or times.

Demerits of Harmonic Mean

- 1 It only works in certain situations.
- 2 It is difficult to understand and requires complicated calculations.

- 3 It cannot be calculated if the value of any item is not known.

Like other averages, the harmonic mean also has its specific uses. This mean is specially used in the computation of average speed of velocity, average price at which an article is sold. The rates which are usually averaged by harmonic mean indicate a relationship between two measuring units which can be reciprocally expressed. For instance, if a train travels 80 kms in 2 hours, then its speed can be expressed as;

$$\text{Speed of the train} = \frac{80 \text{ km}}{2 \text{ hours}} = 40 \text{ km. p. h}$$

Here, the units of the distance covered is km. that of time taken an hour on the other hand, it can be put reciprocally as:

$$\text{Speed of the train} = \frac{2 \text{ hrs}}{80 \text{ km}} = 0.025 \text{ hours per km.}$$

Here the unit of time is in hours and that of distance in a km. Thus, when average of different speeds expressed in km. p.h or miles. p.h or average of prices quoted in units of commodity per rupee is to be computed, the harmonic mean is recommended. However, weighted arithmetic mean, taking speed as X and journey time as weighted W , should be calculated if journey times are given in hours (not in km.p.h) etc.

CHECK YOUR PROGRESS

1. Define Geometric Mean and explain its important characteristics.

.....

.....

.....

2. What is Harmonic Mean? Explain its applications.

.....

.....

.....

3. Distinguish between Arithmetic Mean, Geometric Mean, and Harmonic Mean.

.....

.....

.....

6.5 SUMMARY

In this unit, we have discussed the following concepts

- Various measure of central tendency.
- Definition and calculation of Geometric mean.
- Definition and calculation of Harmonic mean.

6.6 EXERCISES

Multiple Choice Questions (MCQs)

1. The geometric mean of 4 and 16 is:
 - a) 8
 - b) 10
 - c) 12
 - d) 14
2. The harmonic mean of 5 and 20 is:
 - a) 8
 - b) 10

- c) 15
d) 12
3. If the geometric mean between two numbers is 18 and one of the numbers is 9, the other number is:
a) 27
b) 36
c) 18
d) 12
4. The harmonic mean of three positive numbers a, b, c is given by:
a) $\frac{a+b+c}{3}$
b) $\sqrt[3]{abc}$
c) $\frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$
d) $\frac{ab+bc+ca}{a+b+c}$
5. The relationship between Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM) is:
a) $AM \geq GM \geq HM$
b) $GM \geq AM \geq HM$
c) $HM \geq GM \geq AM$
d) $AM = GM = HM$ for unequal numbers

Answer:

1. a) 8. 2. b) 10 3. a) 27 4. c) $\frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$ 5. a) $AM \geq GM \geq HM$

2. Short Answer Questions

6. Calculate the geometric mean of 6, 12, and 24.
7. Calculate the harmonic mean of 3, 6, and 9.

8. Prove that the geometric mean of two positive numbers a and b is $\sqrt{ab}$.
9. If three numbers are in geometric progression and their product is 512, find the numbers if their geometric mean is 8.
10. Calculate the harmonic mean of 8 and 18.

3. Long Answer Questions

11. Derive the formula for the harmonic mean of n positive numbers and explain its significance with an example.
12. Explain the relationship between AM, GM, and HM with proof. Use an example to illustrate the inequality.
13. A car travels three equal distances at speeds of 40 km/h, 60 km/h, and 80 km/h. Calculate the average speed using harmonic mean.
14. The sides of a triangle are in geometric progression. The smallest side is 4, and the largest side is 16. Find area of the triangle and length of the middle side.
15. Three cables are 6 m, 12 m, and 24 m long. Find their geometric and harmonic mean. Discuss the practical significance of both means in real-life applications.

6.7 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., McKean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 7 MEASURES OF DISPERSION

STRUCTURE

- 7.1 Introduction
- 7.2 Objectives
- 7.3 Measures of dispersion
- 7.4 Mean deviation or average deviation
- 7.5 Standard deviation and root mean square deviation
- 7.6 Effect of change of origin and scale on standard deviation
- 7.7 Examples
- 7.8 Combined standard deviation
- 7.9 Summary
- 7.10 Exercises
- 7.11 References

7.1 INTRODUCTION

While measures of central tendency like the mean, median, and mode provide insights into the "center" of a data set, they do not convey the spread or variability of the data. This is where **measures of dispersion** come into play. Dispersion refers to how spread out or scattered the data values are around the central value. Understanding dispersion helps in assessing the reliability and consistency of the data and is crucial for making informed decisions.

7.2 OBJECTIVES

After going through this unit, students will be able to

- Understand the importance of dispersion in statistical analysis.
- Identify different measures of dispersion and their applications.

- Calculate key measures of dispersion, including range, interquartile range (IQR), variance, and standard deviation.
- Interpret the spread of data and its implications for real-world scenarios.
- Evaluate how dispersion relates to data distribution and variability.

7.3 MEASURES OF DISPERSION

The measures of dispersion which are in common use includes:

- 1 Range
- 2 Interquartile Range and Quartile Deviation
- 3 Average Deviation or Mean Deviation
- 4 Standard Deviation

Range

It is the simplest measure of dispersion. It is the variation between the highest and lowest values in a set of observations. It is clear that it is based on two extreme values. It is calculated by the formula

$$\text{Range} = L - S$$

where,

$$\begin{aligned} L &= \text{Largest item} \\ S &= \text{Smallest item} \end{aligned}$$

The coefficient of range is calculated for the relative measure of range, it is given by

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

Range is simple, common and easy to understand. However, it is not a very satisfactory measure because it is much affected by extreme values and not based on all observations. It is not a reliable measure of dispersion.

Uses of Range

- 1 It is widely used in industry to study the variation in quality of items.
- 2 It is used in meteorological department to know that actual temperature and in weather forecasts.
- 3 In stock exchange, the lowest and highest rate of shares are declared daily.

Interquartile Range and Quartile Deviation

This is a measure of "Partial range". It measures the difference between the values of items at median or Q_2 . The difference between the upper and lower quartiles i.e., $Q_3 - Q_1$ is known as the interquartile range and half of this difference is denoted by Q is called the semi-interquartile range or quartile deviation. Symbolically,

$$\text{Interquartile Range} = Q_3 - Q_1$$

where, Q_3 = Third quartile, Q_1 = First quartile

$$\text{and Semi-interquartile Range or Quartile Deviation} = \frac{1}{2}(Q_3 - Q_1)$$

$$\text{Also, coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 1: Find out Quartile Deviation and its coefficient from the following datas:

Class:	4	8	12	16	20	24	28	32	36
	– 8	– 12	– 16	– 20	– 24	– 28	– 32	– 36	– 40
Frequency:	6	10	18	30	15	12	10	6	2

Solution: Let us find the table for cumulative frequencies.

Class	Frequency	c.f	Class	Frequency	c.f
4 – 8	6	6	24 – 28	12	91
8 – 12	10	16	28 – 32	10	101
12 – 16	18	34	32 – 36	6	107
16 – 20	30	64	36 – 40	2	109
20 – 24	15	79			

$$Q_1 N_0 = \left(\frac{N}{4}\right)\text{-th item} = \frac{109}{4} = 27.25$$

27.25 lies in the c.f 34, which belongs to the class 12-16. Thus, 12-16 is the required class. Using the formula

$$\begin{aligned}
 Q_1 &= L_1 + \frac{i}{f} \left(\frac{N}{4} - c \right) \\
 &= 12 + \frac{4}{18} (27.25 - 16) \\
 &= 12 + \frac{4}{18} \times 11.25 = 12 + \frac{45}{18} = 12 + 2.5 \\
 &= 14.5 \\
 &= \left(\frac{3 \times 109}{4} \right)\text{-th item} = \left(\frac{327}{4} \right)\text{-th item} \\
 &= 81.75
 \end{aligned}$$

$$Q_3 N_0 = \left(\frac{3N}{4}\right)\text{-th item} = \left(\frac{3 \times 109}{4}\right)\text{-th item} = \left(\frac{327}{4}\right)\text{-th item}$$

81.75 lies in the c. f 91, which belongs to the class 24 – 28. Thus, 24 – 28 is the required class.

Using the formula

$$\begin{aligned}Q_3 &= L_1 + \frac{i}{f} \left(\frac{3N}{4} - c \right) \\&= 24 + \frac{4}{12} (81.75 - 79) \\&= 24 + \frac{4}{12} \times 2.75 = 24 + \frac{11.00}{12} = 24 + 0.92 \\&= 24.92 \\Q \cdot D &= \frac{Q_3 - Q_1}{2} = \frac{24.92 - 14.5}{2} = \frac{10.42}{2} = 5.21 \\Q.D &= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{24.92 - 14.5}{24.92 + 14.5} = \frac{10.42}{39.42} = 0.264\end{aligned}$$

Coefficient of

Merits of Q.D.

- 1 It can be easily computed.
- 2 It is a better measure of dispersion than the range.
- 3 It has a special use in measuring variation of open end distributions.
- 4 It is useful when extreme values are likely to affect the result.

Demerits of Q.D.

- 1 It does not take into account all the items.
- 2 It is not capable of further algebraic treatment.
- 3 It is affected by the fluctuations of sampling.

7.4 MEAN DEVIATION OR AVERAGE DEVIATION

We have seen that the range and quartile deviation were the positional measures of dispersion and were not based on all the observations. To overcome these limitations, mean deviation is defined. Mean deviation

of a series is the arithmetic mean of the deviations of various items from a measure of central tendency (mean, median or mode) taking all deviations as positive. In other words, mean deviation is the sum of the deviations (taken + ve) from an average divided by the number of items in the series. Symbolically, the mean deviation is expressed as follows:

$$\text{Mean Deviation} = \frac{1}{N} \sum |x - M|$$

where M is the mean or median or mode according as the mean deviation from mean or median or mode is required. The coefficient of mean deviation is obtained by dividing mean deviation by the central value used for calculating it.

Thus,

$$\text{Coefficient of Mean Deviation} = \frac{\text{Mean deviation}}{\text{Mean or median or mode}}$$

For a frequency distribution, if f_i be the frequency of the variate x_i , $i = 1, 2, \dots, N$, then the mean deviation is given by

$$M \cdot D = \frac{1}{N} \sum f_i |x_i - M|, N = \sum f_i$$

Example 2: Ten students obtained the following marks (out of 100).

Calculate mean deviation.

5, 10, 20, 25, 40, 42, 45, 48, 70, 80

Solution: Mean = $\frac{5+10+20+25+40+42+45+48+70+80}{10} = \frac{385}{10} = 38.5$

M.D. about Mean = $\frac{1}{N} \sum |x_i - \text{Mean}|$

$$= \frac{1}{10} [33.5 + 28.5 + 18.5 + 13.5 + 1.5 + 3.5 + 6.5 + 9.5 + 31.5 + 41.5]$$

$$= \frac{188}{10} = 18.8$$

Example 3: Calculate M.D. from median and its coefficient from the following data.

Size x :	10	11	12	13	14
Frequency:	3	12	18	12	3

Solution: Table for Calculation

Size (x)	Frequency (f)	c. f	$ x - Md $	$f x - Md $
10	3	3	2	6
11	12	15	1	12
12	18	33	0	0
13	12	45	1	12
14	3	48	2	6
	$N = 48$			$\Sigma f x - Md = 36$

Median = Value of $\left(\frac{N}{2}\right)$ -th item = value of $\left(\frac{48}{2}\right)$ -th item = 24-th item
 = 12

Mean deviation about median = $\frac{1}{N} \Sigma f|x - Md|$

$$= \frac{1}{48} \times 36 = 0.75$$

Coefficient of M.D. = $\frac{\text{M.D.}}{\text{Median}} = \frac{0.75}{12} = 0.0625$

Example 4: On the basis of the following distribution, compute mean deviation from mean and median.

Marks:	140 – 150	150 – 160	160 – 170	170 – 180	180 – 190	190 – 200
Frequency:	4	6	10	18	9	3

Solution: Computation of M.D. about mean

Marks	M.V. (x)	f	x – 165 = d	fd	$ x - \bar{x} $ – 171.2	$f x - \bar{x} $ – 171.2
140 – 150	145	4	-20	-80	26.2	104.8
150 – 160	155	6	-10	-60	16.2	97.2
160 – 170	165	10	0	0	6.2	62.0
170 – 180	175	18	+10	+180	3.8	68.4
180 – 190	185	9	+20	+180	13.8	124.2
190 – 200	195	3	+30	+90	23.8	71.4
		N = 50		Σfd = 310		$\Sigma f_i x - \bar{x} $ = 528.0

$$\begin{aligned}\bar{x} &= A + \frac{\Sigma fd}{N} \\ &= 165 + \frac{310}{50} = 16.5 + 6.2 = 171.2\end{aligned}$$

$$\text{Mean Deviation (about mean)} = \frac{1}{N} \sum_i f_i |x_i - \bar{x}|$$

$$= \frac{528}{50} = 10.56$$

Computation of M.D. about median

Marks	$M.V(x)$	f	$c.f$	$ x - Md $	$f x - Md $
140 – 150	145	4	4	27.78	111.12
150 – 160	155	6	10	17.78	106.68
160 – 170	165	10	20	7.78	77.80
170 – 180	175	18	38	2.22	39.96
180 – 190	185	9	47	12.22	109.98
190 – 200	195	3	50	22.22	66.66
		N $= 50$			$\sum f x - Md $ $= 512.2$

Median = Value of $\left(\frac{N}{2}\right)$ -th item = Value of $\left(\frac{50}{2}\right)$ th item = Value of 25 -th item.

This belongs to the class 170 – 180. Thus,

$$L_1 = 170, i = 10, f = 18, c = 20$$

Using the formula

$$\begin{aligned}
 \text{Median} &= L_1 + \frac{i}{f} \left(\frac{N}{2} - c \right) \\
 &= 170 + \frac{10}{18} (25 - 20) = 170 + \frac{50}{18} \\
 &= 170 + 2.78 = 172.78
 \end{aligned}$$

$$\begin{aligned}
 M.D. (\text{ about median}) &= \frac{1}{N} \sum_i f_i |x_i - \text{Median}| \\
 &= \frac{512.20}{50} = 10.24
 \end{aligned}$$

Properties of Mean Deviation

Property 1: Mean deviation about median is minimum.

Proof: Let $x_1, x_2, \dots, x_N$ be the N values of a variate x which are arranged in increasing order of magnitude, so that

$$x_1 \leq x_2 \leq x_3 \leq \dots \leq x_N$$

There may be two cases depending upon the situation of N i.e., N may be even or odd. Let us discuss both the cases.

Case I: When N is even say, $N = 2m$, then x_m and x_{m+1} will be the two middle values of the series. Let A be any assumed value of the variate x , then

$$M.D. \text{ about ' } A' = \frac{1}{N} \sum_i |x_i - A|$$

Now considering the deviations of the variate values from A , we see that the sum

$$|x_{2m} - A| + |x_1 - A| \text{ is minimum when } x_1 \leq A \leq x_{2m}$$

Similarly,

$$|x_{2m-1} - A| + |x_2 - A| \text{ is minimum when } x_2 \leq A \leq x_{2m-1}$$

Lastly,

$$|x_{m+1} - A| + |x_m - A| \text{ is minimum when } x_m \leq A \leq x_{m+1}$$

Since each of these sums will be minimum, hence

$$M.D. \text{ about 'A'} = \frac{1}{N} \sum_i |x_i - A|$$

Will be minimum when $x_m \leq A \leq x_{m+1}$

In other words 'A' is the median value.

Case II: When N is odd, say $N = 2m + 1$, then x_{m+1} will be middle value of the variate x . As in the previous case, it can be seen that for any assumed value A of the variate x ,

$$|x_{2m+1} - A| + |x_1 - A| \text{ is minimum when } x_1 \leq A \leq x_{2m+1}$$

Similarly,

$$|x_{2m} - A| + |x_2 - A| \text{ is minimum when } x_2 \leq A \leq x_{2m}$$

Lastly,

$$|x_{m+2} - A| + |x_m - A| \text{ is minimum when } x_m \leq A \leq x_{m+2}$$

Hence,

$$M.D. \text{ about 'A'} = \frac{1}{N} \sum_i |x_i - A|$$

will be minimum when $A = x_{m+1} = \text{Median value}$.

Thus, in both the cases whether N is even or odd we see that $M.D.$ is minimum when deviations are taken from median.

Property 2: Absolute mean deviation about mean is independent of change of origin.

Proof: Let the variate under study is x and if origin is transferred to any point ' a ' such that new variate $d_i = x_i - a$.

Then by definition,

$$\begin{aligned}
 \text{M.D. about mean} &= \frac{1}{N} \sum_i |x_i - \bar{x}| \\
 &= \frac{1}{N} \sum_i \left| x_i - a - \frac{\sum f d_i}{N} \right| \left(\text{as } \bar{x} = a + \frac{\sum f d_i}{N} \right) \\
 &= \frac{1}{N} \sum_i \left| d_i + a - a - \frac{\sum f d_i}{N} \right| \\
 &= \frac{1}{N} \sum_i \left| d_i - \frac{\sum f d_i}{N} \right| \\
 &= \frac{1}{N} \sum_i |d_i - \bar{d}|
 \end{aligned}$$

which is independent of ' a '. Thus, absolute mean deviation about mean is independent of change of origin.

Merits of Mean Deviation

- 1 It is easy and simple to calculate.
- 2 It is based on all the observations.
- 3 It is hardly affected by extreme values.

Demerits of Mean Deviation

- 1 In calculation of M.D., plus and minus signs are ignored, which is non-algebraic. And thus, it is not capable of further algebraic treatment.
- 2 It is not well defined as deviations may be taken from any central value.

7.5 STANDARD DEVIATION AND ROOT MEAN SQUARE DEVIATION

The main drawback of mean deviation is that the deviations are considered as positive only, which is a non-algebraic method. This drawback is removed in standard deviation. The standard deviation is the square root of the mean of the squares of the deviation from the arithmetic mean. This is by far, the best and most frequently used measure of dispersion. Symbolically it is denoted by σ (Greek letter called sigma) and its name i.e. standard deviation was suggested by Karl Pearson. Thus,

$$\text{S.D.} = \sigma = \sqrt{\left[\frac{1}{N} \sum f_i (x_i - \bar{x})^2 \right]}, \quad N = \sum f_i$$

Also its square i.e., σ^2 is called variance.

Illustration 5: Ten students obtained following marks (out of 100).

Calculate standard deviation.

5,10,20,25,40,42,45,48,70,80

$$\text{Solution: Mean}(\bar{x}) = \frac{5+10+20+25+40+42+45+48+70+80}{10} = \frac{385}{10} = 38.5$$

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{1}{N} \sum (x - \bar{x})^2} \\ &= \sqrt{\frac{1}{10} [(5 - 38.5)^2 + (10 - 38.5)^2 + (20 - 38.5)^2 + (25 - 38.5)^2 + (40 - 38.5)^2 \\ &\quad + (42 - 38.5)^2 + (45 - 38.5)^2 + (48 - 38.5)^2 + (70 - 38.5)^2 + (80 - 38.5)^2]} \\ &= \sqrt{\frac{1}{10} [(-33.5)^2 + (-28.5)^2 + (-18.5)^2 + (13.5)^2 + (1.5)^2 + (3.5)^2 + (6.5)^2 \\ &\quad + (9.5)^2 + (31.5)^2 + (41.5)^2]} \\ &= \sqrt{\frac{1}{10} [1122.25 + 812.25 + 342.25 + 182.25 + 2.25 + 12.25 + 42.25 + 90.25 \\ &\quad + 992.25 + 1722.25]} \\ &= \sqrt{\frac{1}{10} [5320.5]} = \sqrt{532.05} = 23.06 \end{aligned}$$

Example 6: Calculate standard deviation from the following data.

Size:	0 – 10	10 – 20	20 – 30	30 – 40
Frequency:	1	2	4	3

Solution:

Size	f	M $\cdot V(x)$	fx	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
0 – 10	1	5	5	-19	361	361
10 – 20	2	15	30	-9	81	162
20 – 30	4	25	100	+1	1	4
30 – 40	3	35	105	+11	121	363
	N $= 10$		Σfx $= 240$			$\Sigma f(x - \bar{x})^2$ $= 890$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{240}{10} = 24$$

$$\begin{aligned} \text{S.D.} &= \sigma = \sqrt{\frac{1}{N} \Sigma f_i(x_i - \bar{x})^2} \\ &= \sqrt{\frac{1}{10} \times 890} = \sqrt{89} = 9.4 \end{aligned}$$

Root Mean Square Deviation

In standard deviation the deviations were taken from the mean $\bar{x}$. If deviations are taken from an arbitrary value say ' a ' instead of $\bar{x}$ we

find root mean square deviation. Thus, root mean square deviation is the positive square root of the mean of the squared deviations from an arbitrary value 'a' and is denoted by S. Thus,

Relation Between σ and S

Consider

$$\begin{aligned}
 \sigma^2 &= \frac{1}{N} \sum_i (x_i - \bar{x})^2 \\
 &= \frac{1}{N} \sum_i (x_i - a + a - \bar{x})^2 \\
 &\quad \text{(Add and subtract 'a')} \\
 &= \frac{1}{N} \left[\sum_i \{(x_i - a) + (a - \bar{x})\}^2 \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 + \sum_i (a - \bar{x})^2 + 2(a - \bar{x}) \sum_i (x_i - a) \right] \quad \left[\text{as } (a - \bar{x}) \text{ is constant} \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 + N(a - \bar{x})^2 + 2(a - \bar{x}) \sum_i (x_i - a) \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 + N(a - \bar{x})^2 + 2(a - \bar{x}) \sum_i x_i - 2Na \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 + N(a - \bar{x})^2 + 2(a - \bar{x})N\bar{x} - 2Na \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 + N(a - \bar{x})^2 + 2(a - \bar{x})N(\bar{x} - a) \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 + N(a - \bar{x})^2 - 2N(a - \bar{x})^2 \right] \\
 &= \frac{1}{N} \left[\sum_i (x_i - a)^2 - N(a - \bar{x})^2 \right] \\
 &= \frac{1}{N} \sum_i (x_i - a)^2 - (a - \bar{x})^2 \\
 \sigma^2 &= S^2 - (a - \bar{x})^2 \text{ is the required relation} \\
 &= S^2 - d^2, \text{ where } d = (a - \bar{x})
 \end{aligned}$$

Now, $\sigma^2 = S^2$ if $(a - \bar{x})^2 = 0$ i.e., $a - \bar{x} = 0$ or $a = \bar{x}$.

Thus the root mean square deviation from the mean is the least of all other mean square deviations.

7.6 EFFECT OF CHANGE OF ORIGIN AND SCALE ON STANDARD DEVIATION

If x is the variate under study which takes values $x_1, x_2, \dots, x_N$, then by definition, variance is given by

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Let origin is shifted to new variable as $d_i = x_i - A$

Then

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (x_i - A + A - \bar{x})^2 \\ &= \frac{1}{N} \left[\sum_i (x_i - A)^2 + \sum_i (A - \bar{x})^2 + 2(A - \bar{x}) \sum_i (x_i - A) \right] \\ &= \frac{1}{N} \left[\sum_i (x_i - A)^2 + N(A - \bar{x})^2 + 2(A - \bar{x}) \sum_i (x_i - A) \right] \\ &= \frac{1}{N} \left[\sum_i d_i^2 + N(A - \bar{x})^2 + 2(A - \bar{x}) \sum_i d_i \right] \end{aligned}$$

Now we know that

$$\bar{x} = A + \frac{\sum_i d_i}{N} \text{ or } A - \bar{x} = \left(\frac{-\sum_i d_i}{N} \right)$$

Using this relation (1) becomes

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \left[\sum_i d_i^2 + N \left(\frac{-\sum_i d_i}{N} \right)^2 + 2 \left(\frac{-\sum_i d_i}{N} \right) \sum_i d_i \right] \\ &= \frac{1}{N} \left[\sum_i d_i^2 + \frac{(\sum_i d_i)^2}{N} - \frac{2}{N} \left(\sum_i d_i \right)^2 \right] \\ &= \frac{1}{N} \left[\sum_i d_i^2 + \frac{(\sum_i d_i)^2}{N} \right] \end{aligned}$$

$$\sigma^2 = \frac{\sum_i d_i^2}{N} - \left(\frac{\sum_i d_i}{N} \right)^2$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N} - \left(\frac{\sum_i d_i}{N} \right)^2}$$

Which is independent of A and thus standard deviation is not affected by the change of origin.

Let us now see the effect of change of origin and scale both, for which consider, a new variable, $u_i = \frac{x_i - A}{h}$

Consider,

$$\begin{aligned}\sigma^2 &= \frac{1}{N} \left[\sum_i (x_i - \bar{x})^2 \right] \\ &= \frac{1}{N} \left[\sum_i (x_i - A + A - \bar{x})^2 \right] \\ &= \frac{1}{N} \left[\sum_i (x_i - A)^2 + N(A - \bar{x})^2 + 2(A - \bar{x}) \sum_i (x_i - A) \right] \\ &= \frac{1}{N} \left[\sum_i h^2 u_i^2 + N(A - \bar{x})^2 + 2(A - \bar{x}) \sum_i h u_i \right] \\ \bar{x} &= A + \frac{\sum_i u_i}{N} \times h \\ A - \bar{x} &= - \left(\frac{h \sum_i u_i}{N} \right)\end{aligned}$$

Now we know that

Using this relation (2) becomes,

$$\begin{aligned}&= \frac{1}{N} \left[h^2 \sum_i u_i^2 + N \left(\frac{-h \sum_i u_i}{N} \right)^2 + 2 \left(\frac{-h \sum_i u_i}{N} \right) h \sum_i u_i \right] \\ &= \frac{1}{N} \left[h^2 \sum_i u_i^2 + h^2 \frac{(\sum_i u_i)^2}{N} - 2h^2 \frac{(\sum_i u_i)^2}{N} \right]\end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{N} \left[h^2 \sum_i u_i^2 - h^2 \frac{(\sum_i u_i)^2}{N} \right] \\
 \sigma^2 &= h^2 \text{ times } \frac{1}{N} \left[\sum_i u_i^2 - \frac{(\sum_i u_i)^2}{N} \right] = h^2 \text{ times } \left[\frac{\sum_i u_i^2}{N} - \left(\frac{\sum_i u_i}{N} \right)^2 \right] \\
 \sigma &= h \cdot \sqrt{\frac{\sum_i u_i^2}{N} - \left(\frac{\sum_i u_i}{N} \right)^2}
 \end{aligned}$$

This is dependent on h and thus, Standard deviation changes proportionally when the scale of the data is altered, indicating that it is affected by changes in scale.

7.7 EXAMPLES

Example 7: For first n natural numbers, calculate the mean and standard deviation.

Solution: Let the numbers are $1, 2, 3, \dots, n$.

$$\text{Mean} = \frac{1 + 2 + 3 + \dots + n}{n} = \frac{n(n+1)}{2 \times n} = \frac{n+1}{2}$$

Also,



$$\begin{aligned}
\text{Variance} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] \\
&= \frac{1}{n} \left[1^2 + 2^2 + \dots + n^2 - n \left(\frac{n+1}{2} \right)^2 \right] \quad \left(\text{as } \bar{x} = \frac{n+1}{2} \right) \\
&= \frac{1}{n} \left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \right] \\
&\quad \left[\text{sum of the squares of } n \text{ natural numbers is } \frac{n(n+1)(2n+1)}{6} \right] \\
&= \frac{n(n+1)}{2n} \left[\frac{(2n+1)}{3} - \frac{(n+1)}{2} \right] \\
&= \frac{(n+1)}{2} \left[\frac{4n+2-3n-3}{6} \right] = \frac{(n+1)}{12} [n-1] = \frac{n^2-1}{12} \\
\text{S.D.} &= \sqrt{\frac{(n^2-1)}{12}}
\end{aligned}$$

Example 8: For the first ten natural numbers, calculate mean and standard deviation. Hence or otherwise find the mean and standard deviation of

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Solution: For first ten natural numbers

$$\begin{aligned}
x &: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \\
\sum x &= 1 + 2 + 3 + \dots + 10 = \frac{10(10+1)}{2} = 55 \\
\sum x^2 &= 1^2 + 2^2 + 3^2 + \dots + 10^2 = \frac{10(10+1)(2 \times 10 + 1)}{6} = 385 \\
\therefore \text{Mean} &= \frac{1}{10} \sum x = \frac{55}{10} = 5.5 \\
\text{S.D.} &= \sqrt{\frac{1}{n} \sum x^2 - \left(\frac{\sum x}{n} \right)^2} \\
&= \sqrt{\frac{385}{10} - (5.5)^2} = \sqrt{38.5 - 30.25} = \sqrt{8.25} = 2.87
\end{aligned}$$

Again given

$x: 0, 1, 2, 3, \dots, 10$

$$\sum x = 0 + 1 + 2 + 3 + \dots + 10 = 55$$

$$\sum x^2 = 0^2 + 1^2 + 2^2 + \dots + 10^2 = 385$$

$$\text{Mean} = \frac{1}{11} \sum x = \frac{55}{11} = 5$$

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{1}{n} \sum x^2 - \left(\frac{\sum x}{n}\right)^2} \\ &= \sqrt{\frac{1}{11} \times 385 - 5^2} = \sqrt{35 - 25} = \sqrt{10} = 3.16 \end{aligned}$$

Example 9: If the mean and s.d. of a variable x are m and σ respectively. Calculate the mean and s.d of $\frac{ax+b}{c}$, where a , b and c are constants.

Solution: Let $y = \frac{ax+b}{c}$

$$\begin{aligned} \therefore \text{Mean} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \left(\frac{ax_i + b}{c} \right) + \frac{1}{n} \left[\frac{1}{c} \left(\sum ax_i + nb \right) \right] \\ &= \frac{1}{n} \left[\frac{1}{c} \left(a \sum x_i + nb \right) \right] = \frac{1}{n} \left[\frac{1}{c} (an\bar{x} + nb) \right] \left(\text{as } \bar{x} = \frac{\sum x_i}{n} \right) \\ &= \frac{a\bar{x}}{c} + \frac{b}{c} = \frac{1}{c} (a\bar{x} + b) = \frac{1}{c} (am + b) \end{aligned}$$

Example 10: (i) A dataset of 10 observations has a mean of 40 and a variance of 36. However, one value was incorrectly recorded as 30 instead of 40. Determine the corrected mean and variance.

(ii) A student calculated the mean and standard deviation of 100 items to be 50 and 5, respectively. Later, it was discovered that two values were wrongly entered as 40 and 50, whereas the correct values should have been 60 and 30. Find the corrected mean and standard deviation.

Solution:

(i) Here $n = 10$, Incorrect Mean $\bar{x} = 40$

Using $\bar{x} = \frac{\sum x}{n}$, Incorrect $\sum x = 10 \times 40 = 400$

Thus, correct $\sum x = 400 - 30 + 40 = 410$

$$\therefore \text{Correct mean} = \frac{\text{correct } \sum x}{n} = \frac{410}{10} = 41$$

Incorrect variance $\sigma^2 = 36$

Using

$$\begin{aligned}\sigma^2 &= \frac{1}{n} \sum (x - \bar{x})^2 \\ &= \frac{1}{n} \left(\sum x^2 - n\bar{x}^2 \right) = \frac{\sum x^2}{n} - \bar{x}^2\end{aligned}$$

Thus,

$$\begin{aligned}36 &= \frac{1}{10} \sum x^2 - (40)^2 \\ 360 &= \sum x^2 - 16000\end{aligned}$$

$\therefore$ Incorrect $\sum x^2 = 16360$

Correct $\sum x^2 = 16360 - (30)^2 + (40)^2$

$$\begin{aligned}&= 16360 - 900 + 1600 \\ \sum x^2 &= 17060\end{aligned}$$

$\therefore$ Correct variance $\sigma^2 = \frac{1}{n} (\text{Correct } \sum x^2) - (\text{Correct } \bar{x})^2$

$$= \frac{1}{10} (17060) - (41)^2 = 1706 - 1681 = 25$$

(ii) Here $n = 100$, Incorrect mean $\bar{x} = 50$

Using $\bar{x} = \frac{\sum x}{n}$, Incorrect $\sum x = 50 \times 100 = 5000$

Thus, Correct $\sum x = 5000 - (40 + 50) + (60 + 30) = 5000$

$$\therefore \text{Correct mean} = \frac{\text{correct } \sum x}{n} = \frac{5000}{100} = 50$$

Incorrect variance $\sigma^2 = 5^2 = 25$

Using

$$\sigma^2 = \frac{1}{n} \sum (x - \bar{x})^2 = \frac{1}{n} \left(\sum x^2 - n\bar{x}^2 \right)$$

$$\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2$$

Deviations from the mean $|x - \bar{x}|$, are

$$nd, (n-1)d, (n-2)d, \dots, (n-2)d, (n-1)d, nd$$

$\therefore$ Mean Deviation from Mean is

$$\begin{aligned} M.D. &= \frac{nd + (n-1)d + (n-2)d + \dots + (n-2)d + (n-1)d + nd}{(2n+1)} \\ &= \frac{d \cdot 2[n + (n-1) + (n-2) + \dots + 1]}{2n+1} = \frac{d \cdot 2 \frac{n(n+1)}{2}}{2n+1} \\ &= \frac{n(n+1)d}{2n+1} \end{aligned}$$

The squares of the deviations from mean are

$$\begin{aligned} &(nd)^2 + [(n-1)d]^2 + [(n-2)d]^2 + \dots + [(n-2)d]^2 + [(n-1)d]^2 + (nd)^2 \\ \sigma^2 &= \frac{d^2 \cdot 2[n^2 + (n-1)^2 + (n-2)^2 + \dots + 2^2 + 1^2]}{2n+1} \\ &= \frac{d^2 \cdot 2n(n+1)(2n+1)}{6(2n+1)} = \frac{n(n+1)d^2}{3} \end{aligned}$$

$$\sigma = \sqrt{\frac{n(n+1)}{3}} \cdot d$$

$$\sigma > M.D. \text{ if } \sqrt{\frac{n(n+1)}{3}} > \frac{n(n+1)}{2n+1}$$

$$\text{or if } \frac{n(n+1)}{3} > \frac{n^2(n+1)^2}{(2n+1)^2}$$

$$\text{or if } (2n+1)^2 > 3n(n+1)$$

$$\text{or if } 4n^2 + 4n + 1 > 3n^2 + 3n$$

$$\text{or if } n^2 + n + 1 > 0 \text{ which is valid for every } n.$$

Hence, $\sigma > M$, for this series.

Example 11: For every discrete distribution, the standard deviation is always greater than or equal to the mean deviation taken from the mean.

Solution: Let $x_1, x_2, \dots, x_n$ be the values of the variate under study with respective frequencies $f_1, f_2, \dots, f_n$. Then we have to prove that

$$\begin{aligned} & S.D. \geq M.D. \cdot \bar{x} \quad (M.D. \cdot \bar{x} \text{ means } M.D. \text{ about mean } \bar{x}) \\ \text{i.e.,} & S.D. \geq M.D. \cdot \bar{x} \\ \text{or} & S.D.^2 \geq (M.D. \cdot \bar{x})^2 \\ \text{or} & \frac{1}{n} \sum_{i=1}^n f_i (x_i - \bar{x})^2 \geq \left[\frac{1}{n} \sum_{i=1}^n f_i |x_i - \bar{x}| \right]^2 \end{aligned}$$

Put

$$\begin{aligned} |x_i - \bar{x}| &= y_i \therefore (x_i - \bar{x})^2 = y_i^2 \\ \frac{1}{n} \sum f_i y_i^2 &\geq \left[\frac{1}{n} \sum f_i y_i \right]^2 \\ \frac{1}{n} \sum f_i y_i^2 - \bar{y}^2 &\geq 0 \end{aligned}$$

or

$$\frac{1}{n} \sum f_i (y_i - \bar{y})^2 \geq 0$$

7.8 COMBINED STANDARD DEVIATION

The s.d. of two series containing n_1 and n_2 members are σ_1 and σ_2 respectively, being measured from their respective means $\bar{x}_1$ and $\bar{x}_2$. If the two series are grouped together as one series of $(n_1 + n_2)$ members. Then the standard deviation, σ , of this series, measured from its mean $\bar{x}$ is given by

$$\sigma^2 = \frac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2$$

Proof: Let S_1^2 and S_2^2 be the mean square deviations of the two series respectively and S^2 be the mean square deviation of the two series taken together. Then if ' a ' be the assumed mean, we have

$$\begin{aligned}
 S^2 &= \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} f(x - a)^2 \\
 &= \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} f(x - a)^2 + \sum_{i=n_1+1}^{n_2} f(x - a)^2 \right] \\
 &= \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} \left[\text{as } S_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} f(x - a)^2 \text{ etc.} \right] \\
 &= \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{(n_1 + n_2)} \\
 &\quad [\text{where } S^2 = \sigma^2 + d^2 \text{ and } d = (\bar{x} - a)] \\
 &= \frac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{(n_1 + n_2)} = \frac{n_1 d_1^2 + n_2 d_2^2}{(n_1 + n_2)}
 \end{aligned}$$

Now

$$d_1 = \bar{x}_1 - a \text{ and } d_2 = \bar{x}_2 - a$$

If ' a ' be the mean of the two combined series i.e., if $a = \bar{x}$, then

$$S^2 = \sigma^2$$

Also

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\begin{aligned}
d_1 &= \bar{x}_1 - \bar{x} = \bar{x}_1 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{n_2(\bar{x}_1 - \bar{x}_2)}{(n_1 + n_2)} \\
d_2 &= \bar{x}_2 - \bar{x} = \bar{x}_2 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{n_1(\bar{x}_2 - \bar{x}_1)}{(n_1 + n_2)} \\
\therefore n_1d_1^2 + n_2d_2^2 &= \frac{n_1n_2^2(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2} + \frac{n_2n_1^2(\bar{x}_2 - \bar{x}_1)^2}{(n_1 + n_2)^2} \\
\therefore &= \frac{n_1n_2(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2} (n_2 + n_1) \\
\therefore \text{From (1)} &\frac{n_1n_2(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)} \\
(\because S^2 = \sigma^2) & \\
\sigma^2 &= \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1n_2}{(n_1 + n_2)} (\bar{x}_1 - \bar{x}_2)^2
\end{aligned}$$

Example 12: The means of two samples, consisting of 50 and 100 observations respectively, are 54.1 and 50.3. Their corresponding standard deviations are 8 and 7. Determine the combined mean and standard deviation for the merged sample of size 150.

Solution: We are given that, $n_1 = 50, n_2 = 100, \bar{x}_1 = 54.1, \bar{x}_2 = 50.3, \sigma_1 = 8$ and $\sigma_2 = 7$. If $\bar{x}$ is the mean of the combined series, then

$$\begin{aligned}
\bar{x} &= \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{50 \times (54.1) + 100(50.3)}{150} \\
&= \frac{2705 + 5030}{150} = \frac{7735}{150} = 51.57
\end{aligned}$$

If σ be the standard deviation of the combined series, then

$$\begin{aligned}
\sigma^2 &= \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1d_1^2 + n_2d_2^2}{n_1 + n_2} \\
d_1 &= \bar{x}_1 - \bar{x} = 54.1 - 51.57 = 2.53 \\
d_2 &= \bar{x}_2 - \bar{x} = 50.3 - 51.57 = -1.27 \\
\sigma^2 &= \frac{50 \times 64 + 100 \times 49}{150} + \frac{50 \times (2.53)^2 + 100 \times (-1.27)^2}{150} \\
&= \frac{3200 + 4900}{150} + \frac{320.045 + 161.29}{150} \\
&= \frac{3200 + 4900 + 320.045 + 161.29}{150} = \frac{8581.335}{150} \\
&= 57.2089; \sigma = 7.56
\end{aligned}$$

CHECK YOUR PROGRESS

1. What is meant by dispersion? Explain its importance in statistical analysis.

.....

.....

.....

2. Explain Range, Quartile Deviation, Mean Deviation, and Standard Deviation.

.....

.....

.....

3. Distinguish between absolute and relative measures of dispersion.

.....

.....

.....

7.9 SUMMARY

In this unit, we have discussed the following concepts

- Measures of dispersion
- Mean deviation or average deviation
- S.D. and RMSD.
- change of origin and change of scale effect on standard deviation
- Combined standard deviation

7.10 EXERCISES

Section A: Multiple Choice Questions (MCQs)

1. **Choose measure of dispersion.**
 - a) Mean
 - b) Median
 - c) Variance
 - d) Mode
2. **What is the formula for calculating the range of a data set?**
 - a) Highest value - Lowest value
 - b) Sum of values / Number of values
 - c) Highest value + Lowest value
 - d) Median – Mode
3. **What does the standard deviation measure?**
 - a) Central tendency
 - b) Spread of the data around the mean
 - c) The most frequent value
 - d) The midpoint of the data
4. **If all the data points are multiplied by a constant k , the standard deviation is:**
 - a) Unchanged
 - b) Divided by k
 - c) Multiplied by k
 - d) Multiplied by k^2
5. **What is the root mean square (RMS) deviation of the data set {2, 4, 4, 6}?**
 - a) 2
 - b) 3

c) 4

d) 5

Section B: Short Answer Questions

6. Calculate the range for the following data set:

15, 22, 30, 35, 40, 50

7. Find the mean deviation about the mean for the data set:

3, 7, 7, 10, 12

8. Calculate the s. d. for the following data set:

5, 10, 15, 20, 25

9. Explain the effect of multiplying each data point by 3 on the standard deviation.

10. Find the combined standard deviation for two groups:

- Group 1: $n_1 = 5$, mean = 10, standard deviation = 2
- Group 2: $n_2 = 7$, mean = 15, standard deviation = 3

Section C: Practical Problems

11. **Mean Deviation Calculation:** Calculate the mean deviation about the mean for the following data set: 2, 4, 6, 8, 10

12. **S.D. for Frequency Distribution:** Calculate the s. d. for the following frequency distribution:

Value (x)	Frequency (f)
2	3
4	5
6	2
8	4

13. **Root Mean Square (RMS) Deviation:** Calculate the root mean square deviation for the data set: 1, 3, 5, 7, 9
14. **Effect of Origin and Scale:** Given a data set with a standard deviation of 4, determine the new standard deviation if:
- Each data point is increased by 5.
 - Each data point is multiplied by 3.
15. **Combined Standard Deviation Calculation:** Two datasets have the following characteristics:
- **Dataset 1:** $n_1 = 6$, mean = 20, standard deviation = 4
 - **Dataset 2:** $n_2 = 4$, mean = 30, standard deviation = 5
- Calculate the combined mean and the combined standard deviation.

Answers and Guidelines for Instructors

Section A (MCQs)

1. c) Variance
2. a) Highest value - Lowest value
3. b) Spread of the data around the mean
4. c) Multiplied by k
5. b) 3

For Sections B and C, students should perform detailed calculations for mean deviation, standard deviation, RMS deviation, and combined standard deviation. These EXERCISESs are designed to reinforce key concepts related to **measures of dispersion** and **the effect of transformations on standard deviation**.

7.11 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 8 SKEWNESS AND KURTOSIS

STRUCTURE

- 8.1 Introduction
- 8.2 Objectives
- 8.3 Skewness
- 8.4 Measures of skewness based on moments
- 8.5 Kurtosis
- 8.6 Summary
- 8.7 EXERCISES
- 8.8 References

8.1 INTRODUCTION

In statistics, understanding how data is distributed is just as important as knowing its average or spread. **Skewness** and **kurtosis** are two measures that help describe the shape of a distribution.

Skewness indicates the asymmetry of data. A distribution can be symmetric, or skewed to the left or right, revealing important information about trends and outliers. **Kurtosis**, on the other hand, measures the "peaked ness" or flatness of a distribution compared to a normal curve, helping us understand the presence of extreme values.

This unit introduces the concepts, types, and formulas for skewness and kurtosis. It also includes graphical interpretation and solved examples to help learners apply these measures to real data sets effectively.

8.2 OBJECTIVES

Students who complete this unit will be capable of

- Understand the concepts of skewness and kurtosis and their role in data analysis.
- Distinguish between positively and negatively skewed distributions.
- Interpret kurtosis to evaluate the sharpness or flatness of a data distribution.
- Calculate skewness and kurtosis for a given data set.
- Apply these measures to assess the normality and shape of real-world data distributions.

8.3 SKEWNESS

By ‘skewness’ of a distribution we mean lack in symmetry or departure from symmetry. We know that a symmetric distribution has its mean, median and mode all equal and the frequency curve is symmetrically situated about the ordinate through the mean, or mode i.e., highest value of the ordinate. Thus, if a distribution is not symmetrical it is said to be skewed. When the distribution has longer tail on right hand side it is said to be positively skewed and when the distribution has longer tail on left hand side, it is said to be negatively skewed. It has been shown below by the different figures.

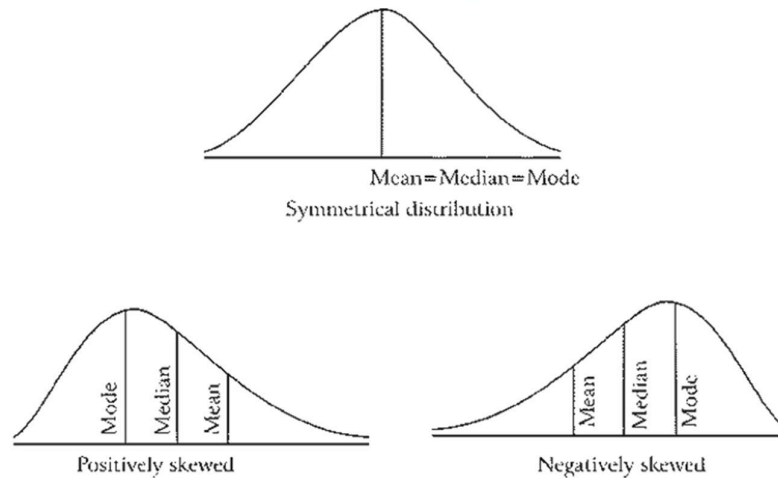


Figure 8.1 Negative, positive and symmetrical skewed distribution

In a moderately skewed distribution, an interesting relationship is located among mean, median and mode. In such type of distributions, it can be proved that the distance between mean and median is approximately one third of the distance between the mean and mode. This relationship can be expressed as follows:

$$\text{Mean} - \text{Median} > \frac{1}{3} (\text{Mean} - \text{Mode})$$

or

$$\text{Mode} \approx 3 \text{ Median} - 2 \text{ Mean}$$

It is clear from the above approximate relationship that if we know any two of the values of the averages, the third value of the average can be determined.

Measures of Skewness

There are four relative measures of skewness, namely,

- 1 Measures of skewness based on moments
- 2 Karl Pearson's coefficient of skewness

3 Bowley's coefficient of skewness

4 Kelly's coefficient of skewness

These measures are pure numbers independent of the units of measurement and can be used to compare two distributions with regards to symmetry.

Let us discuss these measures one by one.

8.4 MEASURES OF SKEWNESS BASED ON MOMENTS

Karl Pearson defined skewness based on moments as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ where } \mu_2 \text{ and } \mu_3 \text{ are 2nd and 3rd moments about mean.}$$

β_1 measures skewness and

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} \text{ we denote } \sqrt{\beta_1} \text{ by } \gamma_1$$

These two coefficients are based on two central moments μ_2 and μ_3 . A distribution is symmetric if

$$\beta_1 = 0 \Rightarrow \mu_3 = 0$$

The greater the value of β_1 , the more skewed is the distribution. Thus, this measure of skewness enables us to know the extent of symmetry in a distribution.

However, the direction of skewness i.e., positive or negative is determined by γ_1 . If $\mu_3 > 0$, the skewness is positive and if $\mu_3 < 0$ then the skewness is negative.

Karl Pearson's Coefficient of Skewness

Karl Pearson's relative measures of skewness is given by the formula:

$$\text{Coefficient of Skewness (S.Kp)} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

This formula cannot, however, be used when the mode is ill-defined. In such cases, we use the following value of the mode for the moderately skewed distribution.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Substituting this value of mode in (1) we get the formula:

$$\text{Coefficient of skewness (S. K}_p\text{)} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

From (1) and (2), S. K_p = 0 if mean = mode = Median

Bowley's Coefficient of Skewness

Bowley suggested the following measure of skewness based on first quartile Q_1 and third quartile Q_3 .

$$\begin{aligned} \text{Coefficient of skewness (S. K}_B\text{)} &= \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{(Q_3 - \text{Median}) + (\text{Median} - Q_1)} \\ &= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \end{aligned}$$

Thus, S. K_B = 0 if $Q_3 - \text{Median} = \text{Median} - Q_1$

Hence, in a symmetrical distribution,

$$\begin{aligned} Q_3 - \text{Median} &= \text{Median} - Q_1 \\ \Rightarrow Q_3 + Q_1 - 2 \text{ Median} &= 0 \end{aligned}$$

Kelly's Coefficient of Skewness

Bowley's coefficient of skewness can be extended by considering two deciles or two percentiles that are located at equal distances on either side of the median. This approach incorporates a wider range of observations than the quartile-based measure. **Kelly** proposed a coefficient of skewness based on

the 10th and 90th percentiles, which are symmetrically positioned around the median. The measure can also be expressed in terms of the first and ninth deciles.

90th percentiles or the first and the 9th deciles. Thus, Kelly's coefficient of skewness ($S.K_K$)

$$= \frac{P_{10} + P_{90} - 2 \text{ Median}}{P_{90} - P_{10}}$$

$$= \frac{D_1 + D_9 - 2 \text{ Median}}{D_9 - D_1}$$

However, this method is not popular in practice. Generally, Karl Pearson's method is used.

8.5 KURTOSIS

Kurtosis is an important characteristic used to describe the shape of a frequency distribution. It indicates the extent to which the frequency curve is relatively peaked or flat around its central region.

Two distributions may have the same mean and variance but still differ in the shape and sharpness of their peaks. Therefore, even when a distribution is symmetrical, its degree of peakedness may vary considerably.

The degree of kurtosis is assessed by comparing the shape of a distribution with that of the normal distribution, which is taken as the standard reference. In this way, kurtosis indicates whether a frequency curve is more sharply peaked or more flat-topped than the normal curve.

- 1 If a curve has much flatness than the normal curve, it is said to be 'Leptokurtic'.
- 2 If a curve has lesser flatness than the normal curve, it is said to be 'Platykurtic'.
- 3 If a curve is equally flat as normal curve, it is said to be 'Mesokurtic'.

The degree of peakedness is known as excess or kurtosis. The three types of curves with different shapes are shown below:

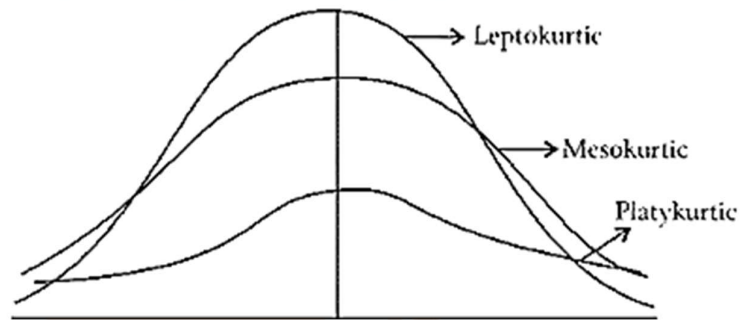


Figure 8.2 Types of curves with different shapes

Measures of Kurtosis

Karl Pearson defined a coefficient known as β_2 as follows:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where μ_2 and μ_4 are respectively the second and fourth moments about mean. The value of β_2 measures the kurtosis. The greater the value of β_2 , the more peaked is the distribution.

For a normal curve, $\beta_2 = 3$. When $\beta_2 > 3$, compared to the normal curve, the curve is more peaked and is called leptokurtic. When $\beta_2 < 3$, compared to the normal curve, the curve is less peaked and is called platykurtic. Hence, when we measure kurtosis, we should find whether the calculated value of β_2 is greater than or less than 3. The measure of β_2 is dimensionless, since both numerator and denominator have the same dimension viz., 4.

Another measure γ_2 is also suggested. It is defined as

$$\gamma_2 = \beta_2 - 3$$

$\beta_2 - 3$ or γ_2 is called the excess of Kurtosis or simply excess, for a normal curve $\gamma_2 = 0$ and so the curve is mesokurtic. If $\gamma_2 > 0$, then the curve is leptokurtic and if $\gamma_2 < 0$, then the curve is platykurtic.

Example 1: Find the skewness and kurtosis (about mean) if first four moments of the distribution about $x = 2$ are 1.0, 2.5, 5.5 and 16.0.

Solution: We are given that $\mu'_1 = 1.0, \mu'_2 = 2.5, \mu'_3 = 5.5$ and $\mu'_4 = 16.0$

$$\begin{aligned}\therefore \mu'_1 &= \frac{1}{N} \sum_i f_i(x_i - 2) = 1.0 \\ \Rightarrow \frac{1}{N} \sum_i f_i x_i - \frac{2}{N} \sum_i f_i &= 1.0 \\ \Rightarrow \bar{x} - 2 &= 1 \\ \Rightarrow \bar{x} &= 3\end{aligned}$$

Thus, mean of the distribution is 3.

Also,

$$\begin{aligned}2.5 &= \frac{1}{N} \sum_i f_i(x_i - 2)^2 \\ &= \frac{1}{N} \sum_i f_i[(x_i - 3) + 1]^2 \\ &= \frac{1}{N} \sum_i f_i(x_i - 3)^2 + \frac{2}{N} \sum_i f_i(x_i - 3) + \frac{1}{N} \sum_i f_i \cdot 1 \\ &= \mu_2 + \frac{2}{N} \sum_i f_i x_i - \frac{6}{N} \sum_i f_i + 1 \\ 2.5 &= \mu_2 + 2 \times 3 - 6 + 1 \\ \mu_2 &= 2.5 - 1 \\ \mu_2 &= 1.5 \\ 5.5 &= \frac{1}{N} \sum_i f_i(x_i - 2)^3 \\ &= \frac{1}{N} \sum_i f_i[(x_i - 3) + 1]^3\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N} \sum_i f_i [(x_i - 3)^3 + 3(x_i - 3)^2 + 3(x_i - 3) + 1] \\
&= \frac{1}{N} \sum_i f_i (x_i - 3)^2 + \frac{3}{N} \sum_i f_i (x_i - 3) + \frac{3}{N} \sum_i f_i (x_i - 3) + \frac{1}{N} \sum_i f_i \\
&= \mu_3 + 3\mu_2 + 3\bar{x} - \frac{9}{N} \sum_i f_i + \frac{1}{N} \sum_i f_i \\
&= \mu_3 + 3\mu_2 + 3\bar{x} - 8 \\
&= \mu_3 + 3 \times 1.5 + 3 \times 3 - 8 \\
&= \mu_3 + 4.5 + 1
\end{aligned}$$

$$\mu_3 = 5.5 - 5.5 = 0$$

Again,

$$\begin{aligned}
16.0 &= \frac{1}{N} \sum_i f_i (x_i - 2)^4 = \frac{1}{N} \sum_i f_i [(x_i - 3) + 1]^4 \\
&= \frac{1}{N} \sum_i f_i [(x_i - 3)^4 + 4(x_i - 3)^3 + 6(x_i - 3)^2 + 4(x_i - 3) + 1] \\
&= \frac{1}{N} \sum_i f_i (x_i - 3)^4 + \frac{4}{N} \sum_i f_i (x_i - 3)^3 + \frac{6}{N} \sum_i f_i (x_i - 3)^2 \\
&\quad + \frac{4}{N} \sum_i f_i x_i - \frac{4 \times 3}{N} \sum_i f_i + \frac{1}{N} \sum_i f_i \\
&= \mu_4 + 4\mu_3 + 6\mu_2 + 4\bar{x} - 2 \\
&= \mu_4 + 4 \times 0 + 6 \times 1.5 + 4 \times 3 - 11 \\
&= \mu_4 + 9.0 + 12 - 11 \\
\mu_4 &= 16.0 - 10 = 6
\end{aligned}$$

Thus, $\mu_1 = 0, \mu_2 = 1.5, \mu_3 = 0$ and $\mu_4 = 6$

$$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0,$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{6}{(1.5)^2} = \frac{6}{2.25} = 2.67$$

$$\gamma_2 = \beta_2 - 3 = 2.67 - 3 = -ve$$

$$\gamma_1 = \sqrt{(\beta_1)} = 0,$$

Since β_1 is zero, we can say that distribution is symmetrical and there is no skewness in the distribution. However, γ_2 is negative we say that the curve of the distribution is platykurtic.

Example 2: Calculate μ_2, μ_3 and μ_4 from the following data and determine whether the distribution is platykurtic, leptokurtic or mesokurtic.

Marks:	0	10	20	30	40	50	60
	– 10	– 20	– 30	– 40	– 50	– 60	– 70
No. of students:	8	12	20	30	15	10	5

Solution: We prepare the following table for the calculation of μ'_1, μ'_2, μ'_3 and μ'_4 about say 35.

Mar ks	No. of students (f_i)	$M.V. (x_i)$	u_i $= \frac{x_i - 35}{10}$	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
0 – 10	8	5	-3	-24	72	-216	648
10 – 20	12	15	-2	-24	48	-96	192
20 – 30	20	25	-1	-20	20	-20	20
30 – 40	30	35	0	0	0	0	0
40 – 50	15	45	+1	+15	15	+15	15
50 – 60	10	55	+2	+20	40	+80	160
60 – 70	5	65	+3	+15	45	+135	405
	$N = 100$			$\Sigma f_i u_i$	$\Sigma f_i u_i^2$	$\Sigma f_i u_i^3$	$\Sigma f_i u_i^4$
				$= -18$	$= 240$	$= -10$	$= 1440$

Now

$$\begin{aligned}
 \mu'_r &= \frac{1}{N} \sum_i f_i (x_i - 35)^r = \frac{1}{N} \sum_i f_i (10u_i)^r \\
 &= \frac{10^r}{N} \sum_i f_i u_i^r
 \end{aligned}$$

Hence,

$$\begin{aligned}\mu'_1 &= \frac{10}{100} \sum_i f_i u_i = \frac{1}{10} \times (-18) = -1.8 \\ \mu'_2 &= \frac{10^2}{100} \sum_i f_i u_i^2 = \frac{100}{100} \times 240 = 240 \\ \mu'_3 &= \frac{10^3}{100} \sum_i f_i u_i^3 = 10 \times (-102) = -1020 \\ \mu'_4 &= \frac{10^4}{100} \sum_i f_i u_i^4 = 100 \times 1440 = 144000\end{aligned}$$

We now calculate the central moments.

$$\begin{aligned}\mu_2 &= \mu'_2 - \mu_1'^2 \\ &= 240 - (1.8)^2 = 240 - 3.24 = 236.76 \\ \mu_3 &= \mu'_3 - 3\mu'_1\mu'_2 + 2\mu_1'^3 = -1020 + 3 \times 1.8 \times 240 + 2 \times (-1.8)^3 \\ &= -1020 + 1296 - 11.664 = 264.336 \\ \mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6\mu_1'^2\mu'_2 - 3\mu_1'^4 \\ &= 144000 - 4 \times 1.8 \times 1020 - 6(-1.8)^2 \times 240 - 3(-1.8)^4 \\ &= 144000 - 7344 + 4665.5 - 31.4928 \\ &= 141290.11\end{aligned}$$

$$\text{Hence, } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{141290.11}{(236.76)^2} = \frac{141290.11}{56055.2976} = 2.52 \text{ (nearly)}$$

Since $\beta_2 < 3$, hence $\gamma_2 = \beta_2 - 3 < 0$

Hence, the distribution is platykurtic.

Example 3: For a symmetric distribution $Q_1 = 30$ and $Q_3 = 50$, find its median and 50th percentile.

Solution: Since distribution is symmetric, the skewness will be zero, and thus by Bowley's coefficient of skewness we have

$$\begin{aligned}S_{K.B.} &= \frac{(Q_1 + Q_3) - 2 \text{ Median}}{Q_3 - Q_1} = 0 \\ \Rightarrow (Q_1 + Q_3) - 2 \text{ Median} &= 0 \Rightarrow (30 + 50) = 2 \text{ Median} \\ \Rightarrow \text{Median} &= \frac{80}{2} = 40\end{aligned}$$

Since 50th percentile is that value which divides the distribution in two parts and median is also the value which divides the distribution in two equal parts. Thus $P_{50} = Q_2 = \text{Median} = 40$.

CHECK YOUR PROGRESS

1. Define skewness and explain its different types.

.....

.....

.....

2. Explain Karl Pearson's and Bowley's measures of skewness.

.....

.....

.....

3. What is kurtosis? Explain the different types of kurtosis.

.....

.....

.....

8.6 SUMMARY

In this unit, we have discussed the following concepts

- Skewness
- Measures of skewness based on moments
- Kurtosis

8.7 EXERCISES

1. Explain the concepts of dispersion and skewness in statistics. Describe the different methods commonly used to measure dispersion and skewness.
2. Define mean deviation, standard deviation, and coefficient of variation. Using the first n natural numbers, calculate their coefficient of variation.
3. What are mean deviation, standard deviation, and coefficient of variation? Prove that the standard deviation represents the minimum root mean square deviation. Why is standard deviation generally preferred to mean deviation as a measure of dispersion?
4. Define mean deviation and standard deviation. Demonstrate that the mean deviation is minimum when deviations are measured from the median. Also prove that the root mean square deviation attains its minimum value when deviations are taken from the arithmetic mean.
5. Explain the major measures of dispersion. Prove that the absolute mean deviation about the arithmetic mean remains unchanged under a change of origin.
6.
 - (a) Demonstrate that variance is independent of a change in origin.
 - (b) Prove that the mean deviation is minimum when it is calculated from the median.
7. Distinguish between range and quartile deviation. Using the following frequency distribution, calculate the quartile deviation:

Class	0	40	80	120	160	200	240
Interval:	– 40	– 80	– 120	– 160	– 200	– 240	– 280
Frequency	394	461	391	334	169	113	148
:							

- 8 (a) Explain why the standard deviation is regarded as superior to other measures of dispersion. What is its chief defect?

(b) Show that mean deviation is least when taken from median.

[B.Sc.2002(Back)]

- 9 Describe use of moments in determining the characteristics of a frequency distribution.
- 10 Explain the terms 'skewness' and 'kurtosis' of a frequency distribution. Describe the different measures of 'skewness' and 'kurtosis'.
- 11 Define moments of a frequency distribution. Show, how would you obtain first four central moments from those about origin?
- 12 Describe the theory of moments of a frequency distribution.
- 13 What is skewness? What purpose does it serve? Discuss Pearson's coefficients of skewness, quartile and percentile coefficient of skewness and moment coefficient of skewness.
- 14 What do you understand by the moments of a frequency distribution? Why only the four moments are generally calculated for a frequency distribution?
- 15 What do you understand by skewness? How is it measured? Distinguish clearly by giving figures between positive and negative skewness.
- 16 What do you mean by moments about mean and about any point? Calculate coefficient of skewness and kurtosis based on moments about mean from the following data:

$x:$	4	14	24	34	44	54	64	74	84	94
$f:$	1	5	12	22	17	9	4	3	1	1

17. What are moments? How are they used in statistics? Find the relation between moments about mean and moments about an arbitrary constant.
18. Write short notes on:

- (a) Charlier Check
- (b) Skewness
- (c) Sheppard's correction
- (d) Coefficient of variation
- (e) Skewness and Kurtosis
- (f) Kurtosis

8.8 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., McKean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

BLOCK III: ELEMENTARY PROBABILITY, RANDOM VARIABLES AND DISTRIBUTION

UNIT 9 ELEMENTARY PROBABILITY

STRUCTURE

- 9.1 Introduction
- 9.2 Objectives
- 9.3 Basic definitions and notations
- 9.4 Probability
- 9.5 Conditional probability
- 9.6 Bays theorem
- 9.7 Summary
- 9.8 Exercises
- 9.9 References

9.1 INTRODUCTION

Probability is the mathematical study of uncertainty and randomness. It provides a framework for analysing situations where outcomes are uncertain, allowing us to quantify the likelihood of different events. In real-world scenarios, such as predicting weather, understanding stock market trends, or making decisions in games of chance, probability plays a critical role. This unit introduces the foundational concepts of probability, which are essential tools in both theoretical and applied contexts.

Probability is at the core of many disciplines, including statistics, data science, engineering, economics, and artificial intelligence. By mastering the basics of probability, students can analyse uncertainty, make informed predictions, and solve problems involving randomness. This unit serves as a foundation for more advanced topics, such as stochastic processes and statistical inference.

Through examples and practical applications, this unit will equip you with the tools to understand and work with uncertainty effectively.

9.2 OBJECTIVES

After completing this unit, students will be able to

- Define sample spaces, events, and apply the axioms and rules of probability.
- Use permutations and combinations to solve probability problems.
- Compute probabilities of dependent events and identify independent events.
- Determine expected values, variances, and interpret their significance.
- Apply probability concepts to practical scenarios in various fields.

9.3 BASIC DEFINITIONS AND NOTATIONS

9.3.1. Random experiment: A random experiment is any experiment with unpredictable or unpredictable results.

OR

A random experiment is an experiment whose outcomes or result is not unique and therefore cannot be predicted with certainty.

Tossing of a coin, head or tail may occur, throwing a die, 1, 2, 3, 4, 5 or 6 may appear and drawing a card from a well shuffled pack of cards are examples of random experiments.

9.3.2. Sample space and sample point: A sample space is the collection of all potential results from a random experiment, and a sample point is an element of a sample space.

If a die is thrown, it will land with any one of its 6 faces pointing upward, resulting in any one of the numbers 1,2,3,4,5,6 appearing on the top face. Hence, the number of each face is a possible result. We write

$$S = \{1,2,3,4,5,6\}$$

The set S is a sample space of throwing a die and 1,2,3,4,5,6 are called sample point.

When two dice are thrown, then the sample space consists of the collection of 36 points:

$$S = \begin{bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{bmatrix}$$

Note. Extensive events or sample space is the collection of all potential results from a single run of a random experiment.

9.3.3. Event: Any subset A of a sample space S is called an event.

Consider the experiment of tossing of a coin, we have

$$S = \{H, T\}$$

Here are 4 subsets of S , ϕ , $\{H\}$, $\{T\}$, $\{H, T\}$. Each subset of S is an event.

9.3.4. Simple event: If an event just has one sample point, it is considered simple. In tossing of a coin, the events $\{H\}$ and $\{T\}$ are simple events.

9.3.5. Compound event: If an event contains many sample points, it is referred to as a compound event. When two coins are tossed once then sample space $(S) = \{HH, HT, TH, TT\}$. Each event of S is compound event.

9.3.6. Complement of an event: The complement of an event A with respect to the sample space S is the set of all elements of S which are not in A .

$$\bar{A} = S - A$$

9.3.7. Favourable events: The total number of outcomes that lead to the occurrence of an event in a trial is known as the number of favourable outcomes for that event. For example, when two dice are rolled, the favourable outcomes for obtaining a sum of 6 are: (1,5), (2,4), (3,3), (4,2), and (5,1) — giving a total of 5 favourable outcomes.

9.3.8. Equally likely events: If there is no reason to favor one event over another, then all events are considered equally likely. For example, Each, of the six faces has an equal chance of appearing when a die is thrown.

9.3.9. Mutually exclusive events: Two events A and B are said to be mutually exclusive events iff $A \cap B = \phi$.

If a die is thrown, we will get either an 'even number' or an 'odd number'. The event A "getting an even number" and event B "getting an odd number". We say that A and B are mutually exclusive events because both the events cannot occur at the same time. We have

$$A = \{2,4,6\} \text{ and } B = \{1,3,5\}$$

Then,

$$A \cap B = \phi$$

In tossing of a coin, we will get either a head or tail. The event A "getting a head" and event B "getting a tail". We say that A and B are mutually exclusive events because both the events cannot occur at the same time.

We have

$$A = \{H\} \text{ and } B = \{T\}$$

Then,

$$A \cap B = \phi$$

Note. 1. Simple events of a sample space are always mutually exclusive.
2. The events, which ensure the required happening is called favourable events.

9.3.10. Odd in favour of an event and odd against an event: Let there are m outcomes favourable to a certain event and n outcomes are not favourable to the event in a sample space, then

$$\text{odd in favour of the event} = \frac{m}{n}$$

$$\text{and odd against the event} = \frac{n}{m}.$$

9.3.11. Permutation: A permutation is an arrangement of objects in a definite order. The number of permutations of n objects used r at a time, denoted by ${}^n P_r$ is

$${}^n P_r = \frac{n!}{(n-r)!}$$

9.3.12. Combination: A combination is a selection of objects without regard to order. The number of combinations of n distinct objects selected r at a time, denoted by ${}^n C_r$ is

$${}^n C_r = \frac{n!}{r!(n-r)!}$$

9.4 PROBABILITY

The chance of happening of an event when expressed quantitatively is called probability. However, we give three definitions of probability:

- (1) Classical or Mathematical definition of probability.
- (2) Statistical or Empirical definition of probability.
- (3) Axiomatic definition of probability.

9.4.1. Classical or Mathematical definition of probability: If an event A can happen m ways out of possible n mutually exclusive and equally likely outcomes of a random experiment then probability of A is

$$P(A) = \frac{\text{Number of favourable cases}}{\text{Total numbers of possible cases}} = \frac{m}{n}$$

The probability of not happening of A is

$$P(\bar{A}) = 1 - \frac{m}{n} = \frac{n - m}{n}$$

Thus, $P(A) + P(\bar{A}) = 1, 0 \leq P(A) \leq 1$ and $0 \leq P(\bar{A}) \leq 1$.

Note. **1.** If $P(A) = 1$, then A is called certain event.

2. If $P(A) = 0$, then A is called impossible event.

3. This definition is fail when outcomes are not equally likely and number of outcomes is infinite.

4. An event's probability is a number that ranges from 0 to 1. If occurrence is certain, its probability is 1. If the event not possible then its probability is 0.

9.4.2. Statistical or Empirical definition of probability: If trial be repeated for a large number of times, say n , under the same conditions, and a certain event A happen on m times then the probability of the event A is

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}$$

where, the limit is unique and finite.

9.4.3. Axiomatic definition of probability: Let S be a finite sample space and A be any event in S , then probability of A is defined by following three conditions.

- (i) For every event A in S , $0 \leq P(A) \leq 1$
- (ii) $P(S) = 1$.
- (iii) If A and B are mutually exclusive events in S , then

$$P(A \cup B) = P(A) + P(B)$$

9.4.4. ELEMENTARY THEOREMS ON PROBABILITY

Theorem 1: If ϕ is an impossible event then $P(\phi) = 0$.

Proof: Since, impossible event contains no sample point, therefore, the certain event S and the impossible event ϕ are mutually exclusive

$$\begin{aligned} \Rightarrow S \cup \phi &= S \\ \Rightarrow P(S \cup \phi) &= P(S) \\ \Rightarrow P(S) + P(\phi) &= P(S) \\ \Rightarrow P(\phi) &= 0. \end{aligned}$$

Theorem 2: The probability of the complementary event $\bar{A}$ of A is given by

$$P(\bar{A}) = 1 - P(A)$$

Proof: Since, A and $\bar{A}$ are mutually disjoint events, so that

$$\begin{aligned} A \cup \bar{A} &= S \\ \Rightarrow P(A \cup \bar{A}) &= P(S) \\ \Rightarrow P(A) + P(\bar{A}) &= P(S) \\ &= 1 \\ \Rightarrow P(\bar{A}) &= 1 - P(A) \end{aligned}$$

Note. If ϕ is the impossible event and S is the sample space then $\bar{S} = \phi$

$$\begin{aligned}\Rightarrow P(\bar{S}) &= P(\phi) \\ P(\bar{S}) &= 1 - P(S) \\ &= 1 - 1 = 0 \\ \Rightarrow P(\phi) &= 0\end{aligned}$$

Theorem 3: For any two events A and B .

- (i) $P(A \cap \bar{B}) = P(A) - P(A \cap B)$
- (ii) $P(\bar{A} \cap B) = P(B) - P(A \cap B)$

Proof: (i) From the Venn diagram, we have

$$A = (A \cap B) \cup (A \cap \bar{B})$$

where, $A \cap B$ and $A \cap \bar{B}$ are disjoint events.

$$\begin{aligned}\Rightarrow P(A) &= P(A \cap B) + P(A \cap \bar{B}) \\ \Rightarrow P(A \cap \bar{B}) &= P(A) - P(A \cap B).\end{aligned}$$

(ii) From the Venn diagram, we have

$$B = (A \cap B) \cup (\bar{A} \cap B)$$

where, $A \cap B$ and $\bar{A} \cap B$ are disjoint events.

$$\begin{aligned}\Rightarrow P(B) &= P(A \cap B) + P(\bar{A} \cap B) \\ \Rightarrow P(\bar{A} \cap B) &= P(B) - P(A \cap B)\end{aligned}$$

Theorem 4: If $B \subset A$, then

- (i) $P(A \cap \bar{B}) = P(A) - P(B)$
- (ii) $P(B) \leq P(A)$
- (iii) $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$.

Proof: (i) If $B \subset A$, then B and $A \cap \bar{B}$ are mutually exclusive events, so that

$$\begin{aligned}
 A &= B \cup (A \cap \bar{B}) \\
 \Rightarrow \\
 \Rightarrow P(A) \\
 \Rightarrow P(A \cap \bar{B}) \\
 &= P(A) - P(B) + P(B)
 \end{aligned}$$

(ii) If $B \subset A$, then by (1), we have

$$\begin{aligned}
 P(A \cap \bar{B}) &= P(A) - P(B) \\
 \Rightarrow P(A \cap \bar{B}) &\geq 0 \\
 \Rightarrow P(A) - P(B) &\geq 0 \\
 \Rightarrow P(A) &\geq P(B)
 \end{aligned}$$

or

$$P(B) \leq P(A).$$

(iii) If $B \subset A$, then by (2), we have

$$P(B) \leq P(A)$$

Since, $A \cap B \subset A$ and $A \cap B \subset B$

$$\Rightarrow P(A \cap B) \leq P(A) \text{ and } P(A \cap B) \leq P(B).$$

Theorem 5: If A and B are two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof: From the Venn diagram, we have

$$A \cup B = A \cup (\bar{A} \cap B)$$

where, A and $\bar{A} \cap B$ are mutually disjoint events.

$$\begin{aligned}
 \Rightarrow P(A \cup B) &= P(A) + P(\bar{A} \cap B) \\
 &= P(A) + P(B) - P(A \cap B) [\because P(\bar{A} \cap B) = P(B) - P(A \cap B)] \dots (1)
 \end{aligned}$$

Note. 1. If A and B are disjoint events, then

$$P(A \cap B) = P(\phi) = 0$$

Then by (1), we have

$$P(A \cup B) = P(A) + P(B)$$

- 2 $P(A \cup B)$ is also written as $P(A + B)$ and $P(A \cap B)$ is also written as $P(AB)$.

Theorem 6: If A, B and C are any three events, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C).$$

Proof: Using the above theorem 1 for two events, we have

$$\begin{aligned} P(A \cup B \cup C) &= P[A \cup (B \cup C)] \\ &= P(A) + P(B \cup C) - P[A \cap (B \cup C)] \\ &= P(A) + [P(B) + P(C) - P(B \cap C)] - P[(A \cap B) \cup (A \cap C)] \\ &= P(A) + P(B) + P(C) - P(B \cap C) - [P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)] \\ &= P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C) \end{aligned}$$

Theorem 7: If $A_1, A_2, A_3, \dots, A_n$ are n mutually exclusive events, The likelihood that one of them will occur is

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n) \\ = P(A_1) + P(A_2) + P(A_3) + \dots + P(A_n). \end{aligned}$$

Proof: Let N be the total number of mutually exclusive, exhaustive and equally likely cases of which m_1 are favourable to A_1 , m_2 are favourable to A_2 , m_3 are favourable to A_3 and so on.

$$\text{Probability of happening of event } A_1 = P(A_1) = \frac{m_1}{N}.$$

$$\text{Probability of happening of event } A_2 = P(A_2) = \frac{m_2}{N}.$$

$$\text{Probability of happening of event } A_n = P(A_n) = \frac{m_n}{N}$$

$\therefore$ Probability of happening of one of the events $A_1, A_2, A_3, \dots, A_n$ is

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n) &= \frac{m_1}{N} + \frac{m_2}{N} + \frac{m_3}{N} + \dots + \frac{m_n}{N} \\ &= P(A_1) + P(A_2) + P(A_3) + \dots + P(A_{tr}). \end{aligned}$$

Example 1: Find the probability a randomly chosen leap year have 53 Sundays.

Solution: We know that, there are 366 days in a leap year, which is equal to 52 full weeks and 2 extra days.

Combinations of these two days are as follows:

- (1) Monday, Tuesday
- (2) Tuesday, Wednesday
- (3) Wednesday, Thursday
- (4) Thursday, Friday
- (5) Friday, Saturday
- (6) Saturday, Sunday
- (7) Sunday, Monday

Of these seven $P(S) = 7$ are likely cases only and last two are favourable i.e. $n(A) = 2$.

The required probability $= \frac{n(A)}{n(S)} = \frac{2}{7}$.

Example 2: A single card is randomly chosen from a standard deck of 52 playing cards. Find the probability that the card will be a king.

Solution: From a pack of 52 cards 1 card can be drawn in 52 ways, i.e. $n(S) = 52$.

Number of ways in which a card can be king = 4
i.e.

$$n(A) = 4$$

The required probability = $\frac{n(A)}{n(S)} = \frac{4}{52} = \frac{1}{13}$.

Example 3: What is the likelihood of rolling a number exceeding 3 on a standard die numbered from 1 to 6?

Solution: There are 6 possible ways in which the die can fall, and all of these there (4,5,6) are favourable to the event required
i.e.

$$\begin{aligned} n(S) &= 6 \text{ and } n(A) = 3 \\ \Rightarrow P(> 3) &= \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}. \end{aligned}$$

Example 4: Three cards are chosen at random from a standard deck of 52 playing cards. What is the probability that the cards drawn are a king, a queen, and a jack(knave)?

Solution: From a deck of 52 cards a draw of 3 can be made in ${}^{52}C_3$ ways.
Thus,

$$n(S) = {}^{52}C_3$$

In a pack of 52 cards are 4 kings, 4 queens and 4 knaves. A king can be drawn in ${}^4C_1$ ways, a queen in ${}^4C_1$ ways and a knave in ${}^4C_1$ ways.
Each of these ways be withdrawn in ${}^4C_1 \times {}^4C_1 \times {}^4C_1$ ways.

Thus,

$$\begin{aligned} n(A) &= {}^4C_1 \times {}^4C_1 \times {}^4C_1 \\ P(A) &= \frac{n(A)}{n(S)} = \frac{{}^4C_1 \times {}^4C_1 \times {}^4C_1}{{}^{52}C_3} \\ &= \frac{4 \times 4 \times 4 \times 1 \times 2 \times 3}{52 \times 51 \times 50} = \frac{16}{5525} \end{aligned}$$

Example 5: In a given race the odds in favour of four horses, A, B, C, D are 1: 3, 1: 4, 1: 5, 1: 6 respectively. Assuming that a dead heat is impossible, find the chance that one of them wins the race.

Solution: Let p_1, p_2, p_3 and p_4 be the probabilities of winning of the horses A, B, C and D respectively, then

$$P(A) = p_1 = \frac{1}{4}, P(B) = p_2 = \frac{1}{5}$$

$$P(C) = p_3 = \frac{1}{6}, P(D) = p_4 = \frac{1}{7}$$

Since, these are mutually exclusive events the chance that one of them wins

$$= P(A) + P(B) + P(C) + P(D)$$

$$= p_1 + p_2 + p_3 + p_4$$

$$= \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} = \frac{319}{420}.$$

Example 6: If from a lottery of 30 tickets, marked 1, 2, 3, ... four tickets be drawn, what is the chance that marked 1 and 2 are among them.

Solution: We have, $n(S) = {}^{30}C_4$ (four tickets can be selected)

When 2 tickets are to be included always, remaining two can be selected in ${}^{28}C_2$ ways, i.e. $n(A) = {}^{28}C_2$

The required chance $P(A) = \frac{n(A)}{n(S)} = \frac{{}^{28}C_2}{{}^{30}C_2} = \frac{2}{145}.$

Example 7: A manufacturer supplies cheap quarter horse power motors in a lot of 25. A buyer before taking a lot, tests a random sample of 5 motors and accepts the lot if they are all good, otherwise he rejects the lot. Find the probability that

(i) he will accept a lot containing 5 defective motors.

(ii) he will reject a lot containing only one defective motor.

Solution: The buyer can choose a random sample of 5 motors out of 25 in ${}^{25}C_5$ ways.

(i) The buyer will accept a lot, if in his sample all the motors are good. That means his sample consists of 5 motors from 20 non-defective motors. So the number of ways of selecting the sample for acceptance of the lot, will be ${}^{20}C_5$

$$\begin{aligned}\text{The required probability} &= \frac{{}^{20}C_5}{{}^{25}C_5} = \frac{\frac{20!}{5!15!}}{\frac{25!}{5!20!}} \\ &= \frac{20 \times 19 \times 18 \times 17 \times 16}{25 \times 24 \times 23 \times 22 \times 21} = \frac{2584}{8855} = 0.292\end{aligned}$$

(ii) When he is rejecting a lot containing one defective motor, so his sample will contain the only defective motor and 4 others which are chosen from the 24 non-defective motors. So, number of ways of selecting this sample is

$$= {}^{24}C_4 \times {}^1C_1$$

$$\text{The required probability} = {}^{24}C_4 = \frac{\frac{24!}{4!20!}}{\frac{25!}{5!20!}} = \frac{5}{5!20!} = \frac{1}{5} = 0.2$$

9.4.5. INDEPENDENT EVENTS

Two events A and B are said to be independent events iff $P(A \cap B) = P(A) \cdot P(B)$.

Example 8: When a die is thrown, let A be the event "number appearing is a multiple of 3 " and B be the event "number appearing is even". We have

$$A = \{3,6\}, B = \{2,4,6\}$$

and

$$P(A) = \frac{2}{6} = \frac{1}{3}$$

$$P(B) = \frac{3}{6} = \frac{1}{2}$$

$$A \cap B = \{6\}$$

$$P(A \cap B) = \frac{1}{6}$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$\frac{1}{6} = \frac{1}{3} \times \frac{1}{2}$$

Hence, A and B are independent events.

9.5 CONDITIONAL PROBABILITY

Let A and B be events such that $P(B) \neq 0$. The conditional probability of A given B , denoted by $P(A/B)$, is defined by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

where, $P(A/B)$ is the probability of occurrence of A given that B has already happened.

Theorem 1: If the two events A and B are independent events, then

$$P(A/B) = P(A) \text{ and } P(B'/A) = P(B)$$

Proof: We know that if A and B are given to be independent events, then

$$P(A \cap B) = P(A) \cdot P(B)$$

we have,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B)$$

9.5.1 MULTIPLICATION THEOREM OF PROBABILITY

The chance of the simultaneous occurrence of two occurrences, when their individual probabilities are known, is determined by multiplying the probability of one event by the conditional probability of the other event, contingent upon the occurrence of the first, i.e.

$$\begin{aligned}P(A \cap B) &= P(A)P(B/A) \\ P(B \cap A) &= P(B) \cdot P(A/B)\end{aligned}$$

Proof: Let out of n outcomes, m_1 be the number of cases favourable to the happening of A .

$$\therefore \text{Probability of happening of the event } A = P(A) = \frac{m_1}{n}.$$

Let m_2 be the number of cases favourable to the happening of B .

$\therefore$ Conditional probability of B , given that A has happened =

$$P(B/A) = \frac{m_2}{m_1}$$

Thus, the number of cases favourable to the happening of both A and B are m_2 out of n .

$$\therefore P(A \cap B) = \frac{m_2}{n} = \frac{m_1}{n} \times \frac{m_2}{m_1} = P(A) \cdot P(B/A).$$

Note. 1. If A and B are independent events, then

$$\begin{aligned}P(B/A) &= P(B) \\ P(A \cap B) &= P(A) \cdot P(B)\end{aligned}$$

2 $P(A \cap B)$ is also written as $P(A \text{ and } B)$ or $P(AB)$.

3 $P(A \cup B)$ is also written as $P(A \text{ or } B)$.

4 If two events A and B are independent, then

(i) A and $\bar{B}$ are independent event.

(ii) $\bar{A}$ and B are independent event.

(iii) $\bar{A}$ and $\bar{B}$ are independent event.

Example 1: A die is rolled twice, and the total of the numbers obtained is found to be 6. Given this, what is the probability that the number 4 appeared at least once?

Solution: Let A = number 4 appears at least once

B = the sum of the numbers appearing is 6

then,

and

$A = \{(4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (1,4), (2,4), (3,4), (5,4), (6,4)\}$

$B = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$

The required probability $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{5}$

Example 2: A die is rolled two times, and it is noted that the sum of the numbers obtained is 7. Given this information, find the conditional probability that the number 2 appeared at least once.

Solution: Let A = getting the number 2 at least once

B = getting 7 as the sum of the numbers on two dice.

Then, $A =$

$\{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (1,2), (3,2), (4,2), (5,2), (6,2)\}$

and

$B = \{(2,5), (3,4), (4,3), (5,2), (6,1), (1,6)\}$

The required probability $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{6} = \frac{1}{3}$.

Example 3: A family has two children. Considering that at least one child is male, what is the probability that both children are male?

Solution: Let the sample space S is given by

$S = \{(b, b), (b, g), (g, b), (g, g)\}$

Let A represent the event that both children are male, and B designate the event that at least one of them is male, assuming all outcomes are equally probable; thus, the required probability is

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Example 4: Ajeet can either take a course in computers or in maths. If Ajeet takes the computer course, then he will receive an A grade with probability $\frac{1}{2}$; if he takes the maths course then he will receive an A grade with probability $\frac{1}{3}$. Ajeet decides to base his decision on the flip of a fair coin. Find the probability that Ajeet will get an A in maths.

Solution: Let C be the event that Ajeet takes maths and A denote the event that he receives an A in whatever course he takes, then the desired probability is

$$\begin{aligned} P(A \cap C) &= P(C)P(A/C) \\ &= \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} \end{aligned}$$

Example 5: Ashu and Ankit appear in an interview for two vacancies in the same post. The probability of Ashu's selection is $\frac{1}{7}$ and that of Ankit's selection is $\frac{1}{5}$. Find the probability that

- (i) both of them will be selected?
- (ii) none of them will be selected?
- (iii) only one of them is selected?

Solution: Let A denote the Ashu is selected and B denote the Ankit is selected then A and B are independent.

1. P (both of them will be selected)

$$= P(A \cap B) = P(A)P(B) = \frac{1}{7} \times \frac{1}{5} = \frac{1}{35}$$

2 P (none of them will be selected)

$$\begin{aligned} &= P(\text{not } A \text{ and not } B) \\ &= P(\text{not } A) \cdot P(\text{not } B) \\ &= [1 - P(A)][1 - P(B)] \\ &= \left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{5}\right) = \frac{6}{7} \times \frac{4}{5} = \frac{24}{35} \end{aligned}$$

3 P (only one of them will be selected)

$$\begin{aligned} &= P('A \text{ and not } B' \text{ or } 'B \text{ and not } A) \\ &= P(A \text{ and not } B) + P(B \text{ and not } A) \\ &= P(A)P(\text{not } B) + P(B)P(\text{not } A) \\ &= P(A)[1 - P(B)] + P(B)[1 - P(A)] \\ &= \frac{1}{7} \times \left(1 - \frac{1}{5}\right) + \frac{1}{5} \times \left(1 - \frac{1}{7}\right) = \frac{1}{7} \times \frac{4}{5} + \frac{1}{5} \times \frac{6}{7} = \frac{10}{35} = \frac{2}{7} \end{aligned}$$

Example 6: The chance that a teacher conducts a surprise test in any given class is $\frac{1}{5}$. If a student misses two classes, what is the probability that they will miss at least one of these tests?

Solution: Let T_1 be the event of I test held on his first day of absence and T_2 be the event of II test held on second day of his absence. Since, T_1 and T_2 are independent events, then the desired probability P (probability that he will miss at least one test)

$$\begin{aligned} P(T_1 \cup T_2) &= P(T_1) + P(T_2) - P(T_1 \cap T_2) \\ &= P(T_1) + P(T_2) - P(T_1)P(T_2) \\ &= \frac{1}{5} + \frac{1}{5} - \frac{1}{5} \times \frac{1}{5} = \frac{2}{5} - \frac{1}{25} = \frac{9}{25} \end{aligned}$$

Example 7: Three groups of children contain 3 girl and 1 boy; 2 girls and 2 boys; 1 girl and 3 boys respectively. One child is selected at random from each group. Find the chance that there selected comprise 1 girl and 2 boys.

Solution: Let A, B and C be the three groups. Given that

	A	B	C
Boys	1	2	3
Girls	3	2	1

Let B_1, B_2, B_3 be the events of selecting a boy from A, B and C group respectively and let G_1, G_2, G_3 be the events of selecting a girl from A, B and C group respectively. Then $B_1, B_2, B_3, G_1, G_2, G_3$ are independent events such that

$$P(B_1) = \frac{1}{4}, P(B_2) = \frac{2}{4} = \frac{1}{2}, P(B_3) = \frac{3}{4}$$

$$P(G_1) = \frac{3}{4}, P(G_2) = \frac{2}{4} = \frac{1}{2}, P(G_3) = \frac{1}{4}$$

Now one girl and 2 boys can be chosen in the following three mutually exclusive ways.

Groups	A	B	C
(I)	Girl	Boy	Boy
(II)	Boy	Girl	Boy
(III)	Boy	Boy	Girl

The required probability = $P(I) + P(II) + P(III)$

$$\begin{aligned}
&= P(G_1 \cap B_2 \cap B_3) + P(B_1 \cap G_2 \cap B_3) + P(B_1 \cap B_2 \cap G_3) \\
&= P(G_1)P(B_2)P(B_3) + P(B_1)P(G_2)P(B_3) + P(B_1)P(B_2)P(G_3) \\
&= \frac{3}{4} \times \frac{1}{2} \times \frac{3}{4} + \frac{1}{4} \times \frac{1}{2} \times \frac{3}{4} + \frac{1}{4} \times \frac{1}{2} \times \frac{1}{4} \\
&= \frac{13}{32}.
\end{aligned}$$

Example 8: A and B are two independent witnesses in a case. The probability that A will speak the truth is x and the probability that B will speak the truth is y . A and B agree in a certain statement. Prove that the probability that this statement is true: $\frac{xy}{1-x-y+2x}$.

Solution: Let E_1 is the event that A and B agree in a statement and E_2 is the event that their statement is correct.

$$\begin{aligned}
\text{Then, } P(E_1) &= xy + (1-x)(1-y) \\
&= 1 - x - y + 2xy
\end{aligned}$$

and

$$P(E_1 \cup E_2) = xy$$

We know that

$$\begin{aligned}
P(E_1 \cup E_2) &= P(E_1)P(E_2/E_1) = P(E_2)P(E_1/E_2) \\
\therefore P(E_2/E_1) &= \frac{P(E_1 \cup E_2)}{P(E_1)} \\
&= \frac{xy}{1 - x - y + 2xy}
\end{aligned}$$

9.6 BAYE'S THEOREM

Let $B_1, B_2, B_3, \dots, B_n$ are mutually exclusive events with $P(B) \neq 0$ ($i = 1, 2, \dots, n$). For any arbitrary event A in sample space S with $P(A) \neq 0$ and for $1 \leq k \leq n$.

$$P(B_k/A) = \frac{P(B_k)P(A/B_k)}{\sum_{i=1}^n P(B_i)P(A/B_i)}$$

Proof: By the definition of conditional probability

$$P(B_k/A) = \frac{P(B_k \cap A)}{P(A)}$$

By the theorem of total probability

$$P(A) = \sum_{i=1}^n P(B_i \cap A) = \sum_{i=1}^n P(B_i)P(A/B_i)$$

Also from the theory of multiplication

$$P(B_k \cap A) = P(B_k)P(A/B_k)$$

Putting the values of $P(A)$ and $P(B_k \cap A)$ from (2) and (3) in equation (1), we get

$$P(B_k/A) = \frac{P(B_k)P(A/B_k)}{\sum_{i=1}^n P(B_i)P(A/B_i)}$$

Example 1: Urn I contains 3 white balls and 4 red balls, while Urn II contains 5 white balls and 6 red balls. A single ball is randomly selected from one of the two urns and turns out to be white. What is the probability that the ball was drawn from Urn I?

Solution: Let B_1 : the ball is drawn from urn I

B_2 : the ball is drawn from urn II

A : the ball is white

To find $P(B_1/A)$.

By Baye's theorem

$$P\left(\frac{B_1}{A}\right) = \frac{P(B_1)P\left(\frac{A}{B_1}\right)}{P(B_1)P\left(\frac{A}{B_1}\right) + P(B_2)P\left(\frac{A}{B_2}\right)} \quad (1)$$

Since, two urn are equally likely to be selected,

$$P(B_1) = P(B_2) = \frac{1}{2}$$

$$P(A/B_1) = \text{Probability of a white ball is drawn from urn I} = \frac{3}{7}$$

$$P(A/B_2) = \text{Probability of a white ball is drawn from urn II} = \frac{5}{11}$$

By (1), we have

$$P(B_1/A) = \frac{\frac{1}{2} \times \frac{3}{7}}{\frac{1}{2} \times \frac{3}{7} + \frac{1}{2} \times \frac{5}{11}} = \frac{33}{68}$$

Example 2: The probability that a person can hit a target is $\frac{3}{5}$ and the probability that another person can hit the same target is $\frac{2}{5}$. But the first person can fire 8 shoots in the time the second person fires 10 shoots. They fire together, what is the probability that the second person shoots the target.

Solution: Let A denote the event of shooting the target and B_1 denote the event that the first person shoots the target and B_2 denote the event that the second person shoots the target. Therefore, $P(B_2/A)$ is the probability that the second person shoots the targets.

$$\text{Now, we have } P(A/B_1) = \frac{3}{5}, P(A/B_2) = \frac{2}{5}$$

It is given that the ratio of the shoots of the first person to those of the second person in the same time is $\frac{8}{10}$ i.e. $\frac{4}{5}$.

$$\text{Thus, we have } P(B_1) = \frac{4}{5} P(B_2)$$

By Baye's theorem, we have

$$\begin{aligned}
 P(B_2/A) &= \frac{P(B_2)P(A/B_2)}{P(B_1)P(A/B_1) + P(B_2)P(A/B_2)} \\
 &= \frac{P(B_2) \cdot \frac{2}{5}}{\frac{4}{5}P(B_1) \cdot \frac{3}{5} + P(B_2) \cdot \frac{2}{5}} = \frac{\frac{2}{5}}{\frac{3}{5} \times \frac{4}{5} + \frac{2}{5}} = \frac{5}{11}
 \end{aligned}$$

CHECK YOUR PROGRESS

1. Define probability and explain its basic concepts.

.....

.....

.....

2. Explain the addition and multiplication rules of probability.

.....

.....

.....

3. Distinguish between mutually exclusive and independent events.

.....

.....

.....

9.7 SUMMARY

In this unit, we have discussed the following concepts

- Basic definitions and notations related to probability
- Probability
- Conditional probability

9.8 EXERCISES

1. An ordinary six-faced die is thrown once. What is the probability of obtaining the number 3?
2. Determine the probability that a randomly selected non-leap year contains 53 Sundays.
3. An integer is selected at random from the first 200 natural numbers. Find the probability that the selected integer is divisible by 6 or 8.
4. A class consists of 66 students, of whom 13 are boys and the remaining students are girls. If one student is selected at random, determine the probability that the selected student is a girl.
5. A bag contains 7 white balls and 12 black balls. If one ball is selected randomly, find the probability that the ball drawn is white.
6. Consider the set of integers from [the numerical range is missing in the source text]. If one integer is selected at random, determine the probability that it is divisible by 3, 5 or 7.
7. A bag contains 9 discs comprising 4 red discs, 3 blue discs and 2 yellow discs. One disc is selected randomly. Calculate the probability that the selected disc is:
 - (i) red
 - (ii) yellow
 - (iii) blue
 - (iv) not blue
7. Probability that a boy will pass an examination is $\frac{2}{5}$ and that for a girl it is $\frac{2}{5}$. What is the probability that at least one of them passes examination?
8. If $P(A) = \frac{3}{5}$ and $P(B) = \frac{1}{3}$, find
 - (i) $P(A \cup B)$, if A and B are mutually exclusive events.
 - (ii) $P(A \cap B)$, if A and B are independent events.

4. In the two dice experiment, if A is the event of getting the sum of numbers on dice as 11 and B is the event of getting a number other than 5 on the first die, find $P(A \cap B)$. Are A and B independent events?
5. If A and B are independent events, then show that
 - (i) $\bar{A}, B$
 - (ii) $A, \bar{B}$ and
 - (iii) $\bar{A}, \bar{B}$ are also independent events.
6. If on average 1 vessel in every 10 is wrecked find the probability that out of 5 vessels expected to arrive, 4 at least will arrive safely.
7. A teacher claims that he could often tell while his students were still in their first year whether they will obtain I, II, III divisions or fail in their final examinations. To demonstrate his claim, he forecasts the fates of 8 students. Find the probability of his being correct in 5 cases.
8. Determine the probability of throwing more than 8 with 3 , perfectly symmetrical dice.
9. A and B in turns toss an ordinary die for a prize of Rs. 44. The first to toss a 'six' wins. If A has first throw, what is his expectation?
10. A bag contains 2 white and 3 black balls. Four persons A, B, C and D in the order named each draw one ball and do not replace it. The person to draw a white ball receives Rs. 200. Determine their expectations.
11. A box containing 2^n tickets among which ${}^nC_i$ bear the number i , ($i = 0, 1, 2, \dots, n$). A group of m tickets is drawn, what is the expectation of the sum of numbers?

12. A person draws 2 balls from a bag containing 3 white and 5 black balls. If he is to receive Rs. 10 for every white ball which he draws and Rs. 1 for every black ball. What is his expectation?
13. Two players of equal skill A and B play a set of games; they leave off playing when A wants 3 points and B wants 2. If the stake is Rs. 16, what share ought each to take?

9.9 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.



UNIT 10 RANDOM VARIABLES AND MATHEMATICAL EXPECTATION

STRUCTURE

- 10.1 Introduction
- 10.2 Objectives
- 10.3 Some important theorems
- 10.4 Random variables and mathematical expectation
- 10.5 Distribution function and their properties
- 10.6 Continuous random variable and probability density
function
- 10.7 Summary
- 10.8 Exercises
- 10.9 References

10.1 INTRODUCTION

In this unit, we explore Random variable, mathematical expectation and Standard Probability Distributions, which are fundamental tools in statistical analysis and probability theory. We will cover both discrete and continuous probability distributions, emphasizing their functions and properties. By understanding these distributions, you will be able to model real-world phenomena, calculate probabilities effectively, and interpret the behaviour of random variables.

10.2 OBJECTIVES

After completing this unit, students will be able to

- Define random variables and mathematical expectation
- Differentiate between discrete and continuous random variables.
- Explain distribution functions and their key properties.

- Construct and interpret discrete distribution functions.
- Understand continuous random variables and probability density functions (PDFs).
- Apply standard probability distributions to solve real-world problems

10.3 SOME IMPORTANT THEOREMS

10.3.1. BINOMIAL THEOREM

Let the probability of the happening of an event in one trial is p and $q = 1 - p$ be the probability that the event fails in one trial. Let us find out the probability of exactly r successes in n trials.

The chance that an event happens at least r times in n trials is given

$$P = p^n + {}^nC_1 p^{n-1} q + {}^nC_2 p^{n-2} q^2 + \dots + {}^nC_r p^r q^{n-r}$$

Hence, the probability that the event will happen exactly r times in n trials is the $(n + 1)^{\text{th}}$ term in the expansion of $(q + p)^n$.

10.3.2. MULTINOMIAL THEOREM

If a dice has f faces marked with $1, 2, 3, \dots, f$, the probability of throwing a total p with n dice is given by

$$\begin{aligned}
 &= \frac{\text{Coefficient of } x^p \text{ in the expansion of } (x^1 + x^2 + x^3 + \dots + x^f)^n}{f^n} \\
 &= \frac{\text{Coeff. of } x^p \text{ in } \frac{(1 - x^f)^n}{x^n(1 - x)^n}}{f^n} \\
 &= \frac{\text{Coeff. of } x^{p-n} \text{ in } (1 - x^f)^n(1 - x)^{-n}}{f^n}
 \end{aligned}$$

Note. If n dice are different with faces $f_1, f_2, f_3, \dots, f_n$, then the required chance is

$$= \frac{\text{Coeff. of } x^p \text{ in } (x^1 + x^2 + \dots + x^{f_1})(x^1 + x^2 + \dots + x^{f_2}) \dots (x^1 + x^2 + \dots + x^{f_n})}{f_1 f_2 f_3 \dots f_n}$$

10.4 RANDOM VARIABLES AND MATHEMATICAL EXPECTATION

10.4.1. RANDOM VARIABLE

A variable, which takes a definite set of values with a definite probability associated with each value is called a random variable.

Hence, if a variable x takes the values $x_1, x_2, x_3, \dots, x_n$ with respective probability $p_1, p_2, p_3, \dots, p_n$ so that $\sum_{i=1}^n p_i = 1$. Then a discrete probability distribution is defined. The function $P(x)$, which has respective probabilities $p_1, p_2, p_3, \dots, p_n$ for $x_1, x_2, x_3, \dots, x_n$ is known as the probability function of x . The variable x in such a case is called the random variable.

10.4.2. EXPECTED VALUE

When X is a discrete variable, which may take n mutually exclusive values $x_i (i = 1, 2, 3, \dots, n)$ and no other, with respective probabilities $p_i (i = 1, 2, 3, \dots, n)$ the expectation of X is

$$E(X) = p_1 x_1 + p_2 x_2 + p_3 x_3 + \dots + p_n x_n = \sum_{i=1}^n p_i x_i$$

when X is a continuous variable, the expectation of x is given by

$$E(X) = \int_{-\infty}^{\infty} x \phi(x) dx$$

where, $\phi(x)$ is the probability density defining the function

Note 1. If X and Y are random variables, then

$$E(X + Y) = E(X) + E(Y)$$

2 If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

3 The expected value of a random variable X , denoted by $E(X)$ or μ , is defined as

$$E(X) \text{ or } \mu = \begin{cases} \sum_{i=-\infty}^{\infty} x_i f(x_i) & (\text{if } X \text{ is discrete}) \\ \int_{-\infty}^{\infty} x f(x) dx & (\text{if } X \text{ is continuous}) \end{cases}$$

4 $\text{Mean}(\bar{X}) = E(X) = \mu'_1$

5 $\mu'_2 = E(X')$

6 $\text{Variance } \sigma^2 = E(X^2) - [E(X)]^2 = \mu'_2 - \mu_1'^2$

7 $\mu'_r = E(X^r)$

8 $E(aX + b) = aE(X) + b$

9 $E\{a\phi(x)\} = aE\{\phi(x)\}$

10 $V(aX + b) = a^2V(X)$, where $V(X)$ is the variance of X .

Example 1: In a single throw, with a pair of dice, what is the chance of throwing doublets or not.

Solution: When throw a pair of dice, we get only 6 doublets namely (1,1); (2,2); (3,3); (4,4); (5, 5); (6, 6).

The chance of throwing a doublet in one throw $= \frac{6}{6^2} = \frac{1}{6}$

And the chance of not throwing a doublet

$$= 1 - \frac{1}{6} = \frac{5}{6}$$

Example 2: A dice is thrown five times. Find the probability of 3 coming up (i) exactly 3 times (ii) at least 3 times.

Solution: The probability of 3 coming up $= \frac{1}{6}$.

Let $p = 1/6$ and $q = 5/6$.

The probability of 3 successes in 5 trials is ${}^5C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$.

Probability of at least 3 successes = Probability of 3 or 4 or 5 successes

$$= {}^5C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 + {}^5C_4 \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right) + {}^5C_5 \left(\frac{1}{6}\right)^5.$$

Example 3: Four dice are thrown, what is the probability that the sum of the number appearing on the dice is 18?

Solution: The probability that the sum of the number appearing on the dice is 18

$$\begin{aligned} &= \frac{\text{Coeff. of } x^{18} \text{ in } (x + x^2 + \dots + x^6)^4}{6^4} \\ &= \frac{\text{Coeff. of } x^{14} \text{ in } (1 + x + \dots + x^5)^4}{6^4} \\ &= \frac{\text{Coeff. of } x^{14} \text{ in } (1 - x^6)^4 (1 - x)^{-4}}{6^4} \\ &= \frac{\text{Coeff. of } x^{14} \text{ in } (1 - 4x^6 + 6x^{12} - \dots) \times (1 + 4x + 10x^2 + \dots + 16x^8 + \dots + 680x^{14} + \dots)}{6^4} \end{aligned}$$

Example 4: A person throws two dice, one the common cube, and the other regular tetrahedron, the number in the lowest face being taken in the case of a tetrahedron. What is the chance that the sum of the numbers thrown is not less than 5.

Solution: We know that the common cube has 6 faces and the regular tetrahedron has 4 faces. Hence, 6×4 is the total number of ways in which the cube and the tetrahedron can fall.

$$\begin{aligned}
 & \text{The favourable ways of getting a sum not less than 5} \\
 &= \text{the sum of the coeff. of } x^5, x^6, \dots, x^{10} \text{ in } (x + x^2 + \dots + x^6)(x + x^2 + x^3 + x^4) \\
 &= \text{the sum of the coeff. of } x^5, x^6, \dots, x^{10} \text{ in } (x^2 + 2x^3 + 3x^4 + 4x^5 + 4x^6 + 4x^7 + 3x^8 + 2x^9 + x^{10}) \\
 &= 4 + 4 + 4 + 3 + 2 + 1 \\
 &\therefore \text{The required chance} = \frac{18}{6 \times 4} = \frac{3}{4}
 \end{aligned}$$

Example 5: Four tickets marked 00, 01, 10, 11 respectively are placed in a bag. A ticket is drawn at random five times, being replaced each time. Find the probability that the sum of the numbers on tickets thus drawn is 23.

Solution: Given that four tickets marked 00,01,10,11 respectively are placed in a bag. The four tickets can be drawn five times in 4^5 ways. The favourable number of ways for obtaining a sum

$$\begin{aligned}
 &= \text{Coeff. of } x^{23} \text{ in } (x^0 + x^1 + x^{10} + x^{11})^5 \\
 &= \text{Coeff. of } x^{23} \text{ in } (1 + x)^5(1 + x^{10})^5 \text{ (on factorization)} \\
 &= \text{Coeff. of } x^{23} \text{ in } (1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5)(1 + 5x^{10} + 10x^{20} + \dots) \\
 &= 100
 \end{aligned}$$

$$\text{Hence, the required probability} = \frac{100}{4^5} = \frac{25}{256}.$$

Example 6: If m things are distributed among a men and b women, show that the chance that the number of things received by men is odd, is $\frac{1}{2} \frac{(b+a)^m - (b-a)^m}{(a+b)^m}$.

Solution: Given that m things are distributed among a men and b women. The probability that men may get a thing is $\frac{a}{a+b}$.

The probability that a women may get a thing is $\frac{b}{a+b}$. If out of m things ' a ' men get only one thing and other things go to the women, then the probability for men is

$$= {}^m C_1 \left(\frac{a}{a+b} \right) \left(\frac{b}{a+b} \right)^{m-1}$$

Similarly, the probability that 'a' men may get three things, five things, ... are respectively

$${}^m C_3 \left(\frac{a}{a+b} \right)^3 \left(\frac{b}{a+b} \right)^{m-3}, {}^m C_5 \left(\frac{a}{a+b} \right)^5 \left(\frac{b}{a+b} \right)^{m-5}, \dots$$

Let A be the event the number of things got by men is odd and A_i denotes the event 'men get i things'

$$i = 1, 3, 5, 7, \dots$$

Thus, the probability that the number of things received by men is odd is,

$$\begin{aligned} P(A) &= P(A_1) + P(A_3) + P(A_5) + \dots \\ &= {}^m C_1 \left(\frac{a}{a+b} \right) \left(\frac{b}{a+b} \right)^{m-1} + {}^m C_3 \left(\frac{a}{a+b} \right)^3 \left(\frac{b}{a+b} \right)^{m-3} + {}^m C_5 \left(\frac{a}{a+b} \right)^5 \left(\frac{b}{a+b} \right)^{m-5} + \dots \\ &= \frac{1}{(a+b)^m} [{}^m C_1 a b^{m-1} + {}^m C_3 a^3 b^{m-3} + {}^m C_5 a^5 b^{m-5} + \dots] \\ &= \frac{1}{(a+b)^m} \left[\frac{1}{2} (b+a)^m - \frac{1}{2} (b-a)^m \right] \\ &= \frac{1}{2} \cdot \frac{(b+a)^m - (b-a)^m}{(a+b)^m} \end{aligned}$$

Example 7: From a bag containing 5 one rupee coins and 3 coins of 20 paisa each, a person is allowed to draw 2 coins indiscriminately. Find the value of his expectation.

Solution: Given that a bag containing 5 one rupee coins and 3 coins of 20 paisa each.

$$\text{Probability of drawing 2 rupees} = \frac{{}^5 C_2}{{}^8 C_2} = \frac{5}{14}$$

Probability of drawing 1 rupee and one 20 paisa coin

$$= \frac{{}^5C_1 \times {}^3C_1}{{}^8C_2} = \frac{15}{28}$$

Probability of drawing two 20 paise coins = $\frac{{}^3C_2}{{}^8C_2} = \frac{3}{28}$.

The values of draws in three cases are Rs 2, 1.20 and 0.40 rupees respectively.

$$\begin{aligned} \therefore \text{Expected value} &= \frac{5}{14} \times 2 + \frac{15}{28} \times 1.20 + \frac{3}{28} \times 0.40 \\ &= \frac{10}{14} + \frac{180}{280} + \frac{12}{280} \\ &= \frac{200 + 180 + 12}{280} \\ &= \frac{392}{280} = \frac{7}{5} = \text{Rs. 1.40} \end{aligned}$$

Example 8: A person draws cards one by one from a pack until he draws all the aces. How many cards he may be expected to draw?

Solution: Let a person has to make n draws for all aces. It means that in $n - 1$ draws, three aces and in n^{th} , one ace. The probability of such an occurrence

$$\begin{aligned} &= \frac{{}^4C_3 \times {}^{48}C_{n-4}}{{}^{52}C_{n-1}} \times \frac{1}{52 - (n - 1)} \\ &= \frac{4 \times 48! (n - 1)! (52 - n + 1)!}{(n - 4)! (52 - n)! 52!} \times \frac{1}{52 - n + 1} \\ &= \frac{4(n - 1)(n - 2)(n - 3)}{49 \times 50 \times 51 \times 52}. \end{aligned}$$

The least square number of draws he has to make is 4 and the maximum number 52. Hence, n range 4 to 52.

The expected number of draws

$$= \sum_{n=1}^{52} n \cdot 4 \frac{(n-1)(n-2)(n-3)}{49 \times 50 \times 51 \times 52}$$

$$= \frac{4}{49 \times 50 \times 51 \times 52} \left[\sum_{n=1}^{52} n^4 - 6 \sum_{n=1}^{52} n^3 - 11 \sum_{n=1}^{52} n^2 - 6 \sum_{n=1}^{52} n \right].$$

Example 9: An urn contains n tickets numbered from 1 to n and m tickets are drawn one by one without replacements from the urn. What is the mathematical expectation of the sum of the numbers drawn?

Solution: Suppose the variable associated with the j^{th} ticket be x

$$j = 1, 2, 3, \dots$$

Now,

$$E(x_j) = \sum p_j x_j = \frac{1}{n} \sum x_j$$

$$= \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

Let S be the sum, then the expected value of the sum

$$E(S) = \sum E(x_j) = m \sum E(x)$$

$$= m \frac{(n+1)}{2}.$$

Example 10: What is the expectation of the number of failures preceding the first success in an infinite series of independent trials with constant probability of successes?

Solution: The probabilities of success in first, second, third, ... trials respectively are

$$p, qp, q^2p, q^3p, \dots$$

Let X be the no. of failures

Then, the expected number of failures preceding the first success

$$\begin{aligned}
 E(X) &= 0 \cdot p + 1 \cdot qp + 2q^2p + \cdots \dots \\
 &= qp[1 + 2q + 3q^2 + \cdots \dots \infty), q < 1 \\
 &= \frac{qp}{(1-q)^2} \\
 &= \frac{q}{p} = \frac{1-p}{p}.
 \end{aligned}$$

$[\because p = 1 - q]$

10.5 DISTRIBUTION FUNCTION AND THEIR PROPERTIES

10.5.1 Distribution Function: Let X be a random variable. Then the function

$$F_X(x) = P(X \leq x) = P\{\omega: X(\omega) \leq x\}, -\infty < x < \infty$$

is called the distribution function of X .

10.5.2 Properties of Distribution Function. We now proceed to derive a number of properties common to all distribution functions.

Property 1. If F is the d.f. of the r.v. X and if $a < b$, then

$$P(a < X \leq b) = F(b) - F(a)$$

Remark. When $P(X = a) = 0$ and $P(X = b) = 0$, all four events $a \leq X \leq b$, $a < X < b$, $a \leq X < b$ and $a < X \leq b$ have the same probability $F(b) - F(a)$.

Property 2. If F is the d.f. of one-dimensional r.v. X , then (i) $0 \leq F(x) \leq 1$, (ii) $F(x) \leq F(y)$ if $x < y$.

In other words, all distribution functions are monotonically non-decreasing and lie between 0 and 1.

Property 3. If F is d.f. of one-dimensional r.v. X , then

$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$$

and

$$F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$$

This gives $F(-\infty) = 0$ and $F(\infty) = 1$.

Remarks. 1. Discontinuities of $F(x)$ are at most countable.

$$2. F(a) - F(a - 0) = \lim_{h \rightarrow 0} P(a - h \leq X \leq a), h > 0$$

$$\therefore F(a) - F(a - 0) = P(X = a)$$

and

$$F(a + 0) - F(a) = \lim_{h \rightarrow 0} P(a \leq X \leq a + h) = 0, h > 0$$

$$\Rightarrow F(a + 0) = F(a)$$

10.5.3. Discrete Random Variable: If a random variable takes at most a countable number of values, it is called a discrete random variable. In other "words, a real valued function defined on a discrete sample space is called a discrete random variable.

10.5.4. Probability Mass Function: Suppose X is a one-dimensional discrete random variable taking at most a countably infinite number of values $x_1, x_2, \dots$. With each possible outcome x_i , we associate a number $p_i = P(X = x_i) = p(x_i)$, called the probability of x_i . The numbers $p(x_i); i = 1, 2, \dots$ must satisfy the following conditions:

$$(i) p(x_i) \geq 0 \forall i, (ii) \sum_{i=1}^{\infty} p(x_i) = 1.$$

This function p is called the probability mass function of the random variable X and the set $(x_i, p(x_i))$ is called the probability distribution (p.d.) of the r.v. X .

Remarks: 1. The set of values which X takes is called the spectrum of the random variable.

2. For discrete random variable, a knowledge of the probability mass function enables us to compute probabilities of arbitrary events. In fact, if E is a set of real numbers, we have

$$P(X \in E) = \sum_{x \in E \cap S} p'(x), \text{ where } S \text{ is the sample space.}$$

Example 1. Toss of coin, $S = (H, T)$. Let X be the random variable defined by

$X(H) = 1$, i.e., $X = 1$, if 'Head' occurs.

$X(T) = 0$, i.e., $X = 0$, if 'Tail' occurs.

If the coin is 'fair' the probability function is given by

$$P((H)) = P([T]) = \frac{1}{2}$$

and we can speak of the probability distribution of the random variable X as

$$P(X = 1) = P((H)) = \frac{1}{2},$$

$$P(X = 0) = P([T]) = \frac{1}{2}$$

10.5.5 Discrete Distribution Function: In this case there are a countable number of points $x_1, x_2, x_3, \dots$ and numbers $p_1 \geq 0, \sum_1^\infty p_1 = 1$ such that $F(X) = \sum_{(i: x_1 \leq x)} p_i$. For example, if x_i is just the integer i , $F(x)$ is a "step function" having jump p_i at i , and being constant between each pair of integers.

Theorem 1. $p(x_j) = P(X = x_j) = F(x_j) - F(x_{j-1})$, where F is the d.f. of X .

Example 2. An experiment consists of three independent tosses of a fair coin. Let

X = The number of heads

Y = The number of head runs,

Z = The length of head runs,

The number of heads that occur together in three-coin flips is the length of a head run, which is defined as the consecutive occurrence of at least two heads.

Find the probability function of (i) X , (ii) Y , (iii) Z , (iv) $X + Y$ and (v) XY and construct probability tables and draw their probability charts.

Solution.

Table 1

S. No.	Elementary event	Random Variables				
		X	Y	Z	$X + Y$	XY
1	HHH	3	1	3	4	3
2	HHT	2	1	2	3	2
3	HTH	2	0	0	2	0
4	HTT	1	0	0	1	0
5	THH	2	1	2	3	2
6	THT	1	0	0	1	0
7	TTH	1	0	0	1	0
8	TTT	0	0	0	0	0

Here sample space is.

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

(i) Obviously X is a r.v. which can take the values 0,1,2, and 3

$$p(3) = P(HHH) = (1/2)^3 = 1/8$$

$$\begin{aligned} p(2) &= P[HHT \cup HTH \cup THH] \\ &= P(HHT) + P(HTH) + P(THH) = 1/8 + 1/8 + 1/8 = 3/8 \end{aligned}$$

Similarly, $p(1) = 3/8$ and $p(0) = 1/8$.

These probabilities could also be obtained directly from the above table 1.

Table 1

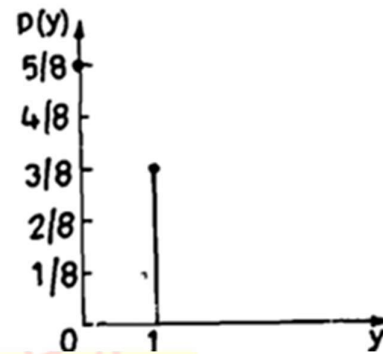
Probability table of X

Values of X (x)	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

Table 3

(ii) Probability Table of Y

Values of Y , (y)	0	1
$p(y)$	5/8	3/8



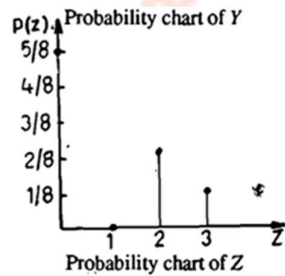
This is obvious from table 1.

(iii) From table 1, we have

Table 4

Probability Table of

Values of Z, (z)	0	1	2	3
$p(z)$	5/8	0	2/8	1/8

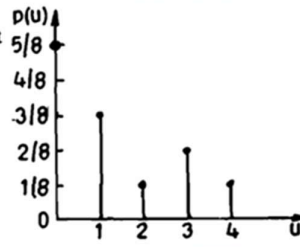


(iv) Let $U = X + Y$. From table 1, we get

Table 5

Probability Table of U

Values of U, (u)	0	1	2	3	4
$p(u)$	1/8	3/8	1/8	2/8	1/8

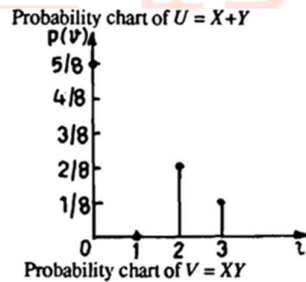


(v) Let $V = XY$

Table 6

Probability Table of V

Values of V, (v)	0	1	2	3
$p(v)$	5/8	0	2/8	1/8



Example 3. A random variable X has the following probability distribution:

$x:$	0	1	2	3	4	5	6	7
$p(x):$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2 + k$

(i) Find k , (ii) Evaluate $P(X < 6)$, $P(X \geq 6)$, and $P(0 < X < 5)$, (iii) If $P(X \leq c) > \frac{1}{2}$, find the minimum value of c , and (iv) Determine the distribution function of X .

Solution. Since $\sum_{x=0}^7 p(x) = 1$, we have

$$\begin{aligned} \Rightarrow k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k &= 1 \\ \Rightarrow 10k^2 + 9k - 1 &= 0 \\ \Rightarrow (10k - 1)(k + 1) = 0 \Rightarrow k &= 1/10 \end{aligned}$$

[$\because k = -1$, is rejected, since probability cannot be negative.]

$$\begin{aligned}
 P(X < 6) &= P(X = 0) + P(X = 1) + \dots + P(X = 5) \\
 &= \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} = \frac{81}{100} \\
 P(X \geq 6) &= 1 - P(X < 6) = \frac{19}{100} \\
 P(0 < X < 5) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 8k' = 4/5
 \end{aligned}$$

(iii) $P(X \leq c) > \frac{1}{2}$. By trial, we get $c = 4$.

(iv)

X	$F_X(x) = P(X \leq x)$
0	0
1	$k = 1/10$
2	$3k = 3/10$
3	$5k = 5/10$
4	$8k = 4/5$
5	$8k + k^2 = 81/100$
6	$8k + 3k^2 = 83/100$
7	$9k + 10k^2 = 1$

10.6 CONTINUOUS RANDOM VARIABLE AND PROBABILITY DENSITY FUNCTION

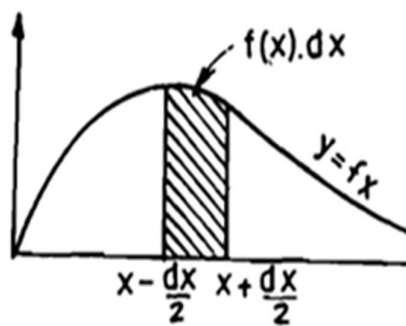
10.6.1 Continuous Random Variable. A continuous random variable is a type of random variable that can take **any real value within a given interval**. It is used to model outcomes where the values form a continuum and are not countable.

Examples of continuous random variables are age, height, weight etc.

10.6.2 Probability Density Function (Concept and Definition).

Consider the small interval $(x, x + dx)$ of length dx round the point x . Let $f(x)$ be any continuous function of x so that $f(x) dx$ represents the probability that X falls in the infinitesimal interval $(x, x + dt)$. Symbolically

$$P(x \leq X \leq x + dx) = f_X(x)dx$$



In the figure, $f(x)dx$ represents the area bounded by the curve $y = f(x)$, the x-axis, and the ordinates at the points x and $x + dx$. The function $f_X(x)$ so defined is the probability density function of random variable X , usually abbreviated as p.d.f. The area under the curve $f(x)dx$, usually written as $dF(x)$, is the probability differential and $y = f(x)$ is known as the probability curve or simply probability curve.

Figure 10.1 PDF

Definition. p.d.f $f_X(x)$ of the r.v. X is defined as:

$$f_X(x) = \lim_{\delta x \rightarrow 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}$$

The probability for a variate value to lie in the interval dx is $f(x)dx$ and hence the probability for a variate value to fall in the finite interval $[\alpha, \beta]$ is:

$$P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx$$

which represents the area between the curve $y = f(x)$, x -axis and the ordinates at $x = \alpha$ and $x = \beta$. Further since total probability is unity, we have $\int_a^b f(x)dx = 1$, where $[a, b]$ is the range of the random variable X . The range of the variable may be finite or infinite.

The probability density function (p.d.f.) of a random variable X usually denoted by $f_X(x)$ or simply by $f(x)$ has the following obvious properties

- (i) $f(x) \geq 0, -\infty < x < \infty$
- (ii) $\int_{-\infty}^{\infty} f(x)dx = 1$
- (iii) The probability $P(E)$ given by

$$P(E) = \int_E f(x)dx$$

is well defined for any event E .

Important Remark. In case of discrete random variable, the probability at a point, i.e., $P(x = c)$ is not zero for some fixed c . However, in case of continuous random variables the probability at a point is always zero, i.e., $P(x = c) = 0$ for all possible values of c . This follows directly from (5.5b) by taking $\alpha = \beta = c$.

This also agrees with our discussion earlier that $P(E) = 0$ does not imply that the event E is null or impossible event. This property of continuous r.v., viz.,

$$P(X = c) = 0, \forall c$$

leads us to the following important result:

$$P(\alpha \leq X \leq \beta) = P(\alpha \leq X < \beta) = P(\alpha < X \leq \beta) = P(\alpha < X < \beta) \dots (5 \cdot 5g)$$

i.e., in case of continuous random variable, it does matter whether we include the end points of the interval from α to β .

However, this result is in general not true for discrete random variables.

10.6.3 Different Central Tendency, Skewness, Kurtosis, and Dispersion Measures for Continuous Probability Distributions: The formulae for these measures in case of discrete frequency distribution can be easily extended to the case of continuous probability distribution by simply replacing $p_i = f_i/N$ by $f(x)dx$, x_i by x and the summation over 'i' by integration over the specified range of the variable X .

Let $f_X(x)$ or $f(x)$ be the *p.d.f.* of a random variable X where X is defined from a to b . Then

(i) Arithmetic mean $= \int_a^b xf(x)dx$

(ii) Harmonic mean. Harmonic mean H is given by

$$\frac{1}{H} = \int_a^b \left(\frac{1}{x}\right) f(x)dx$$

(iii) Geometric mean. Geometric mean G is given by

$$\log G = \int_a^b \log x f(x)dx$$

(iv) μ'_r (about origin) $= \int_a^b x^r f(x)dx \dots (1)$

μ'_r (about the point $x = A$) $= \int_a^b (x - A)^r f(x)dx$

$$\text{and } \mu \text{ (about mean)} = \int_a^b (x - \text{mean})' f(x) dx$$

In particular, we have

$$\mu'_1 \text{ (about origin)} = \text{Mean} = \int_a^b x f(x) dx$$

and

$$\mu'_2 = \int_a^b x^2 f(x) dx$$

Hence

$$\mu_2 = \mu'_2 - \mu_1'^2 = \int_a^b x^2 f(x) dx - \left(\int_a^b x f(x) dx \right)^2$$

From (1), on putting $r = 3$ and 4 respectively, we get the values of μ'_3 and μ'_4 and consequently the moments about mean can be obtained by using the relations:

and

$$\left. \begin{aligned} \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4 \end{aligned} \right\}$$

and hence β_1 and β_2 can be computed.

(v) Median. The point that splits the entire distribution in half is called the median. The median is the point at which the entire area is divided into two equal pieces in the event of a continuous distribution. Thus if M is the median, then

$$\int_a^M f(x) dx = \int_M^b f(x) dx = \frac{1}{2}$$

Thus solving

$$\int_a^M f(x)dx = \frac{1}{2} \text{ or } \int_M^b f(x)dx = \frac{1}{2}$$

for M , we get the value of median.

(vi) Mean Deviation. Mean deviation about the mean μ'_1 is given by

$$MD. = \int_a^b |x - \text{mean}| f(x)dx$$

(vii) Quartiles and Deciles. Q_1 and Q_3 are given by the equations

$$\int_a^{Q_1} f(x)dx = \frac{1}{4} \text{ and } \int_a^{Q_3} f(x)dx = \frac{3}{4}$$

D_i , i th decile is given by

$$\int_a^{D_i} f(x)dx = \frac{i}{10}$$

(viii) Mode. Mode is the value of x for which $f(x)$ is maximum. Mode is thus the solution of

$$f'(x) = 0 \text{ and } f''(x) < 0$$

provided it lies in $[a, b]$.

Example 1. The diameter of an electric cable, say X , is assumed to be a continuous random variable with p.d.f. $f(x) = 6x(1 - x), 0 \leq x \leq 1$.

- (i) Check that above is p.d.f.,
- (ii) Determine a number b such that $P(X < b) = P(X > b)$

Solution. Obviously, for $0 \leq x \leq 1, f(x) \geq 0$

Now

$$\begin{aligned}\int_0^1 f(x)dx &= 6 \int_0^1 x(1-x)dx \\ &= 6 \int_0^1 (x-x^2)dx = 6 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = 1\end{aligned}$$

Hence $f(x)$ is the p.d.f. of r.v. X

(ii)

$$P(X < b) = P(X > b.) \quad (2)$$

$$\begin{aligned}\Rightarrow \int_0^b f(x)dx &= \int_b^1 f(x)dx \\ \Rightarrow 6 \int_0^b x(1-x)dx &= 6 \int_b^1 x(1-x)dx \\ \Rightarrow \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^b &= \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_b^1 \\ \Rightarrow \left(\frac{b^2}{2} - \frac{b^3}{3} \right) &= \left[\left(\frac{1}{2} - \frac{1}{3} \right) - \left(\frac{b^2}{2} - \frac{b^3}{3} \right) \right] \\ \Rightarrow 3b^2 - 2b^3 &= [1 - 3b^2 + 2b^3] \\ \Rightarrow 4b^3 - 6b^2 + 1 &= 0 \\ (2b-1)(2b^2-2b-1) &= 0 \\ \Rightarrow 2b-1 &= 0 \text{ or } 2b^2-2b-1 = 0\end{aligned}$$

Hence $b = 1/2$ is the only real value lying between 0 and 1 and satisfying (2).

Example 2. A continuous random variable X has a p.d.f.

$$f(x) = 3x^2, 0 \leq x \leq 1. \text{ Find } a \text{ and } b \text{ such that}$$

$$(i) \quad P(X \leq a) = P(X > a), \quad \text{and}$$

$$(ii) \quad P(X > b) = 0.05.$$

Solution. (i) Since $P(X \leq a) = P(X > a)$,

each must be equal to $1/2$, because total probability is always one.

$$\therefore P(X \leq a) = \frac{1}{2} \Rightarrow \int_0^a f(x) dx = \frac{1}{2}$$

$$3 \int_0^a x^2 dx = \frac{1}{2} \Rightarrow 3 \left[\frac{x^3}{3} \right]_0^a = \frac{1}{2}$$

$$a^3 = \frac{1}{2} \Rightarrow a = \left(\frac{1}{2} \right)^{\frac{1}{3}}$$

(ii)

$$P(X > b) = 0.05 \Rightarrow \int_b^1 f(x) dx = 0.05$$

$$3 \left[\frac{x^3}{3} \right]_b^1 = \frac{1}{20} \Rightarrow 1 - b^3 = \frac{1}{20}$$

$$b^3 = \frac{19}{20} \Rightarrow b = \left(\frac{19}{20} \right)^{\frac{1}{3}}$$

Example 3. Let X be a continuous, random variate with p.d.f.

$$\begin{aligned} f(x) &= ax, 0 \leq x \leq 1 \\ &= a, 1 \leq x \leq 2 \\ &= -ax + 3a, 2 \leq x \leq 3 \\ &= 0, \text{ elsewhere} \end{aligned}$$

(i) Determine the constant a .

(ii) Compute $P(X \leq 1.5)$.

Solution. (i) Constant ' a ' is determined from the consideration that total probability is unity, i.e.,

$$\begin{aligned}
 & \int_{-\infty}^{\infty} f(x)dx = 1 \\
 \Rightarrow & \int_{-\infty}^0 f(x)dx + \int_0^1 f(x)dx + \int_1^2 f(x)dx + \int_2^3 f(x)dx + \int_3^{\infty} f(x)dx = 1 \\
 \Rightarrow & \int_0^1 axdx + \int_1^2 a \cdot dx + \int_2^3 (-\omega x + 3a)dx = 1 \\
 \Rightarrow & a \left[\frac{x^2}{2} \right]_0^1 + a|x|_1^2 + a \left[-\frac{x^2}{2} + 3x \right]_2^3 = 1 \\
 \Rightarrow & \frac{a}{2} + a + a \left[\left(-\frac{9}{2} + 9 \right) - (-2 + 6) \right] = 1 \\
 \Rightarrow & \frac{a}{2} + a + \frac{a}{2} = 1 \Rightarrow 2a = 1 \Rightarrow a = \frac{1}{2}
 \end{aligned}$$

(ii)

$$\begin{aligned}
 P(X \leq 1.5) &= \int_{-\infty}^{1.5} f(x)dx = \int_{-\infty}^0 f(x)dx + \int_0^1 f(x)dx + \int_1^{1.5} f(x)dx \\
 &= a \int_0^1 xdx + \int_1^{1.5} a \cdot dx \\
 &= a \left[\frac{x^2}{2} \right]_0^1 + a|x|_1^{1.5} = \frac{a}{2} + 0.5a \\
 &= a = \frac{1}{2} \left[\because a = \frac{1}{2}, \text{Parl}(i) \right]
 \end{aligned}$$

Example 4. A probability curve $y = f(x)$ has a range from 0 to ∞ . If $f(x) = e^{-x}$, find the mean and variance and the third moment about mean.

Solution.

$$\begin{aligned}\mu, (\text{rth moment about origin}) &= \int_0^{\infty} x^r f(x) dx \\ &= \int_0^{\infty} x^r e^{-x} dx = \Gamma(r+1) = r!\end{aligned}$$

(Using Gamma Integral)

Substituting $r = 1, 2$ and 3 successively, we get

$$\text{Mean} = \mu'_1 = 1! = 1, \mu'_2 = 2! = 2, \mu'_3 = 3! = 6$$

$$\text{Hence variance} = \mu_2 = \mu'_2 - \mu_1^2 = 2 - 1 = 1$$

and

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1^3 = 6 - 3 \times 2 + 2 = 2$$

CHECK YOUR PROGRESS

1. What is a random variable? Explain its different types.

.....

.....

.....

2. Distinguish between discrete and continuous random variables.

.....

.....

.....

3. Define mathematical expectation and explain its important properties.

10.7 SUMMARY

In this unit, we have discussed the following concepts

- Random variables and mathematical expectation
- Distribution function and their properties
- Continuous random variable and probability density function

10.8 EXERCISES

- 1 (a) A continuous random variable X follows the probability law

$$f(x) = Ax^2, 0 \leq x \leq 1$$

Determine A and find the probability that (i) X lies between 0.2 and 0.5, (ii) X is less than 0.3, (iii) $1/4 < X < 1/2$ and (iv) $X > 3/4$ given $X > 1/2$.

- (b) If a random variable X has the density function

$$f(x) = \begin{cases} 1/4, & -2 < x < 2 \\ 0, & \text{elsewhere.} \end{cases}$$

Obtain (i) $P(X < 1)$,

(ii) $P(|X| > 1)$

(iii) $P(2X + 3 > 5)$

- 2 Are any of the following probability mass or density functions?

Prove your answer in each case.

(a) $f(x) = x; x = \frac{1}{16}, \frac{3}{16}, \frac{1}{4}, \frac{1}{2}$

(b) $f(x) = \lambda e^{-\lambda x}; x \geq 0; \lambda > 0$

(c) $f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 4 - 2x, & 1 < x < 2 \\ 0, & \text{elsewhere,} \end{cases}$

- 3 Prove that the geometric mean G of the distribution $dF = 6(2 - x)(x - 1)dx, 1 \leq x \leq 2$ is given by $6 \log(16G) = 19$.

10.9 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 11 BINOMIAL DISTRIBUTIONS

STRUCTURE

- | | |
|------|---------------------------|
| 11.1 | Introduction |
| 11.2 | Objectives |
| 11.3 | Terminology and notations |
| 11.4 | Binomial distribution |
| 11.5 | Solved examples |
| 11.6 | Bernoulli distribution |
| 11.7 | Summary |
| 11.8 | Exercises |
| 11.9 | References |

11.1 INTRODUCTION

In this unit, we delve into two foundational probability distributions: the Bernoulli distribution and the Binomial distribution. These distributions are essential for modelling situations where outcomes are binary, such as success/failure or yes/no. The Bernoulli distribution represents a single trial, while the Binomial distribution extends this to multiple independent trials. Understanding these distributions will provide you with the tools to model and analyze real-world scenarios involving repeated trials and probability of success.

11.2 OBJECTIVES

After completing this unit, students will be able to

- Define and explain the Bernoulli and Binomial distributions.
- Identify real-world applications where Bernoulli and Binomial distributions are appropriate.

- Calculate probabilities using the Bernoulli distribution for a single trial.
- Compute probabilities and expectations using the Binomial distribution for multiple trials.
- Understand and apply key properties, such as the mean and variance of these distributions.

11.3 TERMINOLOGY AND NOTATIONS

- 1 **Random variable:** A variable, which takes a definite set of values with a definite probability associated with each value is called a random variable.
- 2 **Continuous random variable:** A random variable X is said to be a continuous random variable, if it takes all possible values between its limits. For example, height of a tree, weight of a school child etc.
- 3 **Discrete random variable:** A random variable X is said to be a discrete random variable, if it takes only finite values between its limits, for example, the number of students appearing in a festival consisting of 400 students is a discrete random variable which can assume values other than 0,1,2, ...,400.
- 4 **Continuous probability distribution:** Let X be a random variable which assume values in the interval $[-\infty, \infty]$, then the probability in (a, b) is defined by

$$P(a \leq x \leq b) = \int_a^b f(x) dx$$

where, $f(x)$ is called probability density function which satisfies the following

conditions:

$$(i) f(x) \geq 0 \quad \forall x \in (-\infty, \infty) \quad (ii) \int_{-\infty}^{\infty} f(x) = 1$$

- 5 **Discrete probability distribution:** Let X be a random variable which assume values $x_1, x_2, \dots, x_n$ with probability $p_1, p_2, \dots, p_n$ respectively, then the probability is

$$P(X = x_j) = p(x_j) \text{ or } f(x_j) \text{ for } i = 1, 2, \dots, n$$

where, $p(x_j)$ is called the probability mass function which satisfies the following conditions:

$$(i) f(x) \geq 0 \quad (ii) \sum_{i=1}^n f(x) = 1$$

- 6 **Mean and variance of random variables:** Let X be a random variable which assume values $x_1, x_2, \dots, x_n$ with probability $p_1, p_2, \dots, p_n$ respectively. We denote mean (average or expected value) by μ ($\bar{x}$ or m) or

$$E(X) = \frac{\sum p_i x_i}{\sum p_i} = \sum p_i x_i \quad [\because \sum p_i = 1]$$

And we denote variance by (σ^2 or μ_2 or $\mu_2' - \mu_1'^2$ or $E(X^2) - [E(X)]^2$)

$$= \sum_i p_i (x_i - \bar{x})^2 \text{ or } \sum_i p_i (x_i - m)^2$$

Standard deviation (σ) = $+\sqrt{\text{variance}}$.

11.4 BINOMIAL DISTRIBUTION

The binomial distribution is an important discrete probability distribution that arises from a sequence of independent trials, each of which has only two possible outcomes—success or failure. The distribution is associated with the work of James Bernoulli, who developed the mathematical foundation of repeated independent trials.

Suppose an experiment is performed (n) times independently, and let (X) denote the random variable representing the number of successes obtained in these (n) trials.

$$P(X = r) = {}^nC_r p^r q^{n-r}, \quad r = 0, 1, 2, \dots, n$$

The probability distribution of the random variable X is therefore given by

X	0	1	2	 r	 n
$P(X)$	${}^nC_0 p^0 q^n$	${}^nC_1 p^1 q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	${}^nC_r p^r q^{n-r}$	${}^nC_n p^n q^{n-n}$

Hence, the probability distribution is called the binomial distribution because for $r = 0, 1, 2, \dots, n$, are the probabilities of the successive terms of the binomial expansion of $(q + p)^n$.

Note. 1. The constant n, p, q are called parameters of the distribution.

2. Also denoted $b(r; n, p) = {}^nC_r p^r q^{n-r}, \quad r = 0, 1, 2, \dots, n$

3. For N set of n trial the successes $0, 1, 2, \dots, r, \dots, n$ are given by $N(q + p)^n$, which is called binomial distribution.

11.4.1 CONSTANTS OF BINOMIAL DISTRIBUTION

For the binomial distribution

$$P(X = r) = {}^nC_r P^r q^{n-r}, \quad r = 0, 1, 2, \dots, n$$

$$, \quad p + q = 1.$$

Taking out origin at 0 successes, we have

$$\begin{aligned}
 \mu'_1 = E(X) &= \sum_{r=0}^n r^n C_r p^r q^{n-r} \\
 &= 0 \cdot q^n + 1 \cdot {}^n C_1 p q^{n-1} + 2 \cdot {}^n C_2 p^2 q^{n-2} + 3 \cdot {}^n C_3 p^3 q^{n-3} + \dots \dots \dots + n p^n \\
 &= 0 + n p q^{n-1} + \frac{1}{2} n(n-1) 2 p^2 q^{n-2} + \frac{1}{6} n(n-1)(n-2) 3 p^3 q^{n-3} + \dots \dots + n p^n \\
 &= n p [q^{n-1} + {}^{n-1} C_1 p q^{n-2} + {}^{n-1} C_2 p^2 q^{n-3} + \dots \dots + p^{n-1}] \\
 &= n p (q + p)^{n-1}, \text{ since } q + p = 1 \\
 &= n p
 \end{aligned}$$

$$\therefore \text{Mean } (\bar{X}) = n p$$

$$\begin{aligned}
 \mu'_2 &= E(X^2) = \sum_{r=0}^n r^2 {}^n C_r p^r q^{n-r} \\
 &= \sum_{r=0}^n [r + r(r-1)] {}^n C_r p^r q^{n-r} \\
 &= \sum_{r=0}^n r^n C_r p^r q^{n-r} + \sum_{r=0}^n r(r-1) {}^n C_r p^r q^{n-r} \\
 &= n p + \sum_{r=0}^n r(r-1) \frac{n(n-1)}{r(r-1)} {}^{n-2} C_{r-2} p^r q^{n-r} \\
 &= n p + n(n-1) p^2 \sum_{r=0}^n {}^{n-2} C_{r-2} p^{r-2} q^{n-r} \\
 &= n p + n(n-1) p^2 (q + p)^{n-2}, \text{ since } q + p = 1 \\
 &= n p + n(n-1) p^2 \\
 &= n p [1 + (n-1) p] \\
 &= n p [1 + n p - p] \\
 &= n p [q + n p] \\
 &= n p q + n^2 p^2
 \end{aligned}$$

$$\therefore \text{Variance } (\mu_2) = \mu'_2 - \mu_1'^2 \text{ or } E(X^2) - [E(X)]^2$$

$$\begin{aligned}
 &= n p q + n^2 p^2 - n^2 p^2 \\
 &= n p q
 \end{aligned}$$

$$\text{Variance } (\mu_2 \text{ or } \sigma^2) = n p q$$

and

$$\text{Standard deviation } (\sigma) = \sqrt{n p q}$$

$$\begin{aligned}\mu'_3 &= E(X^3) = \sum_{r=0}^n r^3 {}^nC_r p^r q^{n-r} \\ &= \sum_{r=0}^n [r(r-1)(r-2) + 3r(r-1) + r] {}^nC_r p^r q^{n-r}\end{aligned}$$

On simplification in above similar manner, we have

$$\begin{aligned}&= n(n-1)(n-2)p^3(q+p)^{n-3} + 3n(n-1)p^2(q+p)^{n-2} + np(q+p)^{n-1} \\ &= n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3 = npq(q-p) \\ \mu'_4 &= E(X^4) = \sum_{r=0}^n r^4 {}^nC_r p^r q^{n-r} \\ &= \sum_{r=0}^n \{r(r-1)(r-2)(r-3) + 6r(r-1)(r-2) + 7r(r-1) + r\} {}^nC_r p^r q^{n-r}\end{aligned}$$

On simplification in above similar manner, we have

$$\begin{aligned}&= n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4 = npq[1 + 3(n-2)pq]\end{aligned}$$

Now,

$$\begin{aligned}\beta_1 &= \frac{\mu_3'}{\mu_2'^3} = \frac{[npq(q-p)]^2}{(npq)^3} = \frac{(q-p)^2}{npq} \\ \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{3n^2p^2q^2 + npq(1-6pq)}{n^2p^2q^2} = 3 + \frac{1-6pq}{npq} \text{ (Kurtosis)} \\ \gamma_1 &= \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} \text{ (Skewness)} \\ \gamma_2 &= \beta_2 - 3 = \frac{1-6pq}{npq}\end{aligned}$$

Note. 1. The mean of binomial distribution is greater than the variance, since $0 < q < 1$.

2. If skewness is zero i.e., $\beta_1 = 0$ so $p = q = \frac{1}{2}$.

3. If $p < \frac{1}{2}$, skewness is positive, if $p > \frac{1}{2}$, negative.

11.5 SOLVED EXAMPLES

Example 1: Show that for the binomial distribution $(q + p)^n, \mu_{r+1} = pq \left(nr\mu_{r-1} + \frac{d\mu_r}{dp} \right)$ where, μ_r is the r th moment about the mean.

Hence, obtain μ_2, μ_3 and μ_4 .

Solution: Using the definition of r th moment about mean, we have

$$\begin{aligned}\mu_r &= E[X - E(X)]^r \\ &= \sum_{x=0}^n (x - np)^r {}^nC_x p^x q^{n-x}\end{aligned}$$

Differentiating with respect to p both sides, we get

$$\begin{aligned}\frac{d\mu_r}{dp} &= \sum_{x=0}^n -nr(x - np)^{r-1} {}^nC_x p^x q^{n-x} \\ &\quad + \sum_{x=0}^n (x - np)^r {}^nC_x [xp^{x-1}q^{n-x} - (n-x)p^x q^{n-x-1}] \\ &= -nr \sum_{x=0}^n (x - np)^{r-1} {}^nC_x p^x q^{n-x} \\ &\quad + \sum_{x=0}^n (x - np)^r {}^nC_x \left(\frac{1}{pq} \right) [xq - np + xp] p^x q^{n-x} \\ &= -nr\mu_{r-1} + \frac{1}{pq} \sum_{x=0}^n (x - np)^r {}^nC_x [x(p + q) - np] p^x q^{n-x} \\ &= -nr\mu_{r-1} + \frac{1}{pq} \sum_{x=0}^n (x - np)^{r+1} {}^nC_x p^x q^{n-x} \\ &= -nr\mu_{r-1} + \frac{1}{pq} \mu_{r+1} \\ \Rightarrow \mu_{r+1} &= pq \left(nr\mu_{r-1} + \frac{d\mu_r}{dp} \right)\end{aligned}$$

Putting $r = 1, 2, 3$ in (1), we get

$$\begin{aligned}
\mu_2 &= pq \left(\frac{d\mu_1}{dp} + n \cdot 1 \cdot \mu_0 \right) \\
&= pq(0 + n), \text{ since } \mu_1 = 0, \mu_0 = 1 \\
&= npq \\
\mu_3 &= pq \left(\frac{d\mu_2}{dp} + n \cdot 2\mu_1 \right) \\
&= pq(nq - np + 2n \cdot 0) = npq(q - p) \\
\mu_4 &= pq \left(\frac{d\mu_3}{dp} + n \cdot 3\mu_2 \right) \\
&= pq \left[\frac{d}{dp} \{np(1-p)(1-2p)\} + 3n \cdot npq \right] \\
&= pq \left[\frac{d}{dp} \{np - 3np^2 + 2np^3\} + 3n^2pq \right] \\
&= pq[n - 6np + 6np^2 + 3n^2pq] \\
&= npq - 6np^2q + 6np^3q + 3n^2p^2q^2 \\
&= 3n^2p^2q^2 + npq + 6npq(p^2 - p) \\
&= 3n^2p^2q^2 + npq + 6np^2q(p - 1) \\
&= 3n^2p^2q^2 + npq - 6np^2q^2 \\
&= 3n^2p^2q^2 + npq(1 - 6pq).
\end{aligned}$$

Example 2: A die is tossed thrice. Getting an even number is considered as success. What is variance of the binomial distribution?

Solution: Let p be the probability of getting an even number i.e.,

$$p = \frac{3}{6} = \frac{1}{2} \text{ then } q = 1 - \frac{1}{2} = \frac{1}{2}, n = 3$$

The variance of binomial distribution = npq , where $q = 1 - p$

$$= 3 \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}.$$

Example 3: The mean and variance of a binomial distribution are 4 and 3 respectively. Find the probability of getting exactly six successes in this distribution.

Solution: The mean of binomial distribution

$$= np = 4 \text{ (given)}$$

$$= npq = 3 \text{ (given)}$$

and the variance of binomial distribution Using (1) and (2), we have

$$npq = 3 \Rightarrow 4 \cdot q = 3 \Rightarrow q = \frac{3}{4}$$

$$p = 1 - q = 1 - \frac{3}{4} = \frac{1}{4}$$

by (1),

$$np = 4 \Rightarrow n \times \frac{1}{4} = 4 \Rightarrow n = 16$$

$$\text{The probability of 6 success} = {}^{16}C_6 \left(\frac{1}{4}\right)^6 \left(\frac{3}{4}\right)^{10} = \frac{8008 \times 3^{10}}{4^{16}}.$$

Example 4: In 256 sets of 12 tosses of a coin, in how many cases one can expect 8 heads and 4 tails.

Solution: Let p be the probability of head i.e., $p = \frac{1}{2}$ and q be the probability of tail i.e., $q = \frac{1}{2}$.

$$n = 12, N = 256.$$

The binomial distribution is $N(q + p)^n = 256 \left(\frac{1}{2} + \frac{1}{2}\right)^{12}$

The probability of 8 heads and 4 tails in 12 trial is

$${}^{12}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^4 = \frac{495}{4096}$$

The expected number of such cases in 256 sets

$$= 256 \times \frac{495}{4096} = 30.9 = 31 \text{ (nearly).}$$

Example 5: During war, 1 ship out of 9 was sunk on an average in making a voyage. What was the probability that exactly 3 out of a convoy of 6 ships would arrive safely?

Solution: Let p be the probability of a ship arriving safely i.e.,

$$p = 1 - \frac{1}{9} = \frac{8}{9} \text{ then } q = 1 - p = \frac{1}{9}$$

$$n = 6, N = 1$$

The binomial distribution is

$$N(q + p)^n = \left(\frac{1}{9} + \frac{8}{9}\right)^6$$

The probability that exactly 3 ship arrive safely

$$= {}^6C_3 \left(\frac{8}{9}\right)^3 \left(\frac{1}{9}\right)^3 = 20 \times \frac{512}{9^6} = \frac{10240}{9^6}.$$

Example 6: The probability that a pen manufactured by a company will be defective is $\frac{1}{10}$. If 12 such pens are manufactured, find the probability that

- (i) exactly two will be defective.
- (ii) at least one will be defective.
- (iii) none will be defective.

Solution: Given that the probability of a defective pen $= \frac{1}{10} = 0.1$

Then the probability of a non-defective pen is $= 1 - 0.1 = 0.9$

- (i) The probability that exactly two pens will be defective

$$= {}^{12}C_2 (0.1)^2 (0.9)^{10} = 0.2301$$

- (ii) The probability that at least two will be defective

$$\begin{aligned}
&= 1 - \{ \text{Probability that either none or one is non-defective} \} \\
&= 1 - \{ {}^{12}C_0(0.9)^{12} + {}^{12}C_1(0.1)(0.9)^{11} \} \\
&= 1 - 0.6588 = 0.3412
\end{aligned}$$

(iii) The probability that none will be defective

$$= {}^{12}C_{12}(0.9)^{12} = 0.2833$$

Example 7: Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or six?

Solution: Let p be the chance of getting 5 or 6 with one die then

$$\begin{aligned}
p &= \frac{1}{3}, \\
n &= 6, \quad N - \frac{1}{3} = \frac{2}{3} \\
n &= 729
\end{aligned}$$

Since, dice are in sets of 6 and there are 729 sets.

The binomial distribution is

$$N(q + p)^n = 729 \left(\frac{2}{3} + \frac{1}{3} \right)^6$$

The required number is

$$\begin{aligned}
&= 729 \left\{ {}^6C_3 \left(\frac{1}{3} \right)^3 \left(\frac{2}{3} \right)^3 + {}^6C_4 \left(\frac{1}{3} \right)^4 \left(\frac{2}{3} \right)^2 + {}^6C_5 \left(\frac{1}{3} \right)^5 \left(\frac{2}{3} \right) + {}^6C_6 \left(\frac{1}{3} \right)^6 \right\} \\
&= \frac{729}{3^6} \{160 + 60 + 12 + 1\} = 233
\end{aligned}$$

Example 8: A perfect cubical die is thrown a large number of times in set of 8. The occurrence of 5 or 6 is called a success. In what proportion of sets do you expect 3 successes.

Solution: Let p = the chance of occurrence 5 or 6 with one die = $\frac{2}{6} = \frac{1}{3}$

$$q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}, n = 8$$

The binomial distribution is given by

$$N(q + p)^n = N\left(\frac{2}{3} + \frac{1}{3}\right)^8$$

The number of sets in which 3 successes are expected

$$\begin{aligned} N \left\{ {}^8C_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^5 \right\} &= N \frac{8!}{3! 5!} \cdot \frac{2^5}{3^8} \\ &= N \cdot \frac{8 \times 7 \times 6 \times 5! \times 2^5}{3 \times 2 \times 5! \times 3^8} = N \frac{56 \times 32}{243 \times 27} = N \frac{1792}{6591} \end{aligned}$$

The percentage is $= N \frac{1792}{6591} \times \frac{100}{N} = 27.3\%$.

Example 9: In 800 families with 4 children each how many families would be expected to have (i) 2 boys and 2 girls (ii) at least one boy (iii) no girl (iv) at least two girls? Assuming that equal probabilities for boys and girls.

Solution: Since, the probabilities for boys and girls are equal

Then let p = probability of having a boy = $\frac{1}{2}$

and

$$q = \text{probability of having a girl} = \frac{1}{2}$$

Here $n = 4$, $N = 800$

The binomial distribution is $N(q + p)^n = 800 \left(\frac{1}{2} + \frac{1}{2}\right)^4$.

(i) The expected number of families having 2 boys and 2 girls

$$= 800 {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{800 \times 6}{16} = 300$$

(ii) The expected number of families having at least one boy

$$= 800 \left\{ {}^4C_1 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right) + {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 + {}^4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right) + {}^4C_4 \left(\frac{1}{2}\right)^4 \right\}$$

$$= \frac{800}{2^4} [4 + 6 + 4 + 1] = 750$$

(iii) The expected number of families having no girl

$$= 800 {}^4C_4 \left(\frac{1}{2}\right)^4 = \frac{800}{16} = 50.$$

(iv) The expected number of families having at most two girls

$$= 800 \left\{ {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 + {}^4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right) + {}^4C_4 \left(\frac{1}{2}\right)^4 \right\}$$

$$= \frac{800}{2^4} [6 + 4 + 1] = 550$$

Example 10: The probability that a men aged 60 years will live up to 70 years is 0.65. What is the probability that out of ten men now 60 years, at least 7 would live up to 70 years?

Solution: Given that the probability of a men aged 60 years will live up to 70 years is = 0.65

Then $p = 0.65, q = 1 - p = 1 - 0.65 = 0.35$

$$n = 10$$

Let X be the number of men who live up to 70 years.

The probability that out of 10 men, r men will up to 70 years is

$$P(X = r) = {}^nC_r p^r q^{n-r} = {}^{10}C_r (0.65)^r (0.35)^{10-r}$$

The probability of at least 7 men would live up to 70 years.

$$\begin{aligned}
 P(X \geq 7) &= P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\
 &= {}^{10}C_7(0.65)^7(0.35)^3 + {}^{10}C_8(0.65)^8(0.35)^2 + {}^{10}C_9(0.65)^9(0.35)^1 + {}^{10}C_{10}(0.65)^{10} \\
 &= 0.2523 + 0.1756 + 0.0725 + 0.0135 = 0.5140.
 \end{aligned}$$

Example 11: On weekdays, between 2:00 p.m. and 3:00 p.m., approximately 1 out of every 15 telephone calls is found to be busy. If 6 telephone numbers are selected at random, determine the probability that:

- (i) no more than three of the selected numbers are busy; and
- (ii) at least three of the selected numbers are busy.

Solution: Let p be the probability of a telephone number out of 15 called between 2 p.m. and 3 p.m. on week days is busy, i.e., $p = \frac{1}{15}$ then $q = 1 - \frac{1}{15} = \frac{14}{15}$, $n = 6$, $N = 1$

The binomial distribution is $N(q + p)^n = \left(\frac{14}{15} + \frac{1}{15}\right)^6$

(i) The probability that not more than three will be busy

$$\begin{aligned}
 P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\
 &= {}^6C_0 \left(\frac{14}{15}\right)^6 + {}^6C_1 \left(\frac{14}{15}\right)^5 \left(\frac{1}{15}\right) + {}^6C_2 \left(\frac{14}{15}\right)^4 \left(\frac{1}{15}\right)^2 + {}^6C_3 \left(\frac{14}{15}\right)^3 \left(\frac{1}{15}\right)^3 \\
 &= \frac{(14)^3}{(15)^6} [2744 + 1176 + 210 + 20] \\
 &= \frac{2744 \times 4150}{(15)^6} = 0.9997
 \end{aligned}$$

(ii) The probability that at least three of them will be busy

$$\begin{aligned}
 P(X \geq 3) &= P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) \\
 &= {}^6C_3 \left(\frac{14}{15}\right)^3 \left(\frac{1}{15}\right)^3 + {}^6C_4 \left(\frac{14}{15}\right)^2 \left(\frac{1}{15}\right)^4 + {}^6C_5 \left(\frac{14}{15}\right) \left(\frac{1}{15}\right)^5 + {}^6C_6 \left(\frac{1}{15}\right)^6 \\
 &= 0.005.
 \end{aligned}$$

Example 12: The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to these data:

x	0	1	2	3	4	5	6	7	8	9	10	Total
f	6	20	28	12	8	6	0	0	0	0	0	80

Solution: Given that $n = 10$, $N = 80$ and $\Sigma f = 80$.

The arithmetic mean = $\frac{\Sigma fx}{\Sigma f}$

$$= \frac{1 \times 20 + 2 \times 28 + 3 \times 12 + 4 \times 8 + 5 \times 6 + 6 \times 0 + 7 \times 0 + 8 \times 0 + 9 \times 0 + 10 \times 0}{80}$$

$$= \frac{174}{80}$$

The mean of a binomial distribution = np

$$\therefore np = \frac{174}{80} \Rightarrow p = \frac{174}{80 \times 10} = 0.2175$$

Then

$$q = 1 - p = 0.7825.$$

Hence, the binomial distribution to be fitted to the data is

$$N(q + p)^n = 80(0.7825 + 0.2175)^{10}$$

Thus, the theoretical frequencies are

x	0	1	2	3	4	5	6	7	8	9	10
f	6.9	19.1	24.0	17.8	8.6	2.9	0.7	0.1	0	0	0

11.6 BERNOULLI DISTRIBUTION

The **Bernoulli distribution** is one of the simplest and most fundamental probability distributions in statistics and probability theory. It models the outcomes of a single trial of an experiment with only two possible outcomes: **success (1)** and **failure (0)**. The distribution is named after Swiss mathematician Jacob Bernoulli.

Characteristics of the Bernoulli Distribution:

1. **Outcomes:**

- The trial results in one of two outcomes: success ($x = 1$) or failure ($x = 0$).
- These outcomes are often associated with probabilities $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 \leq p \leq 1$.

2. **Probability Mass Function (PMF):**

The PMF of a Bernoulli random variable X is given by:

$$P(X = x) = p^x(1 - p)^{1-x}, x = 0, 1$$

3. **Expected Value:**

The expected value or mean of a Bernoulli random variable is:

$$E[X] = p.$$

4. **Variance:**

The variance measures the spread of the distribution:

$$\text{Var}(X) = p(1 - p)$$

5. **Special Case:**

- When $p=0.5$, the Bernoulli distribution is symmetric.
- As p approaches 0 or 1, the distribution becomes more skewed.

Applications of the Bernoulli Distribution:

1. **Binary Outcomes:** Used to model processes with two outcomes, such as flipping a coin (heads/tails), passing or failing an exam, or defective/non-defective products.

2. **Building Blocks for Other Distributions:** The Bernoulli distribution forms the basis for other distributions like the Binomial distribution, which models multiple Bernoulli trials.
3. **Decision-Making:** Applied in hypothesis testing, machine learning (classification problems), and decision-making scenarios.

Example: If a coin is flipped and the probability of heads ($X=1$) is 0.6, then the PMF is:

$$P(X = 1) = 0.6, P(X = 0) = 0.4.$$

Here, the mean is 0.6, and the variance is $0.6 \cdot 0.4 = 0.240$.

Understanding the Bernoulli distribution is crucial for working with more complex probabilistic models and concepts in data science, statistics, and various applied fields.

CHECK YOUR PROGRESS

1. Define Binomial Distribution and state its assumptions.

.....

.....

.....

2. Explain the mean and variance of a Binomial Distribution.

.....

.....

.....

3. Discuss the characteristics and applications of Binomial Distribution.

11.7 SUMMARY

In this unit, we have discussed the following concepts

□ **Bernoulli Distribution:**

- Represents a single trial with two possible outcomes: success (1) and failure (0).
- The probability of success is denoted by p , and the probability of failure is $1 - p$.

- **Probability Mass Function (PMF):**

$$P(X = x) = p^x(1 - p)^{1-x}, x = 0, 1$$

- **Mean (Expectation):** $E(X) = p$
- **Variance:** $Var(X) = p(1 - p)$

□ **Binomial Distribution:**

- Represents the number of successes in n independent Bernoulli trials, each with probability p of success.

- **Probability Mass Function (PMF):**

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, 1, 2, \dots, n$$

- **Mean (Expectation):** $E(X) = np$
- **Variance:** $Var(X) = np(1 - p)$

□ **Applications:**

Bernoulli and Binomial distributions are used in quality control,

medical trials, risk analysis, and any scenario involving binary outcomes and repeated trials.

11.8 EXERCISES

- 1 The mean and variance of binomial distribution are 4 and $\frac{4}{3}$ respectively. Find $P(X \geq 1)$.
- 2 The items produced by a firm are supposed to contain 5% defective items. What is the probability that a sample of 8 items will contain less than 2 defective items?
- 3 Find the binomial distribution for which the mean is 4 and variance is 3.
- 4 Bring out the fallacy in the statement:
The mean of a binomial distribution is 3 and variance is 4.
5. If 10% bolts produced by a machine are defective, determine the probability that out of 10 bolts chosen at random (i) 1 (ii) none (iii) at least one (iv) at most two bolts, will be defective.
6. An irregular six-faced die is thrown, and the expectation that in 100 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws would you expect it to give no even number?
7. The incidence of occupational disease in an industry is such that the workmen have a 25% chance of suffering from it. What is the probability that out of six workmen 4 or more will contract the disease?
8. A box contains ' a ' red and ' b ' black balls, ' n ' balls are drawn. Find the expected number of red balls drawn.
9. Assuming that half the population are consumers of rice so that the chance of an individual being a consumer is $\frac{1}{2}$ and assuming that 100 investigators, each take ten individuals to see whether they are consumers, how many investigators do you expect to report that three people or less are consumers?

10. Find the most probable number of heads in 99 tossing of a biased coin, given that the probability of a head in a single tossing is $\frac{3}{5}$.

11. In litters of 4 mice the number of litters which contained 0,1,2,3,4 females were noted. The figures are given in the table below:

Number of female mice	0	1	2	3	4	Total
Number of litters	8	32	34	24	5	103

In the chance of obtaining a female in a single trial is assumed constant, estimate this constant of unknown probability.

12. The probability of a man hitting a target in $\frac{1}{4}$. He fires 7 times. What is the probability of his hitting the target at least twice. How many times must he fires so that the probability of his hitting at least one is greater than $\frac{2}{3}$?

13. Eight coins are tossed at a time, 256 times. Number of heads observed at each throw are recorded and the results are as given below. Find the expected frequency and fit a binomial distribution. What are the theoretical values of the mean and standard deviation? Calculate also the mean and standard deviation of the observed frequencies.

14. The probability that a bomb dropped from a plane strike the target is $\frac{1}{5}$. If six such bombs are dropped find the probability that at least two will strike the target, and exactly two will strike the target.

11.9 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.

4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.



UNIT 12 POISSON AND NORMAL DISTRIBUTIONS

STRUCTURE

- 12.1 Introduction
- 12.2 Objectives
- 12.3 Poisson's distribution
- 12.4 Normal distributions
- 12.5 Summary
- 12.6 Exercises
- 12.7 References

12.1 INTRODUCTION

In this unit, we explore two widely used probability distributions: the Poisson distribution and the Normal distribution. The Poisson distribution models the number of events that occur within a fixed interval of time or space, especially when events happen independently and randomly. The Normal distribution, often called the Gaussian distribution, describes continuous data that tend to cluster around a central value. Understanding these distributions is essential for statistical modelling, hypothesis testing, and real-world applications involving random events or natural variations.

12.2 OBJECTIVES

After completing this unit, students will be able to

- Define and explain the Poisson and Normal distributions.
- Identify scenarios where the Poisson and Normal distributions are applicable.

- Compute probabilities using the Poisson distribution for count data.
- Understand the properties of the Normal distribution and its importance in statistics.
- Apply the Poisson and Normal distributions to solve real-world problems.

12.3 POISSON'S DISTRIBUTION

Poisson distribution was discovered by a French mathematician Simeon Denis Poisson in 1837. Poisson distribution is also a discrete probability distribution of a discrete random variable, which has no upper bound. Poisson distribution is a limiting form of the binomial distribution $(q + p)^n$ under the following conditions:

- (i) $n \rightarrow \infty$, i.e., the number of trials is indefinitely large.
- (ii) $p \rightarrow 0$, i.e., the constant probability of success for each trial is indefinitely small.
- (iii) np is a finite quantity, say m .

Thus, $p = \frac{m}{n}$, $q = 1 - \frac{m}{n}$, where m is a positive real number.

Poisson distribution deals with situations explained below:

- (i) Number of suicides or deaths by heart attack in 1 minute.
- (ii) Number of accidents that take place on a busy road in time t .
- (iii) Number of printing mistakes at each unit of the book.
- (iv) Number of cars passing a certain street in time t .
- (v) Emission of radioactive particles.
- (vi) Number of faulty blades in a packet of 1000.
- (vii) Number of person born blind per year in a certain village.
- (viii) Number of telephone calls received at a particular switch board in 1 minute.

The probability distribution of a random variable X is said to be a Poisson distribution, if the random variable assumes only non-negative values and its probability distribution is given by

$$P(X = r) = \frac{m^r e^{-m}}{r!}, \quad r = 0, 1, 2, \dots$$

where, m is called the parameter of the distribution and $m > 0$

The probability of r successes in a binomial distribution

$$\begin{aligned} P(X = r) &= {}^n C_r p^r q^{n-r} \\ &= \frac{n!}{r!(n-r)!} p^r q^{n-r} = \frac{n(n-1)(n-2) \dots (n-r+1)}{r!} p^r (1-p)^{n-r} \\ &= \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \dots \left(1-\frac{r-1}{n}\right) p^r (1-p)^{n-r} n^r}{r!} \\ &= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \dots \left(1-\frac{r-1}{n}\right) (np)^r (1-p)^{n-r}}{r!} \end{aligned}$$

Taking limit $n \rightarrow \infty, p \rightarrow 0$ such that $np = m$

$$\begin{aligned} &= \lim_{n \rightarrow \infty} \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \dots \left(1-\frac{r-1}{n}\right)}{r!} m^r \left(1-\frac{m}{n}\right)^{n-r} \\ &= \frac{m^r}{r!} \lim_{n \rightarrow \infty} \left(1-\frac{m}{n}\right)^{n-r} = \frac{m^r}{r!} \lim_{n \rightarrow \infty} \left(1-\frac{m}{n}\right)^n \left(1-\frac{m}{n}\right)^{-r} \\ &= \frac{m^r}{r!} e^{-m} \cdot 1 \\ &= \frac{m^r e^{-m}}{r!}, \quad r = 0, 1, 2, \dots \end{aligned}$$

The probability of $0, 1, 2, \dots, r, \dots$ successes are

$$e^{-m}, \frac{e^{-m} m}{1!}, \frac{e^{-m} m^2}{2!}, \frac{e^{-m} m^3}{3!}, \dots, \frac{e^{-m} m^r}{r!}, \dots \text{ respectively.}$$

Note. It should be noted that $\sum_{r=0}^{\infty} P(X = r) = \sum_{r=0}^{\infty} \frac{m^r e^{-m}}{r!} =$

$$e^{-m} \sum_{r=0}^{\infty} \frac{m^r}{r!} = e^{-m} e^m = 1.$$

CONSTANTS OF POISSON DISTRIBUTION

$$\begin{aligned}
 \mu'_1 &= E(X) = \sum_{r=0}^{\infty} r \cdot P(X=r) \\
 &= \sum_{r=0}^{\infty} r \cdot \frac{e^{-m} m^r}{r!} \\
 &= e^{-m} \sum_{r=0}^{\infty} \frac{m^r}{(r-1)!} = e^{-m} \left(m + \frac{m^2}{1!} + \frac{m^3}{2!} + \dots \right) \\
 &= m e^{-m} \left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \dots \right) = m e^{-m} \cdot e^m = m
 \end{aligned}$$

$\therefore$ Mean ($\bar{X}$) = m

$$\begin{aligned}
 \mu'_2 &= E(X^2) = \sum_{r=0}^{\infty} r^2 \cdot P(X=r) \\
 &= \sum_{r=0}^{\infty} \{r(r-1) + r\} \frac{e^{-m} m^r}{r!} \\
 &= \sum_{r=0}^{\infty} \frac{r(r-1)e^{-m} m^r}{r!} + \sum_{r=0}^{\infty} \frac{r \cdot e^{-m} m^r}{r!} \\
 &= e^{-m} \sum_{r=0}^{\infty} \frac{m^r}{(r-2)!} + m \\
 &= e^{-m} \left(m^2 + \frac{m^3}{1!} + \frac{m^4}{2!} + \dots \right) + m \\
 &= m^2 e^{-m} \left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \dots \right) + m \\
 &= m^2 e^{-m} \cdot e^m + m = m^2 + m
 \end{aligned}$$

$$\therefore \text{Variance } (\mu_2) = \mu'_2 - \mu_1^2 \text{ or } E(X^2) - [E(X)]^2$$

$$= m^2 + m - m^2 = m$$

Variance (μ_2 or σ^2) = m

Hence, mean of Poisson distribution = Variance of Poisson distribution

Standard deviation (σ) = $\sqrt{m}$

$$\mu'_3 = E(X^3) = \sum_{r=0}^{\infty} r^3 P(X=r)$$

$$\begin{aligned}
&= \sum_{r=0}^{\infty} \{r(r-1)(r-2) + 3r(r-1) + r\} \frac{e^{-m} m^r}{r!} \\
&= \sum_{r=0}^{\infty} \frac{r(r-1)(r-2)e^{-m} m^r}{r!} + 3 \sum_{r=0}^{\infty} \frac{r(r-1)e^{-m} m^r}{r!} + \sum_{r=0}^{\infty} \frac{r e^{-m} m^r}{r!} \\
&= e^{-m} m^3 \sum_{r=0}^{\infty} \frac{m^{r-3}}{(r-3)!} + 3e^{-m} m^2 \sum_{r=0}^{\infty} \frac{m^{r-2}}{(r-2)!} + e^{-m} \cdot m \sum_{r=0}^{\infty} \frac{m^{r-1}}{(r-1)!} \\
&= mm^3 + 3m^2 + m. \\
\mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \\
&= m \\
\mu_4' &= E(X^4) = \sum_{r=0}^{\infty} r^4 P(X=r) \\
&= \sum_{r=0}^{\infty} \{r(r-1)(r-2)(r-3) + 6r(r-1)(r-2) + 7r(r-1) + r\} \frac{e^{-m} m^r}{r!} \\
&= e^{-m} m \sum_{r=0}^4 \sum_{r=0}^{\infty} \frac{m^{r-4}}{(r-4)!} + 6e^{-m} m^3 \sum_{r=0}^{\infty} \frac{m^{r-3}}{(r-3)!} + 7e^{-m} m^2 \sum_{r=0}^{\infty} \frac{m^{r-2}}{(r-2)!} \\
&\quad + e^{-m} m \sum_{r=0}^{\infty} \frac{m^{r-1}}{(r-1)!} \\
&= m^4 + 6m^3 + 7m^2 + m. \\
\mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 \\
&= (m^4 + 6m^3 + 7m^2 + m) - 4(mm^3 + 3m^2 + m)(m) + 6(mm^2 + m)m - 3m^4 \\
&= 3mm^2 + m \\
\beta_1 &= \frac{\mu_3'^2}{\mu_2'^3} = \frac{m^2}{m^3} = \frac{1}{m} \\
\beta_2 &= \frac{\mu_4}{\mu_2'^2} = \frac{3m^2 + m}{m^2} = 3 + \frac{1}{m} \\
\gamma_1 &= \sqrt{\beta_1} = \frac{1}{\sqrt{m}} \\
\gamma_2 &= \beta_2 - 3 = \frac{1}{m}
\end{aligned}$$

When $m \rightarrow \infty$, $\beta_1 = 0$ and $\beta_2 = 3$

Hence, Poisson distribution is always positively skewed.

Note. $\mu_1 = \mu_1' - \mu_0' m = m - m = 0$, since

$$\mu_n = \mu_n' - {}^nC_1 \mu_{n-1}' \cdot m + {}^nC_2 \mu_{n-2}' m^2 - \dots$$

RECURRENCE FORMULA FOR POISSON DISTRIBUTION

Poisson distribution for r successes, we have

$$P(X = r) = \frac{m^r e^{-m}}{r!}, \quad r = 0, 1, 2, 3, \dots \quad (1)$$

where m is the mean of successes.

$$\text{Also } P(X = r + 1) = \frac{m^{r+1} e^{-m}}{(r + 1)!}, \quad r = 0, 1, 2, \dots \quad (2)$$

divide (2) by (1), we get

$$\begin{aligned} \frac{P(r + 1)}{P(r)} &= \frac{m}{r + 1} \\ \Rightarrow P(r + 1) &= \frac{m}{r + 1} P(r) \end{aligned}$$

This is the recurrence formula for Poisson distribution.

MODE OF THE POISSON DISTRIBUTION

The value of r which gives the greatest probability is the mode of the Poisson distribution. Thus

$$\frac{m^{r-1} e^{-m}}{(r - 1)!} \leq \frac{m^r e^{-m}}{(r)!} \geq \frac{m^{r+1} e^{-m}}{(r + 1)!}$$

Where, $m > r > m - 1$

$$\text{i.e., } m - 1 < r < m$$

Thus, if m is an integer then there are two modes $m - 1$ and m . If m is not an integer then the mode is the integral value between $m - 1$ and m .

Example 1: Show that for the Poisson distribution with mean m

$$\mu_{r+1} = r m \mu_{r-1} + m \frac{d\mu_r}{dm}, \text{ where } \mu_r = \sum_{x=0}^{\infty} (x-m)^r \frac{e^{-m} m^x}{x!}.$$

Solution: Given that $\mu_r = \sum_{x=0}^{\infty} (x-m)^r e^{-m} \frac{m^x}{x!}$ Differentiating with respect to m both sides, we get

$$\begin{aligned} \frac{d\mu_r}{dm} &= \sum_{x=0}^{\infty} (-r)(x-m)^{r-1} \frac{e^{-m} m^x}{x!} + \sum_{x=0}^{\infty} (x-m)^r \cdot \frac{1}{x!} \{e^{-m} \cdot x m^{x-1} - m^x e^{-m}\} \\ &= -r \sum_{x=0}^{\infty} (x-m)^{r-1} \frac{e^{-m} m^x}{x!} + \sum_{x=0}^{\infty} \frac{(x-m)^r}{x!} e^{-m} m^{x-1} \{x-m\} \\ &= -r \mu_{r-1} + \sum_{x=0}^{\infty} (x-m)^{r+1} \frac{e^{-m} m^{x-1}}{x!} \\ &= -r \mu_{r-1} + \frac{1}{m} \sum_{x=0}^{\infty} (x-m)^{r+1} \frac{e^{-m} m^x}{x!} \\ &= -r \mu_{r-1} + \frac{1}{m} \mu_{r+1} \\ \Rightarrow \mu_{r+1} &= r m \mu_{r-1} + m \frac{d\mu_r}{dm} \end{aligned}$$

Example 2: Six coins are tossed simultaneously and the experiment is repeated 6,400 times. Using the Poisson distribution as an approximation to the binomial distribution, determine the approximate probability of obtaining six heads in a single trial.

Solution: Let p be the probability of getting all the six head in a throw of six coins. i.e., $p = \frac{1}{2^6} = \frac{1}{64}$.

Here, $n = 6400$

The mean of Poisson distribution $= np = 6400 \times \frac{1}{64} = 100$

The probability of getting all six heads x times according to Poisson distribution is $\frac{e^{-100} (100)^x}{x!}$.

Example 3: Show that for a Poisson's distribution $\gamma_1 \gamma_2 \sigma m = 1$, when σ and m are the standard deviation and mean respectively.

Solution: In Poisson distribution $\gamma_1 = \frac{1}{\sqrt{m}}$ and $\gamma_2 = \frac{1}{m}$

The standard deviation
and the mean

$$\begin{aligned}(\sigma) &= \sqrt{m} \\ &= m\end{aligned}$$

Now we have

$$\gamma_1 \gamma_2 \sigma m = \frac{1}{\sqrt{m}} \frac{1}{m} \sqrt{m} m = 1.$$

Example 4: If X is a Poisson variate such that $P(X = 2) = 9P(X = 4) + 90P(X = 6)$. Find the mean and variance of X .

Solution: The Poisson distribution is

$$P(X = r) = \frac{m^r e^{-rm}}{r!}, r = 0, 1, 2, \dots, m > 0$$

Given that

$$P(X = 2) = 9P(X = 4) + 90P(X = 6)$$

$$\begin{aligned}\Rightarrow & \frac{m^2 e^{-m}}{2!} = 9 \frac{m^4 e^{-m}}{4!} + 90 \frac{m^6 e^{-m}}{6!} \\ \Rightarrow & \frac{m^2}{2} = \frac{3m^4}{8} + \frac{m^6}{8} \\ \Rightarrow & 4 = 3m^2 + m^4 \Rightarrow mm^4 + 3m^2 - 4 = 0 \\ \Rightarrow & (m^2 - 1)(m^2 + 4) = 0 \Rightarrow (m^2 - 1) = 0 \Rightarrow m = 1\end{aligned}$$

(Taking only real values)

Hence, the mean and variance of X are each equal to 1 .

Example 5: In a Poisson distribution $P(x)$ for $x = 0$ is 0.1. Find the mean given that $\log_e 10 = 2.3026$.

Solution: The Poisson distribution is

$$P(x) = \frac{m^x e^{-m}}{x!}, \quad x = 0, 1, 2, \dots, m > 0$$

Given that for $x = 0$ $P(x) = 0.1$

Then by (1), we have

$$0.1 = \frac{m^0 e^{-m}}{0!} \Rightarrow e^{-m} = 0.1 \Rightarrow e^m = 10$$

$$\Rightarrow m = \log_e 10 = 2.3026$$

Example 6: If 2% electric bulbs manufactured by a company be defective, find the probability of (i) 0 (ii) 1 (iii) 2 (iv) 3 defectives in a lot of 100 bulbs.

Solution: Let p be the probability of electric bulbs manufactured by a company be defective.

i.e.,

$$p = \frac{2}{100}.$$

Given that,

$$n = 100$$

we have,

$$np = 100 \times \frac{2}{100} = 2 = m(\text{mean})$$

Let X be the number of electric bulbs manufactured by a company be defective.

Then,

$$P(X = r) = \frac{m^r e^{-m}}{r!} = \frac{2^r e^{-2}}{r!}, \quad r = 0, 1, 2, 3, \dots$$

$$\begin{aligned}
 \text{(i)} \quad P(X = 0) &= \frac{2^0 e^{-2}}{0!} = \frac{1}{e^2} = \frac{1}{(2.718)^2} = 0.135 \\
 \text{(ii)} \quad P(X = 1) &= \frac{2^1 e^{-2}}{1!} = \frac{2}{e^2} = \frac{2}{(2.718)^2} = 2 \times (0.135) = 0.2706 \\
 \text{(iii)} \quad P(X = 2) &= \frac{2^2 e^{-2}}{2!} = \frac{2}{e^2} = \frac{2}{(2.718)^2} = 2 \times (0.135) = 0.2706 \\
 \text{(iv)} \quad P(X = 3) &= \frac{2^3 e^{-2}}{3!} = \frac{4}{3} \times \frac{1}{e^2} = \frac{4}{3} \times \frac{1}{(2.718)^2} = (1.33) \times (0.135) = 0.18.
 \end{aligned}$$

Example 7: A Poisson distribution has a double mode at $x = 4$ and $x = 5$. Find the probability that x will have either of these values $P(x = 4) = P(x = 5)$.

Solution: The Poisson distribution is

$$P(X = r) = \frac{m^r e^{-m}}{r!}, \quad r = 0, 1, 2, \dots$$

Given that

$$P(x = 4) = P(x = 5)$$

$$\Rightarrow \frac{m^4 e^{-m}}{4!} = \frac{m^5 e^{-m}}{5!} \Rightarrow m = 5$$

We know that, $e^{-5} = 0.006738$

Then

$$P(x = 4) = \frac{5^4 \times 0.006738}{4!}$$

The required probability is

$$\begin{aligned}
 P(x = 4) + P(x = 5) &= 2P(x = 4) \\
 &= 2 \times \frac{5^4 \times 0.006738}{4!} = 0.35
 \end{aligned}$$

Hence, it is easy to show that $P(4) > P(3)$ and also greater than $P(6)$.

Example 8: Fit a Poisson's distribution to the following data and calculate theoretical frequencies:

Deaths	0	1	2	3	4
Frequencies	122	60	15	2	1

Solution: The Poisson distribution is

$$P(X = r) = \frac{m^r e^{-m}}{r!}, r = 0, 1, 2, 3, \dots$$

Example 9: If X is a Poisson variate such that $P(X = 2) = 9P(X = 4) + 90P(X = 6)$. Find the value of mean and variance of X .

Solution: The Poisson distribution is

$$P(X = r) = \frac{m^r e^{-m}}{r!}, r = 0, 1, 2, \dots, m > 0$$

Given that

$$P(X = 2) = 9P(X = 4) + 90P(X = 6)$$

$$\Rightarrow \frac{m^2 e^{-m}}{2!} = 9 \frac{m^4 e^{-m}}{4!} + 90 \frac{m^6 e^{-m}}{6!}$$

$$\Rightarrow \frac{m^2}{2} = \frac{3m^4}{8} + \frac{m^6}{8}$$

$$\Rightarrow 4 = 3m^2 + mm^4 \Rightarrow m^4 + 3m^2 - 4 = 0$$

$$\Rightarrow (mm^2 - 1)(mm^2 + 4) = 0 \Rightarrow (mm^2 - 1) = 0 \Rightarrow m = 1$$

(Taking only real values)

Hence, the mean and variance of X are each equal to 1.

Example 10: In a Poisson distribution $P(x)$ for $x = 0$ is 0.1. Find the mean given that $\log_e 10 = 2.3026$.

Solution: The Poisson distribution is

$$P(x) = \frac{m^x e^{-m}}{x!}, \quad x = 0, 1, 2, \dots, m > 0$$

Given that for $x = 0$ $P(x) = 0.1$

Then by (1), we have

$$\begin{aligned} 0.1 &= \frac{m^0 e^{-m}}{0!} \Rightarrow e^{-m} = 0.1 \Rightarrow e^m = 10 \\ \Rightarrow m &= \log_e 10 = 2.3026. \end{aligned}$$

Example 11: If 2% electric bulbs manufactured by a company be defective, find the probability of (i) 0 (iii) 1 (iii) 2 (iv) 3 defectives in a lot of 100 bulbs.

Solution: Let p be the probability of electric bulbs manufactured by a company be defective.

i.e.,

$$p = \frac{2}{100}.$$

Given that,

$$n = 100$$

we have,

$$np = 100 \times \frac{2}{100} = 2 = m \text{ (mean)}$$

Let X be the number of electric bulbs manufactured by a company be defective.

Then,

$$P(X = r) = \frac{m^r e^{-m}}{r!} = \frac{2^r e^{-2}}{r!}, \quad r = 0, 1, 2, 3, \dots$$

$$(i) P(X = 0) = \frac{2^0 e^{-2}}{0!} = \frac{1}{e^2} = \frac{1}{(2.718)^2} = 0.135$$

$$(ii) P(X = 1) = \frac{2^1 e^{-2}}{1!} = \frac{2}{e^2} = \frac{2}{(2.718)^2} = 2 \times (0.135) = 0.2706$$

$$(iii) P(X = 2) = \frac{2^2 e^{-2}}{2!} = \frac{2}{e^2} = \frac{2}{(2.718)^2} = 2 \times (0.135) = 0.2706$$

$$(iv) P(X = 3) = \frac{2^3 e^{-2}}{3!} = \frac{4}{3} \times \frac{1}{e^2} = \frac{4}{3} \times \frac{1}{(2.718)^2} = (1.33) \times (0.135) = 0.18.$$

Example 12: A Polsson distribution has a double mode at $x = 4$ and $x = 5$. Find the probability that x will have either of these values $P(x = 4) = P(x = 5)$.

Solution: The Poisson distribution is

$$P(X = r) = \frac{m^r e^{-m}}{r!}, r = 0, 1, 2, \dots$$

Given that

$$P(x = 4) = P(x = 5)$$

$$\Rightarrow \frac{m^4 e^{-m}}{4!} = \frac{m^5 e^{-m}}{5!} \Rightarrow m = 5$$

We know that, $e^{-5} = 0.006738$

Then

$$P(x = 4) = \frac{5^4 \times 0.006738}{4!}$$

The required probability is

$$\begin{aligned} P(x = 4) + P(x = 5) &= 2P(x = 4) \\ &= 2 \times \frac{5^4 \times 0.006738}{4!} = 0.35 \end{aligned}$$

Hence, it is easy to show that $P(4) > P(3)$ and also greater than $P(6)$.

Example 13: Fit a Poisson's distribution to the following data and calculate theoretical frequencies:

Deaths	0	1	2	3	4
Frequencies	122	60	15	2	1

Solution: The Poisson distribution is

$$P(X = r) = \frac{m^r e^{-m}}{r!}, r = 0, 1, 2, 3, \dots$$

$$\text{Here, mean} = \frac{\sum f(x)}{\sum f} = \frac{0 \times 122 + 1 \times 60 + 2 \times 15 + 3 \times 2 + 4 \times 1}{122 + 60 + 15 + 2 + 1} = \frac{100}{200} = 0.5 = m$$

Now, $e^{-0.5} = 0.61$ (approximately)

By equation (1), the number of r deaths is given by

$$= 200 \times \frac{(0.5)^r e^{-0.5}}{r!}, r = 0, 1, 2, 3, 4.$$

for $r = 0, 1, 2, 3, 4$ the theoretical frequencies are 121, 61, 15, 2 and 0.

Example 14: If the variance of the Poisson distribution is 2, find the probabilities for $r = 1, 2, 3, 4$ from the recurrence relation of the Poisson distribution. Also find $P(r \geq 4)$.

Solution: We know that for Poisson's distribution mean and variance both are equal.

i. e.,

$$\text{mean} = \text{variance} = 2.$$

Recurrence relation for Poisson distribution

$$P(r+1) = \frac{m}{r+1} P(r)$$

$$\Rightarrow P(r+1) = \frac{2}{r+1} P(r)$$

The Poisson's distribution is

$$P(r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, 3, \dots \dots \quad (1)$$

$$P(0) = e^{-m} = e^{-2} = 0.1353$$

By (1), if $r = 0$,

$$P(1) = \frac{2}{0+1} P(0) = 2 \times (0.1353) = 0.2706$$

if $r = 1$,

$$P(2) = \frac{2}{1+1} P(1) = 0.2706$$

if $r = 2$,

$$P(3) = \frac{2}{2+1} P(2) = \frac{2}{3} \times (0.2706) = 0.1804$$

if $r = 3$,

$$P(4) = \frac{2}{3+1} P(3) = \frac{1}{2} \times (0.1804) = 0.0902$$

Now, we have

$$\begin{aligned} P(r \geq 4) &= P(4) + P(5) + P(6) + \dots \dots \\ &= 1 - [P(0) + P(1) + P(2) + P(3)] \\ &= 1 - [0.1353 + 0.2706 + 0.2706 + 0.1804] \\ &= 1 - 0.8569 = 0.1431 \end{aligned}$$

Example 15: If the probability that an individual suffers a bad reaction from injection of a given serum is 0.001. Determine the probability that out of 2000 individual (i) exactly 3. (ii) more than 2 individuals suffer from bad reaction, (iii) none.

Solution: Let p be the probability that an individual suffers a bad reaction from injection of a given serum i.e.,

$$p = 0.001$$

Since, p is very small so we used Poisson's distribution.
Given that

$$\begin{aligned} n &= 2000 \\ np &= 0.001 \times 2000 = 2 \end{aligned}$$

we have

Let X be the number of individuals who suffer from bad reaction.
Then

$$P(X = r) = \frac{m' e^{-x}}{r!} = \frac{2^r e^{-2}}{r!}$$

(i) The probability for exactly 3

$$P(X = 3) = \frac{2^3 e^{-2}}{3!} = \frac{4}{3e^2} = \frac{4}{3 \times (2.718)^2} = 0.18$$

(ii) The probability for more than 2 individuals suffer from bad reaction

$$\begin{aligned} P(X > 2) &= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - \left[\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \frac{2^2 e^{-2}}{2!} \right] \\ &= 1 - \left[\frac{1}{e^2} + \frac{2}{e^2} + \frac{2}{e^2} \right] = 1 - \frac{5}{e^2} = 1 - \frac{5}{(2.718)^2} = 0.323 \end{aligned}$$

(iii) The probability for none

$$P(X = 0) = \frac{2^0 e^{-2}}{0!} = e^{-2} = \frac{1}{e^2} = \frac{1}{(2.718)^2} = 0.135$$

Example 16: In a certain factory turning Lexi blades, there is a small chance, 0.002 for any blade to be defective. The blades are in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 10,000 packets.

Solution: Let p be the probability that a blade is defective i.e.,

$$p = 0.002$$

Since, p is very small so Poisson's distribution is used. Here, $n = 10$ then

$$m = np = 10 \times 0.002 = 0.02$$

Let X denote the number of defective blades in a packet of 10. Then

$$P(X = r) = \frac{m^r e^{-m}}{r!} = \frac{(0.02)^r e^{-0.02}}{r!}, r = 0, 1, 2, \dots$$

$$e^{-0.02} = 1 - (0.02) + \frac{1}{2}(0.02)^2 + \dots$$

Now the number of packets containing no defective blade is

$$= 10,000 \times 0.9802 = 9802$$

The number of packets containing one defective blade

$$= 10,000 \times m e^{-m}$$

$$= 10,000 \times (0.2) e^{-0.2}$$

$$= 10,000 \times (0.2)(.9802)$$

$$= 196$$

The number of packets containing two defective blades

$$= 10,000 \times m m^2 e^{-m}$$

$$= 10,000 \times (0.2)^2 e^{-0.2}$$

$$= 10,000 \times (0.2)^2 (0.9802)$$

$$= 2 \text{ (nearly).}$$

Example 17: If X and Y be independent Poisson variates, show that the conditional distribution of X , given $X + Y = n$, is binomial.

Solution: Let X and Y be independent Poisson variates with parameters m_1 and m_2 .

$$\begin{aligned} \text{We have, } P\left[X = \frac{r}{X+Y=n}\right] &= \frac{P(X=r, X+Y=n)}{P(X+Y=n)} \\ &= \frac{P(X=r) \times P(Y=n-r)}{P(X+Y=n)} \\ &= \frac{\frac{m_1^r e^{-m_1}}{r!} \times \frac{m_2^{n-r} e^{-m_2}}{(n-r)!}}{\frac{(m_1 + m_2)^n e^{-(m_1+m_2)}}{n!}} = \frac{n!}{r!(n-r)!} \times \frac{m_1^r m_2^{n-r}}{(m_1 + m_2)^n} \end{aligned}$$

Letting

$$\begin{aligned} p &= \frac{m_1}{m_1 + m_2} \text{ and } q = \frac{m_2}{m_1 + m_2} \text{ then} \\ &= {}^n C_r p^r q^{n-r} \quad (0 \leq r \leq n) \end{aligned}$$

Hence, the conditional distribution of X , given $X + Y$, is binomial.

12.4 NORMAL DISTRIBUTIONS

NORMAL DISTRIBUTION

The Normal Distribution is one of the most important continuous probability distributions in statistics. It was first introduced by Abraham de Moivre in 1733 in connection with the approximation of the binomial distribution. Later, its importance was extensively developed in statistical theory and applications.

A normal distribution is a continuous probability distribution that is widely used in statistical analysis. It is particularly important because many naturally occurring measurements and statistical errors tend to

follow, or approximately follow, a normal pattern. For this reason, it is also commonly associated with the law of errors or error distribution.

The normal distribution can be obtained as the limiting form of the binomial distribution when the number of trials n becomes very large, while the probability of success p and the probability of failure q are not very small. Since

$$q = 1 - p,$$

the probability function of the normal distribution is expressed as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where μ represents the **mean** and σ denotes the **standard deviation**.

$$N(m, \sigma^2) = f(x) = y(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-m)^2/\sigma^2}; -\infty < x < \infty, -\infty < m < \infty, \sigma > 0$$

where $m = \text{Arithmetic mean } (\mu)$

$\sigma = \text{Standard deviation}$ are two parameters of the continuous distribution. Equation (1) is also called the normal curve. It is also denoted by $X \sim N(m, \sigma^2)$.

CONSTANTS OF NORMAL DISTRIBUTION

About the origin, we have

$$\begin{aligned} \mu'_n &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} x^n dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{x^n e^{-x^2/2\sigma^2}}{\sigma} dx \end{aligned}$$

Put

$$\begin{aligned}
 \frac{x}{\sigma} = t &\Rightarrow \frac{dx}{\sigma} = dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^n t^n e^{-t^2/2} dt \\
 &= \frac{\sigma^n}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^n e^{-\frac{1}{2}t^2} dt \\
 &= 0, \text{ when } n \text{ is odd } \{ \text{property of definite integrals} \}
 \end{aligned}$$

Hence, all odd moment about the origin is vanish.

i.e.,

$$\mu'_1 = \mu'_3 = \mu'_5 = \dots = 0$$

If n is even, by equation (1), we have

$$\mu'_n = \frac{2}{\sqrt{2\pi}} \frac{1}{\sigma} \int_0^{\infty} x^n e^{-x^2/2\sigma^2} dx$$

Putting,

$$\begin{aligned}
 \frac{x^2}{2\sigma^2} = u &\Rightarrow \frac{x}{\sigma^2} dx = du \\
 &= \frac{2}{\sqrt{2\pi}\sigma} \int_0^{\infty} (2u\sigma^2)^{(n-1)/2} e^{-u} \sigma^2 du \\
 &= \frac{\sigma^n 2^{n/2}}{\sqrt{\pi}} \int_0^{\infty} u^{(n-1)/2} e^{-u} du \\
 &= \frac{\sigma^n 2^{n/2}}{\sqrt{\pi}} \Gamma\left(\frac{n+1}{2}\right)
 \end{aligned}$$

When $n = 2$

$$\mu'_2 = \frac{\sigma^2 2^{2/2}}{\sqrt{\pi}} \Gamma(3/2) = \frac{\sigma^2 \cdot 2}{\sqrt{\pi}} \cdot \frac{1}{2} \sqrt{\pi} = \sigma^2$$

When $n = 4$

$$\mu'_4 = \frac{\sigma^4 2^2}{\sqrt{\pi}} \cdot \Gamma\left(\frac{5}{2}\right) = \frac{\sigma^4 \cdot 4}{\sqrt{\pi}} \cdot \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi} = 3\sigma^4$$

When $n = 0$

$$\mu'_0 = \frac{\sigma^0 2^0}{\sqrt{\pi}} \Gamma \frac{1}{2} = \frac{1.1}{\sqrt{\pi}} \sqrt{\pi} = 1$$

Here mean is at the origin itself, these are also undashed μ'_s i.e., $\mu_2 = \mu'_2$, $\mu_3 = \mu'_3$ etc.

Now,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0; \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$$

$$\gamma_1 = \sqrt{\beta_1} = 0; \gamma_2 = 0$$

Hence, the normal curve has zero kurtosis.

MOMENT ABOUT THE MEAN M

$$\begin{aligned} \mu_{2n+1} &= \int_{-\infty}^{\infty} (x - m)^{2n+1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx \\ &= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\frac{x-m}{\sigma}\right)^{2n+1} \frac{e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}}{\sigma} dx \end{aligned}$$

Putting $\frac{x-m}{\sigma} = t \Rightarrow \frac{dx}{\sigma} = dt$

$$= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} = \int_{-\infty}^{\infty} t^{2n+1} e^{-\frac{1}{2}t^2} dt$$

= 0, since it is integral of an odd function of t .

Now we have

$$\begin{aligned} \mu_{2n} &= \int_{-\infty}^{\infty} (x - m)^{2n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx \\ &= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\frac{x-m}{\sigma}\right)^{2n} \frac{e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}}{\sigma} dx \end{aligned}$$

Putting

$$\begin{aligned}
 \frac{x - m}{\sigma} &= t \Rightarrow \frac{dx}{\sigma} = dt \\
 &= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^{2n} e^{-\frac{1}{2}t^2} dt \\
 &= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (2T)^n e^{-T} \frac{dT}{\sqrt{2T}} \\
 &= \frac{2^{n-1} \sigma^{2n}}{\sqrt{\pi}} \int_{-\infty}^{\infty} T^n e^{-T} dT \\
 &= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_0^{\infty} T^n e^{-T} dT \\
 \Rightarrow \mu_{2n} &= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \Gamma\left(n + \frac{1}{2}\right)
 \end{aligned}$$

Putting $n = n - 1$ in (1), we get

$$\mu_{2n-2} = \frac{2^{n-1} \sigma^{2n-2}}{\sqrt{\pi}} \Gamma\left(n - \frac{1}{2}\right)$$

Now equation (1) divide by (2), we get

$$\begin{aligned}
 \frac{\mu_{2n}}{\mu_{2n-2}} &= \frac{2\sigma^2 \Gamma\left(n + \frac{1}{2}\right)}{\Gamma\left(n - \frac{1}{2}\right)} \quad [\because \Gamma n = (n-1)\Gamma(n-1)] \\
 \Rightarrow &= 2\sigma^2 \left(n - \frac{1}{2}\right) \\
 \Rightarrow \mu_{2n} &= \sigma^2 (2n-1) \mu_{2n-2}
 \end{aligned}$$

which is the recurrence relation for the moments of normal distribution.

AREA UNDER THE NORMAL CURVE

The probability density function of the normal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

If x is a normal random variable with mean m and standard deviation σ , then the random variable $z = \left(\frac{x-m}{\sigma}\right)$ has the normal distribution with mean 0 and the standard deviation 1. The random variable z is called the standard normal random variable.

The probability density function of standard nor-

Putting $t^2 = 2T, t = \sqrt{2T}$

$$\Rightarrow 2t dt = 2dT$$

$$\text{or } dt = \frac{dT}{\sqrt{2T}}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$

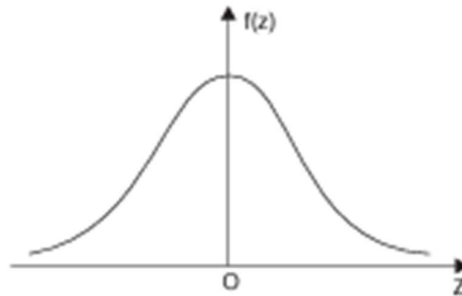


Figure 12.1 Bell shaped curve

Note 1. The graph of $f(z)$ is famous 'bell shaped' curve.

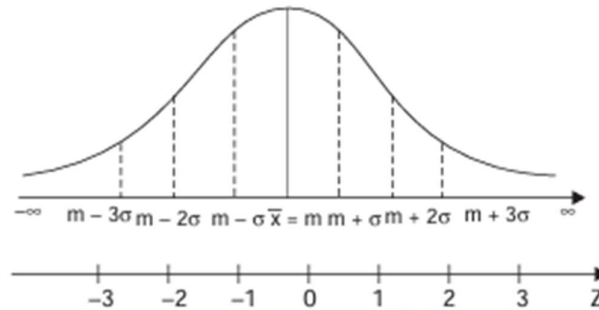
2. If $f(z)$ is the probability density function for the normal distribution, then

$$P(z_1 \leq z \leq z_2) = \int_{z_1}^{z_2} f(z) dz = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

PROPERTIES OF THE NORMAL DISTRIBUTION AND NORMAL CURVE

The normal probability curve is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, -\infty < x < \infty$$



12.2 Normal probability curve

- (1) Mean, Median and Mode of the Normal distribution coincide.
- (2) Area under the normal curve is unity.
- (3) The probability of the continuous random variable $X(X_1 \leq X \leq X_2)$ is denoted by

$$P(X_1 \leq X \leq X_2) = \int_{X_1}^{X_2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-w}{\sigma}\right)^2} dx$$

Where m is the mean and σ is standard deviation.

- (4) The curve is bell-shaped and symmetry about the line $x = m$ i.e., y -axis.
- (5) As x increases then $f(x)$ decreases rapidly, the maximum probability at point $x = \mu$ is given by

$$f(x) \text{ or } [P(x)]_{\max} = \frac{1}{\sqrt{2\pi}\sigma}$$

(6) Since, $f(x)$ is probability density function, so it never be negative, i.e., no portion of the curve lies below the x -axis.

(7) x -axis is an asymptote of the curve.

(8) The point of inflexion of the normal curve are $x = \pm\sigma$ when the origin is taken at the mean.

(9) Mean deviation about mean $= \sqrt{\frac{2}{\pi}}\sigma = \frac{4}{5}\sigma$ (approximate).

(10) Area under the normal curve is distributed as following 68.27% area lies between $m - \sigma$ to $m + \sigma$ or $(\bar{x} - \sigma$ to $\bar{x} + \sigma)$ i.e., between $-1 \leq z \leq 1$

94.45% area lies between $m - 2\sigma$ to $m + 2\sigma$, i.e., between $-2 \leq z \leq 2$

99.73% area lies between $m - 3\sigma$ to $m + 3\sigma$, i.e., between $-3 \leq z \leq 3$

50% area lies in between $-0.745 \leq z \leq 0.745$

99% area lies in between $-2.58 \leq z \leq 2.58$.

Example 1: Prove that for normal distribution the mean deviation from the mean equal to $\frac{4}{5}$ of standard deviation nearly.

Solution: Let m and σ be the mean and standard deviation of the normal distribution respectively.

By definition of mean deviation

$$\begin{aligned}
 &= \int_{-\infty}^{\infty} |x - m| f(x) dx \\
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} |x - m| e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx \\
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \sigma |z| e^{-\frac{1}{2}z^2} \sigma dz \\
 &= \sigma \sqrt{\frac{2}{\pi}} \int_0^{\infty} z e^{-\frac{1}{2}z^2} dz \\
 &= \sigma \sqrt{\frac{2}{\pi}} \left[-e^{-\frac{z^2}{2}} \right]_0^{\infty} \\
 &= \sigma \sqrt{\frac{2}{\pi}} \cdot 1 = 0.7979\sigma = 0.8\sigma \\
 &= \frac{8}{10}\sigma = \frac{4}{5}\sigma.
 \end{aligned}$$

Example 2: If X is a normal variate with mean 12 and standard deviation 4. Find

- (f) $P(X \geq 20)$
- (ii) $P(X \leq 20)$
- (iii) $P(0 \leq X \leq 12)$.

Solution: Given that $m = 12$ and $\sigma = 4$



Let

$$Z = \frac{X - m}{\sigma} = \frac{X - 12}{4}$$

(f) When $X = 20$, $Z = \frac{20-12}{4} = 2$

$$\begin{aligned} P(X \geq 20) &= P(Z \geq 2) \\ &= 0.5 - P(0 \leq Z \leq 2) \\ &= 0.5 - 0.4772 \text{ (Using table)} [\because P(0 \leq Z \leq 2) + P(Z \\ &\geq 2) = 0.5] \\ &= 0.0228. \end{aligned}$$

(ii)

$$\begin{aligned} P(X \leq 20) &= P(Z \leq 2) \\ &= 0.5 + P(0 \leq Z \leq 2) \\ &= 0.5 + 0.4772 \\ &= 0.9772 \end{aligned}$$

(iii) When $X = 0$, $Z = \frac{-1}{4} = -3$ and for $X = 12$, $Z = 0$

$$\therefore P(0 \leq X \leq 12) = P(-3 \leq Z \leq 0) = P(0 \leq Z \leq 3) = 0.4987.$$

Example 3: Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 (inches)². How many soldiers in a regiment of 1000 would you expect to be over 6 feet tall? Given that the area under the standard normal curve between $Z = 0$ and $Z = 0.35$ is 0.1368 ; and between $Z = 0$ and $Z = 1.15$ is 0.3746.

Solution: Given that mean (m) = 68.22, variance (σ^2) = 10.8

and standard deviation (σ) = $\sqrt{10.8}$

We have $Z = \frac{X-m}{\sigma}$

$$\Rightarrow Z = \frac{X - 68.22}{\sqrt{10.8}}$$

$$\Rightarrow Z = \frac{X - 68.22}{3.286}$$

$$\text{When } X = 72, Z = \frac{72-68.22}{3.286} = 1.15$$

$$\text{Now } P(X \geq 72) = P(Z \geq 1.15) = 0.5 - 0.3749 = 0.1251$$

The probability of getting a soldier above six feet is 0.1251. Hence, the number of soldiers who are over 6 feet tall.

$$= 1000 \times 0.1251 = 125.1$$

$$= 125 \text{ (nearly)}$$

Example 4: Assuming the resistance of the resistors to be normally distributed with mean 100 ohms and standard deviation 2 ohms, what percentage of resistors will have resistance between 98 ohms to 102 ohms.

Solution: Given that mean (m) = 100 ohms standard deviation (σ) = 2 ohms

We have

$$Z = \frac{X - m}{\sigma}$$

$$\Rightarrow Z = \frac{X - 100}{2}$$

$$\text{When } X = 102, Z = \frac{102 - 100}{2} = 1$$

$$\text{When } X = 98, Z = \frac{98 - 100}{2} = -1$$

Now,

$$P(98 \leq X \leq 102) = P(-1 \leq Z \leq 1)$$

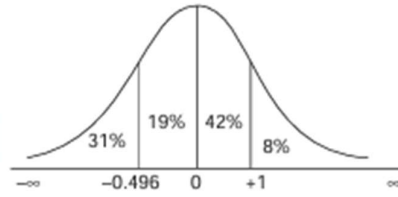
$$= 2P(0 \leq Z \leq 1)$$

$$= 2 \times (0.3413) = 0.6826$$

Hence, 68.26% of resistors will have resistance between 98 ohms to 102 ohms.

Example 5: In a normal distribution 31% items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

Solution: Let m and σ be the mean and standard deviation respectively.



If $x = 45$,

$$Z = \frac{45 - m}{\sigma}$$

$$Z = \frac{64 - m}{\sigma}$$

If $x = 64$,

The area between 0 and $\frac{45-m}{\sigma} = 0.50 - 0.31 = 0.19$

From the table, for the area 0.19,

$$\begin{aligned} Z &= 0.496 \\ \frac{45 - m}{\sigma} &= -0.496 \end{aligned}$$

Area between $Z = 0$ and $Z = \frac{64-m}{\sigma} = 0.5 - 0.08 = 0.42$

From the table, for the area 0.42, $Z = 1.405$

$$\frac{64 - m}{\sigma} = 1.405$$

Solving equation (1) and (2), we get

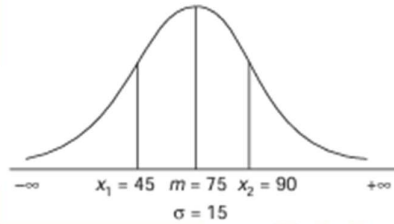
$$m = 50 \text{ and } \sigma = 10.$$

Example 6: The distribution of weekly wages for 500 workers in a factory is approximately normal with the mean and standard deviation of Rs. 75 and Rs. 15. Find the number of workers who receive weekly wages:

- (i) more than Rs. 90
- (ii) less than Rs. 45.

Solution: The normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, \quad -\infty < x < \infty$$



Given that $N = 500$, $m = 75$ and $\sigma = 15$. we have

$$Z = \frac{X - m}{\sigma} = \frac{X - 75}{15}$$

(i) when $X = 90$, $Z = \frac{90}{15} = 1$

$$\begin{aligned} P(X > 90) &= P(Z > 1) = 0.5 - P(0 < Z < 1) \\ &= 0.5 - 0.3413 = 0.1587 \end{aligned}$$

The number of workers who receive weekly wages more than Rs. 90

$$\begin{aligned} &= 0.1587 \times 500 \\ &= 79.35 = 79 \text{ (nearly) workers.} \end{aligned}$$

(ii) When

$$\begin{aligned} X = 45, Z &= \frac{45 - 75}{15} = -2 \\ P(X < 45) &= P(Z < -2) = 0.5 - P(-2 < Z < 0) \\ &= 0.5 - P(0 < Z < 2) \\ &= 0.5 - 0.4772 = 0.0228 \end{aligned}$$

Hence, the number of workers who receive weekly wages more than Rs. 45

$$= 0.0228 \times 500 = 11.4 \text{ i.e., 11 (nearly) workers.}$$

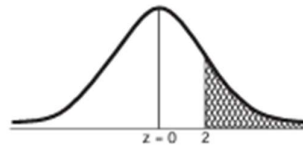
Example 7: The life of army shoes is 'normally' distributed with mean 8 months and standard deviation 2 months. If 5000 pairs are issued how many pair would be expected to need replacement after 12 months? { Given that $P(Z \geq 2) = 0.0288$ and $Z = \left(\frac{x-m}{\sigma}\right)$ }

Solution: Given that

$$\text{Mean } (m) = 8$$

Standard deviation

$$(\sigma) = 2$$



Number of pairs of shoes = 5000

Total month

$$(x) = 12$$

We have

$$Z = \frac{x - m}{\sigma} = \frac{12 - 8}{2} = 2$$

Area $(Z \geq 2) = 0.0288$

The number of pairs whose life is more than 12 months $(Z > 2)$

Replacement after 12 months

$$= 5000 - 114 = 4886 \text{ pairs of shoes.}$$

Example 8: If the height of 300 students is normally distributed with mean 64.5 inches and standard deviation 3.3 inches, how many students have height

- (i) less than 5 feet
- (ii) between 5 feet and 5 feet 9 inches.

Also, find the height below which 99% of the students lie.

Solution: Given that $m = 64.5$, $\sigma = 3.3$ and $N = 300$.

We have

$$Z = \frac{X - m}{\sigma} = \frac{X - 64.5}{3.3}$$

(i) For $X = 60$,

$$Z = \frac{60 - 64.5}{3.3} = -\frac{4.5}{3.3} = -1.36$$

$$\begin{aligned} P(X < 60) &= P(Z < -1.36) = 0.5 - (-1.36 < Z < 0) \\ &= 0.5 - (0 < Z < 1.36) \\ &= 0.5 - 0.4131 \\ &= 0.0869. \end{aligned} \quad \text{(Using$$

table)

The number of student having height less than 5 feet

$$\begin{aligned} &= 0.0869 \times 300 \\ &= 26.07 \text{ i.e., 26 students.} \end{aligned}$$

$$\begin{aligned} P(60 < X < 69) &= P(-1.36 < Z < 1.36) \\ &= 2P(0 < Z < 1.36) = 2(0.4131) = 0.8262 \end{aligned}$$

The number of student having heights lie between 60 inches and 69 inches = $0.8262 \times 300 = 247.86$ i.e., 248 students.

$$\begin{aligned} \text{Now let } P(X < X_1) &= 0.99 \\ \Rightarrow P(Z < Z_1) &= 0.99 \end{aligned}$$

where

$$Z_1 = \frac{X_1 - m}{\sigma}$$

$$\Rightarrow P(0 < Z < Z_1) = 0.99 - 0.5 = 0.49$$

Using the table giving areas under normal curve $Z_1 = 2.33$

$$\begin{aligned}\therefore Z_1 &= \frac{X_1 - m}{\sigma} = \frac{X_1 - 64.5}{3.3} \\ \Rightarrow X_1 &= 64.5 + 3.3 \times 2.33 \\ \Rightarrow X_1 &= 72.18\end{aligned}$$

Hence, 99% of the students have height below 72.18 inches.

The equation of the normal curve fitted to the given data is

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, \quad -\infty \leq x \leq \infty$$

where

$$m = 8.56262, \sigma = 0.0176.$$

Example 9: For $-\infty < x < \infty$, and probability density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

Show that the total probability is 1.

Solution: The total probability

$$\begin{aligned}&= \int_{-\infty}^{\infty} f(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx \\ &\quad \left| \begin{array}{l} \text{Let } \frac{x-m}{\sqrt{2}\sigma} = t \\ \Rightarrow dx = \sqrt{2}\sigma dt \end{array} \right.\end{aligned}$$

$$\begin{aligned}
 &= \frac{\sigma\sqrt{2}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-t^2} dt \\
 &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt \\
 &= \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-t^2} dt
 \end{aligned}$$

Now let

$$\begin{aligned}
 t^2 &= y \text{ or } t = \sqrt{y} \\
 2tdt &= dy \\
 &= \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-y} \frac{dy}{2\sqrt{y}} \\
 &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} y^{-1/2} e^{-y} dy \\
 &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} y^{\frac{1}{2}-1} e^{-y} dy
 \end{aligned}$$

$$\begin{aligned}
 \Rightarrow 2tdt &= dy \\
 &= \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-y} \frac{dy}{2\sqrt{y}} \\
 &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} y^{-1/2} e^{-y} dy \\
 &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} y^{\frac{1}{2}-1} e^{-y} dy \\
 &= \frac{1}{\sqrt{\pi}} \Gamma(1/2) = \frac{1}{\sqrt{\pi}} \times \sqrt{\pi} = 1
 \end{aligned}$$

CHECK YOUR PROGRESS

1. Define Poisson Distribution and explain its main characteristics.

.....

.....

.....

2. Explain the important properties of Normal Distribution.

.....

.....

.....

3. Distinguish between Poisson Distribution and Normal Distribution.

.....

.....

.....

12.5. SUMMARY

In this unit, we have discussed the following concepts

1. Poisson Distribution
2. Parameters and Constants of the Poisson Distribution
3. Recurrence Relation for the Poisson Distribution
4. Determination of the Mode of the Poisson Distribution
5. Normal Distribution
6. Parameters and Constants of the Normal Distribution
7. Moments about the Mean
8. Area under the Normal Curve
9. Important Properties of the Normal Distribution and Normal Curve
10. Numerical Problems Based on Poisson and Normal

12.6 EXERCISES

1. If X be a Poisson variate such that $3P(X = 3) = 4P(X = 4)$, find $P(X = 7)$.

Criticise the following statement:

The mean of a Poisson distribution is 7, while the standard deviation is 6.

2. The probability that a man aged 50 years will die within a year is 0.01125. What is the probability that of 12 such men at least 11 will reach their fifty first birthday?

3. A car-hire-firm has two cars, which it hires, out day by day. The number of demands for a car on each day is distributed as Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused $\{e^{-1.5} = 0.2231\}$.

4. If m is the parameter of a Poisson variate, show that the probabilities that the value of the variate taken at random is even or odd are $e^m \cos hm$ and $e^{-m} \sin hm$.

5. Letters were received in an office on each of 100 days. Assuming the following data to form a random sample from a Poisson's distribution. Fit the distribution and calculate the expected frequencies taking $e^4 = 0.183$.

Number of letters	0	1	2	3	4	5	6	7	8	9	10
Frequencies	1	4	15	22	21	20	8	6	2	0	1

6. The frequency of accidents per shift in a factory is given in the following table. Calculate the mean number of accidents per shift. Find the corresponding Poisson distribution and compare with actual observations.

Accidents per shift	0	1	2	3	4	Total
Frequencies	192	100	24	3	1	320

7. A manufacturer of coffer pins knows that 5 per cent of his product is defective. If he sells coffer pins in boxes of 100 and guarantees that not more than 4 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality? ($e^{-5} = 0.0067$).
8. A telephone switch board handles 600 calls on the average during a rush hour. The board can make a maximum of 20 connections per minute. Use Poisson distribution to estimate the probability that the board will be over during any given minute [$e^{-1} = 0.00004539$].
9. Suppose that a book of 600 pages contains 40 printing mistakes. Assume that these errors are randomly distributed throughout the book and x , the number of errors per page has a Poisson distribution. What is the probability that 10 pages selected at random will be free of errors?
10. Find the probability that at most defective fuses will be found in a box of 200 fuses, if experience show that 2 per cent of such fuses are defective.
11. An insurance company found that only 0.01% of the population is involved in a certain type of accident each year. If its 1000 policy holders were randomly selected from the population. What is the probability that not more than two of its clients are involved in such an accident next year? ($e^{-0.1} = 0.9048$).
12. Red blood cell deficiency may be determined by examining a specimen of the blood under a microscope. Suppose a certain small fixed volume contains on the average 20 red cells for normal persons. Using Poisson distribution, obtain the probability that a specimen from a normal person will contain less than 15 red cells.

Questions related to Normal Distribution

- 1 Let X be a normal variate with mean 30 and standard deviation 5, find the probabilities that (i) $26 \leq X \leq 40$,
(ii) $X \geq 45$ and
(iii) $|X - 30| > 5$

- 2 Students of a class were given an aptitude test. Their marks were found to be normally distributed with mean 60 and standard deviation 5. What percentage of students scored more than 60 marks?
- 3 1000 light bulbs with a mean life of 120 days are installed in a new factory, their length of life is normally distributed with standard deviation 20 days. How many bulbs will expire in less than 90 days?
- 4 On a final examination in mathematics, the mean was 72, and the standard deviation was 15. Determine the standard scores of students receiving grades.
 - (i) 60
 - (ii) 72
 - (iii) 93.
- 5 In a male population of 1000, the mean height is 68.16 inches and standard deviation is 3.2 inches. How many men may be more than 6 feet (72 inches)?
- 6 For a certain normal distribution, the first moment about 10 is 40 and the fourth moment about 50 is 48. What is the arithmetic mean and variance of the normal distribution.
- 7 In a normal distribution whose mean is 2 and standard deviation 3, find a value of the variate such that the probability of the interval from the mean to the value is 0.4115. Find another value such that the probability for the interval from $x = 3.5$ to that value is 0.2307.
- 8 If shells are classified as A, B, C according as the length breadth index as under 75, between 75 and 80, or over 80, find approximately (assuming that distribution is normal) the mean and standard deviation of a series in which A are 58%, B are 38%, and C are 4% being given that

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{1}{2}x^2} dx \text{ then } f(0.20) = 0.08 \text{ and } f(1.75) = 0.46$$

12.7 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.



BLOCK IV: CORRELATION AND REGRESSION ANALYSIS, CURVE FITTING

UNIT 13 CORRELATION; KARL PEARSON'S COEFFICIENT OF CORRELATION

STRUCTURE

- 13.1 Introduction
- 13.2 Objectives
- 13.3 Correlation
- 13.4 Coefficient and measurement of correlation
- 13.5 Karl Pearson's coefficients of correlation
- 13.6 Summary
- 13.7 Exercises
- 13.8 References

13.1 INTRODUCTION

Correlation is a statistical technique used to measure and analyze the strength and direction of the relationship between two variables. Understanding correlation is essential in fields like economics, finance, biology, and social sciences, where determining how one variable affects another is crucial for prediction and decision-making.

Karl Pearson's Coefficient of Correlation (denoted as r) is one of the most widely used methods for measuring linear correlation between two quantitative variables. This coefficient quantifies the degree to which two variables move together. A value of $r = +1$ indicates a perfect positive correlation, $r = -1$ shows a perfect negative correlation, and $r = 0$ suggests no correlation.

Pearson's method not only helps in understanding the relationship between variables but also forms the basis for more complex analyses like regression and hypothesis testing.

13.2 OBJECTIVES

After completing this unit, students will be able to

- Understand the concept of correlation and its significance in data analysis.
- Distinguish between positive, negative, and zero correlation.
- Apply the formula for Karl Pearson's γ to real-world datasets.
- Analyze the value of γ to describe the strength and direction of the relationship between two variables.

13.3 CORRELATION

13.3.1 Frequency distribution: Frequency distribution is a way to represent the large amount of data according to quantitative characteristics frequency distribution helps the management of the bank for announcing new saving policies for the benefit to its customers.

13.3.2 Bivariate Frequency distribution: The study of relationship between two or more variables is one of the most important problems in life. In a bivariate distribution, we find the relationship between the two variables. For example, if we find the heights and weights of 10th class of student, we shall get a bivariate distribution. Here, one variable is related to height and other variable is related to weight.

13.3.3 Correlation: The relationship between two variables such that a change in one is accompanied by change in the other in such a way that

an increase in one is accompanied by an increase or decrease in the other, is called a correlation. For example, the number of policemen and the number of crimes, the number of trains and the number of passengers. If there is no relationship indicated between two variables then they are said to be uncorrelated or statistically independent. For example, marks in statistics of a child and weights of her mother. It is not necessary that one unit of change in one variable is followed by one unit of change in other variable. It is also not essential that the ratio should be the same in all cases.

13.3.4 Positive Correlation: If an increase or decrease in the values of one variable corresponds to an increase or decrease in other variable respectively, then the correlation is said to be positive correlation. For example, age of husbands and wives are known to have a positive correlation. If an increase or decrease in the values of one variable is always followed by a corresponding and proportional increase or decrease in other variable, then the correlation is said to be perfect positive correlation.

13.3.5 Negative Correlation: If an increase or decrease in the values of one variable corresponds to a decrease or increase in other variable respectively, then the correlation is said to be negative correlation. For example, the number of fourth-class workers in a college and the number of duty room in a college. If an increase or decrease in the values of one variable is always followed by a corresponding and proportional decrease or increase in other variable, then the correlation is said to be perfect negative correlation.

13.3.6 Linear and Non-linear Correlation: When the variations in the values of two variables are in a constant ratio, then correlation is said to be linear correlation. This type of relationship is represented by a linear equation of the form $Y = aX + b$. Otherwise, if the ratio of variation in the values of two variables are not constant, then correlation is said to be non-linear correlation.

be non-linear correlation. This type of relationship represented by $y = a + bx + cx^2$ etc.

Note: If the change in one variable has a very little effect in the other variable then there is no correlation.

13.4 COEFFICIENT AND MEASUREMENT OF CORRELATION

The coefficient of correlation measures the degree of relationship between two or more variable. The coefficient of correlation varies between -1 to $+1$. If the coefficient of correlation reaches unity, then it is perfect. The $+1$ indicates perfect positive correlation and -1 indicates perfect negative correlation. Whenever the value of coefficient of correlation is zero then there is no correlation.

Correlation in two or more variables can be studied by the following methods:

- (i) Karl Pearson's coefficients of correlation.
- (ii) Rank correlation or Spearman's coefficient of rank correlation.
- (iii) Scatter diagram or dot diagram.
- (iv) Correlation table.
- (v) Graphic method.
- (vi) Concurrent deviations method.

In this unit, we shall discuss Karl Pearson's coefficient of correlation, Rank correlation and Scatter diagram method only.

13.5 KARL PEARSON'S COEFFICIENTS OF CORRELATION

Karl Pearson (1867-1936), a great statistician and biologist, developed a formula for the calculation of coefficient of correlation. The correlation coefficient between two variables X and Y denoted by γ . defined as

$$\gamma(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (1)$$

If $(x; y); i = 1, 2, 3, \dots, n$ be the set of n pairs of values of variable X and Y then

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \bar{X})(Y - \bar{Y})] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) \end{aligned}$$

where,

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{j=1}^n X_j, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \\ \sigma_X^2 &= E[X - E(X)]^2 = E[X - \bar{X}]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \\ \sigma_Y^2 &= E[Y - E(Y)]^2 = E[Y - \bar{Y}]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2 \end{aligned}$$

Putting these values of $\text{Cov}(X, Y)$, σ_X and σ_Y in (1), we get

$$\gamma(X, Y) = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \right]^{\frac{1}{2}} \left[\frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2 \right]^{\frac{1}{2}}}$$

$$\gamma(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{[\sum_{i=1}^n (x_i - \bar{X})^2]^{\frac{1}{2}} [\sum_{i=1}^n (y_i - \bar{Y})^2]^{\frac{1}{2}}} \quad (2)$$

where $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

Note. 1. $\text{Cov}(X, Y) = \sigma_{XY}$ (also write)

2. Karl Pearson's coefficient of correlation is also called product moment correlation coefficient, because

$$\text{Cov}(X, Y) = E\{\{X - E(X)\}\{Y - E(Y)\}\}$$

3. Simplifying equation (2), we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) &= \frac{1}{n} \sum_{i=1}^n \{x_i y_i - x_i \bar{Y} - \bar{X} y_i + \bar{X} \bar{Y}\} \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{Y} \frac{1}{n} \sum_{i=1}^n x_i - \bar{X} \frac{1}{n} \sum_{i=1}^n y_i + \bar{X} \bar{Y} \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y} \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2 \\ \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2 &= \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2 \end{aligned}$$

Now by (2), we get

$$\gamma(X, Y) = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}}{\sqrt{\left\{ \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2 \right\} \left\{ \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2 \right\}}}$$

Theorem 1: The coefficient of correlation $|\gamma| \leq 1$ i.e., $-1 < \gamma < 1$.

Theorem 2: If X and Y are independent variable, then they are uncorrelated but converse is not true.

Theorem 3: The correlation coefficient is independent of change of origin and scale.

PROBABLE ERROR

The coefficient of correlation also has a probable error. It is that amount when added to and subtracted from the coefficient of correlation would give such limits within which we can reasonably expect the values of coefficient of correlation to vary. The formula for probable error is

$$\text{P.E.} = (0.6745) \times \frac{1 - \gamma^2}{\sqrt{n}}$$

The following rules are observed for probable error:

- (i) If $\gamma < 0.30$, then it is insignificant.
- (ii) If $\gamma < \text{P.E.}$, then it is not at all significant.
- (iii) If $0.4 < \gamma < 0.6$, then correlation coefficient is normal.
- (iv) If $\gamma > 0.9$, then it is highly significant.
- (v) If $\gamma > 6 \text{ P.E.}$, then it is significant i.e., correlation is certain.

CORRELATION COEFFICIENT FOR A BIVARIATE FREQUENCY DISTRIBUTION

For a bivariate frequency distribution

$$\gamma = \frac{\Sigma fUV - \frac{\Sigma fU \Sigma fV}{\Sigma f}}{\left[\left\{ \Sigma fU^2 - \frac{(\Sigma fU)^2}{\Sigma f} \right\} \left\{ \Sigma fV^2 - \frac{(\Sigma fV)^2}{\Sigma f} \right\} \right]^{1/2}}$$

where, f is the frequency of a particular rectangle in the correlation table, and

$$U = \frac{X - \bar{X}}{h}, V = \frac{Y - \bar{Y}}{k}.$$

Example 1: Calculate the coefficient of correlation between X and Y using the following data.

X	-10	-5	0	5	10
Y	5	9	7	11	13

Solution:

X	Y	X^2	Y^2	XY
-10	5	100	25	-50
-5	9	25	81	-45
0	7	0	49	0
5	11	25	121	55
10	13	100	169	130
$\Sigma X = 0$	$\Sigma Y = 45$	$\Sigma X^2 = 250$	$\Sigma Y^2 = 445$	$\Sigma XY = 90$

We have,

$$\bar{X} = \frac{1}{n} \Sigma X = 0,$$

$$\bar{Y} = \frac{1}{n} \Sigma Y = \frac{1}{5} \times 45 = 9$$

The correlation coefficient

$$\begin{aligned} r(X, Y) &= \frac{\frac{1}{n} \Sigma XY - \bar{X} \bar{Y}}{\sqrt{\left\{ \frac{1}{n} \Sigma X^2 - \bar{X}^2 \right\} \left\{ \frac{1}{n} \Sigma Y^2 - \bar{Y}^2 \right\}}} \\ &= \frac{\frac{1}{5} \times 90 - 0 \times 9}{\sqrt{\left\{ \frac{1}{5} \times 250 - (0)^2 \right\} \left\{ \frac{1}{5} \times 445 - (9)^2 \right\}}} \\ &= \frac{18}{\sqrt{50 \times 8}} = \frac{18}{20} = 0.9. \end{aligned}$$

Example 2: Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y).

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution:

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
ΣX = 544	ΣY = 552	ΣX^2 = 37028	ΣY^2 = 38132	ΣXY = 37560

We have

$$\bar{X} = \frac{1}{n} \Sigma X = \frac{544}{8} = 68$$

$$\bar{Y} = \frac{1}{n} \Sigma Y = \frac{552}{8} = 69$$

The correlation coefficient

$$\begin{aligned}
 r(X,Y) &= \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sqrt{\left\{\frac{1}{n}\Sigma X^2 - \bar{X}^2\right\}\left\{\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right\}}} \\
 &= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left\{\frac{1}{8} \times 37028 - (68)^2\right\}\left\{\frac{1}{8} \times 38132 - (69)^2\right\}}} = \frac{4695 - 4692}{\sqrt{4.5 \times 5.5}} = 0.603.
 \end{aligned}$$

Example 3: Calculate the coefficient of correlation between X and Y using the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Solution: Let us assume mean for X be 5 and for Y be 12

X	Y	U $= X - \bar{X}$	V $= Y - \bar{Y}$	U^2	V^2	UV
1	9	-4	-3	16	9	12
2	8	-3	-4	9	16	12
3	10	-2	-2	4	4	4
4	12	-1	0	1	0	0
5	11	0	-1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
n $= 9$		0	0	ΣU^2 $= 60$	ΣV^2 $= 60$	ΣUV $= 57$

$$\begin{aligned}\bar{U} &= \frac{1}{n}\Sigma U = 0, \bar{V} = \frac{1}{n}\Sigma V = 0 \\ \text{Cov}(U, V) &= \frac{1}{n}\Sigma UV - \bar{U}\bar{V} = \frac{1}{9} \times 57 - 0 = \frac{57}{9} = 6.33 \\ \sigma_{U^2} &= \frac{1}{n}\Sigma U^2 - \bar{U}^2 = \frac{1}{9} \times 60 - 0 = \frac{60}{9} = 6.66 \\ \sigma_{V^2} &= \frac{1}{n}\Sigma V^2 - \bar{V}^2 = \frac{1}{9} \times 60 - 0 = \frac{60}{9} = 6.66\end{aligned}$$

The correlation coefficient

$$r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{6.33}{\sqrt{6.66 \times 6.66}} = \frac{6.33}{6.66} = 0.950.$$

Example 4: Calculate the coefficient of correlation from the following data.

X	1	3	5	7	8	10
Y	8	12	15	17	18	20

Solution: Let us assume mean for X be 7 and for Y be 15

X	Y	U = X - $\bar{X}$	V = Y - $\bar{Y}$	U ²	V ²	UV
1	8	-6	-7	36	49	42
3	12	-4	-3	16	9	12
5	15	-2	0	4	0	0
7	17	0	2	0	4	0
8	18	1	3	1	9	3
10	20	3	5	9	25	15
n = 6		-8	0	ΣU^2 = 66	ΣV^2 = 96	ΣUV = 72

$$\begin{aligned}\bar{U} &= \frac{1}{n}\Sigma U = \frac{-8}{6} = -1.33, \bar{V} = \frac{1}{n}\Sigma V = \frac{1}{6} \times 0 = 0 \\ \text{Cov}(U, V) &= \frac{1}{n}\Sigma UV - \bar{U}\bar{V} = \frac{1}{6} \times 72 - (-1.33) \times 0 = 12 \\ \sigma_{U^2} &= \frac{1}{n}\Sigma U^2 - \bar{U}^2 = \frac{1}{6} \times 66 - (-1.33)^2 \\ &= 11 - 1.77 = 9.23 \\ \sigma_{V^2} &= \frac{1}{n}\Sigma V^2 - \bar{V}^2 = \frac{1}{6} \times 96 - (0)^2 = 16\end{aligned}$$

The correlation coefficient

$$\gamma(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{12}{\sqrt{9.23 \times 16}} = \frac{12}{12.15} = 0.9876$$

Example 5: Calculate the value of the Pearson's coefficient of correlation for the following table.

Y	16 – 18	18 – 20	20 – 22	22 – 24
X				
10 – 20	2	1	1	-
20 – 30	3	2	3	2
30 – 40	3	4	5	6
40 – 50	2	2	3	4
50 – 60	-	1	2	2
60 – 70	-	1	2	1

Solution: Let us assume $\bar{X} = 30 - 40$, and $\bar{Y} = 18 - 20$, And we

have $U = \frac{X - (30 - 40)}{10}$ and $V = \frac{Y - (18 - 20)}{2}$

	16- 18	18- 20	20- 22	22- 24	f	U	fU	fU ²	IfV	U 之f fV

10-20	2	1	1		4	-2	-8	16	-1	2
20-30	3	2	3	2	10	-1	-10	10	4	-4
30-40	3	4	5	6	18	0	0	0	14	0
40-50	2	2	3	4	11	1	11	11	9	9
50-60		1	2	2	5	2	10	20	6	12
60-70		1	2	1	4	3	12	36	4	12
f	10	11	16	15	52	Total	15	93	36	31
V	-1	0	1	2	Total					
IV	-10	0	16	30	36					
fV ²	10	0	16	60	86					
ΣfU	-5	3	8	9	15					
VIfU	5	0	8	18	31					

From the table,

$$\Sigma fV = 36$$

$$\Sigma fV^2 = 86$$

$$\Sigma fU = 15$$

$$\Sigma fU^2 = 93$$

$$\Sigma f = 52$$

$$\Sigma fUV = 31.$$

The Pearson coefficient of correlation is

$$\begin{aligned}
 \gamma &= \frac{\Sigma fUV - \frac{\Sigma fU \times \Sigma fV}{\Sigma f}}{\sqrt{\left\{\Sigma fU^2 - \frac{(\Sigma fU)^2}{\Sigma f}\right\}\left\{\Sigma fV^2 - \frac{(\Sigma fV)^2}{\Sigma f}\right\}}} \\
 &= \frac{31 - \frac{15 \times 36}{52}}{\sqrt{\left(93 - \frac{15^2}{52}\right)\left(86 - \frac{36^2}{52}\right)}} = \frac{31 \times 52 - 15 \times 36}{\sqrt{(93 \times 52 - 225)(86 \times 52 - 1296)}} \\
 &= \frac{1072}{\sqrt{14644536}} = \frac{1072}{3826.8} = 0.28.
 \end{aligned}$$

CHECK YOUR PROGRESS

1. Define correlation and explain its importance in statistical analysis.

.....

.....

.....

2. Explain Karl Pearson's coefficient of correlation and its interpretation.

.....

.....

.....

3. Distinguish between positive, negative, and zero correlation.

.....

.....

.....

13.6 SUMMARY

In this unit, we have discussed the following concepts

- Correlation
- Coefficient and measurement of correlation
- Karl Pearson's coefficients of correlation
- Rank correlation
- Scatter diagram or dot diagram

13.7 EXERCISES

- 1 Calculate the coefficient of correlation between X and Y from the table of their values:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

- 2 Calculate the coefficient of correlation between the values of X and Y given below:

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

- 3 Calculate the coefficient of correlation for the following ages of husband and wife:

Husband's age	23	27	28	29	30	31	33	35	36	39
Wife's age	18	22	23	24	25	26	28	29	30	32

- 4 Calculate of coefficient of correlation for X and Y from the following data:

X	45	55	56	58	60	65	68	70	75	80	85
Y	56	50	48	60	62	64	65	70	74	82	90

- 5 From the following data calculate the coefficient of correlation and find its probable error:

Mean Annual birth rate per 1000 of population	35.3	33.5	31.4	30.5	29.3	28.2	26.3	23.6	20.1	19.9	16.7
Mean Annual birth rate per 1000 of population	30.8	19.4	18.9	18.7	17.7	16.0	14.7	14.3	14.4	12.2	12.1

- 6 Find the value of the Pearson's coefficient of correlation for the following table:

Y	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25
X	1	2	-	-	-
4 – 8	-	4	5	8	-
8 – 12	-	-	3	4	-
12 – 16	-	-	-	2	1

7 Find Pearson's coefficient of correlation from the following table:

Y	94.	96.	98.	100.	102.	104.	106.	108.	110.
	5	5	5	5	5	5	5	5	5
X									
29.5	-	-	4	3		4	1	-	1
59.5	1	3	6	18	6	9	2	3	1
89.5	7	3	16	16	4	4	1	-	1
119.	5	9	10	9	2	-	1	2	-
5									
149.	3	5	8	1	-	1	-	-	-
5									
179.	4	2	3	1	-	-	-	-	-
5									
209.	4	4	-	1	-	-	-	-	-
5									
239.	1	1	-	-	-	-	-	-	-
5									

8 Ten students got the following percentage of marks in mathematics and statistics:

No. of students	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Markes in Statistics	84	51	91	60	68	62	86	58	53	47

Find the coefficient of correlation of the above.

13.8 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 14 RANK CORRELATION AND SCATTER DIAGRAM

STRUCTURE

- 14.1 Introduction
- 14.2 Objectives
- 14.3 Rank Correlation
- 14.4 Scatter diagram
- 14.5 Summary
- 14.6 Exercises
- 14.7 References

14.1 INTRODUCTION

Understanding the relationships between variables is a cornerstone of data analysis, making **scatter diagrams** and **correlation** essential tools in statistics and research. This unit introduces these concepts, focusing on their practical applications and how they help reveal patterns and relationships in data.

A **scatter diagram**, also known as a scatter plot, is a graphical representation of data points for two variables. By plotting these points on a Cartesian plane, patterns, trends, or clusters may emerge, providing valuable insights into the data.

Correlation, on the other hand, measures the strength and direction of the relationship between two variables. Ranging from perfect positive (+1) to perfect negative (-1), correlation values help quantify how closely related the variables are and whether changes in one are associated with changes in the other.

14.2 OBJECTIVES

After completing this unit, students will be able to

- Define rank correlation.
- Visualize data effectively using scatter diagrams.
- Analyze and interpret the rank correlation between variables.
- Apply these tools in practical scenarios to make informed decisions based on data.

14.3 RANK CORRELATION

Charles Edward Spearman (1906) a great psychologist and statistician, developed a formula for the calculation of coefficient of correlation, which is called the "Rank correlation coefficient or Spearman's correlation coefficient.

Let us suppose that sometimes the condition is such type which cannot be measured quantitatively, but can be ranked among themselves. For example, it is possible for a class-teacher to arrange his students in ascending or descending order of intelligence, even though intelligence cannot be measured quantitatively. There are so many attributes which are incapable of quantitative measurement such as honesty, character, morality etc.

Suppose $(x_i, y_j), i = 1, 2, 3, \dots, n$ be the ranks of n individuals of a group corresponding to two characteristics A and B . Assuming that no two individuals are equal in either classification, each individual takes the values $1, 2, 3, \dots, n$.

Now,

$$\begin{aligned}\bar{X} &= \frac{1}{n}(1 + 2 + 3 + \dots + n) = \frac{n+1}{2} = \bar{Y} \\ \sigma_X^2 &= \frac{1}{n}\Sigma x_i^2 - \bar{X}^2 = \frac{1}{n}(1^2 + 2^2 + 3^2 + \dots + n^2) - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12} = \sigma_Y^2\end{aligned}$$

Let

$$\begin{aligned}d_i &= x_i - y_i = (x_i - \bar{X}) - (y_i - \bar{Y}) \\ \frac{1}{n}\Sigma d_i^2 &= \frac{1}{n}\Sigma\{(x_i - \bar{X}) - (y_i - \bar{Y})\}^2 \\ &= \frac{1}{n}\{\Sigma(x_i - \bar{X})^2 + \Sigma(y_i - \bar{Y})^2 - 2\Sigma(x_i - \bar{X})(y_i - \bar{Y})\} \\ &= \sigma_x^2 + \sigma_y^2 - 2\text{Cov}(X, Y)\end{aligned}$$

or

$$\begin{aligned}\text{Cov}(X, Y) &= \frac{1}{2}\left[\sigma_x^2 + \sigma_y^2 - \frac{1}{n}\Sigma d_i^2\right] = \frac{1}{2}\left[\frac{n^2-1}{12} + \frac{n^2-1}{12} - \frac{1}{n}\Sigma d_i^2\right] \\ &= \frac{1}{12}(n^2-1) - \frac{1}{2n}\Sigma d_i^2\end{aligned}$$

The correlation coefficient, between two variables X and Y is

$$\gamma(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{12}(n^2-1) - \frac{1}{2n}\Sigma d_i^2}{\sqrt{\frac{n^2-1}{12}} \sqrt{\frac{n^2-1}{12}}}$$

$$\begin{aligned}&= \frac{\frac{(n^2-1)}{12} - \frac{1}{2n}\Sigma d_i^2}{\frac{n^2-1}{12}} \\ \gamma &= 1 - \frac{\sigma \Sigma d_i^2}{n(n^2-1)}\end{aligned}$$

The method of measuring association between two sets of ranks is known as the rank correlation method.

RANK CORRELATION COEFFICIENT FOR REPEATED RANKS

Suppose given the following data:

<i>X</i>	68	64	75	50	64	80	75	40	55	64
<i>Y</i>	62	58	68	45	81	60	68	48	50	70

Here, we see that in *X*-series, 75 occurs 2 times. 64 occurs 3 times and in *Y*-series, 68 occurs 2 times. If any two or more individuals have same value with respect to characteristics *X* and *Y*, then Spearman's rank formula fails. For these common ranking the correct formula for γ , to Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. As in given above example,

For the *X*-series

$$\begin{aligned} &= \frac{2(2^2 - 1)}{12} + \frac{3(3^2 - 1)}{12} = \frac{1}{2} + 2 = \frac{5}{2} \\ &= \frac{2(2^2 - 1)}{12} = \frac{1}{2} \\ \gamma &= 1 - \frac{6\left(\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right)}{n(n^2 - 1)} \end{aligned}$$

14.4 SCATTER DIAGRAM OR DOT DIAGRAM

In this method the values of two variables (series) are plotted on a graph, taking one on *X*-axis and the other along *Y*-axis. There is one dot for one pair of series, there being as many dots on the graph as their pairs. This collection of dots is called a scatter diagram or dot diagram.

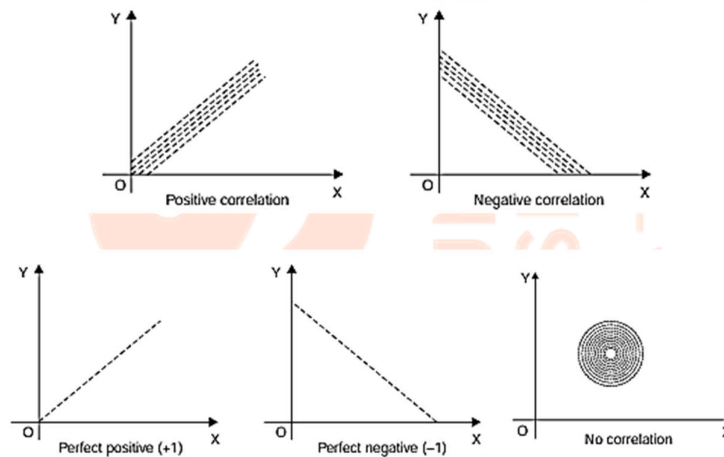


Figure 14.1 Scatter or dot diagram

Example 1: The marks secured by students in mathematics and statistics are given below:

Mathematics (X)	10	15	12	17	13	16	25	14	22
Statistics (Y)	30	42	45	46	33	34	40	35	39

Calculate the rank correlation coefficient.

Solution: Calculation of the coefficient of rank correlation

X	Rank of X R_1	Y	Rank of Y R_2	$R_1 - R_2$ $= d$	$(R_1 - R_2)^2$ $= d^2$
10	9	30	9	0	0
15	5	42	3	2	4
12	8	45	2	6	36
17	3	46	1	2	4
13	7	33	8	-1	1
16	4	34	7	-3	9
25	1	40	4	-3	9

14	6	35	6	0	0
22	2	39	5	-3	9
n = 9		n = 9		$\Sigma d = 10$	$\Sigma C^* = 72$

The coefficient of rank correlation

$$r = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 72}{9(9^2 - 1)} = 1 - \frac{432}{720} = 0.4$$

Example 2: Ten competitors in beauty contest are ranked by three judges in the following order:

First Judge by R_1	1	6	5	10	3	2	4	9	7	8
Second Judge by R_2	3	5	8	4	7	10	2	1	6	9
Third Judge by R_3	6	4	9	8	1	2	3	10	5	7

Use the rank correlation method to discuss which pair of judges have the nearest approach to common tastes in beauty.

Solution: Calculation of the coefficient of rank correlation between the ranking of first, second and third Judges.

R_1	R_2	R_3	R_1 - R_2 = d_1	R_1 - R_3 = d_2	R_2 - R_3 = d_3	d_1^2	d_2^2	d_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36

2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
			Σd_1 = 0	Σd_2 = 0	Σd_3 = 0	Σd_1^2 = 200	Σd_2^2 = 60	Σd_3^2 = 214

Here, $n = 10$.

$$\gamma(1,2) = 1 - \frac{6\Sigma d_1^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33} = -0.212$$

$$\gamma(1,3) = 1 - \frac{6\Sigma d_2^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 90} = 1 - \frac{4}{11} = \frac{7}{11} = 0.636$$

$$\gamma(2,3) = 1 - \frac{6\Sigma d_3^2}{n(n^2 - 1)} = 1 - \frac{6 \times 214}{10 \times 90} = 1 - \frac{214}{165} = -\frac{49}{165} = -0.297$$

Here, we see that the first and third Judges have the nearest approach to their tastes for beauty.

Example 3: Find rank correlation coefficient between x and y .

x	85	91	56	72	95	76	89	51	90	90
y	18.3	20.8	16.9	15.7	19.2	18.1	17.5	14.9	18.9	15.4

Solution: Let us assign the ranks to both the series x and y in order of their magnitude. That is, we assign rank 1 to the highest and rank 2 to the next and so on.

Table showing ranks and calculations

x	y	Rank of x	Rank of y	$d = x - y$	d^2
85	18.3	5	4	1	1
91	20.8	2	1	1	1
56	16.9	9	7	2	4
72	15.7	7	8	-1	1
95	19.2	1	2	-1	1
76	18.1	6	5	1	1
89	17.5	4	6	-2	4
51	14.9	10	10	0	0
59	18.9	8	3	5	25
90	15.4	3	9	-6	36
$n = 10$					$\Sigma d^2 = 74$

$$\begin{aligned}
 r &= 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 74}{10(100 - 1)} \\
 &= 1 - \frac{444}{990} = 1 - 0.45 = 0.55.
 \end{aligned}$$

CHECK YOUR PROGRESS

1. What is a scatter diagram? Explain its uses.

.....

.....

.....

2. Define rank correlation and explain Spearman's rank correlation coefficient.

.....

.....

.....

3. Distinguish between Karl Pearson's correlation and rank correlation.

.....

.....

.....

14.5 SUMMARY

In this unit, we have discussed the following concepts

- Rank correlation
- Scatter diagram or dot diagram

14.6 EXERCISES

1. Ten students got the following percentage of marks in mathematics and statistics:

No. of students	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

Find the coefficient of correlation and rank correlation of the above.

2. Prove that the coefficient of correlation r is independent of change of origin and scale.

3. Find rank correlation coefficient between x and y .

x	1	3	4	5	7	7	10
y	9	13	1	17	21	23	27

4. What is rank correlation? When do you use rank correlation? Compute rank correlation from the following data of marks of 8 students in statistics (y) and accountancy (x).

x :	86	83	91	77	68	85	52	82	57	37
y :	89	92	87	86	83	77	71	63	50	53

5. Obtain the rank correlation for the following data:

x :	68	64	75	50	64	80	75	40	55	60
y :	62	58	68	45	81	60	68	48	50	70

14.7 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., McKean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 15 REGRESSION LINES AND COEFFICIENTS

STRUCTURE

- 15.1 Introduction
 - 15.2 Objectives
 - 15.3 Line of regression
 - 15.4 Standard error of estimate
 - 15.5 Examples
 - 15.6 Summary
 - 15.7 Exercises
 - 15.8 References
-

15.1 INTRODUCTION

Regression analysis offers the means to accomplish the crucial task of predicting one variable based on another in data analysis. Regression lines and regression coefficients, the foundation of statistical analysis and predictive modelling, are introduced in this lesson.

The relationship between two variables in a dataset is represented by a regression line, which is the line of best fit. It illustrates the relationship between changes in a dependent variable (response) and changes in an independent variable (predictor) both mathematically and visually.

This course provides an introduction to more complex predictive analytics subjects and gives you the tools you need to find patterns and make sense of data.

15.2 OBJECTIVES

After completing this unit, students will be able to

- Explain the concept of regression lines and their importance in data analysis.

- Compute and interpret regression coefficients to describe relationships between variables.
- Apply regression models to solve practical problems and make predictions.

15.3 LINE OF REGRESSION

finding the probable value of one variable for a given value of the other, when the two variables are known to be correlated. Thus, if we know that two series relating to supply and price are correlated, we can find out what would be the effect on price, if the supply of commodity is increased or decreased to a particular level.

Regression is the estimation for prediction of unknown values of one variable from known values of another variable. The line which describes the average relationship between two variables is known as line of regression. The equations describing the regression lines are called regression equations. A regression equation is the algebraic expression of regression line.

A regression line is one which shows that the best mean values of one variable corresponding to mean values of the other variable. With two series x and y , if the two regression lines range themselves along two straight lines, the correlation between x and y is linear. If the two straight lines of regression coincide, correlation is perfect. If the two lines cut each other at right angles, correlation is zero.

Let $Y = aX + b$ be the equation of the line of the best fit of x . Changing the origin to $(\bar{X}, \bar{Y})$, where $\bar{X}$ is the mean of x -series and $\bar{Y}$ is the mean of y -series. Let x, y be the deviations from the respective means $\bar{X}$ and $\bar{Y}$.

$$\therefore x = X - \bar{X}, y = Y - \bar{Y}$$

Then, equation $Y = aX + b$ is changed in the form

$$\begin{aligned} y &= ax + b \\ x &= X - \bar{X} \text{ and } y = Y - \bar{Y} \end{aligned}$$

where,

Let (x_c, y) be any point. Then the difference between this point and the above line is

$$y_c - ax_c - b$$

Hence U , the sum of the squares of such distance is

$$U = \Sigma(y - ax - b)^2, \text{ for all values of } c$$

Now for U is minimum, we choose a and b such that

$$\begin{aligned} \frac{\partial U}{\partial a} &= -2\Sigma x(y - ax - b) = 0 \\ \text{and } \frac{\partial U}{\partial b} &= -2\Sigma(y - ax - b) = 0 \\ \text{and } \Rightarrow \Sigma xy - a\Sigma x^2 - b\Sigma x &= 0 \\ \Sigma y - a\Sigma x - nb &= 0 \end{aligned}$$

Since,

$$\Sigma x = 0 = \Sigma y, \text{ we get}$$

$\Rightarrow$

$$\begin{aligned} b &= 0 \text{ and } a = \frac{\Sigma xy}{\Sigma x^2} \\ \Rightarrow a &= \frac{\Sigma xy}{\Sigma x^2} = \frac{\text{Cov}(x, y)}{\sigma_x^2} = \gamma \frac{\sigma_y}{\sigma_x} \end{aligned}$$

where, γ is the coefficient of correlation between x and y .

Hence, the equation to the line of regression

$$y = \gamma \frac{\sigma_y}{\sigma_x} x.$$

$$\Rightarrow Y - \bar{Y} = \gamma \frac{\sigma_y}{\sigma_x} (X - \bar{X}) \text{ where, } x = X - \bar{X} \text{ and } y = Y - \bar{Y},$$

which is called the regression line of Y on X .
Similarly, the regression line of X on Y is

$$X - \bar{X} = \gamma \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

Therefore, if the straight line is so chosen that the sum of squares of deviation parallel to the axis of x is minimum, it is called the line of regression of x on y .

15.4 STANDARD ERROR OF ESTIMATE

The standard error measures the accuracy of estimate.

It is given by

$$E_{jx} = \sqrt{\frac{\Sigma(y - y_c)^2}{n}}$$

where, y is the observed value, y_c is the estimated value and n is the number of pairs of items.

Note.

- 1 The coefficient $\gamma \frac{\sigma_y}{\sigma_x}$ and $\gamma \frac{\sigma_x}{\sigma_y}$ are called the regression coefficient of y on x , and of x on y respectively.
- 2 If the two lines of regression coincide, the correlation between the variables is perfect i.e.,

$$\gamma \frac{\sigma_y}{\sigma_x} = \frac{1}{\gamma} \frac{\sigma_y}{\sigma_x}$$

$$\Rightarrow \gamma^2 = 1$$

$$\Rightarrow \gamma = \pm 1$$

- 3 If the variables x and y are independent i.e., the coefficient of correlation between them is zero.
- 4 Some diagrams for different values of γ or r .

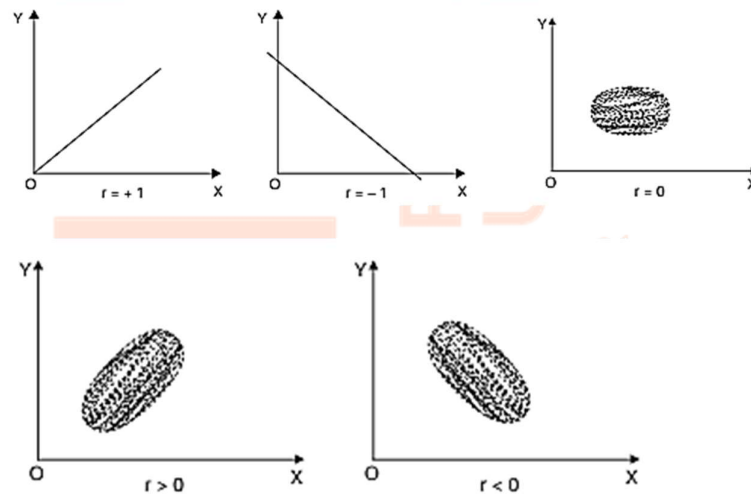


Figure 15.1 Some diagram for different value of γ

- 5 Also denotes $\gamma \frac{\sigma_y}{\sigma_x} = b_{yx}$ and $\gamma \frac{\sigma_x}{\sigma_y} = b_{xy}$

15.5 EXAMPLES

Example 1: If θ be the angle between the lines of regression, show that

$$\tan \theta = \frac{(1 - \gamma^2)}{\gamma} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

Explain the significance of the formula, when $\gamma = 0, \gamma = \pm 1$.

Solution: The regression line of y on x

$$y - \bar{y} = \gamma \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad (1)$$

Here slope

$$m_1 = \gamma \frac{\sigma_y}{\sigma_x} \quad (2)$$

The regression line of x on y

$$x - \bar{x} = \gamma \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

The slope

$$m_2 = \frac{1}{\gamma} \frac{\sigma_y}{\sigma_x} \quad (3)$$

If m_1 and m_2 are the slopes of the lines and θ is the angle between them, we have

$$\begin{aligned} \tan \theta &= \frac{m_2 - m_1}{1 + m_1 m_2} = \frac{\frac{1}{\gamma} \frac{\sigma_y}{\sigma_x} - \gamma \frac{\sigma_y}{\sigma_x}}{1 + \gamma \frac{\sigma_y}{\sigma_x} \cdot \frac{1}{\gamma} \frac{\sigma_y}{\sigma_x}} \\ &= \frac{\frac{\sigma_y}{\sigma_x} \left(\frac{1 - \gamma^2}{\gamma} \right)}{1 + \frac{\sigma_y^2}{\sigma_x^2}} = \left(\frac{1 - \gamma^2}{\gamma} \right) \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}. \end{aligned}$$

when $\gamma = 0$, $\tan \theta = \infty$, $\theta = \frac{\pi}{2}$ i.e., the two lines of regression are perpendicular to each other or variables are uncorrelated.

When $\gamma = \pm 1$, $\tan \theta = 0$, $\theta = 0$ or π i.e., the lines coincide and therefore is a perfect correlation between the two variables x and y .

The estimated value of y is the same for all values of x or vice-versa.

Example 2: For two variables x and y with the same mean, the two regression equations are $y = ax + b$ and $x = \alpha y + \beta$. Show that $\frac{b}{\beta} = \frac{1-a}{1-\alpha}$. Find also the common mean.

Solution: Suppose the common mean is m .

The given two regression equations are

$$y = ax + b \quad (1)$$

$$x = \alpha y + \beta \quad (2)$$

$$y - m = a(x - m) \quad (3)$$

$$x - m = \alpha(y - m) \quad (4)$$

Comparing equation (3) and (4) with (1) and (2), we get

$$b = m(1 - a) \quad (5)$$

$$\beta = m(1 - \alpha) \quad (6)$$

Dividing (5) by (6), we get

$$\frac{b}{\beta} = \frac{1 - a}{1 - \alpha}$$

Also (5) and (6), we have

$$m = \frac{b}{1 - a} = \frac{\beta}{1 - \alpha}$$

Example 3: Prove that the Pearson's coefficient of correlation γ lies between -1 and +1 .

Solution: Let

$$U = \Sigma(y - ax - b)^2$$

For U is minimum when $\frac{\partial U}{\partial a} = 0$ and $\frac{\partial U}{\partial b} = 0$ differentiate partially (1) with respect to a and b respectively, we get and

$$\begin{aligned}\frac{\partial U}{\partial a} &= -2\sum x(y - ax - b) = 0 \\ \frac{\partial U}{\partial b} &= -2\sum (y - ax - b) = 0\end{aligned}$$

Since, $\sum x = 0 = \sum y$, then we get

$$b = 0 \text{ and } a = \frac{\sum xy}{\sum x^2}$$

Now by (1), we have

$$\begin{aligned}U &= \sum (y - ax)^2 = \sum y^2 - 2a\sum xy + a^2\sum x^2 \\ &= \sum y^2 - 2 \frac{\sum xy}{\sum x^2} \cdot \sum xy + \left(\frac{\sum xy}{\sum x^2}\right)^2 \sum x^2 \\ &= \sum y^2 - 2 \frac{(\sum xy)^2}{\sum x^2} + \frac{(\sum xy)^2}{\sum x^2} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} \\ &= \sum y^2 \left[1 - \frac{(\sum xy)^2}{\sum x^2 \sum y^2}\right] \\ &= \sum y^2 [1 - \gamma^2], \text{ where } \gamma = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \\ &= (1 - \gamma^2) \sum y^2\end{aligned}$$

Since, U be the sum of squares then $\sum y^2$ will not be negative.

$$\Rightarrow 1 - \gamma^2 \geq 0 \text{ or } \gamma^2 \leq 1$$

Hence, γ lies between -1 and +1.

Example 4: Show that the coefficient of correlation is the geometric mean between the two regression coefficients.

Solution: The regression coefficient of y on $x = \gamma \frac{\sigma_y}{\sigma_x}$

The regression coefficient of x on $y = \gamma \frac{\sigma_x}{\sigma_y}$

$$\text{Geometric mean} = \sqrt{\gamma \frac{\sigma_y}{\sigma_x} \cdot \gamma \frac{\sigma_x}{\sigma_y}}$$

$= \gamma = \text{coefficient of correlation.}$

Example 5: For a bivariate distribution

$$n = 18, \Sigma X^2 = 60, \Sigma Y^2 = 96, \Sigma X = 12, \Sigma Y = 18, \Sigma XY = 48$$

find the equation of the lines of regression and γ .

Solution: We have,

$$\begin{aligned}\bar{X} &= \frac{1}{n} \Sigma X = \frac{12}{18} = \frac{2}{3} \\ \bar{Y} &= \frac{1}{n} \Sigma Y = \frac{18}{18} = 1 \\ \sigma_X^2 &= \frac{1}{n} \Sigma X^2 - \bar{X}^2 = \frac{60}{18} - \frac{4}{9} = \frac{26}{9} \\ \sigma_Y^2 &= \frac{1}{n} \Sigma Y^2 - \bar{Y}^2 = \frac{96}{18} - 1 = \frac{13}{3} \\ \text{Cov}(X, Y) &= \frac{1}{n} \Sigma XY - \bar{X}\bar{Y} \\ &= \frac{48}{18} - \left(\frac{2}{3}\right)(1) = \frac{48 - 12}{18} = \frac{36}{18} = 2 \\ \gamma(X, Y) &= \frac{\text{Cov}(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{2}{\sqrt{\frac{26}{9} \times \frac{13}{3}}} = \frac{6}{13} \sqrt{\frac{3}{2}}\end{aligned}$$

Now

The regression line of Y on X is

$$\begin{aligned}Y - \bar{Y} &= \gamma \frac{\sigma_Y}{\sigma_X} (X - \bar{X}) \\ Y - 1 &= \frac{6}{13} \sqrt{\frac{3}{2}} \sqrt{\frac{3}{2}} \left(X - \frac{2}{3}\right) \\ Y - 1 &= (0.692)(X - 0.666) \\ Y &= 0.692X - 0.461 + 1 \\ Y &= 0.692X + 0.538\end{aligned}$$

The regression line of X on Y is

$$\begin{aligned}\Rightarrow (X - \bar{X}) &= r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y}) \\ \Rightarrow X - \frac{2}{3} &= \frac{6}{13} \sqrt{\frac{3}{2}} \times \sqrt{\frac{2}{3}} (Y - 1) \\ \Rightarrow X - \frac{2}{3} &= \frac{6}{13} (Y - 1) \\ \Rightarrow X &= 0.46Y - 0.46 + 0.66 \\ \Rightarrow X &= 0.46Y + 0.20.\end{aligned}$$

Example 6: Two random variables have the least square regression lines with equations

$$3x + 2y - 26 = 0 \text{ and } 6x + y - 31 = 0$$

find the mean values and the coefficient of correlation between x and y .

Solution: Given equations are

$$3x + 2y - 26 = 0 \quad (1)$$

$$6x + y - 31 = 0 \quad (2)$$

Since, the two lines of regression pass through the point $(\bar{x}, \bar{y})$, we have

$$3\bar{x} + 2\bar{y} = 26 \quad (3)$$

$$6\bar{x} + \bar{y} = 31 \quad (4)$$

Solving (3) and (4), we get

$$\bar{x} = 4 \text{ and } \bar{y} = 7$$

From (1), we have

From (1), we have $y = -\frac{3}{2}x + 13$ (5)

$\therefore b_{yx} = -\frac{3}{2}$

Now from (2), we have

$$x = -\frac{1}{6}y + \frac{31}{6}$$

$$b_{xy} = -\frac{1}{6}$$

(6)

Multiplying (5) and (6), we get

$$b_{yx}, b_{xy} = \left(-\frac{3}{2}\right) \times \left(-\frac{1}{6}\right)$$

$$\gamma^2 = \frac{1}{4}$$

$$\Rightarrow \gamma = \pm \frac{1}{2} = \pm 0.5$$

Since, both regression coefficient b_{yx} and b_{xy} are negative.

Then

$$\gamma = -0.5.$$

Example 7: Calculate the coefficient of correlation and obtain the line of regression for the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

find also an estimate for Y which would correspond to $X = 6.2$.

Solution: Let $\bar{X} = 4$ and $\bar{Y} = 13$.

X	Y	U $= X - \bar{X}$	V $= Y - \bar{Y}$	U^2	V^2	UV
1	9	-3	-4	9	16	12
2	8	-2	-5	4	25	10
3	10	-1	-3	1	9	3
4	12	0	-1	0	1	0
5	11	1	-2	1	4	-2
6	13	2	0	4	0	0
7	14	3	1	9	1	3
8	16	4	3	16	9	12
9	15	5	2	25	4	10
		$\Sigma U = 9$	$\Sigma V = -9$	ΣU^2 $= 69$	ΣV^2 $= 69$	ΣUV $= 48$

Hence,

$$\begin{aligned}\bar{U} &= \frac{1}{n} \Sigma U = \frac{1}{9} \times 9 = 1, \\ \bar{V} &= \frac{1}{n} \Sigma V = \frac{1}{9} \times (-9) = -1 \\ \text{Cov}(U, V) &= \frac{1}{n} \Sigma UV - \bar{U}\bar{V} = \frac{1}{9} \times 48 - 1(-1) = 5.33 + 1 = 6.33 \\ \sigma_U^2 &= \frac{1}{n} \Sigma U^2 - \bar{U}^2 = \frac{1}{9} \times 69 - 1 = \frac{60}{9} = 6.66 \\ \sigma_V^2 &= \frac{1}{n} \Sigma V^2 - \bar{V}^2 = \frac{1}{9} \times 69 - 1 = \frac{60}{9} = 6.66 \\ \gamma &= \frac{\text{Cov}(UV)}{\sqrt{\sigma_U^2 \sigma_V^2}} = \frac{6.33}{\sqrt{6.66 \times 6.66}} = \frac{6.33}{6.66} = 0.95.\end{aligned}$$

So $\bar{X} = 4$, $\bar{Y} = 13$,

$$\sigma_X = \sigma_Y = \sqrt{\frac{60}{9}}$$

The regression line of Y on X is

$$\begin{aligned}Y - \bar{Y} &= \gamma \frac{\sigma_Y}{\sigma_X} (X - \bar{X}) \\ \Rightarrow Y - 13 &= (0.95) \frac{\sqrt{609}}{\sqrt{609}} (X - 4) \\ \Rightarrow Y - 13 &= (0.95)(X - 4) \\ \Rightarrow Y &= 0.95X - 3.80 + 13 \\ \Rightarrow Y &= 0.95X + 9.20.\end{aligned}$$

Example 8: Show that the A.M. of the coefficient of regression is greater than the coefficient of correlation.

Solution: The regression line of Y on X is

$$Y - \bar{Y} = \gamma \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

The regression line of X on Y is

$$X - \bar{X} = \gamma \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

Now to show that, we have

$$\begin{aligned} \frac{\gamma \frac{\sigma_Y}{\sigma_X} + \gamma \frac{\sigma_X}{\sigma_Y}}{2} &> \gamma \\ \Rightarrow \frac{\sigma_Y^2 + \sigma_X^2}{2\sigma_X\sigma_Y} &> 0 \Rightarrow \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y > 0 \\ (\sigma_Y - \sigma_X)^2 &> 0 \end{aligned}$$

Which is true.

CHECK YOUR PROGRESS

1. Define regression and explain its importance in statistical analysis.

.....

.....

.....

2. Explain the two regression lines and regression coefficients.

.....

.....

.....

3. Distinguish between correlation coefficient and regression coefficient.

.....

.....

.....

15.6 SUMMARY

In this unit, we have discussed the following concepts

- Line of regression
- Standard error of estimate

15.7 EXERCISES

- 1 Find the correlation coefficient and the equation of regression lines for the following values of X and Y :

X	1	2	3	4	5
Y	2	5	3	8	7

- 2 Find $\text{Cov}(X, Y)$, when $\Sigma X = 50, \Sigma Y = -30, \Sigma XY = -115, n = 10$.
- 3 Given the following data:

	X	Y
Arithmetic mean	36	85
Standard deviation	11	8

Correlation coefficient between X and Y is 0.66.

- (a) Find two regression lines.
- (b) Estimate value of X , when $Y = 75$.

4. Find the coefficient of correlation for X and Y from the following data:

X	45	55	56	58	60	65	68	70	75	80	85
Y	56	50	48	60	62	64	65	70	74	82	90

Also find the equation of the lines of regression.

5. The lines of regression for a data are given as under:

$$2y - x - 50 = 0, 3y - 2x - 10 = 0.$$

Show that the regression estimates of y for $x = 150$ is 100, whereas the regression estimates of x corresponding to $y = 100$ is 145. Explain the difference.

6. The variables x and y are connected by the equations $ax + by + c = 0$, show that the correlation between them is -1, if the sign of a and b are like, and +1 if they are different.

7. Heights of fathers and sons are given in the inches

Height of Father (X)	65	66	67	67	68	69	71	73
Height of Son (Y)	67	68	64	68	72	70	69	70

From the two lines of regression and calculate the expected average height of the son when the height of the father is 67.5 inches.

15.8 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 16 CURVE FITTING: METHOD OF LEAST SQUARES

STRUCTURE

- 16.1 Introduction
- 16.2 Objectives
- 16.3 Curve fitting
- 16.4 Method of least squares
- 16.5 Examples
- 16.6 Summary
- 16.7 Exercises
- 16.8 References

16.1 INTRODUCTION

Regression analysis offers the means to accomplish the crucial task of predicting one variable based on another in data analysis. Regression lines and regression coefficients, the foundation of statistical analysis and predictive modelling, are introduced in this lesson.

The relationship between two variables in a dataset is represented by a regression line, which is the line of best fit. It illustrates the relationship between changes in a dependent variable (response) and changes in an independent variable (predictor) both mathematically and visually.

This course provides an introduction to more complex predictive analytics subjects and gives you the tools you need to find patterns and make sense of data.

16.2 OBJECTIVES

After completing this unit, students will be able to

- Recognize the significance of curve fitting in modelling and data analysis.
- Get the equations for the best-fitting lines or curves, use the least squares method.
- Assess the fit's quality and analyze its consequences for patterns and predictions in the data.

16.3 CURVE FITTING

Let $(x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$ are the n numerical values of two variables x (independent) and y (dependent). By the scatter diagram, we get an approximate relation between these two variables called empirical law. Curve fitting means that relationship between two variables in the form of equation of the curve from the given data. In other words, the method of obtaining an equation of best fit is called curve fitting.

Methods of Curve Fitting Following are some methods for fitting a curve:

1. Graphical method
2. Method of least squares
3. Method of group averages
4. Method of moments.

But in this chapter, we shall discuss only method of least squares for straight line and parabola.

16.4 METHOD OF LEAST SQUARE

16.4.1 METHOD OF LEAST SQUARES

Let us consider a set of m observations $(x_i, y_i): i = 1, 2, 3, \dots, m$ of two variables x and y . We wish to fit a polynomial of degree n . Assume that such curve is

$$y = a + bx + cx^2 + \dots + qx^n \text{ of these values.}$$

In order to determine $a, b, c, \dots, q$ such that it represents the curve of best fit. Now we apply the principle of least squares so put $x = x_1, x_2, \dots, x_m$ in (5.1), we get

$$\left. \begin{aligned} y'_1 &= a + bx_1 + cx_1^2 + \dots + qx_1^n \\ y'_2 &= a + bx_2 + cx_2^2 + \dots + qx_2^n \\ &\vdots \\ y'_m &= a + bx_m + cx_m^2 + \dots + qx_m^n \end{aligned} \right\}$$

where $y'_1, y'_2, \dots, y'_m$ are the expected value of y for $x = x_1, x_2, \dots, x_m$ respectively, which are other than observed values $y_1, y_2, \dots, y_m$ of y for $x = x_1, x_2, \dots, x_m$ respectively. In general, expected values and observed values are different, the difference $y_i - y'_i; i = 1, 2, \dots, m$ are called residuals.

Let $A = \Sigma(y_i - y')^2 = \Sigma[y_i - (a + bx_i + cx_i^2 + \dots + qx_i^n)]^2; i = 1, 2, \dots, m$.

The constant $a, b, c, \dots, q$ are chosen as the sum of squares of $y_i - y'_i; i = 1, 2, \dots, m$ is minimum.

Now for maximum and minimum

$$\frac{\partial A}{\partial a} = 0, \frac{\partial A}{\partial b} = 0, \dots, \frac{\partial A}{\partial q} = 0$$

On solving these equations, we get

$$\left. \begin{aligned} ma + b\Sigma x + c\Sigma x^2 + \dots + q\Sigma x^n &= \Sigma y \\ a\Sigma x + b\Sigma x^2 + c\Sigma x^3 + \dots + q\Sigma x^{n+1} &= \Sigma xy \\ a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 + \dots + q\Sigma x^{n+2} &= \Sigma x^2 y \\ &\vdots \\ a\Sigma x^n + b\Sigma x^{n+1} + c\Sigma x^{n+2} + \dots + q\Sigma x^{2n} &= \Sigma x^n y \end{aligned} \right\}$$

These equations are known as normal equations. On solving these normal equations, we get the values of constants $a, b, c \dots \dots q$.

Note 1: If $n = 1$, then the curve to be fitted is a straight line $y = a + bx$ whose normal equations are

$$\begin{aligned} \Sigma y &= ma + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}$$

which can be solved for a and b .

Note 2: If $n = 2$, then the curve to be fitted is a parabola of second degree $y = a + bx + cx^2$ whose normal equations are

$$\begin{aligned} \Sigma y &= ma + b\Sigma x + c\Sigma x^2 \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2 y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}$$

which can be solved for a, b and c .

16.4.2 WORKING RULE TO FIT A STRAIGHT LINE TO GIVEN DATA BY METHOD OF LEAST SQUARES

- (1) Let the equation of straight line be $y = a + bx$.
- (2) From the given data we calculate $\Sigma x, \Sigma y, \Sigma x^2$, and Σxy .
- (3) Putting above values in normal equations

$$\begin{aligned} ma + b\Sigma x &= \Sigma y \\ a\Sigma x + b\Sigma x^2 &= \Sigma xy \end{aligned}$$

- (4) Solving equations (i) and (ii) for a and b .
- (5) Putting the values of a and b in $y = a + bx$ we get equation of straight line of best fit.

16.4.3 WORKING RULE TO FIT A PARABOLA TO THE GIVEN DATA BY METHOD OF LEAST SQUARES

- (1) Let the equation of the parabola be $y = a + bx + cx^2$.
- (2) From the given data we calculate $\Sigma x, \Sigma y, \Sigma x^2, \Sigma x^3, \Sigma x^4, \Sigma xy$ and Σx^2y .
- (3) Putting above values in normal equations

$$\begin{aligned} ma + b\Sigma x + c\Sigma x^2 &= \Sigma y \\ a\Sigma x + b\Sigma x^2 + c\Sigma x^3 &= \Sigma xy \\ a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 &= \Sigma x^2y \end{aligned}$$

- (4) Solving equations (i), (ii) and (iii) for a, b and c .
- (5) Putting the values of a and b in $y = a + bx + cx^2$ we get equation of parabola of best fit.

16.5 EXAMPLES

Example 1. Fit a straight line $y = a + bx$, to the following data regarding x as the independent variable by the method of least squares.

x	0	1	3	6	8
y	1	3	2	4	5

Solution. Let the equation of straight line be $y = a + bx$ and the normal equations are:

$$\begin{aligned} ma + b\Sigma x &= \Sigma y \\ a\Sigma x + b\Sigma x^2 &= \Sigma xy \end{aligned}$$

Now

x	y	x^2	xy
0	1	0	0
1	3	1	3
3	2	9	6
6	4	36	24
8	5	64	40
$\Sigma x = 18$	$\Sigma y = 15$	$\Sigma x^2 = 110$	$\Sigma xy = 73$

Here $m = 5, \Sigma x = 18, \Sigma y = 15, \Sigma x^2 = 110, \Sigma xy = 73$.

Put these values in normal equations

$$5a + 18b = 15 \quad (i)$$

$$18a + 110b = 73 \quad (ii)$$

On solving (i) and (ii), we get

$$a = 1.488, b = 0.420$$

Thus, the equation of straight line $y = 1.488 + 0.42x$.

Example 2. Fit a straight line to the following data regarding x as a independent variable.

x	1	2	3	4	5
y	5	7	9	10	11

Solution. Let equation of straight line be $y = a + bx$ and the normal equations are:

$$\begin{aligned} ma + b\Sigma x &= \Sigma y \\ a\Sigma x + b\Sigma x^2 &= \Sigma xy \end{aligned}$$

Now

x	y	x^2	xy
1	5	1	5
2	7	4	14
3	9	9	27
4	10	16	40
5	11	25	55
$\Sigma x = 15$	$\Sigma y = 42$	$\Sigma x^2 = 55$	$\Sigma xy = 141$

Here $m = 5, \Sigma x = 15, \Sigma y = 42, \Sigma x^2 = 55, \Sigma xy = 141$.

Put these values in normal equations, we have

$$5a + 115b = 42 \quad (i)$$

$$15a + 55b = 141 \quad (ii)$$

On solving (i) and (ii), we get

$$a = 3.9 \text{ and } b = 1.5$$

Thus, the equation of straight line is $y = 3.9 + 1.5x$.

Example 3. Find a least squares straight line for the following data.

x	1	2	3	4
y	3	7	13	21

Solution. Let equation of straight line be $y = a + bx$ and the normal equations are:

$$\begin{aligned} ma + b\Sigma x &= \Sigma y \\ a\Sigma x + b\Sigma x^2 &= \Sigma xy \end{aligned}$$

Now

x	y	x^2	xy
1	3	1	3
2	7	4	14
3	13	9	39
4	21	16	84
$\Sigma x = 10$	$\Sigma y = 44$	$\Sigma x^2 = 30$	$\Sigma xy = 140$

Here $m = 4, \Sigma x = 10, \Sigma y = 44, \Sigma x^2 = 30, \Sigma xy = 140$.

Put these values in normal equations

$$4a + 10b = 44 \quad (i)$$

$$10a + 30b = 140 \quad (ii)$$

On solving (i) and (ii), we get

$$a = -4 \text{ and } b = 6$$

Thus, the equation of straight line is $y = -4 + 6x$.

Example 4. Fit a straight line to the following data regarding x as the independent variable.

x	1	2	3	4	5	6
y	1200	900	600	200	110	50

Solution. Let the equation of straight line be $y = a + bx$ and the normal equations are:

$$\begin{aligned} ma + b\Sigma x &= \Sigma y \\ a\Sigma x + b\Sigma x^2 &= \Sigma xy \end{aligned}$$

Now

x	y	x^2	xy
1	1200	1	1200
2	900	4	1800
3	600	9	1800
4	200	16	800
5	110	25	550
6	50	36	300
$\Sigma x = 21$	$\Sigma y = 3060$	$\Sigma x^2 = 91$	$\Sigma xy = 6450$

Here $m = 6, \Sigma x = 21, \Sigma y = 3060, \Sigma x^2 = 91, \Sigma xy = 6450$.

Put these values in normal equations

$$6a + 21b = 3060 \quad (i)$$

$$21a + 91b = 6450 \quad (ii)$$

On solving (i) and (ii), we get

$$a = 1361.97 \text{ and } b = -243.42$$

Thus, the equation of straight line is $y = 1361.97 - 243.42x$.

Example 5. Fit a polynomial of the second degree to the data points given in the following table.

x	0	1	2
y	1	6	17

Solution. Let $y = a + bx + cx^2$ be a parabola to be fitted for the given data and normal equations are:

$$\begin{aligned} ma + b\Sigma x + c\Sigma x^2 &= \Sigma y \\ a\Sigma x + b\Sigma x^2 + c\Sigma x^3 &= \Sigma xy \\ a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 &= \Sigma x^2y \end{aligned}$$

Now

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	6	1	1	1	6	6
2	17	4	8	16	34	68
Σx $= 3$	Σy $= 24$	Σx^2 $= 5$	Σx^3 $= 9$	Σx^4 $= 17$	Σxy $= 40$	Σx^2y $= 74$

Here $m = 3, \Sigma x = 3, \Sigma y = 24, \Sigma x^2 = 5, \Sigma x^3 = 9, \Sigma x^4 = 17, \Sigma xy = 40, \Sigma x^2y = 74$.

Put these values in normal equations, we have

$$\begin{aligned} 3a + 3b + 5c &= 24 \\ 3a + 5b + 9c &= 40 \\ 5a + 9b + 17c &= 74 \end{aligned}$$

On solving (i), (ii) and (iii), we get $a = 1, b = 2, c = 3$

Thus, the equation of required parabola is $y = 1 + 2x + 3x^2$.

Example 6. Find the values of a, b and c so that $y = a + bx + cx^2$ is the best fit to the data.

x	0	1	2	3	4
y	1	0	3	10	21

Solution. Let $y = a + bx + cx^2$ be a parabola to be fitted for the given data and normal equations are

$$\begin{aligned} ma + b\Sigma x + c\Sigma x^2 &= \Sigma y \\ a\Sigma x + b\Sigma x^2 + c\Sigma x^3 &= \Sigma xy \\ a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 &= \Sigma x^2y \end{aligned}$$

Now

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	0	1	1	1	0	0
2	3	4	8	16	6	12
3	10	9	27	81	30	90
4	21	16	64	256	84	336
Σx = 10	Σy = 35	Σx^2 = 30	Σx^3 = 100	Σx^4 = 354	Σxy = 120	Σx^2y = 438

Here $m = 5, \Sigma x = 10, \Sigma y = 35, \Sigma x^2 = 30, \Sigma x^3 = 100, \Sigma x^4 = 354, \Sigma xy = 120$ and $\Sigma x^2y = 438$

Put these values in normal equations, we have

$$5a + 10b + 30c = 35 \quad (i)$$

$$10a + 30b + 100c = 120 \quad (ii)$$

$$30a + 100b + 354c = 438 \quad (iii)$$

On solving (i), (ii) and (iii), we get $a = 1, b = -3, c = 2$ and equation of parabola

$$y = 1 - 3x + 2x^2.$$

Example 7. Fit a second-degree parabola to the following data x as the independent variable.

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Solution. Let equation of second-degree parabola be $y = a + bx + cx^2$ and the normal equations are

$$\begin{aligned} ma + b\Sigma x + c\Sigma x^2 &= \Sigma y \\ a\Sigma x + b\Sigma x^2 + c\Sigma x^3 &= \Sigma xy \\ a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 &= \Sigma x^2y \end{aligned}$$

Now

x	y	x^2	x^3	x^4	xy	x^2y
1	2	1	1	1	2	2
2	6	4	8	16	12	24
3	7	9	27	81	21	63
4	8	16	64	256	32	128
5	10	25	125	625	50	250
6	11	36	216	1296	66	396
7	11	49	343	2401	77	539
8	10	64	512	4096	80	640
9	9	81	729	6561	81	729
Σx = 45	Σy = 74	Σx^2 = 285	Σx^3 = 2025	Σx^4 = 15333	Σxy = 421	Σx^2y = 2771

Here $m = 9$, $\Sigma x = 45$, $\Sigma y = 79$, $\Sigma x^2 = 285$, $\Sigma x^3 = 2025$, $\Sigma x^4 = 15333$, $\Sigma xy = 421$ and $\Sigma x^2y = 2771$ Put these values in normal equations, we have

$$9a + 45b + 285c = 74 \quad (i)$$

$$45a + 285b + 2025c = 421 \quad (\text{ii})$$

$$285a + 2025b + 15333c = 438 \quad (\text{iii})$$

On solving (i), (ii) and (iii), we get $a = -0.923$, $b = 3.52$, $c = -0.267$

Thus, the equation of parabola is $y = -0.923 + 3.52x - 0.267x^2$.

CHECK YOUR PROGRESS

1. What is curve fitting? Explain its importance in statistical analysis.

.....

.....

.....

2. Explain the principle of the method of least squares.

.....

.....

.....

3. Describe the procedure for fitting a straight-line trend using the least squares method.

.....

.....

.....

16.6 SUMMARY

In this unit, we have discussed the following concepts

- Curve fitting
- Method of least squares

16.7 EXERCISES

- 1 Find a least squares straight line for the following data:

x	1	2	3	4	5	6
y	6	4	3	5	4	2

- 2 Find a least squares straight line for the following data:

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

- 3 Show that the line of best fit to the following data is given by

x	6	7	7	8	8	8	9	9	10
y	5	5	4	5	4	3	4	3	3

- 4 Find the best values of a and b so that $y = a + bx$ fits the data given in the table.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

- 5 Obtain the least squares straight line fit to the following data:

x	0.2	0.4	0.6	0.8	1
$f(x)$	0.447	0.632	0.775	0.894	1

- 6 Fit a linear curve to the data $\{(x, y): (1,1)(2,5)(3,11)(4,8)(5,14)\}$.

- 7 Fit a second-degree parabola to the following data taking x as independent and y as dependent variable.

x	1	2	3	4	5	6	7	8	9
-----	---	---	---	---	---	---	---	---	---

y	2	6	7	8	10	11	11	10	9
-----	---	---	---	---	----	----	----	----	---

- 8 Fit a parabolic curve to the following data x as the independent variable.

x	1	2	3	4	5
y	6	17	34	57	86

- 9 Find the least squares approximation of second degree for the discrete data.

x	-2	-1	0	1	2
$f(x)$	15	1	1	3	19

- 10 Fit a parabola $y = a + bx + cx^2$ in least square sense to the data.

x	10	12	15	23	20
y	14	17	23	25	21



16.8 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

BLOCK V: HYPOTHESIS TESTING AND ANALYSIS OF VARIANCE

UNIT 17 SAMPLING AND TESTING OF HYPOTHESIS

STRUCTURE

- 17.1 Introduction
- 17.2 Objectives
- 17.3 Population or universe
- 17.4 Sampling
- 17.5 Parameters of statistics and standard error
- 17.6 Testing of significance proportions
- 17.7 Testing of significance for single mean
- 17.8 Test of significance for difference of means of two
large samples
- 17.9 Test of significance for the difference of standard
deviations
- 17.10 Summary
- 17.11 Exercises
- 17.12 References

17.1 INTRODUCTION

This unit on "Testing of Hypothesis" presents statistical techniques for drawing conclusions or judgments about populations from sample data. A key idea in statistics, hypothesis testing allows researchers to compare groups, verify hypotheses, and assess the possibility that observed differences were the result of chance. Important subjects like population comprehension, sample techniques, parameters, and errors are covered in this section. In order to examine proportions, means, and standard deviations, other tests of significance are also investigated. These tests

are frequently used in business, research, and daily decision-making, and they serve as the foundation for robust statistical reasoning.

17.2 OBJECTIVES

After completing this unit, students will be able to

- Understand the concepts of population (universe) and sampling.
- Explain parameters, statistics, and the concept of standard error.
- Perform hypothesis tests for proportions.
- Conduct significance tests for single means.
- Apply significance tests to compare means of two large samples.
- Analyze differences in standard deviations using significance tests.
- Draw meaningful conclusions from hypothesis tests in practical scenarios.

17.3 POPULATION OR UNIVERSE

In data analysis and statistics, the term population or universe refers to the complete set of objects, whether animate or inanimate, that are under study. This collection can include individuals, their characteristics, or the outcomes of specific measurable operations. A universe can be classified based on its size and nature.

A finite universe consists of a limited number of members. For example, the set of weights of students in a particular class forms a finite universe. On the other hand, an **infinite universe** includes an unlimited number of elements, such as the pressures recorded at various points in the atmosphere. In some cases, the nature of the universe might be unknown, such as the universe of stars, where it is unclear whether the set is finite or infinite.

Universes can also be categorized based on their existence. An **existent universe** includes tangible, real-world objects, while a **hypothetical universe** consists of all possible outcomes of a specific event. For instance, the collection of heads and tails obtained from tossing a coin an infinite number of times (assuming the coin does not wear out) represents a hypothetical universe. Understanding these distinctions helps provide clarity when defining the scope of a study and analyzing data effectively.

17.4 SAMPLING

Statisticians often face the challenge of analyzing a universe where examining every member is impractical. For instance, determining the average income of a country's population would be an overwhelming task if it required surveying every individual. This leads to the concept of **sampling**, where insights about the whole universe are derived by studying a smaller, manageable subset.

A **sample** is a finite subset of a universe, and its size is referred to as the **sample size**. The process of selecting a sample is called **sampling**, and the **theory of sampling** explores the relationship between a population and its samples. The primary goal of sampling is to extract maximum information about the entire universe by examining only a small portion, thereby saving effort and resources.

Sampling is common in everyday life. For example, we test the quality of sugar or rice by examining a handful from a bag before purchasing. Similarly, a housewife checks the taste of a dish by sampling a spoonful to ensure it is properly cooked. This practical approach underscores the efficiency and utility of sampling in both statistical studies and daily activities.

17.5 PARAMETERS OF STATISTICS AND STANDARD ERROR

The statistical constants of the population such as mean, the variance, etc. are known as the parameters. The statistical concepts of the sample from the members of the sample to estimate the parameters of the population from which the sample has been drawn is known as statistic. Population mean and variance are denoted by μ and σ^2 , while those of the samples are given by $\bar{x}, s^2$.

The standard deviation of the sampling distribution of a statistic is known as the **standard error (S.E.)**. It plays an important role in the theory of large samples and it forms a basis of the testing of hypothesis. If t is any statistic, for large sample.

$z = \frac{t - E(t)}{S.E(t)}$ is normally distributed with mean 0 and variance unity.

For large sample, the standard errors of some of the well-known statistics are listed below:

n -sample size; σ^2 —population variance; s^2 -sample variance; p -population proportion; $Q = 1 - p$; n_1, n_2 -are sizes of two independent random samples.

Number	Statistic	Standard error
1.	$\bar{x}$	$\sigma/\sqrt{n}$
2.	s	$\sqrt{\sigma^2/2n}$
3.	Difference of two sample means $\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
4.	Difference of two sample standard deviation $s_1 - s_2$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
5.	Difference of two sample proportions $p_1 - p_2$	$\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$

6.	Observed sample proportion p	$\sqrt{PQ/n}$
----	--------------------------------	---------------

17.5.1 TEST OF SIGNIFICANCE

An essential part of sampling theory is the study of tests of significance. Such tests provide a basis for deciding, from the results obtained from a sample, whether an observed difference is statistically meaningful or whether it has occurred merely because of chance or sampling fluctuations.

A test of significance is generally used to determine whether:

- (i) the difference between the observed sample statistic and a specified hypothetical value of the population parameter is significant; or
- (ii) the difference between two sample statistics is statistically significant or can reasonably be explained by random variation arising from sampling.

Before conducting a test of significance, a statistical hypothesis is formulated. A hypothesis is a definite statement or assumption concerning a population parameter. The hypothesis initially formulated for testing is known as the Null Hypothesis, represented by

$$H_0.$$

The null hypothesis generally assumes that there is no significant difference or effect in the population.

The hypothesis that is contrary to or complementary to the null hypothesis is known as the **Alternative Hypothesis**, represented by

Alternative hypothesis will be

- (i) $H_1: \mu \neq \mu_0 (\mu > \mu_0 \text{ or } \mu < \mu_0)$ (two tailed alternative hypothesis).
- (ii) $H_1: \mu > \mu_0$ (right tailed alternative hypothesis (or) single tailed).
- (iii) $H_1: \mu < \mu_0$ (left tailed alternative hypothesis (or) single tailed).

Hence alternative hypothesis helps to know whether the test is two tailed test or one tailed test.

17.5.2 CRITICAL REGION

A region corresponding to a statistic t , in the sample space S which amounts to rejection of the null hypothesis H_0 is called as critical region or region of rejection. The region of the sample space S which amounts to the acceptance of H_0 is called acceptance region.

17.5.3 LEVEL OF SIGNIFICANCE

The probability of the value of the variate falling in the critical region is known as level of significance. The probability α that a random value of the statistic t belongs to the critical region is known as the level of significance.

$$P(t \in \omega \mid H_0) = \alpha$$

i.e., the level of significance is the size of the type I error or the maximum producer's risk.

17.5.4 ERRORS IN SAMPLING

The main aim of the sampling theory is to draw a valid conclusion about the population parameters on the basis of the sample results. In doing this we may commit the following two types of errors:

Type I Error. When H_0 is true, we may reject it.

$$P(\text{Reject } H_0 \text{ when it is true}) = P(\text{Reject } H_0 / H_0) = \alpha$$

α is called the size of the type I error also referred to as producer's risk.

Type II Error. When H_0 is wrong, we may accept it

$$P(\text{Accept } H_0 \text{ when it is wrong}) = P(\text{Accept } \frac{H_0}{H_1}) = \beta.$$

β is called the size of the type II error, also referred to as consumer's risk.

The values of the test statistic which separates the critical region and acceptance region are called the critical values or significant values. This value is dependent on (i) the level of significance used and (ii) the alternative hypothesis, whether it is one-tailed or two-tailed.

For larger samples corresponding to the statistic t , the variable $z = \frac{t - E(t)}{S.E(t)}$ is normally distributed with mean 0 and variance 1. The value of z given above under the null hypothesis is known as test statistic.

The critical value of z_α of the test statistic at level of significance α for a two-tailed test is given by

$$p(|z| > z_\alpha) = \alpha$$

i.e., z_α is the value of z so that the total area of the critical region on both tails is α . Since the normal curve is symmetrical, from equation (1), we get

$$p(z > z_\alpha) + p(z < -z_\alpha) = \alpha; \text{ i.e., } 2p(z > z_\alpha) = \alpha; p(z > z_\alpha) = \alpha/2$$

i.e., the area of each tail is $\alpha/2$.

The critical value z_α is that value such that the area to the right of z_α is $\alpha/2$ and the area to the left of $-z_\alpha$ is $\alpha/2$.

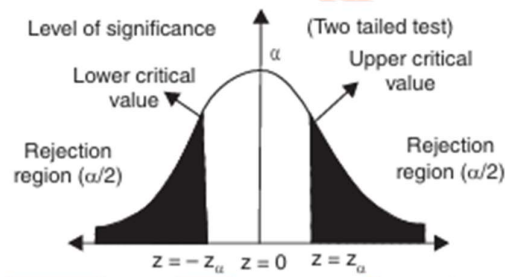


Figure 17.1 Two tailed test

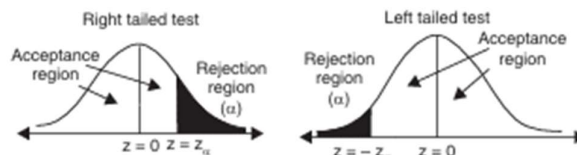


Figure 17.2 Right and left tailed test

In the case of the one-tailed test,

$p(z > z_\alpha) = \alpha$ if it is right-tailed; $p(z < -z_\alpha) = \alpha$ if it is left-tailed.

The critical value of z for a single-tailed test (right or left) at level of significance α is same as the critical value of z for two-tailed test at level of significance 2α .

Using the equation, also using the normal tables, the critical value of z at different levels of significance (α) for both single tailed and two tailed test are calculated and listed below. The equations are

$$p(|z| > z_\alpha) = \alpha; p(z > z_\alpha) = \alpha; p(z < -z_\alpha) = \alpha$$

Table 17.1 Level of significance			
	1%(0.01)	5%(0.05)	10%(0.1)
Two tailed test	$ z_\alpha = 2.58$	$ z = 1.966$	$ z = 0.645$
Right tailed	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left tailed	$z_\alpha = -2.33$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

17.5.5 STEPS IN TESTING OF STATISTICAL HYPOTHESIS

Step 1. Null hypothesis. Set up H_0 in clear terms.

Step 2. Alternative hypothesis. Set up H_1 , so that we could decide whether we should use one tailed test or two tailed test.

Step 3. Level of significance. Select the appropriate level of significance in advance depending on the reliability of the estimates.

Step 4. Test statistic. Compute the test statistic $z = \frac{t - E(t)}{S.E(t)}$ under the null hypothesis.

Step 5. Conclusion. Compare the computed value of z with the critical value z_α at level of significance (α).

If $|z| > z_\alpha$, we reject H_0 and conclude that there is significant difference. If $|z| < z_\alpha$, we accept H_0 and conclude that there is no significant difference.

8.10 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

If the sample size $n > 30$, the sample is taken as large sample. For such sample we apply normal test, as Binomial, Poisson, chi square, etc. are closely approximated by normal distributions assuming the population as normal.

Under large sample test, the following are the important tests to test the significance:

- 1 Testing of significance for single proportion.
- 2 Testing of significance for difference of proportions.
- 3 Testing of significance for single mean.
- 4 Testing of significance for difference of means.
- 5 Testing of significance for difference of standard deviations.

17.5.6 TESTING OF SIGNIFICANCE FOR SINGLE PROPORTION

This test is used to find the significant difference between proportion of the sample and the population. Let X be the number of successes in n independent trials with constant probability P of success for each trial.

$$E(X) = nP; V(X) = nPQ; Q = 1 - P = \text{Probability of failure.}$$

Let $p = X/n$ called the observed proportion of success.

$$E(p) = E(X/n) = \frac{1}{n} E(X) = \frac{np}{n} = p; E(p) = p$$

$$V(p) = V(X/n) = \frac{1}{n^2} V(X) = \frac{1(PQ)}{n} = PQ/n$$

$$\text{S. E. } (p) = \sqrt{\frac{PQ}{n}}; z = \frac{p - E(p)}{\text{S. E. } (p)} = \frac{p - p}{\sqrt{PQ/n}} \sim N(0,1)$$

This z is called test statistic which is used to test the significant difference of sample and population proportion.

NOTE

- 1 The probable limit for the observed proportion of successes is $\pm z_{\alpha} \sqrt{PQ/n}$, where z_{α} is the significant value at level of significance α .
- 2 If p is not known, the limits for the proportion in the population are $p \pm z_{\alpha} \sqrt{pq/n}$, $q = 1 - p$.
- 3 If α is not given, we can take safely 3σ limits.

Hence, the confidence limits for observed proportion p are $p \pm 3 \sqrt{\frac{PQ}{n}}$.

The confidence limits for the population proportion p are $p \pm \sqrt{\frac{pq}{n}}$.

Example 1. A coin is tossed 400 times, resulting in 216 heads. Test whether the coin is fair, i.e., whether the probability of heads is equal to that of tails.

Solution. H_0 : The coin is fair i.e., $P = 0.5$.

H_1 : The coin is not fair (biased); $P \neq 0.5$

Here $n = 400$; X = Number of successes = 216

p = proportion of success in the sample $\frac{X}{n} = \frac{216}{400} = 0.54$.

Population proportion = $0.5 = P$; $Q = 1 - P = 1 - 0.5 = 0.5$.

Under H_0 , test statistic $z = \frac{p-P}{\sqrt{PQ/n}}$

$$|z| = \frac{|0.54 - 0.5|}{\sqrt{\frac{0.5 \times 0.5}{400}}} = 1.6$$

we use a two-tailed test.

Conclusion. Since $|z| = 1.6 < 1.96$

i.e., $|z| < z_{\alpha}$, z_{α} is the significant value of z at 5% level of significance. i.e., the coin is unbiased is $P = 0.5$.

Example 2. A manufacturer claims that only 4% of his products supplied by him are defective. A random sample of 600 products contained 36 defectives. Test the claim of the manufacturer.

Solution. (i) P = observed proportion of success.

i.e., P = proportion of defective in the sample = $\frac{36}{600} = 0.06$

p = proportion of defectives in the population = 0.04

H_0 : $p = 0.04$ is true.

i.e., the claim of the manufacturer is accepted.

H_1 : (i) $P \neq 0.04$ (two tailed test)

(ii) If we want to reject, only if $p > 0.04$ then (right tailed).

$$\text{Under } H_0, z = \frac{p-P}{\sqrt{PQ/n}} = \frac{0.06-0.04}{\sqrt{\frac{0.04 \times 0.96}{600}}} = 2.5.$$

Conclusion. Since $|z| = 2.5 > 1.96$, we reject the hypothesis H_0 at 5% level of significance two tailed.

If H_1 is taken as $p > 0.04$ we apply right tailed test.

$|z| = 2.5 > 1.645(z_\alpha)$ we reject the null hypothesis here also.

In both cases, manufacturer's claim is not acceptable.

Example 3. A machine is producing bolts a certain fraction of which are defective. A random sample of 400 is taken from a large batch and is found to contain 30 defective bolts. Does this indicate that the proportion of defectives is larger than that claimed by the manufacturer if the manufacturer claims that only 5% of his product are defective? Find 95% confidence limits of the proportion of defective bolts in batch.

Solution. Null hypothesis H_0 : The manufacturer claim is accepted i.e.,

$$P = \frac{5}{100} = 0.05$$

$$Q = 1 - P = 1 - 0.05 = 0.95$$

Alternative hypothesis. $p > 0.05$ (Right tailed test).

$$p = \text{observed proportion of sample} = \frac{30}{400} = 0.075$$

Under H_0 , the test statistic

$$z = \frac{p - P}{\sqrt{PQ/n}} \therefore z = \frac{0.075 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{400}}} = 2.2941$$

Conclusion. The tabulated value of z at 5% level of significance for the right-tailed test is

$$z_\alpha = 1.645. \text{ Since } |z| = 2.2941 > 1.645,$$

H_0 is rejected at 5% level of significance. i.e., the proportion of defective is larger than the manufacturer claim.

To find 95% confidence limits of the proportion.

It is given by $p \pm z_{\alpha}\sqrt{PQ/n}$

$$0.05 \pm 1.96 \sqrt{\frac{0.05 \times 0.95}{400}} = 0.05 \pm 0.02135 = 0.07136, 0.02865$$

Hence 95% confidence limits for the proportion of defective bolts are (0.07136, 0.02865).

Example 4. A bag contains defective articles, the exact number of which is not known. A sample of 100 from the bag gives 10 defective articles. Find the limits for the proportion of defective articles in the bag.

Solution. Here p = proportion of defective articles = $\frac{10}{100} = 0.1$;

$$q = 1 - p = 1 - 0.1 = 0.9$$

Since the confidence limit is not given, we assume it is 95%. $\therefore$ level of significance is $5\% z_{\alpha} = 1.96$.

Also, the proportion of population P is not given. To get the confidence limit, we use P and it is given by

$$\begin{aligned} P \pm z_{\alpha}\sqrt{pq/n} &= 0.1 \pm 1.96 \sqrt{\frac{0.1 \times 0.9}{100}} \\ &= 0.1 \pm 0.0588 = 0.1588, 0.0412. \end{aligned}$$

Hence 95% confidence limits for defective articles in the bag are (0.1588, 0.0412).

17.6 TESTING OF SIGNIFICANCE PROPORTIONS

Consider two samples X_1 and X_2 of sizes n_1 and n_2 , respectively, taken from two different populations. Test the significance of the difference between the sample proportion p_1 and p_2 . The test statistic under the null hypothesis H_0 , that there is no significant difference between the two-sample proportion, we have

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } Q = 1 - P.$$

Example 17.6.1. Before an increase in excise duty on tea, 800 people out of a sample of 1000 persons were found to be tea drinkers. After an increase in the duty, 800 persons were known to be tea drinkers in a sample of 1200 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty?

Solution. Here

$$\begin{aligned} n_1 &= 800, n_2 = 1200 \\ p_1 &= \frac{X_1}{n_1} = \frac{800}{1000} = \frac{4}{5}; p_2 = \frac{X_2}{n_2} = \frac{800}{1200} = \frac{2}{3} \\ P &= \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{8}{11}; Q = \frac{3}{11} \end{aligned}$$

Null hypothesis $H_0: p_1 = p_2$, i.e., there is no significant difference in the consumption of tea before and after increase of excise duty.

$$H_1: p_1 > p_2 \text{ (right-tailed test)}$$

The test statistic

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8 - 0.6666}{\sqrt{\frac{8}{11} \times \frac{3}{11} \left(\frac{1}{1000} + \frac{1}{1200}\right)}} = 6.842.$$

Conclusion. Since the calculated value of $|z| > 1.645$ also $|z| > 2.33$, both the significant values of z at 5% and 1% level of significance. Hence H_0 is rejected, i.e., there is a significant decrease in the consumption of tea due to increase in excise duty.

Example 17.6.2. A machine produced 16 defective articles in a batch of 500. After overhauling it produced 3 defectives in a batch of 100. Has the machine improved?

Solution.

$$p_1 = \frac{16}{500} = 0.032; n_1 = 500 \quad p_2 = \frac{3}{100} = 0.03; n_2 = 100$$

Null hypothesis H_0 : The machine has not improved due to overhauling, $p_1 = p_2$.

$$H_1: p_1 > p_2 \text{ (right-tailed)}$$

$$\therefore P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{19}{600} \cong 0.032$$

Under H_0 , the test statistic

$$z = \frac{p_1 - p_2}{\sqrt{PQ \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.032 - 0.03}{\sqrt{(0.032)(0.968) \left(\frac{1}{500} + \frac{1}{100} \right)}} = 0.104.$$

Conclusion. The calculated value of $|z| < 1.645$, the significant value of z at 5% level of significance, H_0 is accepted, i.e., the machine has not improved due to overhauling.

Example 17.6.3. In two large populations, there are 30% and 25%, respectively, of fair-haired people. Is this difference likely to be hidden in samples of 1200 and 900, respectively, from the two populations.

Solution. p_1 = proportion of fair-haired people in the first population = 30% = 0.3; p_2 = 25% = 0.25; Q_1 = 0.7; Q_2 = 0.75.
 H_0 : Sample proportions are equal, i.e., the difference in population proportions is likely to be hidden in sampling.

$$H_1 : p_1 \neq p_2$$

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} = \frac{0.3 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}}} = 2.5376$$

Conclusion. Since $|z| > 1.96$, the significant value of z at 5% level of significance, H_0 is rejected. However, $|z| < 2.58$, the significant value of z at 1% level of significance, H_0 is accepted. At 5% level, these samples will reveal the difference in the population proportions.

17.7 TESTING OF SIGNIFICANCE FOR SINGLE MEAN

The purpose of this test is to determine whether the difference between a **sample mean** and a specified **population mean** is statistically significant or whether it can reasonably be attributed to sampling fluctuations.

Let

$$x_1, x_2, \dots, x_n$$

be a random sample of size n drawn from a large population of size N , having population mean μ and variance σ^2 . The standard error of the sample mean is

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Suppose we want to examine whether a random sample of size n could have been obtained from a population whose mean is μ_0 . In other

words, we test whether the observed difference between the sample mean $\bar{X}$ and the specified population mean μ_0 is significant.

Null and Alternative Hypotheses

The **null hypothesis** is

$$H_0: \mu = \mu_0,$$

which indicates that there is no statistically significant difference between the sample mean and the specified population mean.

The **alternative hypothesis** is

$$H_1: \mu \neq \mu_0,$$

when a two-tailed test is appropriate.

Test Statistic When Population Standard Deviation is Known

If the population standard deviation σ is known, the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

The calculated value of Z is compared with the appropriate critical value at the selected level of significance.

Test Statistic When Population Standard Deviation is Unknown

If the population standard deviation is not known and the sample standard deviation s is used, the test statistic is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

with $n - 1$ degrees of freedom.

Confidence Limits

If the level of significance is α , the critical value determines the confidence limits for the population mean.

For a large sample, the confidence interval is generally expressed as

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

At the **5% level of significance**, the corresponding 95% confidence limits are

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

At the **1% level of significance**, the corresponding 99% confidence limits are

$$\bar{X} \pm 2.576 \frac{\sigma}{\sqrt{n}}.$$

These limits are commonly referred to as confidence limits or fiducial limits.

Example 1

A normal population has a mean of 6.8 and a standard deviation of 1.5. A random sample of 400 observations has a mean of 6.75. Determine whether the difference between the sample mean and the population mean is statistically significant.

Solution:

Given:

$$\mu_0 = 6.8, \sigma = 1.5, n = 400, \bar{X} = 6.75.$$

The hypotheses are:

$$H_0: \mu = 6.8$$

There is no significant difference between the sample mean and the specified population mean.

$$H_1: \mu \neq 6.8$$

There is a significant difference between the two means.

The test statistic is

$$\begin{aligned} Z &= \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{6.75 - 6.8}{1.5/\sqrt{400}} \\ &= \frac{-0.05}{1.5/20} = \frac{-0.05}{0.075} = -0.667. \end{aligned}$$

At the 5% level of significance, the critical value for a two-tailed test is

$$Z_{0.025} = 1.96.$$

Since

$$|Z| = 0.667 < 1.96,$$

the null hypothesis is accepted.

Conclusion: There is no statistically significant difference between the sample mean of 6.75 and the population mean of 6.8 at the 5% level of significance.

Example 2

A random sample of 900 observations has a mean of 3.4 cm. Can it reasonably be regarded as a sample from a large population having a mean of 3.2 cm and a standard deviation of 2.3 cm?

Solution:

Given:

$$n = 900, \bar{X} = 3.4, \mu_0 = 3.2, \sigma = 2.3.$$

We formulate:

$$H_0: \mu = 3.2$$

against

$$H_1: \mu \neq 3.2.$$

Using the two-tailed test,

$$\begin{aligned} Z &= \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{3.4 - 3.2}{2.3/\sqrt{900}} \\ &= \frac{0.2}{2.3/30} = \frac{0.2}{0.07667} \approx 2.61. \end{aligned}$$

At the 5% level of significance, the critical value is

$$Z = 1.96.$$

Since

$$2.61 > 1.96,$$

the null hypothesis is rejected.

Conclusion: The difference is statistically significant at the 5% level. Therefore, the sample cannot reasonably be regarded as having been drawn from a population with mean 3.2 cm.

Example 3

The mean weight obtained from a random sample of 100 observations is 64 g. The population standard deviation is 3 g. Test the hypothesis that the population mean weight is 67 g at the 5% level of significance. Also determine the 95% confidence limits for the population mean.

Solution:

Given:

$$n = 100, \bar{X} = 64 \text{ g}, \sigma = 3 \text{ g}, \mu_0 = 67 \text{ g}.$$

The hypotheses are:

$$H_0: \mu = 67$$

and

$$H_1: \mu \neq 67.$$

The test statistic is

$$\begin{aligned} Z &= \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{64 - 67}{3/\sqrt{100}} \\ &= \frac{-3}{0.3} = -10. \end{aligned}$$

At the 5% level of significance, the critical value for a two-tailed test is

$$\pm 1.96.$$

Since

$$|-10| > 1.96,$$

the null hypothesis is rejected.

Conclusion: There is a statistically significant difference between the sample mean and the hypothesised population mean of 67 g. Hence, the hypothesis that the population mean is 67 g is not supported by the sample evidence.

95% Confidence Limits

The confidence limits are

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

Therefore,

$$\begin{aligned} & 64 \pm 1.96 \left(\frac{3}{10} \right) \\ & = 64 \pm 0.588. \end{aligned}$$

Thus, the 95% confidence limits are

$$63.412 \text{ g and } 64.588 \text{ g.}$$

Therefore, the population mean is estimated to lie between **63.412 g and 64.588 g** with 95% confidence.

$$H_1: \mu \neq 67 \text{ (Two-tailed test).}$$

Under $H_0, z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{64 - 67}{3 / \sqrt{100}} = -10 \therefore |z| = 10.$

Conclusion. Since the calculated value of $|z| > 1.96$, the significant value of z at 5% level of significance, H_0 is rejected, i.e., the sample is not drawn from the population with mean 67.

To find 99% confidence limits, given by

$$\bar{x} \pm 2.58\sigma / \sqrt{n} = 64 \pm 2.58(3 / \sqrt{100}) = 64.774, 63.226$$

Example 4. The average score in mathematics of a sample of 100 students was 51 with a standard deviation of 6 points. Could this have been a random sample from a population with average scores 50?

Solution. Here $n = 100, \bar{x} = 51, s = 6, \mu = 50; \sigma$ is unknown.

H_0 : The sample is drawn from a population with mean 50, $\mu = 50$

$$H_1: \mu \neq 50$$

Under

$$H_0, z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{51 - 50}{6/\sqrt{100}} = \frac{10}{6} = 1.6666.$$

Conclusion. Since $|z| = 1.666 < 1.96, z_\alpha$ the significant value of z at 5% level of significance, H_0 is accepted, i.e., the sample is drawn from the population with mean 50.

17.8 TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS OF TWO LARGE SAMPLES

Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 , and variance σ_1^2 . Let $\bar{x}_2$ be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2 . The test statistic is given by $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

Under the null hypothesis that the samples are drawn from the same population where $\sigma_1 = \sigma_2 = \sigma$, i.e., $\mu_1 = \mu_2$ the test statistic is given by $z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

Note 1. If σ_1, σ_2 are not known and $\sigma_1 \neq \sigma_2$ the test statistic in this case is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- 2 If σ is not known and $\sigma_1 = \sigma_2$. We use $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$ to calculate σ ;

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Example 1. The average income of persons was 210 with a standard deviation of 10 in a sample of 100 people. For another sample of 150 people, the average income was 220 with a standard deviation of 12. The standard deviation of incomes of the people of the city was 11. Test whether there is any significant difference between the average incomes of the localities.

Solution. Here $n_1 = 100, n_2 = 150, \bar{x}_1 = 210, \bar{x}_2 = 220, s_1 = 10, s_2 = 12$.

Null hypothesis. The difference is not significant, i.e., there is no difference between the incomes of the localities.

$$\begin{aligned} H_0: \bar{x}_1 &= \bar{x}_2, H_1: \bar{x}_1 \neq \bar{x}_2 \\ \text{Under } H_0, z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{210 - 220}{\sqrt{\frac{10^2}{100} + \frac{12^2}{150}}} = -7.1428 \therefore |z| = 7.1428 \end{aligned}$$

Conclusion. As the calculated value of $|z| > 1.96$, the significant value of z at 5% level of significance, H_0 is rejected i.e., there is significant difference between the average incomes of the localities.

Example 2. Intelligence tests were given to two groups of boys and girls.

	Mean	Standard deviation	Size
Girls	75	8	60
Boys	73	10	100

Examine if the difference between mean scores is significant.

Solution. Null hypothesis H_0 : There is no significant difference between mean scores, i.e., $\bar{x}_1 = \bar{x}_2$.

$$H_1: \bar{x}_1 \neq \bar{x}_2$$

$$\text{Under the null hypothesis } z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{75 - 73}{\sqrt{\frac{8^2}{60} + \frac{10^2}{100}}} = 1.3912.$$

Conclusion. As the calculated value of $|z| < 1.96$, the significant value of z at 5% level of significance, H_0 is accepted i.e., there is no significant difference between mean scores.

17.9 TEST OF SIGNIFICANCE FOR THE DIFFERENCE OF STANDARD DEVIATIONS

If s_1 and s_2 are the standard deviations of two independent samples, then under the null hypothesis $H_0: \sigma_1 = \sigma_2$, i.e., the sample standard deviations don't differ significantly, the statistic

$$z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}, \text{ where } \sigma_1 \text{ and } \sigma_2 \text{ are population standard deviations.}$$

When population standard deviations are not known, then $z =$

$$\frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}.$$

Example: Random samples drawn from two countries gave the following data relating to the heights of adult males:

	Country A	Country B
Mean height (in inches)	67.42	67.25
Standard deviation	2.58	2.50
Number in samples	1000	1200

- (i) Is the difference between the means significant?
(ii) Is the difference between the standard deviations significant?

Solution. Given: $n_1 = 1000, n_2 = 1200, \bar{x}_1 = 67.42; \bar{x}_2 = 67.25, s_1 = 2.58, s_2 = 2.50$

Since the samples size are large we can take $\sigma_1 = s_1 = 2.58; \sigma_2 = s_2 = 2.50$.

(i) Null hypothesis: $H_0: \mu_1 = \mu_2$ i.e., sample means do not differ significantly.

Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$ (two tailed test)

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67.42 - 67.25}{\sqrt{\frac{(2.58)^2}{1000} + \frac{(2.50)^2}{1200}}} = 1.56.$$

Since $|z| < 1.96$ we accept the null hypothesis at 5% level of significance. (ii) We set up the null hypothesis.

$H_0: \sigma_1 = \sigma_2$ i.e., the sample standard deviations do not differ significantly. Alternative hypothesis: $H_1: \sigma_1 \neq \sigma_2$ (two tailed)

$\therefore$ The test statistic is given by

$$z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}$$

($\because \sigma_1 = s_1, \sigma_2 = s_2$ for large samples)

$$\sqrt{\frac{(2.58)^2}{2 \times 1000} + \frac{(2.50)^2}{2 \times 1200}} = \frac{0.08}{\sqrt{\frac{6.6564}{2000} + \frac{6.25}{2400}}} = 1.0387.$$

Since $|z| < 1.96$ we accept the null hypothesis at 5% level of significance.

CHECK YOUR PROGRESS

1. Define sampling and explain its importance in statistical investigation.

.....

.....

.....

2. Distinguish between census method and sampling method.

.....

.....

.....

3. What is a statistical hypothesis? Explain the concepts of null and alternative hypotheses.

.....

.....

.....

17.10 SUMMARY

In this unit, we explored the essential aspects of hypothesis testing. We began by understanding the concepts of populations and sampling, followed by learning about parameters, statistics, and standard error. We then delved into various tests of significance: for proportions, single means, differences in means of two large samples, and differences in standard deviations. Mastering these concepts allows for robust statistical analysis and decision-making, forming a foundation for research and data-driven conclusions.

17.11 EXERCISES

1. Vegetarian and Non-Vegetarian Population
2. A random sample of 1,000 people was selected from a city. Among them, 540 were found to be vegetarian and the remaining 460 were non-vegetarian. Test whether vegetarian and non-vegetarian food habits are equally prevalent in the city at:
 - (i) [level of significance missing]
 - (ii) [level of significance missing]
3. Proportion of Smokers
4. Out of a random sample of 600 men selected from a large city, 325 were found to be smokers. Based on this sample evidence, examine whether it can be concluded that more than half of the men in the city are smokers.
5. Defective Bolts
6. A random sample of 500 bolts was selected from a large consignment, of which 65 were found to be defective. Estimate the percentage of defective bolts in the entire consignment.
7. Male and Female Births
8. During one week, a hospital recorded 475 female births and 525 male births. Test whether these observations are consistent with the hypothesis that male and female children are born in equal proportions.
9. Bad Apples
10. A random sample of 400 apples was selected from a large basket, and 40 of them were found to be defective. Estimate the proportion of bad apples in the basket and determine the limits within which the population percentage is expected to lie.
11. Comparison of Male Birth Rates
12. In Town A, there were 956 births, of which 52.5% were male. In Towns A and B together, there were 1,406 births, and the

proportion of male births was 0.496. Determine whether there is a statistically significant difference between the proportions of male births in the two towns.

13. Opinion of Male and Female Students

14. In a university referendum, 850 male students and 560 female students participated in voting. Among them, 500 men and 320 women voted in favour of the proposal. Test whether there is a significant difference in the opinions of male and female students at the 1% level of significance.

15. Confidence Interval for Population Mean

16. A random sample of 200 observations drawn from a large population has a mean of 50 and a standard deviation of 9. Determine the [confidence level missing] confidence interval for the population mean.

17. Sample Size for Quality Control

18. A particular type of electric bulb has a guaranteed average life of 1,000 hours, with a standard deviation of 125 hours. A sampling plan is to be designed so that [specified proportion missing] of the bulbs do not fall below the guaranteed average life by more than [specified amount missing]. Determine the minimum sample size required.

19. Mean Height of College Students

20. The heights of college students in a city are normally distributed with a standard deviation of 6 cm. A random sample of 1,000 students has an average height of 158 cm. Test the hypothesis that the population mean height is 160 cm.

21. Comparison of Wheat Yields

22. Two random samples of 1,000 and 2,000 wheat farms have average yields of 2,000 kg and 2,050 kg, respectively. The variance of wheat yield in the country may be assumed to be 100 kg². Examine whether the difference between the average yields of the two samples is statistically significant.

23. Comparison of Heights of Soldiers and Sailors
24. A random sample of 6,400 soldiers has a mean height of 67.85 inches and a standard deviation of 2.56 inches. Another sample of 1,600 sailors has a mean height of 68.55 inches and a standard deviation of 2.52 inches. Determine whether the data provide sufficient evidence to conclude that sailors are, on average, taller than soldiers.

17.12 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 18 Z-TEST, T-TEST AND F-TEST

STRUCTURE

- 18.1 Introduction
- 18.2 Objectives
- 18.3 Test of significance of small samples
- 18.4 t-Test: Student's t-Distribution
- 18.5 Test I: t-Test of significance of the mean of a random sample
- 18.6 TEST II: t-Test for difference of means of two small samples
- 18.7 F-Test
- 18.8 Z-Test
- 18.9 Summary
- 18.10 Exercises
- 18.11 References

18.1 INTRODUCTION

In statistical analysis, it is often necessary to draw conclusions about entire populations based on limited data. When sample sizes are small, standard large-sample techniques may not yield accurate results, requiring specialized approaches for hypothesis testing. This unit dedicated to, *Test of Significance for Small Samples*, covers three key tools—Z-test, T-test, and F-test—that are essential for evaluating claims and making comparisons in such cases.

The Z-test is used when population parameters, like the mean and standard deviation, are known, enabling hypothesis testing for population means or proportions. The T-test is ideal for small samples where the population standard deviation is unknown, making it useful for comparing means between two independent groups or paired data.

The F-test focuses on comparing variances between two samples and serves as a foundation for more advanced methods, such as ANOVA.

By mastering these techniques, learners will develop critical skills to perform accurate statistical analyses, validate hypotheses, and make informed decisions, even with limited data.

18.2 OBJECTIVES

After completing this unit, students will be able to

- Grasp the fundamentals of hypothesis testing and significance levels.
- Identify the appropriate use of Z-test, T-test, and F-test based on data attributes.
- Acquire skills to conduct these tests and interpret the outcomes accurately.
- Enhance the ability to apply these methods to analyze real-world scenarios.
- Recognize the assumptions and limitations inherent to each test.

18.3 TEST OF SIGNIFICANCE OF SMALL SAMPLES

In statistics, a **test of significance** is used to determine whether the observed data significantly differ from what is expected under a null hypothesis. When the **sample size is small** (typically $n < 30$), special methods are used because the normal distribution may no longer provide a reliable approximation.

Key Characteristics of Small Sample Tests:

1. **Population Standard Deviation Unknown:**

- For small samples, the population standard deviation (σ) is usually unknown and is estimated from the sample itself.

2. Use of Student's t-distribution:

- Instead of the normal distribution, the **t-distribution** is used when dealing with small samples.
- The **Student's t-test** accounts for the additional uncertainty due to the smaller sample size.

3. Degrees of Freedom:

- The t-distribution depends on the **degrees of freedom** ($v = n - 1$).
- As the sample size increases, the t-distribution approaches the standard normal distribution.

18.4 t-TEST: STUDENT'S t-DISTRIBUTION

This *t*-distribution is used when sample size is ≤ 30 and the population S.D. is not known.

t-statistic is defined as $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n - 1, f)$ d.f-degrees of freedom where

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

18.4.1 The *t* -Table

The *t*-table given at the end is the probability integral of *t*-distribution. The *t*-distribution has different values for each d.f. and when the d.f. are infinitely large, the *t*-distribution is equivalent to normal distribution

and the probabilities shown in the normal distribution tables are applicable.

18.4.2 APPLICATIONS OF *t*-DISTRIBUTION

The **t-distribution** is commonly used when dealing with **small samples** and the **population standard deviation is unknown**. It is especially important in inferential statistics where we estimate population parameters or test hypotheses based on limited data.

Some of the key applications of the **t-distribution** include:

Testing the significance of

1. a Sample Mean

- Used to determine whether the sample mean $\bar{X}$ differs significantly from a known or hypothesized population mean μ .
- Example: Testing if the average test score of a class differs from the national average.

2. Between Two Sample Means

- Applied to check whether two independent sample means are significantly different from each other.
- Example: Comparing the average performance of two different teaching methods.

3. a Paired Sample

- Also known as the **paired t-test**, this checks the mean difference between two related groups (e.g., before-and-after studies).
- Example: Measuring weight before and after a diet program.

4. a Correlation Coefficient

- Determines whether an observed **partial** or **multiple correlation coefficient** significantly differs from zero.
- This helps assess the strength and reliability of relationships between variables.

5. Confidence Interval Estimation

- The t-distribution is used to construct confidence intervals for population means when sample sizes are small and σ is unknown.

18.4.3 CRITICAL VALUE OF t

The critical value or significant value of t at level of significance α degrees of freedom γ for two tailed test is given by

$$\begin{aligned}P[|t| > t_{\gamma}(\alpha)] &= \alpha \\P[|t| \leq t_{\gamma}(\alpha)] &= 1 - \alpha\end{aligned}$$

The significant value of t at level of significance α for a single tailed test can be determined from those of two-tailed test by referring to the values at 2α .

18.5 TEST I: t-TEST OF SIGNIFICANCE OF THE MEAN OF A RANDOM SAMPLE

To determine if, in situations when the population's variance is unknown, the mean of a sample taken from a normal population differs noticeably from a given value.

H_0 : There is no significant difference between the sample mean $\bar{x}$ and the population mean μ , i.e., we use the statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad \text{where } \bar{X} \text{ is mean of the sample.}$$
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{with d.f. } (n-1)$$

At a given level of significance α_1 and d.f. $(n-1)$. We refer to t -table t_α (two-tailed or one-tailed). If calculated t value is such that $|t| < t_\alpha$ the null hypothesis is accepted. $|t| > t_\alpha$, H_0 is rejected.

18.5.1 FIDUCIAL LIMITS OF POPULATION MEAN

If t_α is the table of t at level of significance α at $(n-1)$ degrees of freedom.

$$\left| \frac{\bar{X} - \mu}{s/\sqrt{n}} \right| < t_\alpha \quad \text{for acceptance of } H_0.$$
$$\bar{x} - t_\alpha s/\sqrt{n} < \mu < \bar{x} + t_\alpha s/\sqrt{n}$$

95% confidence limits (level of significance 5%) are $\bar{X} \pm t_{0.05}s/\sqrt{n}$.

99% confidence limits (level of significance 1%) are $\bar{X} \pm t_{0.01}s/\sqrt{n}$.

NOTE Instead of calculating s , we calculate S for the sample.

Since $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

$$\therefore S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

$$\therefore (n-1)s^2 = nS^2$$

Example 1. A random sample of size 16 has 53 as mean. The sum of squares of the derivation from mean is 135. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.

Solution. H_0 : There is no significant difference between the sample mean and hypothetical population mean.

$$H_0: \mu = 56; H_1: \mu \neq 56 \text{ (Two-tailed test)}$$

$$t: \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1 \text{ difference})$$

Given:

$$\begin{aligned} \bar{X} &= 53, \mu = 56, n = 16, \Sigma(X - \bar{X})^2 = 135 \\ s &= \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}} = \sqrt{\frac{135}{15}} = 3; t = \frac{53 - 56}{3/\sqrt{16}} = -4 \\ |t| &= 4. \text{d.f.u.} = 16 - 1 = 15. \end{aligned}$$

Conclusion. $t_{0.05} = 1.753$.

Since $|t| = 4 > t_{0.05} = 1.753$, the calculated value of t is more than the table value. The hypothesis is rejected. Hence, the sample mean has not come from a population having 56 as mean.

95% confidence limits of the population mean

$$= \bar{X} \pm \frac{s}{\sqrt{n}} t_{0.05} = 53 \pm \frac{3}{\sqrt{16}} (1.725) = 51.706; 54.293$$

99% confidence limits of the population mean

$$= \bar{X} \pm \frac{s}{\sqrt{n}} t_{0.01} = 53 \pm \frac{3}{\sqrt{16}} (2.602) = 51.048; 54.951.$$

Example 2. The lifetime of electric bulbs for a random sample of 10 from a large consignment gave the following data:

Item	1	2	3	4	5	6	7	8	9	10
Life in ' 000'hrs.	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average lifetime of a bulb is 4000 hrs?

Solution. H_0 : There is no significant difference in the sample mean and population mean. i.e., $\mu = 4000$ hrs.

Applying the t -test: $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(10 - 1 \text{ difference})$

X	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6
$X - \bar{X}$	-0.2	0.2	-0.5	-0.3	0.8	-0.6	-0.5	-0.1	0	1.2
$(X - \bar{X})^2$	0.04	0.04	0.25	0.09	0.64	0.36	0.25	0.01	0	1.44

$$\bar{X} = \frac{\Sigma X}{n} = \frac{44}{10} = 4.4 \quad \Sigma(X - \bar{X})^2 = 3.12$$

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{3.12}{9}} = 0.589; t = \frac{4.4 - 4}{\frac{0.589}{\sqrt{10}}} = 2.123$$

For $\gamma = 9, t_{0.05} = 2.26$.

Conclusion. Since the calculated value of t is less than table $t_{0.05} \therefore$

The hypothesis $\mu = 4000$ hrs is accepted, i.e., the average lifetime of bulbs could be 4000 hrs.

Example 3. A sample of 20 items has mean 42 units and standard deviation 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units.

Solution. H_0 : There is no significant difference between the sample

mean and the population mean. i.e., $\mu = 45$ units

$H_1: \mu \neq 45$ (Two tailed test)

Given: $n = 20, \bar{X} = 42, S = 5; \gamma = 19$ difference

$$s^2 = \frac{n}{n-1} S^2 = \left[\frac{20}{20-1} \right] (5)^2 = 26.31 \therefore s = 5.129$$

$$\text{Applying } t\text{-test } t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{42 - 45}{5.129/\sqrt{20}} = -2.615; |t| = 2.615$$

The tabulated value of t at 5% level for 19 d.f. is $t_{0.05} = 2.09$.

Conclusion. Since $|t| > t_{0.05}$, the hypothesis H_0 is rejected, i.e., there is significant difference between the sample mean and population mean. i.e., the sample could not have come from this population.

18.6 TEST II: t-TEST FOR DIFFERENCE OF MEANS OF TWO SMALL SAMPLES (FROM A NORMAL POPULATION)

This test is used to test whether the two samples

$x_1, x_2, \dots, x_{n_1}, y_1, y_2, \dots, y_{n_2}$ of sizes n_1, n_2 have been drawn from two normal populations with mean μ_1

and μ_2 respectively, under the assumption that the population variance are equal. ($\sigma_1 = \sigma_2 = \sigma$).

H_0 : The samples have been drawn from the normal population with means μ_1 and μ_2 , i.e., $H_0: \mu_1 = \mu_2$.

Let $\bar{X}, \bar{Y}$ be their means of the two samples.

Under this H_0 the test of statistic t is given by

$$t = \frac{(\bar{X} - \bar{Y})}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2 \text{ difference})$$

NOTE: 1. If the two sample's standard deviations s_1, s_2 are given, then we have

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}.$$

2. If $n_1 = n_2 = n$, $t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2 + s_2^2}{n-1}}}$ can be used as a test statistic.

3. If the pairs of values are in some way associated (correlated) we can't use the test statistic as given in

4. In this case, we find the differences of the associated pairs of values and apply for single mean i.e., $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ with degrees of freedom $n - 1$.

The test statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

$t = \frac{\bar{d}}{s/\sqrt{n-1}}$, where $\bar{d}$ is the mean of paired difference.

i.e.,

$$d_i = x_i - y_i$$

$\bar{d}_i = \bar{X} - \bar{Y}$, where (x_i, y_i) are the paired data $i = 1, 2, \dots, n$.

Example 1. Two samples of sodium vapor bulbs were tested for length of life and the following results were obtained:

	Size	Sample mean	Sample S.D.

Type I	8	1234 hrs	36 hrs
Type II	7	1036 hrs	40 hrs

Is the difference in the means significant to generalize that Type I is superior to Type II regarding length of life?

Solution. $H_0: \mu_1 = \mu_2$ i.e., two types of bulbs have same lifetime.

$H_1: \mu_1 > \mu_2$ i.e., type I is superior to Type II.

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{8(36)^2 + 7(40)^2}{8 + 7 - 2} = 1659.076$$

$$s = 40.7317$$

$$\text{The } t\text{-statistic } t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1234 - 1036}{40.7317 \sqrt{\frac{1}{8} + \frac{1}{7}}}$$

$$= 18.1480 \sim t(n_1 + n_2 - 2 \text{ difference})$$

$t_{0.05}$ at difference 13 is 1.77 (one tailed test).

Conclusion. Since calculated $|t| > t_{0.05}$, H_0 is rejected, i.e. H_1 is accepted. $\therefore$ Type I is definitely superior to Type II.

where $\bar{X} = \sum_{i=1}^{n_1} \frac{X_i}{n_1}$, $\bar{Y} = \sum_{j=1}^{n_2} \frac{Y_j}{n_2}$; $s^2 = \frac{1}{n_1 + n_2 - 2} \left[\sum (X_i - \bar{X})^2 + \sum (Y_j - \bar{Y})^2 \right]$ is an unbiased estimate of the population variance σ^2 . t follows t -distribution with $n_1 + n_2 - 2$ degrees of freedom.

Example 2. Samples of sizes 10 and 14 were taken from two normal populations with standard deviation 3.5 and 5.2. The sample means were found to be 20.3 and 18.6. Test whether the means of the two populations are the same at 5% level.

Solution. $H_0: \mu_1 = \mu_2$ i.e., the means of the two populations are the same.

$$H_1: \mu_1 \neq \mu_2.$$

Given

$$\begin{aligned}\bar{X} &= 20.3, \bar{X}_2 = 18.6; n_1 = 10, n_2 = 14, s_1 = 3.5, s_2 = 5.2 \\ s^2 &= \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{10(3.5)^2 + 14(5.2)^2}{10 + 14 - 2} = 22.775 \therefore s = 4.772 \\ t &= \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.3 - 18.6}{\left(\sqrt{\frac{1}{10} + \frac{1}{14}}\right) 4.772} = 0.8604\end{aligned}$$

The value of t at 5% level for 22 difference is $t_{0.05} = 2.0739$.

Conclusion. Since $|t| = 0.8604 < t_{0.05}$ the hypothesis is accepted, i.e., there is no significant difference between their means.

Example 3. The height of 6 randomly chosen sailors in inches is 63, 65, 68, 69, 71, and 72. Those of 9 randomly chosen soldiers are 61, 62, 65, 66, 69, 70, 71, 72, and 73. Test whether the sailors are, on average, taller than soldiers.

Solution. Let X_1 and X_2 be the two samples denoting the heights of sailors and soldiers.

Given the sample size $n_1 = 6, n_2 = 9, H_0: \mu_1 = \mu_2$, i.e., the means of both populations are the same.

$$H_1: \mu_1 > \mu_2 \text{ (one tailed test)}$$

Calculation of two sample means:

X_1	63	65	68	69	71	72
$X_1 - \bar{X}_1$	-5	-3	0	1	3	4
$(X_1 - \bar{X}_1)^2$	25	9	0	1	9	16

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = 68; \Sigma(X_1 - \bar{X}_1)^2 = 60$$

X_2	61	62	65	66	69	70	71	72	73
X_2	-	-	-	1.66	1.34	2.34	3.34	4.34	5.34
$- \bar{X}_2$	6.6	5.66	2.66						
	6								
$(X_2$	44.	32.0	7.07	2.75	1.79	5.47	11.1	18.8	28.5
$- \bar{X}_2)$	36	35	56	56	56	56	556	356	156

$$\bar{X}_2 = \frac{\Sigma X_2}{n_2} = 67.66; \Sigma(X_2 - \bar{X}_2)^2 = 152.0002$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2]$$

$$= \frac{1}{6 + 9 - 2} [60 + 152.0002] = 16.3077 \therefore s = 4.038$$

$$\text{Under } H_0, t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{68 - 67.666}{4.0382 \sqrt{\frac{1}{6} + \frac{1}{9}}}$$

$$= 0.3031 \sim t(n_1 + n_2 - 2 \text{ difference})$$

The value of t at 10% level of significance ($\because$ the test is one-tailed) for 13 difference is 1.77.

Conclusion. Since $|t| = 0.3031 < t_{0.05} = 1.77$, the hypothesis H_0 is accepted.

There is no significant difference between their average.

The sailors are not, on average, taller than the soldiers.

Example 4. A certain stimulus administered to each of 12 patients resulted in the following increase in blood pressure: 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure?

Solution. To test whether the mean increase in blood pressure of all patients to whom the stimulus is administered will be positive, we have to assume that this population is normal with mean μ and standard deviation σ which are unknown.

$$H_0: \mu = 0; H_1: \mu_1 > 0$$

The test statistic under H_0

$$\begin{aligned} t &= \frac{\bar{d}}{s/\sqrt{n-1}} \sim t(n-1 \text{ degrees of freedom}) \\ \bar{d} &= \frac{5 + 2 + 8 + (-1) + 3 + 0 + 6 + (-2) + 1 + 5 + 0 + 4}{12} = 2.583 \\ s^2 &= \frac{\sum d^2}{n} - \bar{d}^2 = \frac{1}{12} [5^2 + 2^2 + 8^2 + (-1)^2 + 3^2 + 0^2 + 6^2] \\ &= 8.744 \therefore s = 2.9571 \\ t &= \frac{\bar{d}}{s/\sqrt{n-1}} = \frac{2.583}{2.9571/\sqrt{12-1}} = \frac{2.583\sqrt{11}}{2.9571} \\ &= 2.897 \sim t(n-1 \text{ difference}) \end{aligned}$$

Conclusion. The tabulated value of $t_{0.05}$ at 11 difference is 2.2 .

$\therefore |t| > t_{0.05}$, H_0 is rejected.

i.e., the stimulus does not increase the blood pressure. The stimulus in general will be accompanied by an increase in blood pressure.

Example 5. Memory capacity of 9 students was tested before and after a course of meditation for a month. State whether the course was effective or not from the data below (in same units):

Before	10	15	9	3	7	12	16	17	4
After	12	17	8	5	6	11	18	20	3

Solution. Since the data are correlated and concerned with the same set of students we use paired t -test.

H_0 : Training was not effective $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$ (Two-tailed test).

Before training (X)	After training (Y)	$d = X - Y$	d^2
10	12	-2	4
15	17	-2	4
9	8	1	1
3	5	-2	4
7	6	1	1
12	11	1	1
16	18	-2	4
17	20	-3	9
4	3	1	1
		$\Sigma d = -7$	$\Sigma d^2 = 29$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-7}{9} = -0.7778; s^2 = \frac{\Sigma d^2}{n} - (\bar{d})^2 = \frac{29}{9} - (-0.7778)^2 = 2.617$$

$$t = \frac{\bar{d}}{s/\sqrt{n-1}} = \frac{-0.7778}{\sqrt{2.6172/\sqrt{8}}} = \frac{-0.7778 \times \sqrt{8}}{1.6177} = -1.359$$

The tabulated value of $t_{0.05}$ at 8 difference is 2.31.

Conclusion. Since $|t| = 1.359 < t_{0.05}$, H_0 is accepted, training was not effective in improving performance.

18.7 F-TEST

In statistical analysis, when testing the significance of the difference between the means of two samples, it is generally assumed that the samples are drawn from populations with equal variances. The F-test is used to verify this assumption. Specifically, the F-test helps determine

whether two independent estimates of population variance differ significantly. If the variances are found to be equal, it justifies the use of the t-test for comparing sample means. Therefore, before applying the t-test, it is essential to conduct an F-test to assess the equality of population variances.

Let the variance of two samples with sizes n_1 and n_2 be s_1^2 and s_2^2 . based on these samples the estimate of the population variance is $s_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$ and $s_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$. The d.f. of these estimates are $v_1 = n_1 - 1$, $v_2 = n_2 - 1$.

To determine whether the sample variances s_1^2 and s_2^2 differ significantly—or whether the samples can be considered as coming from the same population or from two populations with equal variance σ^2 —we formulate the null hypothesis:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2,$$

which states that the population variances are equal, implying that the two independent estimates do not show a significant difference.

To test this hypothesis, we use the F-test by computing the statistic:

$$F = \frac{s_1^2}{s_2^2},$$

where $s_1^2 > s_2^2$, so the larger variance is placed in the numerator to ensure that $F \geq 1$. This test helps assess whether the observed difference in sample variances is statistically significant.

Conclusion:

If the computed value of the F-statistic is greater than the critical value $F_{0.05}$, corresponding to $(n_1 - 1, n_2 - 1)$ degrees of freedom from the F-distribution table, then the difference in variances is considered

statistically significant at the 5% level. This implies that the sample variances likely do not arise from populations with equal variance.

On the other hand, if the test statistic does not exceed the critical value, we infer that the samples could plausibly originate from two normal populations with the same variance.

Assumptions of the F-test:

- Each sample must come from a population that follows a normal distribution.
- The samples should be selected randomly and must be independent of each other.
- The population variances are assumed to be equal, or at least the ratio $\frac{\sigma_1^2}{\sigma_2^2} \geq 1$. Therefore, the larger sample variance is placed in the numerator when calculating the F-statistic.

Applications of the F-test:

1. To check whether two independent samples are drawn from normal populations with the same variance σ^2 .
2. To determine whether two independent estimates of population variance are consistent (homogeneous) with each other.

Example 1. From 2 normal populations two random samples drawn are as follows:

A	17	27	18	25	27	29	13	17
B	16	16	20	27	26	25	21	

Examine if both samples have been taken from a single normal population.

Solution. Testing whether two independent samples come from the same population involves:

- (i) checking the equality of means through a t-test, and
- (ii) verifying the equality of variances using an F-test.

Since the *t*-test assumes that the sample variances are equal, we shall first apply the F-test.

F-test. 1. Null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ i.e., the population variance do not differ significantly.

Alternative hypothesis. $H_1: \sigma_1^2 \neq \sigma_2^2$

Test statistic: $F = \frac{s_1^2}{s_2^2}$, (if $s_1^2 > s_2^2$)

Computations for s_1^2 and s_2^2

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
17	-4.625	21.39	16	-2.714	7.365
27	5.735	28.89	16	-2.714	7.365
18	-3.625	13.14	20	1.286	1.653
25	3.375	11.39	27	8.286	68.657
27	5.735	28.89	26	7.286	53.085
29	7.735	54.39	25	6.286	39.513
13	-8.625	74.39	21	2.286	5.226
17	-4.625	21.39			

$$\begin{aligned}\bar{X}_1 &= 21.625; n_1 = 8; \Sigma(\bar{X}_1 - \bar{X}_1)^2 = 253.87 \\ \bar{X}_2 &= 18.714; n_2 = 7; \Sigma(X_2 - \bar{X}_2)^2 = 182.859 \\ s_1^2 &= \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{253.87}{7} = 36.267 \\ s_2^2 &= \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{182.859}{6} = 30.47 \\ F &= \frac{s_1^2}{s_2^2} = \frac{36.267}{30.47} = 1.190\end{aligned}$$

Conclusion. The critical value of F at the 5% significance level for d.f. $v_1 = 7$ and $v_2 = 6$ is 4.21. Since the calculated F-value is less than this tabulated value, we do not reject the null hypothesis H_0 . This indicates that the variability in the two populations is not significantly different.

For the t-test, the null hypothesis is $H_0: \mu_1 = \mu_2$, which states that the population means are equal.

Alternative hypothesis. $H_1: \mu_1 \neq \mu_2$

Test of statistic

$$\begin{aligned}s^2 &= \frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{253.87 + 182.859}{8 + 7 - 2} = 33.594 \\ \therefore s &= 5.796 \\ t &= \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{21.625 - 18.714}{5.796 \sqrt{\frac{1}{8} + \frac{1}{7}}} = 0.9704 \sim t(n_1 + n_2 - 2) \text{ difference}\end{aligned}$$

Conclusion. At the 5% significance level, the critical value of t for 13 d.f. is 2.16. Since the calculated t-value is less than the tabulated value, we accept the null hypothesis H_0 , indicating that there is no significant difference between the population means, i.e., $\mu_1 = \mu_2$. Therefore, we conclude that the two samples may have been drawn from the same normal population.

Example 2. Two independent sample of sizes 7 and 6 had the following values:

Sample A	28	30	32	33	31	29	34
Sample B	29	30	30	24	27	28	

Determine whether the samples have been taken from normal populations having the same variance.

Solution. H_0 : The variances are equal. i.e., $\sigma_1^2 = \sigma_2^2$ i.e., the samples have been drawn from normal populations with same variance.

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Under null hypothesis, the test statistic $F = \frac{s_1^2}{s_2^2} (s_1^2 > s_2^2)$

Computations for s_1^2 and s_2^2

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
28	-3	9	29	1	1
30	-1	1	30	2	4
32	1	1	30	2	4
33	2	4	24	-4	16
31	0	0	27	-1	1
29	-2	4	28	0	0
34	3	9			
		28			26

$$\bar{X}_1 = 31, n_1 = 7; \Sigma(X_1 - \bar{X}_1)^2 = 28$$

$$\bar{X}_2 = 28, n_2 = 6; \Sigma(X_2 - \bar{X}_2)^2 = 26$$

$$s_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{28}{6} = 4.666; s_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{26}{5} = 5.2$$

$$F = \frac{s_2^2}{s_1^2} = \frac{5.2}{4.666} = 1.1158. (\because s_2^2 > s_1^2)$$

Conclusion. The critical value of F at the 5% significance level for degrees of freedom $v_1 = 5$ and $v_2 = 6$ is 4.39. Since the calculated F-value exceeds the tabulated value, the null hypothesis H_0 is **rejected**, indicating a significant difference between the variances. Therefore, the samples do **not** appear to be from populations with equal variance.

18.8 Z-TEST

18.8.1 Fisher's z-distribution: In G.W. Snedecor's F-distribution with (v_1, v_2) d.f., if we put

$$F = \exp(2Z) \Rightarrow Z = \frac{1}{2} \log_e F \quad (1)$$

The distribution of Z becomes

$$\begin{aligned} g(z) &= p(F) \cdot \left| \frac{dF}{dz} \right| \\ &= \frac{(v_1/v_2)^{v_1/2}}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{(e^{2z})^{(v_1/2)-1} 2e^{2z}}{\left[1 + \frac{v_1}{v_2} e^{2z}\right]^{(v_1+v_2)/2}} \\ g(z) &= 2 \frac{\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}}}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{e^{v_1 z}}{\left[1 + \frac{v_1}{v_2} e^{2z}\right]^{\frac{(v_1+v_2)}{2}}}; -\infty < z < \infty \quad (2) \end{aligned}$$

which is the probability function of Fisher's z-distribution with (v_1, v_2) d.f. The tables of significant values z_0 of z which will be exceeded in random sampling with probabilities 0.05 and 0.01, i.e., $P(z > z_0) = 0.05$ and $P(z > z_0) = 0.01$ corresponding to various d.f. (v_1, v_2) were published by Fisher (c.f. Statistical Methods for Research Workers) in 1925. From these tables, Snedecor (1934-38) by using (1) deduced the tables of significant values of the variance ratio which he denoted by F in honour of Prof: R.A. Fisher.

Remark. With the help of relation (1), all the applications of F -distribution may be regarded as the applications of z -distribution also.

18.8.2 Moment Generating Function of z -distribution:

$$\begin{aligned} M_Z(t) &= E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} g(z) dz \\ &= \int_0^{\infty} F^{t/2} f(F) dF \quad [\because e^{2z} = F] \end{aligned}$$

Since μ'_r (about origin) for F -distribution is $\int_0^{\infty} F^r f(F) dF$, we can find m.g.f. of the z -distribution by putting $r = t/2$ in the expression for μ'_r for F -distribution.

Hence

$$\begin{aligned} M_Z(t) &= \left(\frac{v_2}{v_1}\right)^{t/2} \cdot \frac{\Gamma[(v_1 + t)/2] \Gamma[(v_2 - t)/2]}{\Gamma(v_1/2) \Gamma(v_2/2)} \\ K_Z(t) &= \log M_Z(t) \\ &= \frac{t}{2} [\log v_2 - \log v_1] + \log \Gamma[(v_1 + t)/2] \\ &\quad + \log \Gamma[(v_2 - t)/2] - \log \Gamma(v_1/2) - \log \Gamma(v_2/2) \\ &\Rightarrow K_Z(t) = \log M_Z(t) \end{aligned}$$

Using Stirling's approximation for $n!$, when n is large, viz.,

$$\begin{aligned}
\lim_{n \rightarrow \infty} \Gamma(n+1) &= \lim_{n \rightarrow \infty} n! \approx \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} \\
\Rightarrow \log \Gamma(n+1) &= \left(n + \frac{1}{2}\right) \log n - n + \log \sqrt{2\pi}, \text{ we get} \\
\kappa_1 = \mu_1 &= \frac{1}{2} \left(\frac{1}{v_2} - \frac{1}{v_1} \right) \\
\kappa_2 = \mu_2 &= \frac{1}{2} \left(\frac{1}{v_2} + \frac{1}{v_1} + \frac{1}{v_1^2} + \frac{1}{v_2^2} \right) \\
\kappa_3 = \mu_3 &= \frac{1}{2} \left[\left(\frac{1}{v_2^2} - \frac{1}{v_1^2} \right) + \left(\frac{1}{v_2^3} + \frac{1}{v_1^3} \right) \right] \\
\kappa_4 = \mu_4 - 3\mu_2^2 &= \frac{1}{v_1^3} + \frac{1}{v_2^3} + 3 \left(\frac{1}{v_1^4} + \frac{1}{v_2^4} \right),
\end{aligned}$$

whence β_1 and β_2 can be found.

Remark. z-distribution tends to normal distribution with mean $\frac{1}{2} \left(\frac{1}{v_2} - \frac{1}{v_1} \right)$ and variance $\frac{1}{2} \left(\frac{1}{v_1} + \frac{1}{v_2} \right)$, as v_1 and v_2 become large.

18.8.3 Fisher's z-transformation: To test the significance of an observed sample correlation coefficient from an uncorrelated bivariate normal population, t -test is used. But in random sample of size n_i from a bivariate normal population in which $\rho \neq 0$, Prof. R.A. Fisher proved that the distribution of ' r ' is by no means normal and in the neighbourhood of $\rho = \pm 1$, its probability curve is extremely skewed even for large n . If $\rho \neq 0$, Fisher suggested the following transformation

-

$$Z = \frac{1}{2} \log_e \frac{1+r}{1-r} = \tanh^{-1} r \quad (3)$$

and proved that even for small samples, the distribution of Z is approximately normal with mean

$$\xi = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} = \tanh^{-1} \rho \quad (4)$$

and variance $1/(n-3)$ and for large values of n , say > 50 , the approximation is fairly good. z-transformation has the following applications in Statistics.

(I) To test if an observed value of ' r ' differs significantly from a hypothetical value ρ of the population correlation coefficient.

H_0 : There is no significant difference between r and ρ . In other words, the given sample has been drawn from a bivariate normal population with correlation coefficient ρ .

If we take

$$Z = \frac{1}{2} \log_e \{(1+r)/(1-r)\} \text{ and } \xi = \frac{1}{2} \log_e \{(1+\rho)/(1-\rho)\}$$

then under H_0 ,

$$Z \sim N\left(\xi, \frac{1}{n-3}\right) \Rightarrow \frac{Z - \xi}{\sqrt{1/(n-3)}} \sim N(0,1)$$

Thus if $(Z - \xi)\sqrt{(n-3)} > 1.96$, H_0 is rejected at 5% level of significance and if it is greater than 2.58 , H_0 is rejected at 1% level of significance, where Z and ξ are defined in (3) and (4).

Remark: Z defined in equation (3) should not be confused with the Z used in Fisher's z-distribution.

Example 1. A correlation coefficient of 0.72 is obtained from a sample of 29 pairs of observations.

(i) Can the sample be regarded as drawn from a bivariate normal population in which true correlation coefficient is 0.8?

(ii) Obtain 95% confidence limits for ρ in the light of the information provided by the sample.

Solution. (i) H_0 : There is no, significant difference between $r = 0.72$; and

$\rho = 0.80$, i.e., the sample can be regarded as drawn from the bivariate normal population with $\rho = 0.8$.

Here

$$\begin{aligned} Z &= \frac{1}{2} \log_6 \left(\frac{1+r}{1-r} \right) = 1.1513 \log_{10} \left(\frac{1+r}{1-r} \right) \\ &= 1.1513 \log_{10} 6.14 = 0.907 \\ \xi &= \frac{1}{2} \log_0 \left(\frac{1+\rho}{1-\rho} \right) = 1.1513 \log_{10} \frac{(1+0.8)}{(1-0.8)} \\ &= 1.1513 \times 0.9541 = 1.1 \\ \text{S.E. (Z)} &= \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{26}} = 0.196 \end{aligned}$$

Under H_0 , the test statistic is

$$\begin{aligned} U &= \frac{Z - \xi}{1/\sqrt{n-3}} \sim N(0,1) \\ U &= \frac{(0.907 - 1 \cdot 100)}{0.196} = -0.985 \end{aligned}$$

Since $|U| < 1.96$, it is not significant at 5% level of significance and H_0 may be accepted. Hence the sample may be regarded as coming from a bivariate normal population with $\rho = 0.8$.

(ii) 95% confidence limits for ρ on the basis of the information supplied by the sample, are given by

$$\begin{aligned} |U| &\leq 1.96 \\ |Z - \xi| &\leq 1.96 \times \frac{1}{\sqrt{n-3}} = 1.96 \times 0.196 \\ \Rightarrow |0.907 - \xi| &\leq 0.384 \\ \Rightarrow 0.907 - 0.384 &\leq \xi \leq 0.907 + 0.384 \\ \Rightarrow 0.523 &\leq \xi \leq 1.291 \\ \Rightarrow 0.523 &\leq \frac{1}{2} \log_e \left(\frac{1+\rho}{1-\rho} \right) \leq 1.291 \\ \Rightarrow &\leq 0.1513 \log_{10} \left(\frac{1+\rho}{1-\rho} \right) \leq 1.291 \end{aligned}$$

$$\Rightarrow \frac{0.523}{1.1513} \leq \log_{10} \left(\frac{1+\rho}{1-\rho} \right) \leq \frac{1.291}{1.1513}$$

$$\Rightarrow \frac{0.523}{1.1513} \leq \log_{10} \left(\frac{1+\rho}{1-\rho} \right) \leq \frac{1.291}{1.1513} \quad (5)$$

Now $\log_{10} \left(\frac{1+\rho}{1-\rho} \right) = 0.4543^\circ$ and $\log_{10} \left(\frac{1+\rho}{1-\rho} \right) = 1.1213$

$$\Rightarrow \frac{1+\rho}{1-\rho} = \text{Antilog}(0.4543) = 2.846 \quad \Rightarrow \frac{1+\rho}{1-\rho} = \text{Antilog}(1.1213) = 13.22$$

$$\Rightarrow \rho = \frac{2.846 - 1}{2.846 + 1} = \frac{1.846}{3.846} = 0.4799 \quad \rho = \frac{13.22 - 1}{13.22 + 1} = \frac{12.22}{14.22} = 0.86$$

Hence, substituting in (5) we get $0.48 \leq \rho \leq 0.86$

(2) To test the significance of the difference between two independent sample correlation coefficients. Let r_1 and r_2 be the sample correlation coefficients observed in two independent samples of sizes n_1 and n_2 respectively then

$$Z_1 = \frac{1}{2} \log_e \left(\frac{1+r_1}{1-r_1} \right) \text{ and } Z_2 = \frac{1}{2} \log_e \left(\frac{1+r_2}{1-r_2} \right)$$

Under the null hypothesis H_0 : that sample correlation coefficients dQ , not differ significantly, i.e., the samples are drawn from the same bivariate normal population or from different populations with same correlation coefficient ρ , (say), the statistic

$$Z = \frac{(Z_1 - Z_2) - E(Z_1 - Z_2)}{S.E.(Z_1 - Z_2)} \sim N(0,1)$$

Now

$$E(Z_1 - Z_2) = E(Z_1) - E(Z_2) = \xi_1 - \xi_2 = 0$$

$$\left[\because \xi_1 = \xi_2 = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} \text{ (under } H_0) \right]$$

and

$$\begin{aligned} (Z_1 - Z_2), &= \sqrt{V(Z_1) + V(Z_2)} \\ \text{S.E.} &= \sqrt{\left\{ \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right\}} \end{aligned}$$

[Covariance term vanishes since samples are independent]

[Covariance term vanishes since samples are independent]

Under H_0 , the test statistic is

$$Z = \frac{Z_1 - Z_2}{\sqrt{\left\{ \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right\}}} \sim N(0,1)$$

By comparing this value with 1.96 or 2.58, H_0 may be accepted or rejected at 5% and 1% level of significance respectively.

(3) To obtain pooled estimate of ρ . Let $r_1, r_2, \dots, r_k$ be observed correlation coefficients in k -independent samples of sizes $n_1, n_2, \dots, n_k$ from a bivariate normal population. The problem is to combine these estimates of ρ to get a pooled estimate for the parameter.

If we take $Z_i = \frac{1}{2} \log_e \left(\frac{1+r_i}{1-r_i} \right); i = 1, 2, \dots, k$ then $Z_i; i = 1, 2, \dots, k$ are independent normal variates with variances $\frac{1}{(n_i-3)}; i = 1, 2, \dots, k$ and common mean

$$\xi = \frac{1}{2} \log_e \left(\frac{1 + \rho}{1 - \rho} \right)$$

The weighted mean (say $\bar{Z}$) of these Z_i 's is given by

$$\bar{Z} = \frac{\sum_{i=1}^k w_i Z_i}{\sum_{i=1}^k w_i}$$

where w_i is the weight of Z_i .

Now $\bar{Z}$ is also an unbiased estimate of ξ , since

$$E(\bar{Z}) = \frac{1}{\sum w_i} \left[E \sum_{i=1}^k w_i Z_i \right] = \frac{1}{\sum w_i} \left[\sum w_i E(Z_i) \right] = \frac{1}{\sum w_i} \left[\sum_i w_i \xi \right] = \xi$$

$$\text{and } V(\bar{Z}) = \frac{1}{(\sum w_i)^2} V \left[\sum w_i Z_i \right] = \frac{1}{(\sum w_i)^2} \left[\sum w_i^2 V(Z_i) \right]$$

The weights w'_i s, ($i = 1, 2, \dots, n$) are so chosen that $\bar{Z}$ has minimum variance."

-In order that $V(\bar{Z})$ is minimum for variations in w_i , we should have

$$\begin{aligned} \frac{\partial}{\partial w_i} V(\bar{Z}) &= 0; i = 1, 2, \dots, k \\ \Rightarrow \frac{(\sum w_i)^2 2w_i V(Z_i) - [\sum_i w_i^2 V(Z_i)] 2(\sum_i w_i)}{(\sum w_i)^4} &= 0 \\ \Rightarrow w_i V(Z_i) &= \frac{\sum w_i^2 V(Z_i)}{\sum w_i}, \text{ a constant.} \end{aligned}$$

$$\therefore w_i \propto \frac{1}{V(Z_i)} = (n_i - 3); i = 1, 2, \dots, k \quad (6)$$

Hence the minimum variance estimate of ξ is given by

$$\bar{Z} = \frac{\sum_{i=1}^k w_i Z_i}{\sum_{i=1}^k w_i} = \frac{\sum_{i=1}^k (n_i - 3) Z_i}{\sum_{i=1}^k (n_i - 3)} \text{ [On using (6)]}$$

and the best estimate of ρ is then given by

$$\bar{Z} = \frac{1}{2} \log_e \frac{1 + \rho}{1 - \rho} \Rightarrow \hat{\rho} = \tanh \left[\frac{\sum (n_i - 3) Z_i}{\sum (n_i - 3)} \right]$$

Minimum variance of $\bar{Z}$ is given by

$$\begin{aligned} iV(\bar{Z})\}_{\min} &= \frac{\sum \left\{ (n_i - 3)^2 \left(\frac{1}{n_i - 3} \right) \right\}}{\{\sum (n_i - 3)\}^2} = \frac{\sum (n_i - 3)}{\{\sum (n_i - 3)\}^2} \\ &= \frac{1}{\sum_{i=1}^k (n_i - 3)}. \end{aligned}$$

CHECK YOUR PROGRESS

1. Explain the meaning and applications of the Z-test.

.....

.....

.....

2. What is a t-test? Explain its different applications.

.....

.....

.....

3. Explain the F-test and distinguish it from the Z-test and t-test.

.....

.....

.....

18.9 SUMMARY

In this unit, we have discussed the following concepts

- Test of significance of small samples
- t-Test: Student's t-Distribution
- Test I: t-Test of significance of the mean of a random sample
- Test II: t-Test for difference of means of two small samples
- F-Test
- Z-Test

18.10 EXERCISES

- 1 Ten individuals are chosen at random from a normal population of students and their scores found to be 63,63,66,67,68,69,70,70,71,71. In the light of these data discuss the suggestion that mean score of the population of students is 66.
- 2 The following values gives the lengths of 12 samples of Egyptian cotton taken from a consignment: 48,46,49,46,52,45,43,47,47,46,45,50. Test if the mean length of the consignment can be taken as 46.
- 3 A sample of 18 items has a mean 24 units and standard deviation 3 units. Test the hypothesis that it is a random sample from a normal population with mean 27 units.
- 4 A filling machine is expected to fill 5 kg of powder into bags. A sample of 10 bags gave the following weights: 4.7,4.9,5.0,5.1,5.4,5.2,4.6,5.1,4.6, and 4.7. Test whether the machine is working properly.
- 5 The mean life of 10 electric motors was found to be 1450 hrs with a standard deviation of 423 hrs. A second sample of 17 motors chosen from a different batch showed a mean life of 1280 hrs with a standard deviation of 398 hrs. Is there a significant difference between means of the two samples?
- 6 The scores obtained by a group of 9 regular course students and another group of 11 part time course students in a test are given below:

Regula	5	6	6	5	6	5	6	6	5
r:	6	2	3	4	0	1	7	9	8

Part	6	7	7	6	6	5	7	6	7	6	6
time:	2	0	1	2	0	6	5	4	2	8	6

Examine whether the scores obtained by regular students and part time students

differ significantly at 5% and 1% level of significance.

- 7 A group of 10 boys fed on diet A and another group of 8 boys fed on a different diet B recorded the following increase in weight (kgs):

Diet A: 5 6 8 1 12 4 3 9 6 10

Diet B: 2 3 6 8 10 1 2 8

Does it show the superiority of diet A over the diet B?

18.11 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 19 CHI-SQUARE TEST AND GOODNESS OF FIT

STRUCTURE

- 19.1 Introduction
- 19.2 Objectives
- 19.3 Chi-square (χ^2) test overview
- 19.4 The χ^2 distribution
- 19.5 χ^2 independence test
- 19.6 Summary
- 19.7 Exercises
- 19.8 References

19.1 INTRODUCTION

The "*Chi-Square Test and Goodness of Fit*" unit explores the Chi-Square (χ^2) test, a powerful statistical method for analyzing categorical data. This test assesses whether there is a significant difference between actual and predicted frequencies, making it valuable for examining relationships between categorical variables and evaluating how well a model fits the data. The unit delves into the Chi-Square distribution, the test of independence, and practical applications in real-world contexts.

19.2 OBJECTIVES

After completing this unit, students will be able to

- Grasp the core principles of the Chi-Square (χ^2) test.
- Describe the Chi-Square distribution and its key properties.
- Conduct a Chi-Square test to assess the independence of categorical variables.

- Use the Chi-Square test to analyze the goodness of fit for a theoretical distribution.
- Interpret Chi-Square test outcomes in real-world applications.

19.3 CHI-SQUARE (χ^2) TEST OVERVIEW

When a coin is tossed 200 times, the theoretical considerations lead us to expect 100 heads and 100 tails. But in practice, these results are rarely achieved. The quantity χ^2 (a Greek letter, pronounced as chi-square) explains the extent to which theory and observation diverge. The actual and anticipated frequencies are exactly the same for $\chi = 0$. The value of χ^2 increases with the difference between the actual and predicted frequencies. Thus this test relation between theory and observation.

If $O_i (i = 1, 2, \dots, n)$ is a set of actual frequencies and $E_i (i = 1, 2, \dots, n)$ is the related set of predicted frequencies, then, χ^2 is defined as

$$\chi^2 = \sum_{i=1}^n \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

where $\sum O_i = \sum E_i = N$ (total frequency) and degrees of freedom (difference)

$$= (n - 1).$$

Note: (i) If $\chi^2 = 0$, the observed and theoretical frequencies agree exactly.

Note: (ii) If $\chi^2 > 0$ they do not agree exactly.

19.3.1 Degrees of Freedom

To compare the calculated value of χ^2 with the critical value from the chi-square table, we must first determine the degrees of freedom. Suppose we are to choose four numbers that add up to 50. We can freely

select any three of them, but the fourth is automatically fixed as 50 minus the sum of the chosen three. Although we start with four numbers, the condition that their total must be 50 limits our independent choices to three. This means there is one restriction, resulting in $4 - 1 = 3$ degrees of freedom. If there are two such constraints, our independent choices are reduced further, giving $4 - 2 = 2$ degrees of freedom.

In general, the degrees of freedom are calculated by subtracting the number of independent constraints from the total number of observations. This is expressed as:

$$v = n - k,$$

where v represents the degrees of freedom, n is the number of observations, and k is the number of independent constraints.

Note: (i) For a $p \times q$ contingency table (p columns and q rows), $v = (p - 1)(q - 1)$

Note: (ii) In the case of a contingency table, the expected or predicted frequency of any class

$$= \frac{\text{Total of rows in which it occurs} \times \text{Total of columns in which it occurs}}{\text{Total number of observations}}$$

19.3.2 Applications

1. Test of Independence in Contingency Tables

Used to determine whether two categorical variables are independent or associated.

Example: Testing whether gender and voting preference are related.

2. Goodness-of-Fit Test

Used to check how well an observed frequency distribution matches an expected distribution.

Example: Testing if a die is fair by comparing the observed and expected frequencies of outcomes.

3. Test of Homogeneity

Used to compare the distribution of a categorical variable across several populations.

Example: Comparing customer satisfaction ratings across different stores of a retail chain.

4. Genetics and Heredity Studies

Used to verify the genetic ratios predicted by Mendelian inheritance.

Example: Testing whether observed phenotypic ratios in offspring match expected Mendelian ratios.

5. Market Research and Consumer Behaviour

Helps determine if preferences or behaviours differ across demographic groups.

Example: Testing whether product preference is related to age group or income level.

6. Public Health and Epidemiology

Used to analyze the association between risk factors and health outcomes.

Example: Testing the relationship between smoking status and incidence of lung disease.

19.3.3 Conditions for Applying χ^2 Test

Here are the **key conditions** for applying the **Chi-square test** effectively:

1. Data Should Be in Frequency Form

- The data must be counts or frequencies, not percentages, proportions, or measurements.
- Example: Number of people preferring different brands, not the average rating of each brand.

2. Observations Must Be Independent

- Each observation should belong to one and only one category.
- No individual or item should be counted in more than one cell of the table.

3. Expected Frequency Should Be Sufficiently Large

- Generally, each expected frequency should be **at least 5**.
- If some expected frequencies are less than 5, categories may need to be combined.

4. Sample Size Should Be Large

- The Chi-square test is an approximation based on the assumption of a large sample size.
- A small sample can lead to inaccurate results.

5. Categories Must Be Mutually Exclusive and Exhaustive

- Each item must fall into one and only one category.
- All possible outcomes should be represented in the table.

6. Random Sampling

- The data should be collected through a random sampling method to ensure valid inference.

Note: If any one of the theoretical frequencies is less than 5, then we apply a correction given by F Yates, which is usually known as 'Yates correction for continuity', we add 0.5 to the cell frequency which is less than 5 and adjust the remaining cell frequency suitably so that the marginal total is not changed.

19.4 THE χ^2 DISTRIBUTION

For large sample sizes, the sampling distribution of χ^2 can be closely approximated by a continuous curve known as the chi-square distribution. The probability function of χ^2 distribution is given by

$$f(\chi^2) = c(\chi^2)^{(v/2-1)}e^{-\chi^2/2}$$

where $e = 2.71828$, v = number of degrees of freedom; c = a constant depending only on v .

The degrees of freedom are represented symbolically by the letter v and they are derived using the formula $v = n - k$, where k is the number of independent restrictions.

The number of degrees of freedom is typically one fewer than the number of classes when fitting a binomial distribution; however, when fitting a Poisson distribution, the number of degrees of freedom is two fewer than the number of classes because the Poisson distribution's parameter is determined by the arithmetic mean and the total frequency. Since we utilise the total frequency, mean, and standard deviation to fit a normal curve, the number of degrees of freedom is three less than the number of classes.

χ^2 TEST AS A TEST OF GOODNESS OF FIT

χ^2 test enables us to ascertain how well the theoretical distributions such as Binomial, Poisson or Normal etc. fit empirical distributions, i.e.,

distributions obtained from sample data. If the calculated value of χ^2 is less than the table value at a specified level (generally 5%) of significance, the fit is regarded as satisfactory, i.e., the oscillations of simple sampling are the reason for the discrepancy between real and predicted frequencies. The fit is considered unsatisfactory if the computed value of χ^2 exceeds the value in the table.

Example 1. The following table gives the number of accidents that took place in an industry during various days of the week. Test if accidents are uniformly distributed over the week.

Day	Mon	Tue	Wed	Thu	Fri	Sat
Number of accidents	14	18	12	11	15	14

Solution. Null hypothesis H_0 : The accidents are uniformly distributed over the week.

Under this H_0 , the expected frequencies of the accidents on each of these days = $\frac{84}{6} = 14$

Observed frequency O_i	14	18	12	11	15	14
Expected frequency E_i	14	14	14	14	14	14
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} = \frac{30}{14} = 2.1428$$

Conclusion. Table value of χ^2 at 5% level for (6 – 1 = 5 d.f.) is 11.09.

Since the computed value of $\chi^2 < 11.09$, H_0 is accepted, the accidents are uniformly distributed over the week.

Example 2. A die is thrown 270 times and we get the following results:

Number on the die	1	2	3	4	5	6
Frequency	40	32	29	59	57	59

Test whether the die is biased or not.

Solution. Null hypothesis H_0 : Die is unbiased.

Under this H_0 , the expected frequencies for each digit is $\frac{270}{6} = 46$. To find the value of χ^2

O_i	40	32	29	59	57	59
E_i	46	46	46	46	46	46
$(O_i - E_i)^2$	36	196	289	169	121	169

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{980}{46} = 21.30.$$

Conclusion. Tabulated value of χ^2 at 5% level of significance for $(6 - 1 = 5)$ d.f. is 11.09. Since the computed value of $\chi^2 = 21.30 > 11.09$ the tabulated value, H_0 is rejected.
i.e., die is biased.

Example 3. The distribution of digits in randomly selected numbers from a phone book is shown in the following table:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Check to see whether the numbers in the directory may be assumed to occur equally often.

Solution. Null hypothesis H_0 : The digits taken in the directory occur equally frequently.

i.e., there is no significant difference between the observed and expected frequency.

Under H_0 , the expected frequency is given by $= \frac{10,000}{10} = 1000$

To find the value of χ^2

O_i	1026	1107	997	996	1075	1107	933	972	964	853
E_i	1000	1000	1000	1000	1000	1000	1107	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	11449	4489	784	1296	21609

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{58542}{1000} = 58.542.$$

Conclusion. The tabulated value of χ^2 at 5% level of significance for 9 difference is 16.919. Since the computed value of $\chi^2 > 16.919$, H_0 is rejected.

Not all of the numbers in the directory appear at the same frequency.

Example 4. The following data shows how many male and female births occurred among 800 households with four children:

Number of male births	0	1	2	3	4
Number of female births	4	3	2	1	0
Number of families	32	178	290	236	94

Determine if the facts support the hypothesis that the probability of a male birth is equal to that of a female birth and that the Binomial rule is true, namely $p = q = 1/2$.

Solution. H_0 : The data are consistent with the hypothesis of equal probability for male and female births, i.e., $p = q = 1/2$.

We use Binomial distribution to calculate theoretical frequency given by:

$$N(r) = N \times P(X = r)$$

where N is the total frequency. $N(r)$ is the number of families with r male children:

$$P(X = r) = {}^nC_r p^r q^{n-r}$$

where p and q are probability of male and female births, n is the number of children.

$$\begin{aligned} N(0) &= \text{Number of families with 0 male children} = 800 \times {}^4C_0 \left(\frac{1}{2}\right)^4 \\ &= 800 \times 1 \times \frac{1}{2^4} = 50 \end{aligned}$$

$$\begin{aligned} N(1) &= 800 \times {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = 200; N(2) = 800 \times {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 \\ &= 300 \end{aligned}$$

$$\begin{aligned} N(3) &= 800 \times {}^4C_3 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = 200; N(4) = 800 \times {}^4C_4 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 \\ &= 50 \end{aligned}$$

Observed frequency O_i	32	178	290	236	94
Expected frequency E_i	50	200	300	200	50
$(O_i - E_i)^2$	324	484	100	1296	1936

$\frac{(O_i - E_i)^2}{E_i}$	6.48	2.42	0.333	6.48	38.72
-----------------------------	------	------	-------	------	-------

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} = 54.433.$$

Conclusion. Table value of χ^2 at 5% level of significance for $5 - 1 = 4$ difference is 9.49.

Since the computed value of $\chi^2 > 9.49$, H_0 is rejected.

The data are not consistent with the hypothesis that the Binomial law holds and that the chance of a male birth is not equal to that of a female birth.

Note: Since the fitting is Binomial, the d.f.

$$v = n - 1 \text{ i.e., } v = 5 - 1 = 4.$$

Example 5. Verify whether Poisson distribution can be assumed from the data given below:

Number of defects	0	1	2	3	4	5
Frequency	6	13	13	8	4	3

Solution. H_0 : Poisson fit is a good fit to the data.

$$\text{Mean of the Poisson distribution} = \frac{\sum f_i x_i}{\sum f_i} = \frac{94}{47} = 2$$

To fit a Poisson distribution we require m . Parameter $m = \bar{x} = 2$.

By Poisson distribution the frequency of r success is

$$N(r) = N \times e^{-m} \cdot \frac{m^r}{r!}, N \text{ is the total frequency.}$$

$$N(0) = 47 \times e^{-2} \cdot \frac{(2)^0}{0!} = 6.36 \approx 6; N(1) = 47 \times e^{-2} \cdot \frac{(2)^1}{1!} = 12.72 \approx 13$$

$$N(2) = 47 \times e^{-2} \cdot \frac{(2)^2}{2!} = 12.72 \approx 13; N(3) = 47 \times e^{-2} \cdot \frac{(2)^3}{3!} = 8.48 \approx 9$$

$$N(4) = 47 \times e^{-2} \cdot \frac{(2)^4}{4!} = 4.24 \approx 4; N(5) = 47 \times e^{-2} \cdot \frac{(2)^5}{5!} = 1.696 \approx 2.$$

X	0	1	2	3	4	5
O_i	6	13	13	8	4	3
E_i	6.36	12.72	12.72	8.48	4.24	1.696
$\frac{(O_i - E_i)^2}{E_i}$	0.2037	0.00616	0.00616	0.02716	0.0135	1.0026

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} = 1.2864$$

Conclusion. The computed value of $\chi^2 = 1.2864$. At 5% level of significance tabulated value of χ^2 for $\gamma = 6 - 2 = 4$ d.f. is 9.49. Since the calculated value of $\chi^2 < 9.49$. H_0 is accepted i.e., Poisson distribution provides a good fit to the data.

Example 6. According to the hypothesis, the percentage of beans in each of the four groups— G_1, G_2, G_3 , and G_4 —should equal 9: 3: 3: 1. In a 1600-bean trial, the four groups' numbers were 882, 313, 287, and 118. Does the theory make sense in light of the actual results?

Solution. H_0 : The experimental result validates the theory. The theoretical frequency can be computed as follows:

$$E(G_1) = \frac{1600 \times 9}{16} = 900; E(G_2) = \frac{1600 \times 3}{16} = 300;$$

$$E(G_3) = \frac{1600 \times 3}{16} = 300; E(G_4) = \frac{1600 \times 1}{16} = 100$$

To computed the value of χ^2 .

Observed frequency O_i	882	313	287	118
Expected frequency E_i	900	300	300	100
$\frac{(O_i - E_i)^2}{E_i}$	0.36	0.5633	0.5633	3.24

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} = 4.7266.$$

Conclusion. The table value of χ^2 at 5% level of significance for 3 difference is 7.815. Since the computed value of $\chi^2 < 7.815$. Hence H_0 is accepted and the experimental results support the theory.

19.5 χ^2 INDEPENDENCE TEST

A very useful application of χ^2 test is to investigate the relationship between trials or attributes which can be classified into two or more categories.

The sample data set out into two-way table, called contingency table. Let us consider two attributes A and B divided into r classes $A_1, A_2, A_3, \dots, A_r$, and B divided into s classes $B_1, B_2, B_3, \dots, B_s$. If $(A_i), (B_j)$ represents the number of persons possessing the attributes A_i, B_j respectively, $(i = 1, 2, \dots, r, j = 1, 2, \dots, s)$ and $(A_i B_j)$ represent the number of persons possessing attributes A_i and B_j . Also, we have $\sum_{i=1}^r A_i = \sum_{j=1}^s B_j = N$ where N is the total frequency. The contingency table for $r \times s$ is given below:

B\A	A ₁	A ₂	A ₃	... A _r	Total
B ₁	(A ₁ B ₁)	(A ₂ B ₁)	(A ₃ B ₁)	 (A _r B ₁)	B ₁
	(A ₁ B ₂)	(A ₂ B ₂)	(A ₃ B ₂)	 (A _r B ₂)	B ₂
	(A ₁ B ₃)	(A ₂ B ₃)	(A ₃ B ₃)	 (A _r B ₃)	B ₃
B ₂					
					
	(A ₁ B ₈)	(A ₂ B ₈)	(A ₃ B ₈)	 (A _r B ₈)	(B ₈)
B ₃	(A ₁)	(A ₂)	(A ₃)	 (A _r)	N
.....					
B _s					
Total					

H_0 : Both the attributes are independent. i.e., A and B are independent under the null hypothesis, we calculate the expected frequency as follows: $P(A_i)$ = Probability that a person possesses the attribute

$$A_i = \frac{(A_i)}{N} \quad i = 1, 2, \dots, r$$

$$P(B_j) = \text{Probability that a person possesses the attribute } B_j = \frac{(B_j)}{N}$$

$P(A_i B_j)$ = Probability that a person possesses both attributes A_i and B_j

$$= \frac{(A_i B_j)}{N}$$

If $(A_i B_j)_0$ is the expected number of persons possessing both the attributes A_i and B_j

$$\begin{aligned}(A_i B_j)_0 &= NP(A_i B_j) = NP(A_i)(B_j) \\ &= N \frac{(A_i)}{N} \frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N}\end{aligned}$$

($\because$ A and B are independent)

$$\text{Hence } \chi^2 = \sum_{i=1}^r \sum_{j=1}^s \left[\frac{[(A_i B_j) - (A_i B_j)_0]^2}{(A_i B_j)_0} \right]$$

which is distributed as a χ^2 variate with $(r-1)(s-1)$ degrees of freedom.

Note:

- 1 For a 2×2 contingency table where the frequencies are $\frac{a}{b}$, $\frac{c}{d}$, χ^2 can be computed from independent frequencies as

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$$

- 2 If the contingency table is not 2×2 , then the formula for computing χ^2 as given in Note 1, can't be used. Hence, we have another formula for calculating the expected frequency

$$(A_i B_j)_0 = \frac{(A_i)(B_j)}{N} \text{ i.e.,}$$

expected frequency in each cell is =

$$\frac{\text{Product of column total and row total}}{\text{whole total}}.$$

- 3 If $\frac{a|b}{c|d}$ is the 2×2 contingency table with two attributes, $Q =$

$$\frac{ad-bc}{ad+bc} \text{ is called the coefficient of association. If the attributes are independent then } \frac{a}{b} = \frac{c}{d}.$$

- 4 Yates's Correction. In a 2×2 table, if the frequencies of a cell are small, we make Yates's correction to make χ^2 continuous. Decrease by $\frac{1}{2}$ those cell frequencies which are greater than expected frequencies, and increase by $\frac{1}{2}$ those which are less

than expectation. This will not affect the marginal columns.

This correction is known as Yates's correction to continuity.

After Yates's correction $\chi^2 = \frac{N(bc - ad - \frac{1}{2}N)^2}{(a+c)(b+d)(c+d)(a+b)}$ when $ad - bc < 0$

$$\chi^2 = \frac{N(ad - bc - \frac{1}{2}N)^2}{(a+c)(b+d)(c+d)(a+b)} \text{ when } ad - bc > 0.$$

Example 1. What are the expected frequencies of 2×2 contingency tables given below:

(i)

a	b
c	d

(ii)

2	10
6	6

Solution.

Observed frequencies

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$a + b + c + d = N$

Observed frequencies

→

Expected frequencies

$\frac{(a + c)(a + b)}{a + b + c + d}$	$\frac{(b + d)(a + b)}{a + b + c + d}$
$\frac{(a + c)(c + d)}{a + b + c + d}$	$\frac{(b + d)(c + d)}{a + b + c + d}$

Expected frequencies

(ii)

2	10	12
6	6	12
8	16	24

→

$\frac{8 \times 12}{24} = 4$	$\frac{16 \times 12}{24} = 8$
$\frac{8 \times 12}{24} = 4$	$\frac{16 \times 12}{24} = 8$

Example 2. From the following table regarding the color of eyes of father and son test if the color of son's eye is associated with that of the father.

		Eye color of son	
		Light	Not light
Eye color of father	Light	471	51
	Not light	148	230

Solution. Null hypothesis H_0 : The color of son's eye is not associated with that of the father, i.e., they are independent.

Under H_0 , we calculate the expected frequency in each cell as

$$= \frac{\text{Product of column total and row total}}{\text{Whole total}}$$

Expected frequencies are:

Eye color of father \ Eye color of son	Light	Not light	Total
Light	$\frac{619 \times 522}{900} = 359.02$	$\frac{289 \times 522}{900} = 167.62$	522
Not light	$\frac{619 \times 378}{900} = 259.98$	$\frac{289 \times 378}{900} = 121.38$	378
Total	619	289	900

$$\begin{aligned}\chi^2 &= \frac{(471 - 359.02)^2}{359.02} + \frac{(51 - 167.62)^2}{167.62} + \frac{(148 - 259.98)^2}{259.98} + \frac{(230 - 121.38)^2}{121.38} \\ &= 261.498\end{aligned}$$

Conclusion. Tabulated value of χ^2 at 5% level for 1 difference is 3.841.

Since the calculated value of $\chi^2 >$ tabulated value of χ^2 , H_0 is rejected. They are dependent, i.e., the color of son's eye is associated with that of the father.

Example 3. The following table gives the number of good and bad parts produced by each of the three shifts in a factory:

	Good parts	Bad parts	Total
Day shift	960	40	1000
Evening shift	940	50	990
Night shift	950	45	995
Total	2850	135	2985

Test whether or not the production of bad parts is independent of the shift on which they were produced.

Solution. Null hypothesis H_0 : The production of bad parts is independent of the shift on which they were produced.

The two attributes, production and shifts are independent.

$$\text{Under } H_0, \chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \left[\frac{[(A_i B_j)_o - (A_i B_j)_e]^2}{(A_i B_j)_e} \right]$$

Calculation of expected frequencies

Let A and B be the two attributes namely production and shifts. A is divided into two classes A_1, A_2 and B is divided into three classes B_1, B_2, B_3 .

$$(A_1 B_1)_e = \frac{(A_1)(B_1)}{N} = \frac{(2850) \times (1000)}{2985} = 954.77$$

$$(A_1 B_2)_e = \frac{(A_1)(B_2)}{N} = \frac{(2850) \times (990)}{2985} = 945.226$$

$$(A_1 B_3)_e = \frac{(A_1)(B_3)}{N} = \frac{(2850) \times (995)}{2985} = 950$$

$$(A_2 B_1)_e = \frac{(A_2)(B_1)}{N} = \frac{(135) \times (1000)}{2985} = 45.27$$

$$(A_2 B_2)_e = \frac{(A_2)(B_2)}{N} = \frac{(135) \times (990)}{2985} = 44.773$$

$$(A_2 B_3)_e = \frac{(A_2)(B_3)}{N} = \frac{(135) \times (995)}{2985} = 45$$

To calculate the value of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
(A ₁ B ₁)	960	954.77	27.3529	0.02864
(A ₁ B ₂)	940	945.226	27.3110	0.02889
(A ₁ B ₃)	950	950	0	0
(A ₂ B ₁)	40	45.27	27.7729	0.61349
(A ₂ B ₂)	50	44.773	27.3215	0.61022
(A ₂ B ₃)	45	45	0	0
				1.28126

Conclusion. The tabulated value of χ^2 at 5% level of significance for 2 degrees of freedom $(r - 1)(s - 1)$ is 5.991. Since the calculated value of χ^2 is less than the tabulated value, we accept H_0 , i.e., the production of bad parts is independent of the shift on which they were produced.

CHECK YOUR PROGRESS

1. Define the Chi-square test and explain its major applications.

.....

.....

.....

2. Explain the concept of goodness of fit and its statistical significance.

.....

.....

.....

3. State the assumptions and limitations of the Chi-square test.

.....

.....

.....

19.6 SUMMARY

In this unit, we explored the Chi-Square test and its applications in hypothesis testing for categorical data. We learned about the Chi-Square distribution and how to use the Chi-Square test to assess the independence of categorical variables. Additionally, we covered the use of the Chi-Square test for evaluating the goodness of fit of a model. These concepts are essential for analyzing categorical data and drawing valid conclusions in research and practical applications.

19.7 EXERCISES

- 1 The following table gives the frequency of occupancy of the digits 0,1, ...,9 in the last place in four logarithms of numbers 10-99. Examine if there is any peculiarity.

Digits:	0	1	2	3	4	5	6	7	8	9
Frequency:	6	16	15	10	12	12	3	2	9	5

- 2 The sales in a supermarket during a week are given below. Test the hypothesis that the sales do not depend on the day of the week, using a significant level of 0.05.

Days:	Mon	Tues	Wed	Thurs	Fri	Sat
Sales:	65	54	60	56	71	84

- 3 A survey of 320 families with 5 children each revealed the following information:

Number of boys:	5	4	3	2	1	0
Number of girls:	0	1	2	3	4	5
Number of families:	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

- 4 In the accounting department of bank, 100 accounts are selected at random and estimated for errors. The following results were obtained:

Number of errors:	0	1	2	3	4	5	6
Number of accounts:	35	40	19	2	0	2	2

Does this information verify that the errors are distributed according to the Poisson probability law?

19.8 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

UNIT 20 INTRODUCTION TO ANALYSIS OF VARIANCE

STRUCTURE

- 20.1 Introduction
- 20.2 Objectives
- 20.3 Testing homogeneity of means. Analysis of variance
- 20.4 Analysis of Variance for one-way classification
- 20.5 Analysis of Variance for two-way classification.
- 20.6 Tests of significance based on Fisher's z transformation
- 20.7 Summary
- 20.8 Exercises
- 20.9 References

20.1 INTRODUCTION

ANOVA compares the means of three or more groups to discover whether they vary significantly. ANOVA may evaluate many datasets simultaneously, making it a valuable tool for experimental design and data processing. T-tests can only compare two groups.

ANOVA helps determine whether observed differences are due to genuine impacts or random fluctuation by splitting the variance into sources. Biology, economics, psychology, and engineering use this strategy to analyse how variables affect outcomes.

The unit introduces one-way and two-way ANOVA, its assumptions, and its applications. Students will learn to analyse complicated datasets, evaluate findings, and make meaningful inferences from multi-group comparisons in this subject.

20.2 OBJECTIVES

This unit allows students to

- Understand the purpose and fundamental principles of Analysis of Variance (ANOVA).
- Recognize the significance of the F-distribution and t-distribution in hypothesis testing.
- Understand the assumptions and conditions required for performing ANOVA.
- Gain the skills to analyze and interpret ANOVA results, along with their practical implications

20.3 TESTING HOMOGENEITY OF MEANS. ANALYSIS OF VARIANCE

Consider a set of N values of the variate x arranged into h rows and k columns so that $N = hk$. Let x_{ij} represent the value of the member belonging to i th row and j th columns, $\bar{x}_j$ the mean of the j th column and $\bar{x}_i$, the mean of the i th row. The data is then arranged as follows:



	1st col.	2nd col.		jth col.		kth col.	Row totals	Row means
1st row	x_{11}	x_{12}		x_{1j}		x_{1k}	T_1	$\bar{x}_1$
2nd row	x_{21}	x_{22}		x_{2j}		x_{2k}	T_2	$\bar{x}_2$
.....								
jth row	x_{j1}	x_{j2}		x_{jj}		x_{jk}	T_j	$\bar{x}_j$
.....								
hth row	x_{h1}	x_{h2}		x_{hj}		x_{hk}	T_h	$\bar{x}_h$
Column totals	T_1	T_2		T_j		T_k	T	$\bar{x}_{..} = \frac{T}{N}$
Column means	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_j$		$\bar{x}_k$	$\bar{x}_{..}$	

We regard the data as classified into h classes, each row corresponding to a class according to some factors of classification. Thus, each row may represent the price of commodity obtained from k shops in a district and the different rows may correspond to different districts or the data may represent marks obtained by h groups of students, each group consisting of k students.

The object of the analysis of variance would be to break up the total variation into components due to each of the factors and then compare them by the F test.

If we assume that each group is homogeneous within itself, then the Variation within itself is due to chance or random causes. If the data is homogeneous with respect to the factor of classification, i.e., if there is no effect due to this factor then the variation between classes would also be due to random causes and the two variations would be of the same order.

20.4 ANALYSIS OF VARIANCE FOR ONE-WAY CLASSIFICATION

Let

$$x_{ij} = \mu_i + e_{ij}$$

x_{ij} stands for the j th observation in the i th set, μ_i is the effect of fertilizer number i and e_{ij} is the chance error which is due to chance variation in the j th observation of the i th set. We shall prefer a general variation in the experiment i.e. we shall take $\mu_i = \mu + \alpha_i$ where μ is the model for our and α_i is equal to the deviation. Thus, we have

$$x_{ij} = \mu + \alpha_i + e_{ij}$$

where $i = 1, 2, \dots, h; j = 1, 2, \dots, k$, with

Assumptions

1. e_{ij} , s are independently and identically normally distributed with mean zero and variance σ^2 i.e. e_{ij} 's are $N(0, \sigma^2)$.
2. α 's are constants with the restriction $\sum_{i=1}^h \alpha_i = 0$.
3. μ is constant

Null hypothesis. The null hypothesis to be set up is

$$H_0: \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_k$$

or equivalently, $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_i = \dots = \alpha_k$.

Principle of Least Squares will be helpful in estimating the parameters μ and $\alpha_i, i = 1, 2, \dots, h$. Hence if

$$L = \sum_{i=1}^h \sum_{j=1}^k e_{ij}^2 = \sum_{i=1}^h \sum_{j=1}^k (x_{ij} - \mu - \alpha_i)^2 = \sum_{i=1}^h \sum_{j=1}^k (x_{ij} - \mu_i)^2 \quad (1)$$

The normal equations for determining the μ_i 's are

$$\frac{\partial L}{\partial \mu_i} = 0 \Rightarrow \sum_{j=1}^k (x_{ij} - \mu_i) = 0$$

$$\Rightarrow \sum_{j=1}^k x_{ij} - \sum_{j=1}^k \mu_i = 0 \text{ i.e., } T_i = k\mu_i \text{ where } T_i = \sum_{j=1}^k x_{ij}$$

Thus, the estimated value of μ_i is obtained by $\hat{\mu}_i = \frac{1}{k} T_i$.

$$\text{Also } \hat{\alpha}_p - \hat{\alpha}_q = (\hat{\mu}_p - \mu) - (\hat{\mu}_q - \mu) = \hat{\mu}_p - \hat{\mu}_q = \frac{1}{k} (T_p - T_q)$$

$$\therefore \text{Minimum value of } L = \sum (x_{ij} - \hat{\mu}_i)^2$$

$$= \sum_{i,j} \left(x_{ij} - \frac{T_i}{k} \right)^2 = \sum_{i,j} x_{ij}^2 - \frac{1}{k} \sum_{i=1}^h T_i^2$$

Hence

$$S^2 = \sum_{i=1}^h \sum_{j=1}^k \left(x_{ij} - \frac{T_i}{k} \right)^2$$

is called the measure of variation due to variation within set.

If we set up the hypothesis that the difference among the observed value may be attributed to chance variation, then the model reduces to

$$x_{ij} = \mu + e_i$$

$$\therefore L_0 = \sum (x_{ij} - \mu)^2$$

therefore, the normal equation for estimating μ , is

$$0 = \frac{\partial L_0}{\partial \mu} = \sum_{i,j} (x_{ij} - \mu)$$

$$\therefore \hat{\mu} = \frac{1}{hk} \sum_{i,j} x_{ij} = \frac{1}{hk} \bar{x}$$

$\therefore$ Under $H_0: S_0^2 = \sum_{i,j} \left(x_{ij} - \frac{\bar{x}}{hk}\right)^2$

where S_0^2 is called measure of the total variation in the data.

$\therefore S_0^2 - S^2$ is called the measure of the variation due to the hypothesis H_0 and is given by

$$\begin{aligned} S_0^2 - S^2 &= \sum_{i,j} \left(x_{ij} - \frac{\bar{x}}{hk}\right)^2 - \sum_{i,j} \left(x_{ij} - \frac{\bar{x}_i}{k}\right)^2 \\ &= \sum_i \left(\frac{\bar{x}_i^2}{k}\right) - \frac{\bar{x}^2}{hk}. \end{aligned}$$

But

$$S_0^2 = S + (S_0^2 - S^2)$$

i.e. Total variation = Within set variation + Between set variation

$$\text{Again, } E\left[\frac{S_0^2 - S^2}{h-1}\right] = \sigma^2, E\left[\frac{S^2}{hk-h}\right] = \sigma^2.$$

Hence $(S_0^2 - S^2)/(h-1)$ and $S^2/(hk-h)$ are unbiased estimates of σ^2

Also $(S_0^2 - S^2)/\sigma^2$ is distributed as χ^2 with $(h-1)d.f.$ i.e. χ_{h-1}^2 and S^2/σ^2 is distributed as χ^2 with $h(k-1)d.f.$ i.e. $\chi_{h(k-1)}^2$

Also χ_{h-1}^2 and $\chi_{h(k-1)}^2$ are independent χ^2 variates.

Hence

$$\frac{(S_0^2 - S^2)/(h-1)}{S^2/h(k-1)} = F_{h-1, h(k-1)}$$

We know that the three unbiased estimates of σ^2 are not independent. However, it can be shown that the second and third estimates are independent and their ratio can be treated by the F test

$$F = \frac{\sum k(\bar{x}_i - \bar{x})^2}{\sum \sum (x_{ij} - \bar{x}_i)^2} \times \frac{N-h}{h-1} \text{ with } v_1 = h-1, v_2 = N-h.$$

The above results are expressed conveniently in the analysis of variance table given on the below.

Table 20.1 Analysis of variance table

Source of variation	D.F.	Sum of squares	Mean square
Between class means	$h - 1$	$\sum_i k(\bar{x}_i - \bar{x})^2$	$\sum_i \frac{k(\bar{x}_i - \bar{x})^2}{(h - 1)}$
Within classes	$N - h$	$\sum_i \sum_j (x_{ij} - \bar{x})^2$	$\sum_i \sum_j \frac{(x_{ij} - \bar{x})^2}{N - h}$
Total	$N - 1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2$	

Note 1. It will be noted that the mean squares are obtained by dividing the S.S. by the corresponding *d. f.* For the practical calculation of S.S., we may note first that all the S.S. are independent of change of origin and any convenient origin may, therefore, be chosen.

Note 2. Secondly, all the S.S. can be expressed in terms of row totals T_i , column total T_j and grand total T .

$$\text{Thus } \sum_{ij} \sum_j (x_{ij} - \bar{x})^2 = \sum_{if} \sum_{ij}^2 - 2\bar{x} \sum_{ij} \sum_{ij} + N\bar{x}^2 = \sum_{ij} \sum_{ij}^2 - \frac{T^2}{N} \quad (2)$$

$$\text{and } \sum_i k(\bar{x}_i - \bar{x})^2 = k \sum \left(\frac{T_i}{k} - \frac{T}{hk} \right)^2 = \sum_i \frac{T_i^2}{k} - \frac{T^2}{N},$$

Now $\sum \sum (x_{ij} - \bar{x}_i)^2$ can be obtained by subtracting (3) from (2). The term T^2/N is usually called the correction factor on account of measurement being taken from the mean.

Analysis of variance Table

Source of variation	D.F.	S.S.	M.S.
Between varieties	9	160.54	17.84
Within varieties	20	43.33	2.17
Total	29	203.87	

$$F = \frac{17 \cdot 84}{2 \cdot 17} = 8 \cdot 22, v_1 = 9, v_2 = 20$$

This is highly significant at 1% level.

20.5 ANALYSIS OF VARIANCE FOR TWO WAY CLASSIFICATION

Suppose the data are divided according to both rows and columns. Thus, if we study the effects of two types of treatments, say fertilizers and varieties, then the observation corresponding to the i th treatment of one type and the j th treatment of the other type may be denoted by x_{ij} .

Consider the model $x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, 2, \dots, h; j = 1, 2, \dots, k$ Where α_i is the effect due to the i th fertilizer, β_j is the effect due to the the variety and e_{ij} is the chance or random error and are $N(0, \sigma^2)$ i.e. $E(e_{ij}) = 0, \text{Var}(e_{ij}) = \sigma^2$. With the restrictions

$$\sum_{i=1}^h \alpha_i = 0, \sum_{j=1}^k \beta_j = 0$$

By the principle of least squares,

$$P = \sum_{i=1}^h \sum_{j=1}^k e_{ij}^2 = \sum_{i=1}^h \sum_{j=1}^k (x_{ij} - \mu - \alpha_i - \beta_j)^2$$

$$= \sum_{i=1}^h \sum_{j=1}^k (x_{ij} - \mu - \beta_j)^2 \quad (1)$$

Normal equations for estimation of the constants μ_i and β_j are:

$$0 = \frac{\partial P}{\partial \mu_i} = -2 \sum_{j=1}^k (x_{ij} - \mu_i - \beta_j) \quad (2)$$

and

$$0 = \frac{\partial P}{\partial \beta_j} = -2 \sum_{i=1}^h (x_{ij} - \mu_i - \beta_j) \quad (3)$$

Equation (2) gives

$$\hat{\mu}_i = \frac{1}{k} \sum_{j=1}^k x_{ij} = \frac{1}{k} \bar{x}_i$$

and (3) gives

$$\hat{\beta}_j = \frac{1}{h} \sum_{i=1}^h x_{ij} = \frac{1}{h} \bar{x}_j$$

Also

$$\hat{\alpha}_p - \hat{\alpha}_q = \hat{\mu}_p - \hat{\mu}_q = \frac{\bar{x}_p}{k} - \frac{\bar{x}_q}{k}, \hat{\beta}_p - \hat{\beta}_q = \frac{\bar{x}_p}{h} - \frac{\bar{x}_q}{h}$$

$$\begin{aligned} \therefore \sum_{i=1}^h \sum_{j=1}^k (x_{ij} - \bar{x})^2 &= \sum_{ij} \left(x_{ij} - \frac{\bar{x}_i}{k} - \frac{\bar{x}_j}{h} + \frac{\bar{x}}{hk} \right)^2 \\ &= k \sum_{i=1}^h (\bar{x}_i - \bar{x})^2 + h \sum_{j=1}^k (\bar{x}_j - \bar{x})^2 + \sum_{i=1}^h \sum_{j=1}^k (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2 \end{aligned}$$

As in the last section, on the basis of normality of distribution of the variates and the null hypothesis of the homogeneity of both the row and column factors of classification the above four sums on being divided respectively by $hk - 1, h - 1, k - 1$ and $(h - 1)(k - 1)$ give unbiased estimates of the population variance σ^2 and by using the theorem due to Cochran it can be shown that estimates given by S.S. the between row 'means' and also those given 'between column means' are independent of the estimate given by the last error, ie., S.S. The analysis of variance

Table Analysis of Variance

Source of variation	D.F.	S.S	M.S	F
Between rows	$h - 1$	$\sum_i k(\bar{x}_i - \bar{x}_{..})^2$	S.S/D.F.	
Between columns	$k - 1$	$\sum_j k(\bar{x}_j - \bar{x}_{..})^2$		
Error	$(h - 1)(k - 1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}_{..})^2$		
Total	$hk - 1$	$\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2$		

20.6 TESTS OF SIGNIFICANCE BASED ON FISHER'S Z TRANSFORMATION

Fisher's z transformation

$$z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right) \equiv 1.1513 \log_{10} \left(\frac{1+r}{1-r} \right)$$

have been discussed in last unit. Fisher has shown that for samples of a large size n , z is distributed asymptotically normally about mean ξ , where

$$\xi = \frac{1}{2} \log_e \left(\frac{1 + \rho}{1 - \rho} \right) = 1.1513 \log_{10} \left(\frac{1 + \rho}{1 - \rho} \right)$$

with variance $1/(n - 3)$. This transformation forms the basis of the tests of significance discussed below, but will be obvious that unlike the tests of the previous two sections the present tests will hold for large samples only.

20.6.1 To test whether r differs significantly from ρ .

Given a random sample $(x_1, y_1), \dots, (x_n, y_n)$ from a bivariate normal population, we have to test the hypothesis that the population correlation coefficient is ρ . To do this, we calculate z and ξ by applying Fisher's transformation and then calculate the asymptotically standard normal variate

$$\frac{z - \xi}{1/\sqrt{n - 3}} \quad (1)$$

If the absolute value of this statistic exceeds 1.96 the difference is significant at 5% level.

Example 1. A correlation coefficient of 0.7 is discovered in a sample of 28 pairs. Apply z transformation to find out if this differs significantly (a) from 0 (b) from 0.5.

Solution. (a)

$$\rho = 0, \quad \xi = 0$$

$$\text{Also } z = 1.1513 \log_{10} \left(\frac{1 + .7}{1 - .7} \right) = 0.87$$

$$\therefore t = \frac{0.87 - 0}{1/\sqrt{28 - 3}} = 4.35 > 1.96$$

The hypothesis of zero correlation is discredited and we conclude there is a correlation in the population.

$$(b) \rho = 0.5, \xi = 1.1513 \log_{10} \left(\frac{1+0.5}{1-0.5} \right) = 0.55, t = \frac{0.87-0.50}{1/\sqrt{28-3}} = 0.6 < 1.96$$

20.6.2: To determine the significance of two independent correlation differences

We have to test the hypothesis of two samples of sizes n_1, n_2 giving correlation coefficients r_1, r_2 may be regarded as drawn from the sample population or from two populations with the same correlation coefficient. If the hypothesis is correct, the statistic

$$t = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \quad (2)$$

where

$$z_1 = \frac{1}{2} \log_e \left(\frac{1+r_1}{1-r_1} \right), z_2 = \frac{1}{2} \log_e \left(\frac{1+r_2}{1-r_2} \right)$$

is asymptotically a standard normal variate. If absolute value of this statistic is greater than 1.96, the difference will be significant at 5% level.

Example 2. The correlation coefficients between temperatures of rice and breakage percentage calculated from two samples of 12 and 16 are 0.8912 and 0.8482 respectively. Do the two estimates differ significantly?

Solution. Here $n_1 = 12, n_2 = 16, r_1 = .8912, r_2 = .8482$

$$\therefore z_1 = 1.4276, z_2 = 1.2496$$

Now

$$\frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{1.4276 - 1.2496}{\sqrt{0.1880}} = 0.41$$

which is not at all significant at 5% level.

20.6.3 Combination of estimates of a correlation coefficient

Correlation coefficients $r_1, r_2, \dots, r_k$ are obtained from samples of sizes $n_1, n_2, \dots, n_k$ from the same bivariate normal population with correlation coefficient ρ , then $z_1, z_2, \dots, z_k$ are independent estimates of the like quantity $\zeta = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right)$. To find a pooled estimate with the minimum weighted average of $z_1, z_2, \dots, z_k$ and to get an estimate with the minimum variance we use weights inversely proportional to the variance.

The estimate is accordingly $\bar{z} = \frac{\sum (n_i - 3) z_i}{\sum (n_i - 3)}$ and from this we can get an estimate for ρ from the relation $\rho = \tanh \bar{z}$. Then since the variates z_i are approximately normally distributed about ξ with variance $1/(n_i - 3)$, the sum $\sum (n_i - 3)(z_i - \bar{z})^2$ is distributed approximately as χ^2 with $(k - 1)df$. The significance can be ascertained from the table of χ^2 . Not significance shows that the estimate r_i of correlation in the population may be regarded as homogeneous.

Example 3. Combine the two estimates of r in Example 2 to get a pooled estimate.

$$\begin{aligned} \bar{z} &= \frac{(n_1 - 3)z_1 + (n_2 - 3)z_2}{(n_1 - 3) + (n_2 - 3)} = \frac{12.8484 + 16.2448}{9 + 13} = \frac{29.0932}{22} \\ &= 1.3224. \end{aligned}$$

The corresponding estimate of ρ is 0.8674.

CHECK YOUR PROGRESS

1. What is Analysis of Variance (ANOVA)? Explain its purpose.

.....

.....

.....

2. Explain the basic assumptions underlying Analysis of Variance.

.....

.....

.....

3. Distinguish between one-way and two-way Analysis of Variance.

.....

.....

.....

20.7 SUMMARY

In this unit, we have discussed the following concepts

- Testing homogeneity of means. ANOVA
- ANOVA for one-way classification
- ANOVA for two-way classification.

20.8 EXERCISES

1. In a sample of 30 sets of values from a normal population $R_{2(31)}$ is found to be 0.5. Show that this is significant of correlation in the population between x_2 and x_1, x_3 . ($F = 4 \cdot 5, v_1 = 2, v_2 = 27$)

2. From a sample of size 20, $R_{1(23)}$ is found to be 5433. Show that this is significant at 1% level.
3. Show that the 95% fiducial limits for ζ are $z \pm 1.96\sqrt{n-3}$. (Example 1)
4. Prove that the 95% fiducial limits for ρ in the above example 1 are 0.45 and 0.85.
5. Independent sample of 21, 30, 39, 26, 35 pair of values give correlation coefficients 0.39, 0.61, 0.43, 0.54, 0.48. Prove that the pooled estimate for ρ is 0.5 nearly and the value of χ^2 is 1.4 approx.
6. Two independent samples of sizes 20 and 25 give correlation coefficients 60 and .80. Show that these are not significantly different. Show also that the weighted average value of $r = .7267$.
7. Show that in EXERCISES of 815.4 .2 the 95% fiducial limits for r_1 from the given value of r_2 are 0.47 and 0.93.

20.9 REFERENCES

1. Spiegel, M. R. (1972). *Theory and Problems of Statistics*. Schaum Publishing Company.
2. Gupta S. C. and Kapoor V. K. (2020). *Fundamentals of Mathematical Statistics*. S. Chand & Sons.
3. Hogg R. V., Mckean J. and Craig A. T. (2013). *Introduction to Mathematical Statistics*, Pearson.
4. Miller I. and Miller M. (1999). *John E. Freund's Mathematical Statistics with Applications*. Pearson.
5. Rohatgi V. K. and Saleh A. K. Md. E. (2015) *An Introduction to Probability and Statistics*. Wiley.
6. Kapoor J. N. and Saxena H. C. (2020) *Fundamentals of Mathematical Statistics*. S. Chand & Sons.

TABLE 1: Significant values $t_v(\alpha)$ of t -distribution (Two Tail Areas)
 $[|t| > t_v(\alpha)] = \alpha$

difference (v)	Probability (Level of significance)					
	0.50	0.10	0.05	0.02	0.01	0.001
1	1.00	6.31	12.71	31.82	63.66	636.62
2	0.82	0.92	4.30	6.97	6.93	31.60
3	0.77	2.32	3.18	4.54	5.84	12.94
4	0.74	2.13	2.78	3.75	4.60	8.61
5	0.73	2.02	2.57	3.37	4.03	6.86
6	0.72	1.94	2.45	3.14	3.71	5.96
7	0.71	1.90	2.37	3.00	3.50	5.41
8	0.71	1.80	2.31	2.90	3.36	5.04
9	0.70	1.83	2.26	2.82	3.25	4.78
10	0.70	1.81	2.23	2.76	3.17	4.59
11	0.70	1.80	2.20	2.72	3.11	4.44
12	0.70	1.78	2.18	2.68	3.06	4.32
13	0.69	1.77	2.16	2.65	3.01	4.22
14	0.69	1.76	2.15	2.62	2.98	4.14
15	0.69	1.75	2.13	2.60	2.95	4.07
16	0.69	1.75	2.12	2.58	2.92	4.02
17	0.69	1.74	2.11	2.57	2.90	3.97
18	0.69	1.73	2.10	2.55	2.88	3.92
19	0.69	1.73	2.09	2.54	2.86	3.88
20	0.69	1.73	2.09	2.53	2.85	3.85
21	0.69	1.72	2.08	2.52	2.83	3.83
22	0.69	1.72	2.07	2.51	2.82	3.79
23	0.69	1.71	2.07	2.50	2.81	3.77
24	0.69	1.71	2.06	2.49	2.80	3.75
25	0.68	1.71	2.06	2.49	2.79	3.73
26	0.68	1.71	2.06	2.48	2.78	3.71
27	0.68	1.70	2.05	2.47	2.77	3.69
28	0.68	1.70	2.05	2.47	2.76	3.67
29	0.68	1.70	2.05	2.46	2.76	3.66
30	0.68	1.70	2.04	2.46	2.75	3.65
∞	0.67	1.65	1.96	2.33	2.58	3.29

TABLE 2: F-Distribution
Values of F for F-Distributions with 0.05 of the Area in The Right Tail

Degrees of freedom for numerator																			
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	40	60	120	∞	
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.88	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.29	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.64	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51

TABLE 3: CHI-SQUARE DISTRIBUTION
Significant Values χ^2 (α) of Chi-Square Distribution Right Tail Areas for
Given Probability α , $P = P_r(\chi^2 > \chi^2(\alpha)) = \alpha$ And ν Degrees of Freedom
(difference)

Degrees of freedom (ν)	Probability (Level of significance)						
	0 = .99	0.95	0.50	0.10	0.05	0.02	0.01
1	.000157	.00393	.455	2.706	3.841	5.214	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	1.145	4.351	9.236	11.070	13.388	15.086
6	.872	2.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	3.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.669
10	2.558	3.940	9.340	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	27.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	4.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	15.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	32.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.65	41.566	44.314
26	12.198	15.379	25.336	35.363	38.885	41.856	45.642
27	12.879	16.151	26.336	36.741	40.113	41.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.933	18.493	29.336	40.256	43.773	47.962	50.892

NOTE For degrees of freedom (ν) greater than 30, the quantity $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ may be used as a normal variate with unit variance.