BASIC STATISTICS **MT(N) - 222**

BECHELOR OF SCIENCE
(FIFTH SEMESTER)

DEPARTMENT OF MATHEMATICS
SCHOOL OF SCIENCES
UTTARAKHAND OPEN UNIVERSITY
HALDWANI, UTTARAKHAND
263139

COURSE NAME: BASIC STATISTICS

COURSE CODE: MT(N) 222

Department of Mathematics
School of Science
Uttarakhand Open University
Haldwani, Uttarakhand, India,
263139

Editor

Dr. Deepak Kumar Sharma
Assistant Professor (AC)
Department of Mathematics
Uttarakhand Open University
Haldwani, Uttarakhand

Unit Writer	Block	Units
1. **Dr. Kamlesh Bisht** Department of Mathematics Uttarakhand Open University Haldwani	**I, II**	**1, 2, 5 & 6**
2. **Prof. Arvind Bhatt** Department of Mathematics Uttarakhand Open University Haldwani	**II**	**3 & 4**
3. **Dr. Deepak Kumar Sharma** Department of Mathematics Uttarakhand Open University Haldwani	**III, IV & V**	**7 to 14**

Course Title and Code : BASIC STATISTICS (MT(N) 222)
Copyright : Uttarakhand Open University
Edition : 2025
Published By : Uttarakhand Open University, Haldwani, Nainital- 263139

CONTENTS: MT(N) - 222

COURSE INFORMATION

The present self-learning material **"Basic Statistics"** has been designed for **B.Sc. (Fifth Semester)** learners of Uttarakhand Open University, Haldwani. This course is divided into 14 units of study. This Self Learning Material is a mixture of Five Blocks.

First block is **Basics and Data,** in this block Different types of data, Tables, charts, histograms, frequency distributions, Measures of association defined Clearly.

Second block is **Probability,** in this block Probability concepts, conditional probability, Bayes theorem defined clearly.

Third block is **Moment Generating Function,** in this block Probability distributions – random variable, expected value and variance, Moments, Moment generating function, Characteristic function are defined.

Fourth block is **Discrete Distribution,** in this block Discrete distributions Binomial distribution, Poisson distribution are defined.

Fifth block is **Continuous Distribution,** in this block Uniform distribution, Normal distribution are defined.

Adequate number of illustrative examples and exercises have also been included to enable the leaners to grasp the subject easily.

COURSE NAME: BASIC STATISTICS

COURSE CODE: MT(N)-222

BLOCK - I: BASICS AND DATA

Unit-1: INTRODUCTION OF STATISTICS

CONTENTS:

1.1 *INTRODUCTION*

Statistics is a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data. It provides tools and techniques for making informed decisions in the presence of uncertainty. Statistics is widely used in various fields such as economics, biology, psychology, education, business, and government to analyze trends, test hypotheses, and draw conclusions based on numerical data. It is broadly classified into two main types: **Descriptive Statistics**, which summarizes and describes the features of a data set, and **Inferential Statistics**, which uses sample data to make predictions or generalizations about a population. Through statistical methods, researchers and analysts can convert raw data into meaningful information for decision-making and problem-solving.

1.2 *OBJECTIVE*

After reading this unit you will be able to:

- To define statistics and understand its meaning as both data and a method.
- To explain the origin and development of statistics as a scientific discipline.
- To distinguish between descriptive and inferential statistics**.**
- To understand the importance and scope of statistics in various fields such as economics, sociology, business, and science.
- To identify basic statistical terms such as data, population, sample, variable, and frequency.
- To learn the characteristics and limitations of statistics as a tool for decision-making.
- To explore the functions of statistics, including data collection, organization, analysis, interpretation, and presentation.
- To create awareness about the misuse and fallacies in the application of statistical information.
- To build a foundation for further study and application of statistical methods in real-world problems.
- To encourage critical thinking when using and interpreting statistical data.

1.3 *DEFINITIONS*

The goal of a definition is to properly outline a subject's meaning, boundaries, and scope. If there are numerous definitions that highlight one or more of its components, this might not be the case. Therefore, in order to get a thorough understanding of the subject, a student needs examine at least some of them critically. These have been categorised under two primary headings: "statistics as methods" and "statistics as data."

1.3.1 STATISTICS AS DATA

In one of its fundamental meanings, **statistics refers to data**—specifically, numerical data that represent facts or pieces of information collected through observation, surveys, experiments, or administrative records. These data can describe a wide range of phenomena such as population size, income levels, exam scores, rainfall amounts, or product sales. When we say "statistics as data," we are emphasizing the **raw figures or measurements** that are collected and later used for analysis, interpretation, and decision-making. These data are typically organized in tables, charts, or graphs and serve as the foundation for statistical analysis.

Characteristics of Statistics as Data

Here are the key characteristics of statistics when understood as **data**:

1. **Numerical in Nature**:
   Statistics typically involve numbers that represent information (e.g., 75 kg, 120 students, ₹50,000 income).
2. **Aggregated Information**:
   Statistics represent group or collective data, not individual cases (e.g., average marks of a class, not one student).
3. **Collected for a Purpose**:
   Data are collected systematically to answer specific questions or solve problems.
4. **Affected by Multiplicity of Causes**:
   Statistical data are influenced by several factors at once (e.g., rainfall depends on geography, season, climate).
5. **Comparable and Classified**:
   Good statistical data can be organized into categories and compared across different groups or time periods.
6. **Condensed Form**:
   Statistics summarize large volumes of data into simple forms like averages, percentages, or graphs.
7. **Subject to Variation**:
   Data values can vary across individuals, time, or place (e.g., prices, height, income).
8. **Used for Analysis and Interpretation**:
   Statistics are not just numbers—they are used to draw conclusions and make informed decisions.
9. **Representative of a Population**:
   Good statistics reflect the characteristics of the entire group (population) being studied, often via samples.
10. **Require Proper Interpretation**: The usefulness of data depends on how accurately it is interpreted and what context it is used in.

1.3.2 STATISTICS AS A METHOD

When we talk about **statistics as a method**, we are referring to the systematic techniques and procedures used to **collect, organize, analyze, interpret, and present data**. This aspect of statistics focuses on the **tools and processes** that help in making decisions, drawing conclusions, and solving problems based on data.

Key Features of Statistics as a Method

1. **Collection of Data**: Involves planning how, when, and where to gather relevant data OR it means methods include surveys, experiments, observations, and records.
2. **Organization of Data**: Data are arranged in a structured form like tables, charts, or frequency distributions. Also, it helps in making raw data readable and understandable.
3. **Presentation of Data**: Data are presented visually through graphs, pie charts, histograms, and bar diagrams for better clarity.
4. **Analysis of Data**: Involves applying statistical tools like mean, median, standard deviation, correlation, regression, etc., to examine patterns and relationships.
5. **Interpretation of Data**: Drawing meaningful conclusions from the analyzed data and understanding what the results imply.

6. **Forecasting and Decision-Making**: Using statistical results to predict future trends and support planning, policy-making, or business strategies.

Hence, statistics as a method is the science of dealing with data in a logical and systematic way from collection to interpretation to assist in reasoning, problem-solving, and informed decision-making in various fields like economics, education, health, and social sciences.

1.4 *SUBJECT MATTER OF STATISTICS*

The subject matter of statistics broadly covers two main aspects:

1. **Statistics as Data (Descriptive Statistics):** This aspect deals with the collection, organization, presentation, and summarization of data**.**

Key Components:

- **Collection of Data**: Gathering information through surveys, experiments, or observation.
- **Organization of Data**: Arranging data into tables or classifications.
- **Presentation of Data**: Using charts, graphs, or diagrams.
- **Summarization**: Calculating measures like mean, median, mode, variance, etc.

2. **Statistics as a Method (Inferential Statistics):** This aspect focuses on drawing conclusions or making predictions based on data, often collected from samples.

Key Components:

- **Estimation**: Predicting population parameters (e.g., population mean) using sample statistics.
- **Hypothesis Testing**: Making decisions or judgments about population characteristics.
- **Correlation and Regression**: Studying relationships between variables.
- **Probability Theory**: Assessing uncertainty and randomness in data.

The subject matter of statistics includes both the data-handling aspect (descriptive) and the analytical aspect (inferential). Together, they enable the study of uncertainty, variability, and decision-making in various fields like economics, business, biology, psychology, education, and more.

1.5 *LAWS OF STATISTICS*

In statistics, several **basic laws or principles** guide the behavior of data and the methods used to interpret them. These are not "laws" in the legal or fixed scientific sense but are general principles based on logic and empirical observation.

1. Law of Statistical Regularity

- **Meaning**: A randomly selected sample is likely to exhibit the same characteristics as the population.
- **Implication**: Reliable conclusions can be drawn from well-chosen random samples.

2. Law of Inertia of Large Numbers (Also called Law of Large Numbers)

- **Meaning**: As the size of the sample increases, the results become more accurate and stable.
- **Implication**: Larger samples tend to produce more reliable and consistent estimates of population parameters.

3. Law of Persistence of Small Numbers (Less commonly used)

- **Meaning**: Small samples may not represent the population accurately and are more prone to random errors.
- **Implication**: Caution must be used when drawing conclusions from small samples.

4. Law of Probability

- **Meaning**: This forms the basis for inferential statistics, expressing the likelihood of the occurrence of an event.
- **Implication**: Probability helps to estimate confidence, error, and risk in decision-making based on sample data.

5. Law of Validity

- **Meaning**: The results obtained using statistical methods are valid only when proper procedures are followed.
- **Implication**: The reliability of statistical conclusions depends on proper data collection, unbiased sampling, and correct analysis.

6. Law of Consistency

- **Meaning**: Under similar conditions, statistical results should remain stable over repeated observations.
- **Implication**: This supports the reliability of statistical models.

7. Law of Homogeneity

- **Meaning**: Statistical methods are applicable only when the data are uniform in nature.
- **Implication**: The population or group under study should be similar or comparable in characteristics.

The laws of statistics provide foundational rules for conducting accurate, meaningful, and reliable statistical analysis. They ensure that the data collected are valid, and the conclusions drawn are scientifically sound.

1.6 *FUNCTIONS OF STATISTICS*

Statistics plays a vital role in various fields by performing several important functions. Below are the major **functions of statistics**:

1. Data Collection

- Statistics provides **systematic methods** for collecting relevant data.
- Tools: Surveys, questionnaires, experiments, observations.

2. Data Organization

- After collection, statistics helps in **arranging data** in meaningful forms.
- Tools: Tables, frequency distributions, classification, coding.

3. Data Presentation

- Makes complex data **easier to understand** through visual representation.
- Tools: Graphs, pie charts, bar diagrams, histograms, etc.

4. Data Summarization

- Condenses large data sets into **simple summaries**.
- Tools: Averages (mean, median, mode), percentages, ratios.

5. Data Analysis

- Helps to **explore patterns, relationships**, and variations in data.
- Tools: Correlation, regression, standard deviation, variance.

6. Interpretation and Conclusion

- Helps in **drawing meaningful conclusions** and making decisions based on data.
- Supports policy-making, planning, and forecasting.

7. Prediction and Forecasting

- Uses past data to **predict future outcomes**.
- Used in economics, business, weather forecasting, etc.

8. Decision-Making

- Provides a **scientific basis** for decision-making under uncertainty.
- Used in government, business, health, education, etc.

9. Testing Hypotheses

- Evaluates assumptions or claims about a population using **sample data**.
- Uses statistical tests like t-test, chi-square test, etc.

10. Control and Planning

- Helps in **monitoring quality**, setting goals, and planning operations.
- Widely used in industries for quality control and performance measurement.

The main functions of statistics include collecting, organizing, presenting, analyzing, and interpreting data for the purpose of understanding, decision-making, and forecasting across various domains.

1.7 *IMPORTANCE OF STATISTICS*

Statistics is essential in almost every field of study and practical life. Its importance lies in its ability to **turn data into meaningful information** for decision-making, analysis, and forecasting.

1. In Decision-Making

- Helps individuals, businesses, and governments make informed decisions using data analysis.
- Example: Companies decide product pricing based on market statistics.

2. In Planning and Policy Formulation

- Provides a **scientific basis** for planning and forming public policies.
- Example: Government budgets, education planning, and health programs rely on statistical data.

3. In Scientific Research

- Essential in designing experiments, analyzing results, and validating hypotheses.
- Ensures the accuracy and reliability of research findings.

4. In Economics and Business

- Used to study demand, supply, inflation, production, and market trends.
- Businesses use statistics to analyze sales, customer behavior, and performance.

5. In Forecasting

- Enables prediction of future trends based on past data.
- Example: Weather forecasting, stock market trends, and economic growth.

6. **In Education**

- Helps in evaluating student performance, curriculum effectiveness, and institutional growth.
- Educational surveys and research are based on statistical tools.\

7. **In Health and Medicine**

- Vital in clinical trials, disease tracking, and public health research.
- Example: COVID-19 spread analysis, vaccine effectiveness.

8. **In Quality Control**

- Used in industries to maintain product standards and improve processes.
- Example: Six Sigma, control charts.

9. **In Social Sciences**

- Assists in studying social issues like poverty, unemployment, literacy, etc.
- Social researchers use statistics to validate social theories.

Statistics is important because it provides objective, data-driven insights that guide understanding, problem-solving, and decision-making across every area of modern life from science and business to governance and daily living.

1.8 *LIMITATION OF STATISTICS*

While **statistics** is a powerful tool for understanding and analyzing data, it has several **limitations** that must be recognized:

1. **Does Not Reveal Individual Information**

- Statistics deals with **aggregates**, not individuals.
- It can describe the average income of a group but not a specific person's income.

2. **Cannot Establish Causation**

- Statistics shows **association or correlation**, but not direct cause and effect.
- Example: Ice cream sales and drowning incidents may be correlated, but one does not cause the other.

3. **Depends on Proper Data Collection**

- If the data collected are **biased, incomplete, or inaccurate**, the results will be misleading.

4. Can Be Misused or Manipulated

- Statistics can be **intentionally or unintentionally misused** to support false claims.
- Misleading graphs, selective data, or wrong averages can distort truth.

5. Requires Expertise

- Statistical methods need technical knowledge for correct application and interpretation.
- Improper use can lead to incorrect conclusions.

6. Limited to Quantitative Data

- Statistics primarily deals with quantitative (numerical) data, and may not fully capture qualitative aspects like emotions, opinions, or values.

7. Cannot Function Without Assumptions

- Statistical models often depend on certain assumptions (e.g., normal distribution, independence), which may not always hold true in real situations.

8. Results May Not Always Be Conclusive

- Statistical analysis may give probable conclusions**,** not absolute certainty.
- There's always a margin of error or uncertainty in interpretation.

Statistics is a useful but not foolproof tool**.** Its effectiveness depends on proper data, correct methods, and sound interpretation. Misuse or misunderstanding can lead to false or misleading conclusions**.**

1.9 *DISTRUST OF STATISTICS*

The phrase "distrust of statistics" refers to the skepticism or doubt people often feel toward statistical results. This distrust arises not because statistics are inherently false, but because of the possibility of misuse, misinterpretation, or manipulation of data. There are following reasons for distrust of statistics

1. Misuse of Statistical Methods

- Inappropriate techniques or faulty calculations can lead to wrong conclusions.
- Example: Using the mean when the median is more appropriate in skewed data.

2. Manipulation of Data

- Statistics can be presented selectively to support a particular viewpoint.
- Example: Showing only favorable survey results while ignoring the rest.

3. Lack of Understanding

- Many people lack statistical literacy, leading them to misunderstand graphs, averages, or probability.
- This causes fear or rejection of statistical findings.

4. Misleading Presentation

- Data can be distorted using tricky graphs, scales, or percentages.
- Example: A graph with a broken axis can exaggerate changes.

5. Biased or Inaccurate Data

- If the original data are biased, outdated, or collected improperly, the results will be unreliable.
- Example: Survey with a small or non-representative sample.

6. Conflicting Results

- Different studies on the same issue may show contradictory results, confusing the public.

7. Overgeneralization

- Drawing broad conclusions from limited or specific data often leads to false impressions.

Distrust of statistics arises when people sense that data are being misused, misunderstood, or manipulated. To overcome this distrust, it's essential to apply statistical methods properly, ensure transparency, and educate people on how to read and interpret data responsibly.

1.10 *FALLACIES IN STATISTICS*

Statistical fallacies are errors or misleading practices in the collection, analysis, interpretation, or presentation of data. These fallacies often lead to **incorrect conclusions**, and they can occur either unintentionally or deliberately. The common Fallacies in Statistics are as follows:

1. Faulty Generalization

- **Meaning**: Drawing broad conclusions from limited or non-representative data.
- **Example**: Surveying only 10 students and claiming it represents all students in a university.

2. Misleading Averages

- **Meaning**: Using the wrong measure of central tendency (mean, median, or mode) to support a claim.
- **Example**: A company advertises the "average salary" as ₹60,000 (mean), while most employees earn only ₹30,000 (median).

3. Improper Sampling

- **Meaning**: Drawing conclusions from a biased or unrepresentative sample.
- **Example**: Surveying only urban residents to understand national healthcare opinions.

4. Confusing Correlation with Causation

- **Meaning**: Assuming that because two variables are correlated, one causes the other.
- **Example**: Saying increased ice cream sales cause more drowning incidents (ignoring the real factor: summer).

5. Cherry-Picking Data

- **Meaning**: Selecting only favorable data while ignoring data that contradicts your point.
- **Example**: Showing only months when a product sold well, while hiding poor sales months.

6. Using Percentages Incorrectly

- **Meaning**: Misinterpreting or exaggerating changes using percentages.
- **Example**: "Profits increased by 200%!"—without revealing the original profit was very small.

7. Ambiguous Graphs

- **Meaning**: Manipulating visual presentation (scales, axes, colors) to exaggerate or hide data trends.
- **Example**: A bar graph that starts the y-axis at 90 instead of 0 to exaggerate small differences.

8. Overlooking the Margin of Error

- **Meaning**: Ignoring uncertainty or variability in estimates, especially in surveys.
- **Example**: Claiming one political candidate is more popular based on a 1% lead, when the margin of error is ±3%.

9. Post Hoc Fallacy

- **Meaning**: Assuming that if one event follows another, the first caused the second.

- **Example**: "After the new law passed, crime decreased — so the law must be the reason," without considering other factors.

Fallacies in statistics undermine the reliability of conclusions and can mislead readers or decision-makers. Recognizing these fallacies is essential for critical thinking and responsible use of statistics.

Check your progress

Problem 1: What are the characteristics of statistic as data?

Problem 2: What are the limitations of statistics?

Problem 3: What are the characteristic of statistics?

1.11 *SUMMARY*

The Introduction of Statistics unit provides a foundational understanding of statistics as a discipline. It begins by defining statistics as both a science and art of collecting, organizing, presenting, analyzing, and interpreting numerical data. The term "statistics" originates from the Latin word *"status"*, meaning a political state, reflecting its historical use in state affairs. The chapter explains the dual meaning of statistics: as numerical data (statistics as data) and as a method (statistics as a method). It highlights the main types of statistics: descriptive statistics, which deal with summarizing data, and inferential statistics, which help in making predictions or generalizations about a population based on a sample.

The chapter also explores the characteristics, importance, and limitations of statistics, its functions in various fields, and common fallacies and misuse of statistics, emphasizing the need for proper interpretation.

1.12 *GLOSSARY*

- Statistics
- Statistics as a data
- Statistics as a method
- Limitation of statistics
- Distrust of statistics
- Fallacies in statistics

1.13 *REFERENCES*

- S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
- Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
- J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.

1.14 *SUGGESTED READINGS*

- https://www.wikipedia.org.
- A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
- R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
- Jim Pitman, (1993), Probability, Springer-Verlag.
- OpenAI. ChatGPT (GPT-4 model) [Large language model]. https://chat.openai.com/
- https://archive.nptel.ac.in/courses/111/105/111105090

1.15 *TERMINAL QUESTIONS*

Long answer type questions

1: Define statistics as data and statistics as a method with suitable examples?

2: What are the limitations of statistics?

3: What are the characteristics of statistics?

Short answer type questions

1: What are the fallacies of statistics?

2: What are the importance of statistics?

Objective type questions:

1. What is the primary purpose of statistics?

A) To collect numerical data only

B) To organize data in books

C) To collect, analyze, interpret, and present data

D) To memorize facts

2. Statistics deals with:
A) Individual facts only
B) Qualitative data only
C) Aggregate of facts
D) Historical data

3. Which of the following is NOT a function of statistics?
A) Data interpretation
B) Forecasting
C) Guessing future outcomes without data
D) Summarizing data

4. The term "statistics" is used in how many senses?
A) One
B) Two
C) Three
D) Four

5. Which of the following is a characteristic of statistical data?
A) Exact and certain
B) Always qualitative
C) Uniform and consistent
D) Affected by multiple factors

6. What is the first step in the statistical method?
A) Analysis of data
B) Interpretation
C) Data collection
D) Presentation of data

7. Statistics is most useful when data are:
A) Individual values
B) Random numbers
C) Systematically organized
D) Not measurable

8. Which measure is used to summarize large data sets?
A) Table
B) Mean
C) Pie Chart
D) Sample

9. Which of the following is a limitation of statistics?
A) It helps in planning
B) It can be used in all fields
C) It deals with qualitative data
D) It may be misused

10. The field of statistics is mainly concerned with:
A) Political science
B) Collection and analysis of numerical data
C) Writing essays
D) Drawing maps

Fill in the blanks questions:

1: The word "statistics" is derived from the Latin word "**_____**" meaning a political state.

2: Statistics is a branch of mathematics that deals with the collection, classification, analysis, and interpretation of **_____**.

3: The two main types of statistics are **_____** statistics and **_____** statistics.

4: A numerical fact or figure that helps in understanding and describing a phenomenon is called a **_____**.

5: The process of arranging data in a systematic manner is called **_____**.

6: The group of individuals or items under study is called a **_____**.

7: A portion selected from the population to represent the entire group is called a **_____**.

8: A common graphical representation of data is the **_____**.

9: Statistics as a subject helps in making **_____** decisions based on data.

True and False questions

1: Statistics is only useful for mathematicians and has no real-life applications.

2: A population includes all items or individuals under consideration in a statistical study.

3: Descriptive statistics involve drawing conclusions or making inferences from data.

4: A sample is always larger than a population.

5: Statistics helps in presenting complex data in a simple and understandable form.

6: The word "statistics" is derived from the Greek word "statikos".

7: The primary function of statistics is to guess outcomes without data.

8: Data collected directly from respondents is known as primary data.

9: Inferential statistics is used to make predictions about a population based on a sample.

10: Pie charts and bar graphs are examples of statistical tools used for data presentation.

1.16 *ANSWERS*

Answer of objective questions

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

Answer of fill in the blanks

1:	Status	**2:**	Data	**3:**	Descriptive, Inferential
4:	Statistic	**5:**	Classification	**6:**	Population
7:	Sample	**8:**	Bar graph	**9:**	Informed

Answer of True and False:

1:	False	**2:**	True	**3:**	False
4:	False	**5:**	True	**6:**	False
7:	False	**8:**	True	**9:**	True
10:	True				

Unit-2: TYPES OF DATA

CONTENTS:

2.1 *INTRODUCTION*

In statistics, data refers to the information collected for analysis, interpretation, and decision-making. To analyze data effectively, it is important to understand its type and nature. The types of data unit introduce the classification of data based on various characteristics such as measurement levels, source, and structure. Broadly, data is classified into qualitative (categorical) and quantitative (numerical) types. Qualitative data describes attributes or categories (like gender, color), whereas quantitative data represents measurable quantities (like height, age, income). Quantitative data is further divided into discrete and continuous types. Recognizing the type of data is crucial because it determines the statistical tools and techniques used in analysis. This chapter lays the foundation for understanding how to collect, organize, and interpret data effectively for any statistical study.

2.2 *OBJECTIVE*

After reading this unit you will be able:

1. To understand the meaning and importance of data in statistical analysis.
2. To classify data into major types: qualitative (categorical) and quantitative (numerical).
3. To differentiate between discrete and continuous data under quantitative data.
4. To recognize nominal, ordinal, interval, and ratio scales of measurement.
5. To identify appropriate data types for various statistical tools and techniques.
6. To develop skills in collecting and organizing data based on its type.
7. To analyze the suitability of different data types in real-world research problems.
8. To avoid errors in interpretation and analysis by correctly identifying data types.

2.3 *PLANNING OF STATISTICAL SURVEY*

Planning a statistical survey is systematic processes of designing a study to collect, analyzes and interprets data to draw meaningful conclusions. A well-planned survey ensures reliable, valid and unbiased results**.** The planning involves several important steps:

1. **Define the Objectives Clearly**: Decide what you want to investigate or measure.

Example: Studying the employment rate among graduates.

2. **Define the Population**: Identify the entire group about which information is needed.

Example: All graduates in a city.

3. **Choose the Type of Data Required**

- Determine if qualitative or quantitative data is needed.

- Choose primary or secondary data sources accordingly.

4. **Decide the Method of Data Collection**
   - **Primary data**: Collected firsthand (e.g., interviews, questionnaires).
   - **Secondary data**: Taken from published sources (e.g., reports, records).
5. **Select the Sampling Method**: Choose between methods like random sampling, stratified sampling, or systematic sampling to represent the population accurately.
6. **Prepare Data Collection Tools**: Design questionnaires, interview schedules, or observation checklists.
7. **Conduct a Pilot Survey**: Test the tools on a small group to detect flaws or confusion.
8. **Collect the Data**: Gather information using the chosen tools and methods.
9. **Edit and Organize the Data**: Check for errors, inconsistencies, or incomplete entries.
10. **Analyze the Data**: Use statistical methods to process and interpret the data.
11. **Present the Results**: Display data through tables, graphs, charts, and statistical summaries.
12. **Draw Conclusions and Make Recommendations**: Based on the findings, conclusions are drawn to inform decisions or actions.

2.4 *COLLECTION OF DATA*

Collection of data is a crucial step in the statistical process. It involves gathering information to study a problem, test a hypothesis, or make informed decisions. The accuracy, reliability, and relevance of statistical results heavily depend on how the data is collected. Statistical data are generally of two kinds:

(a) Primary data collection **(b)** Secondary data collection

2.4.1 *PRIMARY DATA COLLECTION*

Primary data refers to the original data collected first hand by the investigator specifically for the purpose of a particular study or survey. This type of data is fresh, original, and tailored to the research objective, making it highly reliable and relevant.

Characteristics of Primary Data

- Collected directly from the source.
- Specific to the purpose of the study.
- Requires time, effort, and cost.
- Usually more accurate and up-to-date.

Methods of Collecting Primary Data

1. **Personal Interview Method:** The investigator asks questions face-to-face which is best for collecting detailed and qualitative data.

2. **Telephone Interview:** Data collected over the phone which is quick and low-cost**,** suitable for remote areas.
3. **Questionnaire Method:** A list of questions (structured or unstructured) is sent to respondents which can be mailed or conducted online**.**
4. **Observation Method:** The investigator observes and records behaviors or events without interaction which is useful in natural settings (e.g., studying consumer behavior in a store).
5. **Schedule Method:** Similar to a questionnaire, but filled by a trained enumerator which is used when respondents are illiterate or need assistance**.**
6. **Experimental Method:** Data collected through controlled experiments**,** often in scientific or psychological studies.

2.4.2 *SECONDARY DATA COLLECTION*

Secondary data refers to the data that has already been collected, compiled, and published by someone else for a different purpose. It is used when direct data collection is not possible or when time and resources are limited.

Characteristics of Secondary Data

- Pre-existing and readily available.
- Collected by others, not by the current investigator.
- May not be perfectly suited to the current objective.
- Needs critical evaluation before use.

Sources of collection of secondary data: There are three sources of collection of secondary data as follows:

1. **Published Sources**

- Government publications (e.g., Census, Economic Survey)
- Research journals and reports
- Books and magazines
- Newspapers
- Reports from international organizations (e.g., WHO, IMF, UN)

2. **Unpublished Sources**

- Company records
- Personal diaries and letters
- Internal government reports
- Research dissertations and theses

3. Online and Electronic Sources

- Databases (e.g., JSTOR, Google Scholar)
- Government and organizational websites
- Statistical agencies (e.g., NSSO, RBI, CSO)

Advantages of Secondary Data

- Cost-effective and time-saving
- Useful for long-term or historical analysis
- Helps in preliminary understanding of a subject
- Often includes large sample sizes

Limitations of Secondary Data

- May be outdated or irrelevant
- Risk of bias or manipulation
- May not match the specific needs of the current study
- Lack of control over data quality

Precautions While Using Secondary Data

- Check authenticity and reliability of the source.
- Ensure data relevance to the current problem.
- Verify accuracy, completeness, and timeliness.
- Understand the objective and method of original data collection.

Difference between primary and secondary data:

Basis of Comparison	Primary Data	Secondary Data
Meaning	Data collected directly by the investigator for a specific purpose.	Data already collected and published by others.
Source	Original source – collected firsthand.	Pre-existing source – obtained from books, reports, etc.
Purpose	Collected for a specific, current research problem.	Collected for another purpose, not the current study.
Time and Cost	Time-consuming and expensive to collect.	Time-saving and economical, as it's already available.
Accuracy and	Generally more accurate and relevant.	Less accurate, may not suit the

Basis of Comparison	Primary Data	Secondary Data
Reliability		specific study.
Control Over Data	Full control over method of collection.	No control over how the data was collected.
Examples	Surveys, interviews, experiments, observations.	Census reports, company records, published articles.

2.5 *METHODS OF COLLECTING PRIMARY DATA*

Primary data is collected firsthand by the investigator for a specific purpose. There are several methods to collect such data, depending on the nature of the study, objectives, and available resources.

1. Personal Interview Method

- **Description**: The investigator asks questions directly to the respondent face-to-face.
- **Best for**: In-depth, qualitative data.
- **Advantage**: High accuracy and flexibility.
- **Limitation**: Time-consuming and costly.

2. Telephone Interview Method

- **Description**: Questions are asked over a phone call.
- **Best for**: Quick surveys, especially over large geographical areas.
- **Advantage**: Saves time and travel cost.
- **Limitation**: Limited to people with phone access and may lack depth.

3. Questionnaire Method

- **Description**: A set of written questions is sent to respondents (via mail, email, or online forms).
- **Best for**: Large-scale studies.
- **Advantage**: Economical and easy to distribute.
- **Limitation**: Low response rate; answers may be incomplete or misunderstood.

4. Observation Method

- **Description**: The investigator observes and records behavior or events without direct interaction.
- **Best for**: Studying real-time activities (e.g., consumer behavior).
- **Advantage**: Avoids respondent bias.
- **Limitation**: Cannot observe thoughts or attitudes.

5. Schedule Method

- **Description**: Like a questionnaire, but **filled by a trained enumerator** rather than the respondent.
- **Best for**: Illiterate or non-cooperative respondents.
- **Advantage**: More reliable data.
- **Limitation**: Requires trained staff and is more expensive.

6. Experimental Method

- **Description**: Data is collected under **controlled conditions**, usually in laboratories or field experiments.
- **Best for**: Scientific and social science research.
- **Advantage**: High level of accuracy and control.
- **Limitation**: Not always practical or ethical in real-world scenarios.

Summary Table:

Method	Key Feature	Suitable For
Personal Interview	Face-to-face interaction	Detailed, individual responses
Telephone Interview	Oral, remote communication	Quick feedback
Questionnaire	Written questions answered by respondents	Large populations
Observation	Watching and recording behavior	Natural settings
Schedule	Investigator fills responses for respondents	Illiterate or rural populations
Experimental	Data from controlled experiments	Scientific analysis

2.6 *DRAFTING OR FRAMING OF THE QUESTIONNAIRE*

A questionnaire is a structured set of questions used to collect information from respondents. Proper drafting of a questionnaire is essential to ensure that the data collected is reliable, valid, and unbiased. A well-framed questionnaire leads to better responses and helps achieve the objectives of the study effectively.

Principles of Framing a Good Questionnaire

1. **Clear Objectives**
- Know exactly what information is needed.
- Each question should serve a specific purpose.
2. **Simplicity of Language**
- Use clear, simple, and direct language.
- Avoid jargon, technical terms, or complex sentences.
3. **Logical Order of Questions**
- Arrange questions in a logical and natural sequence.
- Start with general questions, then move to specific or sensitive ones.
4. **Brevity**
- Keep the questionnaire short and focused.
- Only include essential questions.
5. **Avoid Leading or Biased Questions:** Do not influence the respondent with your wording.
- Example (✗): "Don't you think online classes are boring?"
  Better (✓): "What is your opinion on online classes?"
6. **Use Closed and Open-Ended Questions Wisely**
- **Closed-ended**: Multiple-choice, yes/no — easier to analyze.
- **Open-ended**: Allow free responses — useful for in-depth insights.
7. **Use Mutually Exclusive Options:** Choices should not overlap.
   Example: Age groups (15–20, 21–25, 26–30)
8. **Include Demographic Questions:** Collect background information such as age, gender, income, etc., if needed.
9. **Pre-test (Pilot Survey):** Test the questionnaire on a small group to identify problems or confusion.
10. **Ensure Anonymity and Confidentiality:** Encourage honest responses by assuring respondents their data is private.

Example: A simple questionnaire section on the topic mobile phone usage

1. What is your age group?
   ☐ 15–20 ☐ 21–25 ☐ 26–30 ☐ 31 and above
2. Do you use a smartphone?
   ☐ Yes ☐ No
3. How many hours do you use your phone daily?
   ☐ <1 hour ☐ 1–3 hours ☐ 4–6 hours ☐ 7+ hours
4. What do you primarily use your phone for? ________________________ (open-ended)

2.6.1 *PRECAUTIONS REQUIRED IN THE USE OF A QUESTIONNAIRE*

A questionnaire is a widely used tool for collecting data, but its effectiveness depends on how carefully it is designed, distributed, and interpreted. If not used properly, it can lead to inaccurate, incomplete, or biased data. Hence, certain precautions must be taken to ensure its success.

Important Precautions to Follow

1. Clarity of Questions

- Use simple and unambiguous language.
- Avoid technical terms, slang, or complex grammar.
- Ensure that every respondent interprets the question the same way.

2. Avoid Leading and Biased Questions

- Do not influence the respondent's answer.
- Frame questions neutrally.
- ❌"Do you agree that online education is better?"
  ✅"What is your opinion on online education?"

3. Logical and Systematic Order

- Arrange questions in a natural sequence.
- Start with easy and general questions, then proceed to specific or sensitive ones.

4. Keep It Short and Relevant

- Avoid making the questionnaire too long, which may discourage responses.
- Include only questions that are essential to the objective.

5. Pilot Testing (Pre-testing)

- Conduct a trial run with a small group.
- Helps detect any errors, confusion, or ambiguity in the questions.

6. Ensure Anonymity and Confidentiality

- Assure respondents that their answers will be kept private.
- This encourages honest and accurate responses.

7. Use Both Closed and Open-Ended Questions Appropriately

- Closed-ended questions (e.g., Yes/No, multiple choice) are easy to analyze.
- Open-ended questions should be limited to avoid lengthy analysis.

8. Avoid Double-Barreled Questions

- Do not ask two things in one question.
- ❌"Do you use your phone for calls and internet?"
  ✅Separate into two questions.

9. Include Clear Instructions

- Tell the respondent how to answer, tick boxes, choose options, etc.

10. Avoid Too Many Personal or Sensitive Questions

- Such questions can make respondents uncomfortable or lead to non-response.

11. **Use Appropriate Language and Cultural Sensitivity**

- Be respectful of the respondent's language, customs, and values.

2.6.2 *MODEL 1: A STUDY ON THE SOCIO-ECONOMIC AND EDUCATIONAL PROBLEMS FACED BY COLLEGE STUDENTS*

Dear Friend,

As the academic year draws to a close, we understand that you may have encountered certain challenges during your time at the college. This survey is being conducted to gather valuable insights into the socio-economic and educational problems faced by students.

The primary objective is to collect constructive suggestions that can help improve the functioning and overall environment of the institution. We assure you that this survey is completely neutral and impartial—it holds no prejudice toward any individual, group, or institution.

This initiative is entirely in the interest of the student community, and your honest feedback will play an important role in enhancing the college experience for current and future students.

We kindly request you to complete this questionnaire in the spirit of cooperation, with the intent of supporting the institution where you have lived, learned, and grown.

We look forward to your prompt and sincere response.

Yours faithfully,

1: Objectives of the Survey

- To identify the socio-economic background of college students.
- To assess the financial challenges faced by students during their education.
- To analyze the educational difficulties, such as lack of resources or support.
- To understand how family income, education level, and social factors affect academic performance.
- To recommend solutions for supporting students from disadvantaged backgrounds.

2: Target Population

- All undergraduate and postgraduate students of the college.

3: Type of Data to be Collected

- Primary data, collected directly from students.

- Qualitative and quantitative data, including income levels, access to resources, academic performance, etc.

4: Method of Data Collection

- Questionnaire Method (distributed in-person or online).
- Optional: Interviews for deeper understanding.

5: Sample Questionnaire Outline

Section A: Personal & Socio-Economic Background

1. Age: ______

2. Gender: ☐ Male ☐ Female ☐ Other

3. Family's Monthly Income:
☐ < ₹10,000 ☐ ₹10,001–₹25,000 ☐ ₹25,001–₹50,000 ☐ > ₹50,000

4. Father's Education Level:
☐ No Formal Education ☐ Primary ☐ Secondary ☐ Graduate ☐ Postgraduate

5. Type of Residence: ☐ Urban ☐ Rural

Section B: Educational Problems

1. Do you have access to a personal study space at home? ☐ Yes ☐ No

2. Do you have regular access to the internet for academic work? ☐ Yes ☐ No

3. Do you find college textbooks or materials too expensive? ☐ Yes ☐ No

4. How many hours per week do you spend on self-study? ______

5. Are you satisfied with the support received from faculty? ☐ Yes ☐ No

Section C: General Opinions and Suggestions

1. What are the main difficulties you face in your studies? ____________________

2. What kind of support do you think the college should provide? _______________

6: Tools for Analysis

- Tabulation, bar graphs, pie charts and percentages.
- Cross-tabulation to find links between income level and academic performance.

7: Expected Outcomes

- Understanding the extent of socio-economic and educational issues.
- Identifying students in need of financial or academic assistance.
- Providing data to improve college support programs (like scholarships, counseling, etc.).

2.6.3 *MODEL 2: QUESTINNAIRE ON PRACTICAL TRAINING FOR ARTICLED CLERKS*

Section A: Personal Details

1. Name (Optional): ______________________________
2. Age: _______
3. Gender: ☐ Male ☐ Female ☐ Other
4. Course Enrolled: ☐ CA ☐ CS ☐ CMA ☐ Other: ____________
5. Current Year of Articleship: ☐ 1st ☐ 2nd ☐ 3rd
6. Name of Training Firm/Organization (Optional): ______________________

Section B: Nature and Scope of Training

7. What is the size of the firm where you are training?
   ☐ Proprietorship ☐ Partnership ☐ Mid-sized firm ☐ Big 4/large corporate
8. What kind of work are you mostly involved in? (Tick all that apply)
   ☐ Audit ☐ Taxation ☐ Compliance ☐ Accounting ☐ ROC work ☐ Client handling
   ☐ Others: _____________
9. Are you getting exposure to practical aspects of your syllabus?
   ☐ Yes, extensively ☐ Somewhat ☐ Not at all
10. Are you rotated among different departments during training?
    ☐ Yes ☐ No
11. Does your principal/senior guide you during assignments?
    ☐ Always ☐ Sometimes ☐ Rarely ☐ Never
12. Are you allowed to visit client sites or interact with clients?
    ☐ Frequently ☐ Occasionally ☐ Never

Section C: Duration and Supervision

13. Do you maintain a regular work schedule (e.g., 9 to 5)?
    ☐ Yes ☐ No
    If no, what are your usual hours? ____________
14. Are your working hours compliant with ICAI/ICSI guidelines?
    ☐ Yes ☐ No ☐ Not sure
15. How would you rate the supervision and support provided by your senior/principal?
    ☐ Excellent ☐ Good ☐ Average ☐ Poor

Section D: Learning & Skill Development

16. Have you improved your technical knowledge during articleship?
    ☐ Significantly ☐ Moderately ☐ Slightly ☐ Not at all
17. Have you developed soft skills (communication, teamwork, etc.)?
    ☐ Yes ☐ No ☐ To some extent

18. Are you encouraged to take part in discussions, reviews, or meetings?
☐ Yes ☐ No

19. Are you allowed time and support for exam preparation?
☐ Sufficient time ☐ Limited time ☐ No time given

Section E: Challenges and Suggestions

20. What are the major challenges you faced during your training?

21. What improvements would you suggest in the articleship/practical training system?

Declaration (Optional)

I declare that the above responses are true to the best of my knowledge and provided in good faith to improve the practical training system.

Signature (Optional): ________________
Date: ________________

2.7 *SOURCES OF SECONDARY DATA*

Secondary data refers to information that has already been collected, processed, and made available by someone else for purposes other than the current research. It is a cost-effective and time-saving way to gather background or supportive information.

1. Published Sources: These are publicly available and often come from reliable and official institutions. The main sources of published sources are:

Government Publications

- Census reports
- Economic surveys
- Statistical abstracts
- Reports by ministries (e.g., Health, Education, Finance)
- RBI, SEBI, NSSO publications (India-specific)

International Organizations

- United Nations (UN)
- World Bank
- International Monetary Fund (IMF)
- World Health Organization (WHO)
- UNESCO, FAO, etc.

Research and Academic Publications

- Journals and periodicals
- Theses and dissertations
- Conference proceedings
- University reports

Magazines and Newspapers

- Articles and reports related to business, economy, and society
- Editorial columns and news analyses

Books

- Reference books, textbooks, and handbooks on various subjects

2. Unpublished Sources: These are not publicly available but may be accessed with permission. The main sources of unpublished sources are:

Company and Business Records

- Internal reports
- Sales records
- Employee data
- Financial statements

Private Institutions or NGOs

- Survey results
- Project reports
- Case studies

Government Departments

- Unpublished internal surveys
- Communication between departments

3. Electronic and Online Sources: The main sources of electronic and online sources are:

Digital Databases and Websites

- Government portals (e.g., data.gov.in, indiastat.com)
- Research databases (e.g., JSTOR, Google Scholar, ResearchGate)
- Online newspapers and journals
- Organizational websites (e.g., WHO.org, RBI.org.in)

Statistical Databases

- World Bank Data Portal
- IMF eLibrary
- UN Data
- Eurostat
- Statista

2.7.1 *SCRUTINY OF SECONDARY DATA*

Scrutiny of secondary data means carefully examining and evaluating the data collected from other sources to ensure that it is reliable, relevant, accurate, and suitable for the current research purpose. Since secondary data was collected for a different objective, it must be scrutinized before use.

Key Points in Scrutinizing Secondary Data

1. **Reliability of the Source**
- Is the data published by a credible and authoritative source**?**
- Government and reputed international organizations (e.g., UN, RBI, WHO) are considered trustworthy.
- Avoid using data from unknown or biased sources**.**

2. **Suitability of the Data**
- Is the data relevant to your current research problem?
- Check if the definitions, classification, and units used match your requirements**.**

3. **Adequacy of the Data**
- Is the amount or scope of data sufficient for your study?
- Ensure that the sample size, time period, and coverage are appropriate.

4. **Accuracy and Consistency**
- Cross-check data with other sources for consistency.
- Look for errors, outliers, or contradictions in the dataset.

5. **Timeliness or Currency**
- Is the data recent enough to be valid for your analysis?
- Outdated data may not reflect current trends or conditions.

6. **Objective and Bias of the Original Data Collection**
- Understand the purpose for which the data was originally collected.
- Watch for intentional or unintentional bias based on who collected it and why.

7. **Methodology Used**
- What methods were used to collect the data?
- Was it through sampling, surveys, or estimation?
- This affects data reliability.

Importance of Scrutiny

- Ensures valid and sound conclusions.
- Prevents misleading interpretations.
- Helps in selecting the most appropriate data for research.

2.8 *UNIT OF INQUIRY*

The Unit of Inquiry refers to the individual element or entity from which data is collected during a statistical investigation. It is the basic unit about which information is gathered and analyzed.

Definition: Unit of Inquiry is the person, object, event, or group about which data is collected for the purpose of statistical analysis.

Examples:

Study Objective	Unit of Inquiry
Studying the literacy rate in a village	Individual person
Analyzing production in factories	Each factory
Surveying satisfaction among college students	Each student
Studying crop yield in different regions	Each region or farm
Studying family income in a city	Each household

2.8.1 *PROCESSING OF DATA*

Processing of data refers to the systematic organization, editing, coding, classification, and tabulation of raw data so that it becomes meaningful, accurate, and ready for analysis. It is a crucial step in any statistical investigation after the collection of data.

Stages in the Processing of Data

1. **Editing**

- Involves checking the completeness, consistency, and accuracy of the collected data.
- Helps to remove errors, omissions, or irrelevant entries.
- Two types:
  - **Field Editing**: Done immediately after collection.
  - **Central Editing**: Done after all data has been collected.

2. Coding

- Assigning numerical or symbolic codes to responses (especially in open-ended or categorical questions).
- Example: Gender – Male = 1, Female = 2, Other = 3
- Makes data easier to input and analyze using statistical tools.

3. Classification

- Grouping data into homogeneous categories or classes.
- Can be based on characteristics (like age, income) or attributes (like yes/no, urban/rural).
- Two types:
  - **Quantitative Classification** – based on measurable characteristics.
  - **Qualitative Classification** – based on descriptive characteristics.

4. Tabulation

- Presenting the data in systematic tables with rows and columns.
- Helps in summarizing large amounts of data in a clear and concise manner.
- Two types:
  - **Simple Tabulation**: One characteristic at a time.
  - **Complex Tabulation**: Two or more characteristics at once.

5. Transcription / Data Entry

- Entering the cleaned and coded data into a computer or spreadsheet for statistical analysis.

Check your progress

Problem 1: With the help of real-life examples, explain the significance of classifying data into appropriate types in research and data analysis.

Problem 2: State any two uses of identifying the type of data in statistics.

2.9 *SUMMARY*

The "Types of Data" units in statistics categorizes data into different forms based on their nature and level of measurement. Broadly, data is classified into qualitative (categorical) and quantitative (numerical) types. Qualitative data includes nominal data (labels without order, like gender or nationality) and ordinal data (categories with a meaningful order, like education level). Quantitative data includes discrete data (countable values, like number of students) and continuous data (measurable values, like height or weight). Based on measurement levels, data can also be classified as nominal, ordinal, interval and ratio**,** with ratio being the highest level as it allows all mathematical operations including a true zero. Understanding these types is essential for choosing appropriate statistical tools and analysis methods.

2.10 *GLOSSARY*

- Data
- Statistical survey
- Primary data
- Secondary data
- Drafting of questionnaire
- Unit of inquiry

2.11 *REFERENCES*

- S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
- Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
- J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.

2.12 *SUGGESTED READINGS*

- https://www.wikipedia.org.
- A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
- R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
- Jim Pitman, (1993), Probability, Springer-Verlag.
- OpenAI. ChatGPT (GPT-4 model) [Large language model]. https://chat.openai.com/
- https://archive.nptel.ac.in/courses/111/105/111105090

2.13 *TERMINAL QUESTIONS*

Long answer type questions

1. Explain the different types of data in statistics with suitable examples.

2. What are the four levels of measurement in statistics? Explain each level in detail with examples.

3. Compare and contrast qualitative and quantitative data. How are they collected, represented, and used in analysis?

4. Discuss the importance of understanding data types in statistical analysis.

5. Describe the characteristics of nominal, ordinal, interval, and ratio scales. Which types of statistical operations are appropriate for each?

Short answer type questions

1. What is nominal data? Give two examples.
2. Define ordinal data with a suitable example.
3. Differentiate between qualitative and quantitative data.
4. Give two real-life examples each of discrete and continuous data.
5. Explain the meaning of interval scale. How is it different from ratio scale?
6. Why can't we calculate the mean of nominal data?
7. Which type of data scale allows all mathematical operations? Give one example.
8. What is the main characteristic of ordinal data?

Objective type questions:

1. Which of the following is NOT a type of data in statistics?
a) Qualitative data
b) Quantitative data
c) Experimental data
d) Ordinal data

2. Which type of data represents names or labels without any order?
a) Ordinal
b) Ratio
c) Nominal
d) Interval

3. Which of the following types of data is classified as quantitative?
a) Gender
b) Age
c) Religion
d) Nationality

4. Ordinal data differs from nominal data because it:
a) Uses numbers
b) Can be measured on a scale
c) Has a meaningful zero
d) Has a logical order

5. Which of the following is an example of continuous data?
a) Number of cars
b) Number of siblings
c) Height of a person
d) Number of houses

6. In which level of measurement is addition and subtraction possible but not multiplication or division?
a) Ratio
b) Interval

c) Nominal
d) Ordinal

7. Which type of data uses categories that can be ranked but not measured?
a) Nominal
b) Ordinal
c) Ratio
d) Continuous

8. Data that can take only specific values (like whole numbers) is called:
a) Continuous
b) Discrete
c) Nominal
d) Interval

9. Which of the following variables is measured on a ratio scale?
a) Temperature in Celsius
b) Age
c) Political party
d) Educational level

10. Which type of data is suitable for computing mean, median, and mode?
a) Nominal
b) Ordinal
c) Ratio
d) Categorical

Fill in the blanks questions:

1. Data that consists of names, labels, or categories is called ___________ data.
2. Data that can be ranked or ordered but does not have meaningful differences between values is known as ___________ data.
3. ___________ data can take any value within a given range and is measured, not counted.
4. The data type that has a true zero point and allows all mathematical operations is called ___________ scale.
5. The ___________ level of measurement has equal intervals but no true zero point.
6. The number of students in a class is an example of ___________ data.
7. Gender, nationality, and blood type are examples of ___________ data.
8. The weight of a person is an example of ___________ data.
9. Ordinal data has a meaningful order but not a consistent ___________ between values.
10. ___________ data represents variables that can be counted and have exact values, like the number of cars.

True and False questions

1. Nominal data can be arranged in a meaningful order.
2. Ordinal data shows both classification and order.
3. Interval data has a true zero point.

4. The number of books on a shelf is an example of discrete data.
5. Continuous data is always counted and not measured.
6. The ratio scale allows for the calculation of all arithmetic operations including multiplication and division.
7. The nominal scale is the highest level of data measurement.
8. Temperature in Fahrenheit is an example of interval data.
9. Ordinal data can be used to compute the mean of the data.
10. Qualitative data can be either nominal or ordinal.

2.14 *ANSWERS*

Answer of objective questions

1:	c	**2:**	c	**3:**	b	**4:**	d
5:	c	**6:**	b	**7:**	b	**8:**	b
9:	b	**10:**	c				

Answer of fill in the blanks

1:	Nominal	**2:**	Ordinal	**3:**	Continuous
4:	Ratio	**5:**	Interval	**6:**	Discrete
7:	Nominal	**8:**	Continuous	**9:**	Difference
10:	Discrete				

Answer of True and False:

1:	False	**2:**	True	**3:**	False
4:	True	**5:**	False	**6:**	True
7:	False	**8:**	True	**9:**	False
10:	True				

COURSE NAME: BASIC STATISTICS

COURSE CODE: MT(N)-222

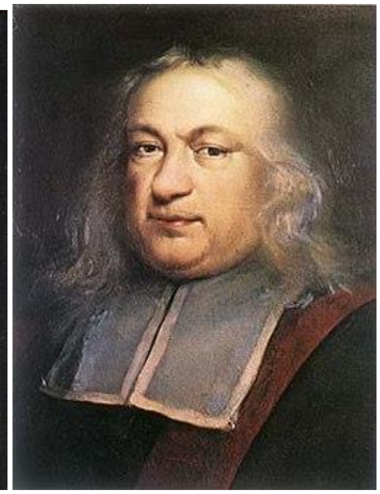
BLOCK - II: PROBABILITY

UNIT-3 PROBABILITY

CONTENTS:

3.1. INTRODUCTION

Dear learners, present unit are a presentation of basic concepts of probability. In previous units we have studied the concepts of Basic Statistics. Now, first we start with the history of probability *Galileo (1564-1642)*, an Italian mathematician, was the first researcher solved the some problems related to the theory of dice in gambling. But the basic foundation of the mathematical theory of probability was explained by two French mathematicians, *B. Pascal (1623-62)* and *P. Fermat (1601-65)*. *Chevalier-De-Mere* put a question in front of Pascal. The famous "problem of points" posed by De-Mere to Pascal is: "Two persons play a game of chance. The person who first gains a certain number of points win the stake. They stop playing before the game is completed. How is the stake to be decided on the basis of the number of points each has won?" The two mathematicians, after a lengthy correspondence between themselves ultimately solved this problems correspondence laid the first foundation of the science of probability.



Blaise Pascal (France) **Pierre de Fermat** (France)
(19 June 1623 – 19 August 1662) (17 August 1601 -12 January 1665)

Fig 3.1

Ref: https://www.wikipedia.org/

Next researcher in this field was *J. Bernouli* (1654-1705) whose "Treatise on Probability" was published posthumously by his nephew *N. Bernouli* in 1713. *De-Movire* (1667-1754) also did. Other main contributors are: *T. Bayes* (Inverse probability), *P. S. Laplace* (1749-1827) who after extensive research over a number of years finally published "Theoric analytique des probabilities" in 1812. In addition to these, other outstanding contributors are *Levy*, *Mises* and *R. A. Fisher*. Russian mathematician also have made very valuable contributions to the modern theory of probability. Chief contributors, to mention only a few of them and *Chebyshev's* (1821-1922); who founded the Russian School of Statisticians; *A. Markov* (1856-1922); Liapounoff (Central Limit Theorem); *A. Khintchine* (Law of large numbers) and *A.Kolmogorov*, who axiomised the calculus of probability.

3.2. OBJECTIVES

After studying this unit learner will be able to:

1. Describe the notion of probability.
2. Explain the basic terminology, which are used in the definition of probability.
3. Evaluate the probability related to basic random experiments.
4. Solved the examples and theorems related to probability.

3.3. BASIC TERMINOLOGY

In this section we shall explain the various terms which are used in the definition of probability under different approaches.

i. Random Experiment, Outcome and Trial. Consider an experiment whose outcome is not unique, but it has possibility of several outcomes. This type of experiment is known as **random experiment.** The result of a random experiment will be called an **outcome.** The random experiment can be repeated under identical conditions. Each repetition/ performance is called a **trial.**

Examples.

a. Rolling a dice is a random experiment and each rolling of this dice is a trial. A die is a small cube used in gambling. On its six faces, dots are marked as

Fig 3.3.1
Ref: https://google.com

Plural of die is dice. The outcome of throwing a die is the number of dots on its uppermost face. In a trial of a random experiment, any outcome among the all possible outcomes is known as elementary event. Tossing a coin give a outcomes as "head" or "tail". Here outcome head is an elementary event, similarly outcome tail is also an elementary event.

b. A pack of card consists of four suits called Spades, Hearts, Diamonds and Clubs. Each consists of 13 cards, of which nine card are numbered from 2 to 10, an ace, king, a queen and jack (or knave). Spades and clubs are black – faced cards, while hearts and diamonds are red faced cards. A standard pack of cards contains a total of 52 cards. This is calculated by multiplying the number of suits by the number of cards in each suit: $4\times13=52$. The cards are divided into 4 suits: Spades: 13 cards (black), Hearts: 13 cards (red), Diamonds: 13 cards (red), Clubs: 13 cards (black).

Each suit contains the following cards:

- Numbered cards: 2,3,4,5,6,7,8,9,10 (9 cards)
- Face cards: Jack, Queen, King (3 cards)
- Ace (1 card)

- The face cards are also known as court cards.
- Spades and Clubs are black suits.
- Hearts and Diamonds are red suits.
- The total number of black cards is 26 (Spades and Clubs).
- The total number of red cards is 26 (Hearts and Diamonds).

Fig 3.3.2
Ref: https://google.com

ii. Sample Space. The collection of all possible outcomes in a random experiments is known as **sample space**. It is usually denoted as "S". In probability theory, the sample space of an experiment or random trial is the set of all possible outcomes or results of that experiment. A sample space is usually denoted using set notation, and the possible ordered outcomes, or sample points, are listed as elements in the set.

a. Rolling a dice and observing the number showing on dice, the sample space is:

S= {1, 2, 3, 4, 5, 6}.

b. Rolling two dices simultaneously and observing the number showing on dices, the sample space is:

S={(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)}.

c. Rolling three coins simultaneously and observing the face of coins, the sample space is:

S = {HHH HHT, HTH, THH, , HTT, THT, TTH, TTT }.

Remark 1. A sample space can be infinite countable or uncountable set. For example consider a random experiment "tossing a coin until a head comes". Then sample space

S = {H, TH, TTH, TTTH,…}.

Remark 2. Any subset of sample space is known as event. It is usually denoted by capital alphabet A, B, E, E_1, E_2 ,…

Remark 3. A sample space is called discrete if "S" is countable, otherwise it is called continuous sample space.

Example 6. Let the random experiment be "rolling two dices simultaneously". Then the sample space is:

S=$\{$(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)$\}$.

Now consider the following events, E_1 is event that sum of both face is six, E_2 is event that sum of both face is seven, E_3 is event that sum of both face is either six or seven.

Then

E_1=$\{$(1,5), (2,4), (3,3), (4,2), (5,1)$\}$,

E_2=$\{$(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$\}$

and

$E_3 = E_1 \cup E_2$=$\{$(1,5),(1,6), (2,4), (2,5), (3,3), (3,4), (4,2), (4,3), (5,1), (5,2), (6,1)$\}$.

We assign a numerical value to finite discrete sample space and its events as follows

$n(S)$ = Total number of elements in sample space S.

$n(E)$ =Total number of elements in event E

CHECK YOUR PROGRESS

CYP1. Let one card drawn from a pack of 52 cards and E_1be the event that card is red and E_2be the event that card is black ace. Then write total number of elements in $E_1$and E_2.

CYP2. What is the set of all possible outcomes of a random experiment called?

a) Event
b) Sample Space
c) Trial
d) Result

CYP3. Which of the following is an example of a random experiment?

a) Measuring the height of a building.

b) Flipping a coin.
c) Calculating the area of a circle.
d) Observing the sunset.

CYP4. Which of the following is not a characteristic of a random experiment?

a) It is repeatable under identical conditions.
b) The outcome is unpredictable.
c) The outcome is always the same.
d) It can be repeated arbitrarily often.

iii. Exhaustive Events and Favourable events. The total number of possible outcomes of a random experiment is known as the **exhaustive events**. Two events are exhaustive when their union is equal to the sample space.

The number of cases **favourable** to an event in a trial is the number of outcomes which entail the happening of the event.

Example 7. When rolling a six-sided die, the events 1, 2, 3, 4, 5, and 6 are collectively exhaustive, because they encompass the entire range of possible outcomes.

Example 8. When rolling a die and want to get an even number, the favorable events would be rolling a 2, 4, or 6.

iv. Mutually Exclusive Events. Events are said to be **mutually exclusive or in compatible** if the happening of any one of them precludes the happening of all the others. In probability, two events are mutually exclusive (or disjoint) if they cannot occur at the same time.

This means their intersection is empty, and the probability of both events happening simultaneously is zero.

Example 9. Flipping a coin and getting both heads and tails is impossible, making those events mutually exclusive.

v. Equally Likely Events. Equally likely events in probability are events that have an equal chance of occurring.

If events A and B are **equally likely,** it means that $P(A) = P(B)$.

Example 10. In throwing an unbiased die, all the six faces are equally likely to come.

vi. Independent Events. Two events are said to be **independent** when the occurrence of one event does not affect the probability of the occurrence of the other event.

Example 11. Flipping a coin and rolling a die are independent events. The outcome of flipping a coin does not influence the outcome of rolling a die. The outcome of one event (e.g., flipping a coin) does not affect the outcome of the other event (e.g., rolling a die).

The independent probability, means that the probability of one event happening does not change based on whether the other event has occurred or not.

CHECK YOUR PROGRESS

CYP5. Which of the following is NOT an example of mutually exclusive events?

a. Flipping a coin and getting heads or tails.
b. Rolling a die and getting an even number or an odd number.
c. Drawing a card from a deck and getting an ace and a king.
d. Choosing a student from a class and having them be both male and female.

CYP6. Mutually Exclusive events ________

a Contain all sample points
b Contain all common sample points
c Does not contain any common sample point
d Does not contain any sample point

3.4. PROBABILITY

If a random experiment or a trial results in $'n'$ exhaustive, mutually exclusive and equally likely outcomes, out of which $'m'$ are favourable to the occurrence of an event E, then the probability $'p'$ of occurrence (or happening) of E, usually denoted by $P(E)$, is given by;

$$p = P(E) = \frac{\text{Number of favourable cases}}{\text{Total number of exhaustive cases}} = \frac{m}{n} \ldots\ldots\ldots\ldots\ldots \textbf{(3.4.1)}$$

Total number of ways of happening of event E is number of cases favourable to the event E and also called number of sampling points in E similarly total number of outcomes is number of sample points in E and also called exhaustive number of cases.

- Two unbiased dice are thrown.

  Find the probability that:

  i. Both the dice show the same number.
  ii. The first die shows 6,
  iii. The total of the numbers on the dice is 8,
  iv. The total of the numbers on the dice is greater than 8,
  v. The total of the numbers on the dice is 13, and
  vi. The total of the numbers on the dice is any number from 2 to 12, both inclusive.

Solution. In a random throw of two dice, since each of the six faces of one die can be associated with each of six faces of the other die, the total number of cases is $6 \times 6 = 36$, as given below:

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Here, the expression say, (i, j) means that the first die shows the number i and the second die shows the number j. Obviously, $(i, j) \neq (j, i)$ if $i \neq j$. Therefore exhaustive number of cases $(n) = 36$.

Solution (i) the favourable cases that both the dice show the same number are: $(1,1), (2,2), (3,3), (4,4), (5,5)$ and $(6,6), i.e., m = 6$.

Therefore probability that the two dice show the same number $= \frac{6}{36} = \frac{1}{6}$.

Solution (ii) the favourable cases the first die shows 6 are :
$(6,1),(6,2),(6,3),(6,4),(6,5)$ and $(6,6), i.e., 6$ in all.

Therefore probability that the first dice shows $'6' = \frac{6}{36} = \frac{1}{6}$.

Solution (iii) the favourable cases to getting a total of 8 on the two dice are:
$(2,6),(3,5),(4,4),(5,3),(6,2)$, $i.e., m = 5$.

Therefore probability that the total numbers on two dice is $'8' = \frac{5}{36}$.

Solution (iv) the favourable cases to getting a total of more than 8 are:
$(3,6),(6,3),(4,5),(5,4),(4,6),(6,4),(5,5),(5,6),(6,5),(6,6),)$
$, i.e., m = 10$.

Therefore probability that the total numbers on two dice is greater than
$'8' = \frac{10}{36} = \frac{5}{18}$.

In solution (v) we have to find out the probability of a events, the total of the numbers on the dice is 13.

This is impossible because the total is maximum $6 + 6 = 12$. Therefore, the required probability is 0.

In a solution (vi) the probability is 1, as the total of the numbers on the two dice certainly from 2 to 12. The given event is called a **certain event**. (This means there is a 100% chance that the event will occur).

Remarks 1. $0 \leq P(E) \leq 1$.

Remarks 2. The non – happening of the event E is called the complementary event of E and is denoted by $\bar{E}$ or E^c.

$$P(E) + P(E^c) = 1 \text{ or } P(E) = 1 - P(E^c).$$

Remarks 3. Probability $'p'$ of the happening of an event is also known as the probability of success the probability $'q'$ of the non-happening of the event as the probability of failure.

i.e., $p + q = 1$.

Remarks 4. If $P(E) = 1$, E is called a certain event and if $P(E) = 0$, E is called an impossible event.

3.4.1. LIMITATIONS

Probability requires ideal conditions, Probability does not predict the future probability models can only provide estimates of the likelihood of rain, but they cannot predict the exact amount of rainfall or the location where it will occur. Probability models can be misinterpreted, leading to wrong conclusions and decisions.

3.4.2. TRANSLATION OF EVENTS IN SET THEORY OPERATION

Let S be a sample space, $E_1$and $E_2$are events. Then there is formulation of new events using$E_1$and E_2. For this consider the following table:

S. No.	Event	Meaning
1.	E_1^c	Event that E_1 does not occur
2.	$E_1 \cup E_2$	Event that E_1 or E_2 occur
3.	$E_1 \cap E_2$	Event that both E_1 and E_2 occur
4.	$E_1 \cap E_2^c$	Event that E_1 occur but E_2does not occur
5.	$E_1 \Delta E_2 = (E_1 \cap E_2^c) \cup (E_1^c \cap E_2)$	Event that exactly one of the events E_1 or E_2 occurs
6.	$E_1^c \cap E_2^c$	Event that none of the events E_1 or E_2 occur
7	$E_1 \cap E_2 = \emptyset$	Events that both E_1 and E_2 are mutually exclusive.
8	Universal set S	Sample space

3.4.3. *AXIOMATIC PROBABILITY*

A purely mathematical definition of probability cannot give us the actual value of $P(A)$, the the probability of occurrence of the event A and this must be considered as a function defined on all events.

Probability function: Let S be a sample space and Ω be σ-field of events.
Then the set function $P: \Omega \rightarrow [0,1]$ is said to be probability function, if it satisfies the following conditions:

i. $P(A) \geq 0$, for every events A in Ω, (Axiom of non-negativity).
ii. $P(S) = 1$,(Axiom of certainty).
iii. If $E_i, i = 1,2,3, \dots n$ are mutually disjoint events ($E_i \cap E_j, i \neq j$), then

$$P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(A_i).\text{(Axiom of additivity)}.$$

Here (S, Ω, P) is called probability space.
Let us consider a random experiment whose sample space S finite number of points, say N, given by:
$S = \{e_1, e_2 \dots\dots e_N\}$, are obviously $e_1, e_2 \dots\dots e_N$ are mutually disjoint events.

Let us now suppose that all the outcomes of random experiment are equally likely so the probability of each of the singleton events $\{e_i\}$ is $\frac{1}{N}$, $i.e.$,
$P(e_i) = \frac{1}{n}; i = 1,2, \dots\dots, n$

Further, let an event A consist of k distinct points of S, say $e_{i1}, e_{i2}, \dots\dots e_{ik}$.

Then by the axiom of additivity, we get:

$$P(A) = \sum_{j=1}^{k} P(e_{ij}) = \sum_{j=1}^{k} \frac{1}{N} = \frac{k}{N}.$$

$$P(A) = \frac{\text{Number of sample points in } A}{\text{Number of sample points in } S}$$
$$= \frac{\text{n(A)}}{\text{n(S)}}$$
$$= \frac{\text{Number of cases favourable to the event } A}{\text{Exhaustive number of cases}}$$

Which is the result obtained by the classical definition of probability.

Hence, the classical probability may be regarded as a special case of axiomatic probability.

Theorem 1. Probability of the impossible event is zero,

i.e., $P(\emptyset) = 0$ ……(1)

Proof. Impossible event contains no sample point and hence the certain event S and the impossible event $\emptyset$ are mutually exclusive.

Therefore $S \cup \emptyset = S$ it implies that $P(S \cup \emptyset) = P(S)$.

Hence, using Axiom of Additivity,

we get

$$P(S) + P(\emptyset) = P(S).$$

It implies that $P(\emptyset) = 0$.

Remark : The means of $P(A) = 0$, does not imply that A is necessarily an empty set. In practice, probability '0' is assigned to the events which are so rare that they happen only once in a lifetime.

Theorem 2. Let E be an event. Then probability of E^c is:

$$P(E^c) = 1 - P(E) \text{.................................(2)}$$

Proof: Since

$$(E \cup E^c) = S,$$

Therefore

$$P(E \cup E^c) = P(S).$$

This implies that

$$P(E) + P(E^c) = 1.$$

And hence,

$$P(E^c) = 1 - P(E).$$

Theorem . $P(E_1 \cap E_2^c) = P(E_1) - P(E_1 \cap E_2)$……(3)

Proof: Since $(E_1 \cap E_2^c)$ and $(E_1 \cap E_2)$ are disjoint sets whose union is E_1.

Therefore,

$$P(E_1 \cap E_2^c) + P(E_1 \cap E_2) = P(E_1).$$

And hence,

$$P(E_1 \cap E_2^c) = P(E_1) - P(E_1 \cap E_2).$$

Theorem . (Probability of union of two events):

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

…………………………..(4)

Proof: Since $(E_1 \cup E_2)$ can be written as disjoint union of E_1 and $E_1^c \cap E_2$.

Therefore

$$P(E_1 \cup E_2) = P(E_1) + P(E_1^c \cap E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

Theorem (For n events):

$$P\left(\bigcup_{i=1}^{n} E_i\right) \geq \sum_{i=1}^{n} P(E_i) - \sum_{1 \leq i < j \leq n} P\left(E_i \cap E_j\right)$$

………..(5)

3.5. *EXAMPLES*

Example 12: A random arrangement of the letters E,N,G,I,N,E,E,R,I,N,G is done. Find the probability of vowels appear together.

Solution: In the letters E, N, G, I, N, E, E, R, I, N, G there are 3 E's, 2 I's, 2 G's and 3 N's.
Therefore total number of different combination formed is:

$$\frac{11!}{3! \times 2! \times 2! \times 3!}$$

Now total number of different combination formed so that vowels appear together is:

$$\frac{7! \times 5!}{3! \times 2! \times 2! \times 3!}$$

Thus the required probability is:

$$\frac{7! \times 5!}{11!} = \frac{1}{66}$$

Example 13: A letter of the English alphabet is chosen at random. Calculate probability that the letter so chosen:

i. is a vowel,
ii. precedes m and is a vowel,
iii. follows m and is a vowel.

Solution: The sample space of the experiment is:

$$S = (a, b, c, d \ldots \ldots, x, y, z), n(S) = 26.$$

i. Let $E_1 = \{a, e, i, o, u\}; n(E_1) = 5$.

Therefore, $P(E_1) = \frac{n(E_1)}{n(S)} = \frac{5}{26}$.

ii. Let E_2 be the event that the letter precedes m and is a vowel. Then:

$$E_2 = \{a, e, i\}; n(E_2) = 3.$$

Therefore, $P(E_2) = \frac{n(E_2)}{n(S)} = \frac{3}{26}$.

iii. Let E_3 be the event that the letter precedes m and is a vowel. Then:

$$E_2 = \{o, u\}; n(E_3) = 2.$$

Therefore, $P(E_3) = \frac{n(E_2)}{n(S)} = \frac{2}{26} = \frac{1}{13}$.

Example 14: From the numbers 1,2,3, ... ,8 two different digits are chosen randomly without replacement. What is the probability that the sum of the digits will be equal to five.

Solution: Total number of ways of choosing two digits from the numbers 1 to 8 is: 8×7=56.

Total ways of choosing two digits so that sum of the digits will be equal to five are : (1,4),(2,3),(3,2) and (4,1).

Hence total number of ways of choosing such that sum will be five is 4.

Thus the required probability is: $\frac{4}{56} = \frac{1}{14}$.

Example 15: From a pack of 52 cards, seven cards are drawn randomly. Find the probability of four cards will be black and three cards will be red?

Solution: Total number of ways of drawing seven card from52 cards is: $\binom{52}{7}$.

Now total number of ways of choosing seven card such that four cards will be black and three cards will be red is $\binom{26}{4} \times \binom{26}{3}$.

Thus the required probability is: $\frac{\binom{26}{4} \times \binom{26}{3}}{\binom{52}{7}} = \frac{1625}{5593}$. .

Example 16: Let(S, Ω, P) is a probability space and $\Phi \in \Omega$ be empty set. Then find $P(\Phi)$.

Solution: We know that Φ and S are trivially disjoint sets.

Therefore,

$$1 = P(S) = P(S \cup \Phi) = P(S) + P(\Phi) = 1 + P(\Phi)$$

this gives $P(\Phi) = 0$.

Example 17: Let $S = \{1,2,3, \dots, n, \dots\}$ be a sample space and Ω be power set of S and for $E \in \Omega, P(E) = \sum_{n \in E} \frac{1}{2^n}$.
Then show that (S, Ω, P) is a probability space.

Solution*:* Since sum of positive real number is positive,
therefore $P(E) \geq 0$, for every events E in Ω and

$$P(S) = \sum_{1 \leq n \leq \infty} \frac{1}{2^n} = 1.$$

As we know that in a convergent series rearrangement does not alter the series sum, therefore third axiom follows.

Example 18: An integer is chosen at random from two hundred digits. What is the probability that the integer is divisible by 6 or 8.

Solution: The sample space of the random experiment is:

$$S = \{1,2,3, \dots \dots \dots .., 199, 200\} \Rightarrow n(S) = 200.$$

The event A: “integer chosen is divisible by 6” has the sample points given by :

$A = \{6,12,18, \dots \dots .., 198\} \Rightarrow n(A) = \frac{198}{6} = 33.$

Therefore $P(A) = \frac{n(A)}{n(B)} = \frac{33}{200}$.

Similarly the event B: ‘integer chosen is divisible by 8” has the sample points given by :

$A = \{8,16,24, \dots \dots .., 200\} \Rightarrow n(B) = \frac{200}{8} = 25.$

Therefore $P(B) = \frac{n(A)}{n(B)} = \frac{25}{200}$.

The LCM of 6 and 8 is 24.

Hence, a number is divisible by both 6 and 8, if it is divisible by 24.

Therefore

$$A \cap B = \{24, 48, 72, \dots \dots, 192\} \Rightarrow n(A \cap B)$$
$$= \frac{192}{24} = 8 \implies P(A \cap B) = \frac{8}{200}.$$

Hence, the required probability is: from equation ***(5)*** $P(A \cup B) = \frac{1}{4}$.

Example 19:A card is drawn from a pack of 52 cards. Find the probability of getting a king or a red card.

Solution: Let us define the following events :

A: the card drawn is a king,
B: the card drawn is heart,
C: the card drawn is red card.
Then $A, B$and C are not mutually exclusive.

$(A \cap B)$: the card drawn is the king of hearts it implies that $n(A \cap B) = 1$.
$(B \cap C) = B$: the card drawn is a heart (since $B \subset C$) it implies $n(B \cap C) = 13$.
$C \cap A$:the card drawn is the king of hearts it implies that $n(A \cap B \cap C) = 1$.

Therefore,

$$P(A) = \frac{n(A)}{n(B)} = \frac{4}{52};$$
$$P(B) = \frac{13}{52};$$
$$P(C) = \frac{26}{52};$$
$$P(A \cap B) = \frac{1}{52};$$
$$P(B \cap C) = \frac{13}{52};$$
$$P(C \cap A) = \frac{2}{52};$$
$$P(A \cap B \cap C) = \frac{1}{52}.$$

The required probability of getting a king or heart or a red card is given by:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$
$$= \frac{4}{52} + \frac{13}{52} + \frac{26}{52} - \frac{1}{52} - \frac{13}{52} - \frac{2}{52} + \frac{1}{52} = \frac{28}{52} = \frac{7}{13}.$$

Example 20: An MBA applies for a job in two firms X and Y. The probability of his being selected in firm X is 0.7 and being rejected at Y is 0.5. The probability of his being selected in firm X is 0.7 and being rejected at Y is 0.5. The probability of at least one of his applications being rejected is 0.6. What is probability that he will be selected that he will be selected in one of the firms?

Solution: Let A and B denote the events that the person is selected in firms X and Y respectively.
Then in the usual notations, we are given:
$P(A) = 0.7 \Rightarrow P(\bar{A}) = 1 - 0.7 = 0.3,$
$P(\bar{B}) = 0.5 \Longrightarrow P(B) = 1 - 0.5 = 0.5$,
and $P(\bar{A} \cup \bar{B}) = 0.6 = P(\bar{A}) + P(\bar{B}) - P(\bar{A} \cap \bar{B}).$
The probability that the persons will be selected in one of the two firms X and Y is given by:

$$P(A \cup B) = 1 - P(\bar{A} \cup \bar{B})$$
$$= 1 - \{P(\bar{A}) + P(\bar{B}) - P(\bar{A} \cup \bar{B})\}$$
$$= 1 - (0.3 + 0.5 - 0.6) = 0.8.$$

CHECK YOUR PROGRESS

CYP7. Let one card drawn from a pack of 52 cards and E_1be the event that card is red and E_2 be the event that card is black ace. Then write total number of elements in $E_1$and E_2.

CYP8. What is the probability that a leap year contains 53 Monday.

CYP9. A box contains 1 red, 4 white and 5 black balls. Two ball drawn out of this box simultaneously. What is the probability that one ball is black and one ball is white?

CYP10. Let S be a sample space, E_1, E_2 and E_3 are three arbitrary events. Find expression for the events noted below, in the context of;
(i) Two and more occur. **(ii)** None Occurs.

CYP11: If something has probability 1,000%, it is sure to happen. T/F

CYP12: If something has probability 90%, it can be expected to happen about nine times as often as its opposite. T/F

3.6. SUMMARY

In present unit we have briefly described the Introduction of probability and Basic Terminology: Random Experiment, Outcome, Trial, Sample Space, Exhaustive Events, Favourable events, Mutually Exclusive Events Equally Likely Events, Independent Events. We also defined the definition of Probability, Limitations of probability, Translation of Events in Set Theory Operation, Axiomatic Probability, Examples and problems related with probability.

3.7. GLOSSARY

i. **Set:** In mathematics, a set is a well-defined collection of distinct objects, called elements, that can be numbers, letters, symbols, or even other sets.

ii. **Event:** In probability, an event is a set of outcomes of a random experiment to which a probability is assigned.

iii. **Trial:** In probability, a trial refers to a single performance of a random experiment. It's the individual execution of an experiment, where the outcome is uncertain and can be one of several possibilities.

iv. **Sample space:** In probability, the sample space is the set of all possible outcomes of a random experiment.

v. **Probability:** In statistics, probability is a numerical measure that represents the likelihood of an event occurring.

vi. **Mutually exclusive**: In statistics, mutually exclusive events are events that cannot occur simultaneously. If two events are mutually exclusive, the probability of both happening at the same time is zero.

vii. **Permutations:** A permutation is an arrangement of objects in a definite order. The number of permutations of n objects taken all at a time, denoted by the symbol n_{P_n} is given by $n_{P_n} = n!$.
The number of permutations of n objects taken r at a time, where 0 < r ≤ n, denoted by , n_{P_r} is given by

$$n_{P_r} = \frac{n!}{(n-r)!}$$

viii. **Combinations:** A combination is a selection of some or all of a number of different objects where the order of selection is immaterial. The number of selections of r objects from the given n objects is denoted by

$$n_{C_r} = \frac{n!}{r!(n-r)!}.$$

3.8. REFRENCES

1. S. C. Gupta and V. K. Kapoor, (2020), *Fundamentals of mathematical statistics*, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), *Schaum's Outline: Introduction to Probability and Statistics*, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold , (2003), *Introduction to Probability and Statistics (4th Edition)*, Tata McGraw-Hill.
4. https://www.wikipedia.org.

3.9. REFRENCES

1. A.M. Goon,(1998), *Fundamental of Statistics (7th Edition)*, 1998.
2. R.V. Hogg and A.T. Craig, (2002), *Introduction to Mathematical Statistics*, MacMacMillan, 2002.
3. Jim Pitman, (1993), *Probability*, Springer-Verlag.
4. https://archive.nptel.ac.in/courses/111/105/111105090

3.10. TERMINAL QUESTIONS

TQ1: Consider a group of 3 men and 2 women and 4 children. From this group four persons are chosen at random.

What is the probability of exactly two of them will be children.

TQ2: What is the probability of a random arrangement of the letters U, N, I, V, E, R, S, I, T, Y, such that two I's do not appear together.

TQ 3: LetE_1, E_2, ... , E_n are n events. Then show that

$$P(\bigcup_{1\leq i\leq n} E_i) \leq \sum_{1\leq i\leq n} P(E_i)$$

TQ 4: LetE_1, E_2 are twoevents. Then show that

$$P(E_1 \cap E_2) \geq 1 - P(E_1^c) - P(E_2^c)$$

TQ 5: Let $E_1, E_2, \ldots, E_n$ are n events. Then show that

$$P(\bigcap_{1\le i\le n} E_i) \ge (\sum_{1\le i\le n} P(E_i)) - (n-1).$$

TQ6: Define the Probability in your words?
……………………………………………………………………………………………….
…………………………………………………………………………………………………
…………………………………………………………………………………………………
TQ7: Application of Probability in daily life?
…………………………………………………………………………………………………
…………………………………………………………………………………………………
TQ8: The probability that a learner passes a Physics test is $\frac{2}{3}$ and the probability that He passes both a Physics test and an English test is $\frac{14}{45}$. The probability that he passes at least one test is $\frac{4}{5}$.
What is the probability that he passes the English Test?

TQ9: Three newspapers A, B and C are published in a Haldwani, Uttarakhand. It is Estimated from a survey that of the adult population : 20% read A, 16% read B, 14% read C, 8% read both A and B, 5% read both A and C, 4% read both B and C, 2%read all three.
Find what percentage read at least one of the papers?

3.11. ANSWERS

Answer of Check your progress Questions:-

CYP 1: $n(E_1) = 26$ and $n(E_2) = 2$.

CYP 2: b, sample Space.

CYP 3: b, flipping a coin.

CYP4: c, the outcome is always the same.

CYP5: d, choosing a student from a class and having them be both male and female

CYP6: c, does not contain any common sample point.

CYQ7: $n(E_1) = 26$ and $n(E_2) = 2$.

CYQ8: Probability that a leap year contains 53 Monday is 2/7.

CYQ9: Probability that one ball is black and one ball is white, is 20/45=4/9.

CYQ10: (i)$\left(E_1 \cap E_2 \cap \overline{\acute{E}_3}\right) \cup (\overline{E_1} \ \cap \ E_2 \cap E_3) \cup (E_1 \cap \overline{E_2} \cap E_3)$:

(ii)$\left(\overline{E_1} \ \cap \ \overline{E_2} \cap \overline{\acute{E}_3}\right)$ or $\overline{E_1 \cup E_2 \cup E_3}$

CYQ11: False. Nothing can have probability 1000 percent -- probability values are restricted to zero to one, or to zero percent to 100 percent. Note that $1.00 = 100$ percent, just as one dollar $= 100$ cents.

CYQ12: True. If something has probability 90 percent, it can be expected to happen about nine times as often as its opposite. Depends on whether "something" and "its opposite" are the only alternatives -- if they are mutually exclusive and collectively exhaustive, then yes, at least as the expected value of a sequence of random trials.

Answer of Terminal Questions:-

TQ1: 10/21.

TQ2: 4/5.

TQ8: $\frac{4}{9}$.

TQ9: 0.35.

UNIT-4 CONDITIONAL PROBABILITY

CONTENTS:

4.1. INTRODUCTION

Dear learners, present unit are a presentation of basic concepts of conditional probability. In previous units we have studied the concepts of Probability.

Conditional probability is known as the possibility of an event or outcome happening, based on the existence of a previous event or outcome. It is deliberated by multiplying the probability of the preceding event by the renewed probability of the succeeding, or conditional, event. Since we discussed earlier, the probability $P(A)$ of an event A represents the likelihood that a random experiment will result in an outcome in the set A relative to the sample space S of the random experiment. However, quite often, while evaluating some event probability, earlier we have some information emerging from the experiment.

For example, if we have initial information that the outcome of the random experiment must be in a set B of S, then this information must be used to reassess the likelihood that the outcome will also be in B. This reassess probability is indicate by $P\left(\frac{A}{B}\right)$ and is read as the conditional probability of the event A, given that the event B has already happened.

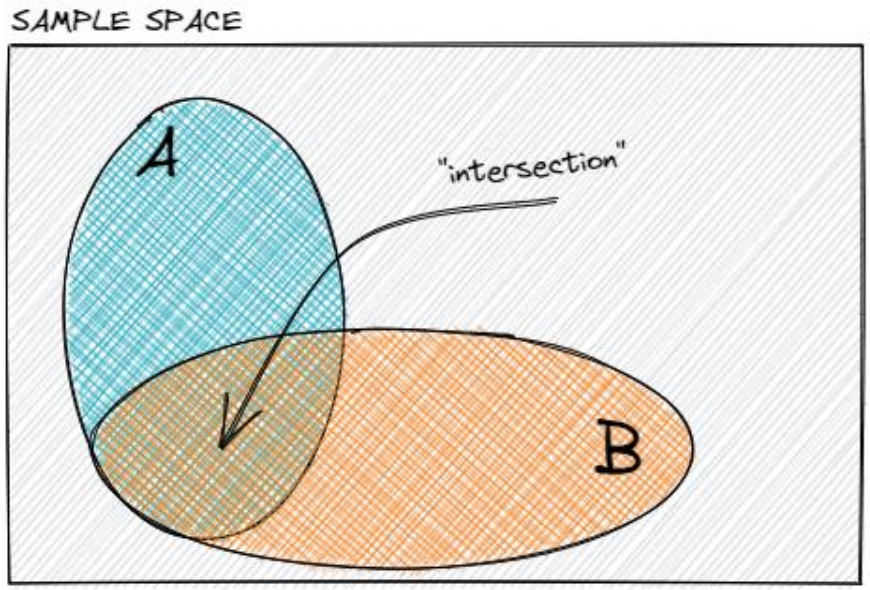


Fig 2.1.1.https://towardsdatascience.com/deriving-bayes-theorem-the-easy-way-59f0c73496db

Conditional probability clarifies by the following illustration.
Let us consider a random experiment of drawing a card from a pack of cards.
Then the probability of happening of the event A: "The card drawn is a king", is given by:$P(A) = \frac{4}{52} = \frac{1}{13}$.
Now suppose that a card is drawn and we are informed that the drawn card is red.
How does this information affect the likelihood of the event A?
Obviously, if the event B: 'The card is red', has happened, the event 'Black card' is not possible.

Hence the probability of the event A must be computed relative to the new sample space 'B' which consists of 26 sample points(red cards only),
i.e., $n(B) = 26$.
Among these 26 red cards, there are two (red) kings so that $n(A \cap B) = 2$.
Hence, the required probability is given by:
$P(A/B) = n(A \cap B)/n(B) = 2/26 = 1/13$.

4.2. INTRODUCTION

After studying this unit learner will be able to:

- **i.** Explain the notion of Conditional probability.
- **ii.** Discuss the independent events.
- **iii.** Analyse the concept of Bays' Theorem.
- **iv.** Solve the problem related to Conditional probability and Bays' Theorem.

4.3. MULTIPLICATION THEOREM OF PROBABILITY

Theorem 4.3.1. For two events A and B,

$$P(A \cap B) = P(A).P(B/A), P(A) > 0$$
$$= P(B).P\left(\frac{A}{B}\right), P(B) > 0 \quad \ldots \ldots \ldots \ldots \ldots \textbf{(4.3.1)}$$

where $P(B/A)$ represents conditional probability of occurrence of B when the event A has already happened and $P(A/B)$ is the conditional probability of happening of A, given that B has already happened.

Proof. In the usual notation we know that

$P(A) = (n(A))/(n(S))$,
$P(B) = (n(B))/(n(S))$
and $P(A \cap B) = n(A \cap B)/(n(S))$ ……………**(4.3.2)**

For the conditional event A/B, the favorable outcomes must be one of the sample points of B, i.e., for the event A/B, the sample space is B and out of the $n(B)$ sample points, $n(A \cap B)$ pertain to the occurrence of the event A.

Therefore, $P(A/B) = n(A \cap B)/n(B)$.
Using equation number (4.3.2),

$$P(A \cap B) = \frac{n(B)}{n(S)} \times \frac{n(A \cap B)}{n(B)}$$
$$= P(B).P(A/B) \ldots\ldots \textbf{(4.3.3)}$$

Similarly, we get from (4.3.2)

$$P(A \cap B) = \frac{n(A)}{n(S)} \times \frac{n(A \cap B)}{n(A)}$$
$$= P(A).P(B/A) \ldots \ldots \ldots \ldots \textbf{(4.3.4)}$$

From (4.3.3) and (4.3.4) we get the result (4.3.1).

Therefore "the probability of the simultaneous occurrence of two events A and B is equal to the product of the probability of one of these events and the conditional probability of the other, given that the first one has occurred.

Remark 4.3.1: The conditional probability $P(B/A)$ and $P(A/B)$ are defined if and only if $P(A) \neq 0$ and $P(B) \neq 0$, respectively.

Remark 4.3.2: For $P(B) > 0, P(A/B) \leq P(A)$.

Remark 4.3.3: The conditional probability $P(A/B)$ is not defined if $P(B) = 0$.

Remark 4.3.4:$P(B/B) = 1$.

4.4. INDEPENDENT EVENTS

Consider the experiment of throwing two dice, say die 1 and die 2. It is obvious that the occurrence of a certain number of dots on the die 1 has nothing to do with a similar event for the die 2. The two are quite independent of each other, so to say. But suppose, the two dice were connected with a piece of thread before being thrown. The situation changes. This time the two events are not independent in as much as that the uppermost face of one die will have something to do in causing a particular face of the other die to be uppermost; and the shorter the thread, the more is this influence or dependence. Similarly, if we draw two cards from pack of cards in succession, then the results of the two draws are in independent if the cards are drawn with replacement (i.e., if the first card drawn in placed back in the pack before drawing the second card) and the results of the two draws are not independent if the cards are drawn without replacement.

Two or more events are said to be independent if the happening or non-happening of any one of them, does not, in any way, affect the happening of others.

An event A is said to be independent (or statistically independent) of another event B,

If the conditional probability of A given B,

i.e., $P(A/B)$ is equal to the unconditional probability of B,

i.e., if $P\left(\frac{A}{B}\right) = P(A)$ ………………**(4.4.1)**

It is important that $P(B) \neq 0$.

Similarly, If the conditional probability of Agiven B,

i.e., $P(B/A)$ is equal to the unconditional probability of B,

i.e., if $P(B/A) = P(B)$…………………………..**(4.4.2)**

In this case $P(A) \neq 0$.

4.4.1.THEOREMS

Theorem 4.5.1. If the events A and B are such that $P(A) \neq 0$, $P(B) \neq 0$ and A is independent of B, then B is independent of A.

Proof. Since the event A is independent of B.

Therefore,

$$P\left(\frac{A}{B}\right) = P(A)$$

$$\Rightarrow \frac{P(A \cap B)}{P(B)} = P(A)$$

$$\Rightarrow P(A \cap B) = P(A)P(B).$$

Therefore,

$$\frac{P(B \cap A)}{P(A)} = P(B)$$

since $P(A) \neq 0$ and $A \cap B = B \cap A$.

It implies that

$P(B/A) = P(B)$

it implies that B is independent of A.

- A is independent of B and B is independent of A it means A and B are independent.
- For any event Ain S, (i) A and the null event $\emptyset$ are independent also A and S are independent.

Theorem 4.5.2. (Multiplication Theorem of Probability for Independent Events). If the A and B are two events with positive probabilities are such that $P(A) \neq 0$, $P(B) \neq 0$

then A and B are independent if and only if

$P(A \cap B) = P(A)P(B)$**…………(2.4.3)**

Proof. Since,

$P(A \cap B) = P(A).P\left(\frac{B}{A}\right) = P(B).P\left(\frac{A}{B}\right)$; $P(A) \neq 0$, $P(B) \neq 0$

…………………………………………………(4.4.4)

A is independent of B and B is independent of A then,

$P\left(\frac{A}{B}\right) = P(A)$ and $P(B/A) = P(B)$

……………………………………………(4.4.5)

From (4.4.3) and (4.4.4), we get $P(A \cap B) = P(A)P(B)$, as required.

Conversely, if (4.4.3) holds,

$$\frac{P(A \cap B)}{P(B)} = P(A)$$

…………………………………………………..(4.4.6)

$$\Rightarrow P\left(\frac{A}{B}\right) = P(A)$$

$$\frac{P(A \cap B)}{P(B)} = P(B)$$

$$\Rightarrow P\left(\frac{B}{A}\right) = P(B)$$

……………………………………………………(4.4.7)

(4.4.6) implies that $A$and B are independent events.

Hence, for independent events A and B,the probability that both of these occur simultaneously is the product of their respective probabilities.

This Rule is known as the Multiplication Rule of Probability.

Theorem 4.4.3. For n events$A_1, A_2, A_3 \dots \dots \dots \dots A_n$ we have

$P(A_1 \cap A_2 \cap \dots \dots \cap A_n) = P(A_1)P(A_2/A_1)P(A_3/(A_1 \cap A_2)) \dots \dots P(A_n/A_1 \cap A_2 \cap \dots \dots \dots \dots \dots \dots \dots \dots \cap A_{n-1} \dots \dots \dots \dots \dots \dots \dots \dots \dots . (\mathbf{4.4.7})$

Where $P(A_i/A_j \cap A_k \quad \cap \dots \dots \cap A_l$ represents the conditional probability of the event A_i, $A_{k,} \dots . A_l$ have already happened.

Theorem4.4.4. Necessary and sufficient condition for independence of$n$ events$A_1, A_2, A_3 \dots \dots \dots \dots A_n$is that the probability of their simultaneous happening is equal to the product of their respective probabilities, i.e.,

$P(A_1 \cap A_2 \cap \dots . . \cap A_n) = P(A_1)\, P(A_2) P(A_3) \dots \dots \dots . P(A_n)$………**(4.4.8)**

Theorem 4.4.5. For a fixed B with $P(B) > 0, P(A/B)$is probability fuction.

Theorem 4.4.6. For any three events A, B and C,

$P((A \cup B)\ /\ C) = P(A\ /\ C) + P(B\ /\ C) - P((A \cap B)\ /\ C)$…………**(4.4.9)**

Theorem 4.4.7. For any three events A, B and C,

$P((A \cap \bar{B})\ /\ C) + \ P((A \cap B)\ /\ C) = P(A\ /\ C)$……………………….**(4.4.10)**

Theorem 4.4.8. For any three events A, B and C defined on the sample space Ssuch that $B \subset C$ and $P(A) > 0, P(B/A) \leq P(C/A)$……………………………**(4.4.11)**

Theorem 4.4.9. If A and B are independent events, then

(i) A and $\bar{B}$ are independent events. (ii) $\bar{A}$ and B are independent events (iii) $\bar{A}$ and $\bar{B}$ are independent events………………………………………………………. **(4.4.12)**

4.4.2. PAIRWISE INDEPENDENT EVENTS

Consider n events $A_1, A_2, A_3 \dots \dots \dots \dots A_n$ defined on the same sample space so that $P(A_i) > 0; i = 1,2, \dots \dots ., n$.

These events are said to be pairwise independent if every pair of two events is independent in the sense of the definition given in Multiplication Theorem of Probability for Independent Events. The events $A_1, A_2, A_3 \dots \dots \dots \dots A_n$ are said to be pairwise independent if and only if :

$$P(A_i \cap A_j) = P(A_i)P(A_j)\ ,\quad i \neq j = 1,2, \dots \dots \dots n \dots \dots \dots \textbf{(4.4.13)}$$

In particular A_1, A_2, A_3 are said to be pairwise independent if and only if :

$$P(A_1 \cap A_2) = P(A_1)P(A_2),$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3),$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3)$$

$$\dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \textbf{(4.4.14)}$$

4.4.3. MUTUALLY INDEPENDENT EVENTS

Let S denote the sample space for a number of events. The events in S are said to be mutually independent if the probability of the simultaneous occurrence of (any) finite number of them is equal to the product of their separate probabilities.

The events $A_1, A_2, A_3 \dots \dots \dots \dots A_n$ in a sample space S are said to be mutually independent if:

$$P(A_{i1} \cap A_{i2} \cap \dots \dots \cap A_{ik}) = P(A_{i1})P(A_{i2}) \dots \dots \dots P(A_{ik})\ ,$$

$$k = 2,3, \dots \dots \dots n$$

$$\dots \dots \dots \dots \dots \dots \dots \dots \dots (\mathbf{4.4.15})$$

Hence, the events are mutually independent if they are independent by pairs, and by triplets, and by quadruples, and so on.

Theorem 4.4.10. If $A,\ B,\ C$ are mutually independent events then $A \cup B$ and C are also independent.

Theorem 4.4.11 If A, B and C are random events in a sample space and if $, B$ and C are pairwise independent and A is independent of $B \cup C$, then A, B and C are mutually independent.

4.4.4. SOLVED EXAMPLES

Example 1. If $A \cap B = \emptyset$, then show that

$$P(A) \leq P(\bar{B}).$$

Solution: Since $A = (A \cap B) \cup (A \cap \bar{B})$

$$= \emptyset \cup (A \cap \bar{B}) = A \cap \bar{B} \text{ (Since } A \cap B = \emptyset).$$

$$A \subseteq \bar{B} \Rightarrow P(A) \leq P(\bar{B}).$$

Since $A \cap B = \emptyset$,

we have $A \subseteq \bar{B}$,

which implies that $P(A) \leq P(\bar{B})$.

Example 2. Let $A$and B be two events such that

$P(A) = \frac{3}{4}$ and $P(A) = \frac{5}{8}$,

show that

$(a) P(A \cup B) \geq \frac{3}{4}$,

and

$(b) \frac{3}{8} \leq P(A \cap B) \leq \frac{5}{8}$.

Solution: We have

$A \subset (A \cup B) \Rightarrow P(A \cap B) \leq P(B) = \frac{5}{8}$.…………………..(4.4.3.1)

Also $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq 1 \Rightarrow \frac{3}{4} + \frac{5}{8} - 1 \leq P(A \cap B)$

Therefore,

$\frac{6+5-8}{8} \leq P(A \cap B) \Rightarrow \frac{3}{8} \leq P(A \cap B)$ … … … … … … … … … (4.4.3.2).

From (4.4.3.1) and (4.4.3.2)

$$\frac{3}{8} \leq P(A \cap B) \leq \frac{5}{8}.$$

Example 3. For any two events A and B,

$$P(A \cap B) \leq P(A) \leq P(A \cup B) \leq P(A) + P(B).$$

Proof. Since we know that $A = (A \cap \bar{B}\,) \cup (A \cap B)$.
So by the definition of probability

$$P(A) = P[(A \cap \bar{B}\,) \cup (A \cap B)] = P(A \cap \bar{B}) + P(A \cap B).$$

Now $P[(A \cap \bar{B}\,)] \geq 0$.
Therefore,

$$P(A) \geq P(A \cap B) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4.4.3.3).$$

Similarly, $P(B) \geq P(A \cap B)$ it implies that $P(B) - P(A \cap B) \geq 0$.
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$……………………..(4.4.3.3)
Therefore, $P(A \cup B) \geq P(A)$
it implies that $P(A) \leq P(A \cap B)$……………..(4.4.3.4)
Also, $P(A \cup B) \leq P(A) + P(B)$……………………………(4.4.3.5)

Hence, from (4.4.3.3), (4.4.3.4) and (4.4.3.5) we get,

$$P(A \cap B) \leq P(A) \leq P(A \cup B) \leq P(A) + P(B).$$

Example4. The odds against Manager X settling the wage dispute with the workers are $8:6$ and odds in favour of manager Y setting the same dispute are $14:16$.

i. What is the chance that neither settles the dispute, if they both try, independently of each other?
ii. What is the probability that dispute will be settled?

Solution: Let A be the event that the manager X will settle the dispute and B be the event that the manager Y will settle the dispute. Then clearly,

$$P(\bar{A}) = \frac{8}{8+6} = \frac{4}{7}$$

it implies that

$$P(A) = 1 - P(\bar{A}) = \frac{6}{14} = \frac{3}{7}.P(\bar{B}) = \frac{14}{14+16} = \frac{7}{15}$$

it implies that $P(A) = 1 - P(B) = \frac{16}{14+16} = \frac{8}{15}$.

i. The required probability that neither settles the dispute is given by:

$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = \frac{4}{7} \times \frac{8}{15} = \frac{32}{105}$.

(Since A and B are independent it implies that $\bar{A}$ and $\bar{B}$ are also independent events).

ii. The dispute will be settled if at least one of the managers X and Y settles the dispute. Hence the required probability is given by:

$P(A \cup B)$= Probability [At least one of X and Y settles the dispute] .

= 1- Probability [None settles the dispute]

= 1 –$P(\bar{A} \cap \bar{B})$

$= 1 - \frac{32}{105}$

$= \frac{73}{105}$.

Example5. A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each other.

Solution: The required event E that 'in a draw of 4 balls from the box at random there is at least one ball of each colour', can materialise in the following mutually disjoint ways:

i. 1 red, 1 white and 2 black balls;
ii. 2 red, 1 white and 1 black balls;
iii. 1 red, 2 white and 1 black balls.

Hence by addition theorem of probability, the required probability is given by:

$$P(E) = P(i) + P(ii) + P(iii)$$
$$= \frac{\binom{6}{1} \times \binom{4}{1} \times \binom{5}{2}}{\binom{15}{4}} + \frac{\binom{6}{2} \times \binom{4}{1} \times \binom{5}{1}}{\binom{15}{4}} + \frac{\binom{6}{1} \times \binom{4}{2} \times \binom{5}{1}}{\binom{15}{4}}$$
$$= \frac{1}{\binom{15}{4}}[6 \times 4 \times 10 + 15 \times 4 \times 5 + 6 \times 6 \times 5]$$
$$= \frac{4!}{15 \times 14 \times 13 \times 12}(240 + 300 + 180)$$
$$= \frac{24 \times 720}{15 \times 14 \times 13 \times 12} = 0.5275.$$

Example 6. Data on readership of a certain magazine show that the proportion of 'male' readers under 35 is 0.40 and over 35 is 0.20. If the proportion of readers under 35 is

0.70, find the proportion of subscribers that are "females over 35 years". Also calculate the probability that a randomly selected male subscriber is under 35 years of age.

Solution. Let us define the following events:

A: Reader of the magazine is a male.
B: Reader of the magazine is over 35 years of age.
Then in usual notations, we are given:

$P(A \cap B) = 0.20,$
$P(A \cap \bar{B}) = 0.40$

and

$P(\bar{B}) = 0.70 \Rightarrow P(B) = 0.30.$

i. The proportion of subscribers that are 'female overover35 years' is:

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$
$$= 0.30 - 0.20 - 0.10$$
$$= 0.10.$$

ii. The probability that a randomly selected male subscribers is under 35 years is:

$$P\left(\frac{\bar{B}}{A}\right) = \frac{P(A \cap \bar{B})}{P(A)}$$
$$= \frac{0.40}{0.60}$$
$$= \frac{2}{3}.$$

[Since, $P(A) = P(A \cap B) + P(A \cap \bar{B}) = 0.20 + 0.40 = 0.60$]

CHECK YOUR PROGRESS

CYP1. An urn contains 4 tickets numbered 1,2,3,4 and another contains 6 tickets numbered 2, 4, 6, 7, 8, 9. If one of the two urns is chosen at random and a ticket is drawn at random from the chosen urn, find the probabilities that the ticket drawn bears the number (i) 2 or 4, (ii) 3, (iii) 1 or 9.

CYP2. From a city population, the probability of selecting (i) a male or a smoker is $\frac{7}{10}$, (ii) a male smoker is $\frac{2}{5}$, and (iii) a male, if a smoker is already selected is $\frac{2}{3}$.

Find the probability of selecting (a) a non-smoker, (b) a male (c) a smoker, if a male is first selected.

CYP3. If A and B are two events such that $(A \cup B) = \frac{5}{6}, P(A \cap B) = \frac{1}{3}, P(B) = \frac{1}{2}$, then the events A and B are

(i) Dependent (ii) Independent (iii) Mutually exclusive (iv) None of these.

CYP4. For two events A and $P(B) = 0.4, P(B) = p, P(A \cup B) = .6$.
Then p equals

(i) 0.2 when A and B are mutually disjoint
(ii) 0.2 when A and B are independent
(iii) Not determined in any case
(iv) 0.2 when A and B are dependent.

4.5. BAYES' THEOREM

Bayes' theorem which was given by Thomas Bayes, a British Mathematician, in 1763, provides a means for making these probability calculations.

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

THOMAS BAYES
1701-1761
Ref: https://en.wikipedia.org/wiki/Thomas_Bayes

Bayes theorem gives the probability of an "event" with the given information on "tests". There is a difference between "events" and "tests".

For example there is a test for liver disease, which is different from actually having the liver disease, i.e. an event. Rare events might be having a higher false positive rate. To prove the Bayes' theorem, use the concept of conditional probability formula.

Theorem 4.5.1.Bayes' Theorem: If $E_1, E_2, E_3, \dots \dots \dots \dots \dots \dots .., E_n$ are mutually disjoint events with

$$P(E_i) \neq 0, (i = 1,2, \dots \dots \dots \dots, n),$$

then for any arbitrary event A which is subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have

$$\mathrm{P}\,(E_i/A) = \frac{\mathrm{P}(\mathrm{E_i})\mathrm{P}(A/E_i)}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P}(\mathrm{E_i})\mathrm{P}(A/E_i)}$$

$$= \frac{\mathrm{P}(\mathrm{E_i})\mathrm{P}(A/E_i)}{\mathrm{P(A)}}; \mathrm{i} = 1,2, \dots \dots .. \mathrm{n} \dots$$

$$\dots\dots\dots\dots\dots\dots..(4.5.1)$$

Proof. Since$A \subset \bigcup_{i=1}^{n} E_i$.

Since by distributive law we have,

$$A = A \cap (\bigcup_{i=1}^{n} E_i) = \bigcup_{i=1}^{n}(A \cap E_i).$$

Since $(A \cap E_i) \subset E_i, (i = 1,2, \dots ..., n)$ are mutually disjoint events, we have by addition theorem of probability:

$$P(A) = P\left\{\bigcup_{i=1}^{n}(A \cap E_i)\right\} = \sum_{i=1}^{n} P\{E_i\}\ P(A/E_i)$$
…………(4.5.2)

By multiplication theorem of probability:

Also we have $P(A \cap E_i) = P(A)P(E_i/A)$

it implies that

$$P\left\{\frac{E_i}{A}\right\} = \frac{P(A \cap E_i)}{P(A)}$$

$$= \frac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)}$$
[from (4.5.2)].

Remark 4.5.1.

$P(E_1),P(E_2)\ldots..P(E_n)$	'Prior probabilities.'
$P\left(\frac{A}{E_i}\right), i = 1,2, \ldots \ldots n$	'likelihoods'
$P\left\{\frac{E_i}{A}\right\}, i = 1,2,3, \ldots \ldots .. n$	'Posterior probabilities.'

Remark 2.5.2.

If $E_1, E_2, E_3, \ldots \ldots \ldots \ldots \ldots \ldots ., E_n$ constitute a disjoint partition of the sample space S and $P(E_i) \neq 0, (i = 1,2, \ldots \ldots \ldots \ldots, n)$, then for any arbitrary event A in S we have,

$$P(A) = \sum_{i=1}^{n} P(E_i)P\left(\frac{A}{E_i}\right) \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (4.5.3)$$

Remark 2.5.3. limitations of Bayes Theorem: The Bayesian approach has no general way to represent and handle the uncertainty within the background knowledge and the prior probability function. This is a serious limitation of Bayesianism, both in theory and in application.

Theorem 4.5.2.Bayes' Theorem for Future Events. The probability of the materialisation of another event C, given

$(C/A \cap E_1), P(C/A \cap E_2), \ldots\ldots\ldots\ldots P(C/A \cap E_n)$ is given by:

$$\text{P (C / A)} = \frac{\sum_{i=1}^{n} P(E_i)P(A/E_i)P(C/E_i \cap A)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)}$$

……………………..(4.5.4)

4.6. SOLVED EXAMPLES

Example 4.6.1. Suppose that a product is produced in three factoriesX, Y and Z. It is known that factory X produces thrice as many items as factory Y, and that factories Y and Z produce the same number of items. Assume that it is known that 3 per cent of the items produced by each of the factories X and Z are defective while 5 per cent of those manufactured by factory Y are defective. All the items produced in three factories are stoked, and an item of product is selected at random.

i. What is the probability that this item is defective?

ii. If an item selected at random is found to be defective, what is the probability that it was produced by factory X, Y and Z respectively?

Solution. Let the number of items produced by each of the factories Y and Z be X, Y and n. Then the number of items produced by the factory X is $3n$. Let E_1, E_2 and E_3 denote the events that the items are produced by factory X, Y and Z respectively and let A be the event of the item being defective. Then we have,

$$P(E_1) = \frac{3n}{3n + n + n} = 0.6;$$

$$P(E_2) = \frac{n}{5n} = 0.2$$

and

$$P(E_2) = \frac{n}{5n} = 0.2.$$

Also,

$P(A/E_1) = P(A/E_3) = 0.03$

and

$P(A/E_2) = 0.05$(Given).

i. The probability that an item selected at random from the stock is defective is given by:

$$P(A) = P\left\{\bigcup_{i=1}^{3}(A \cap E_i)\right\}$$

$$= \sum_{i=1}^{3} P\{E_i\} \ \mathrm{P}(A/E_i)$$

$$= P(E_1)(A/E_1) + P(E_2)(A/E_2) + P(E_3)(A/E_3)$$

$$= 0.6 \times 0.03 + 0.2 \times 0.05 + 0.2 \times 0.03$$

$$= 0.034.$$

ii. By Bayes' Rule, the required probabilities are given by:

$$\mathrm{P}\,(E_1/A) = \frac{\mathrm{P(E_1)P}(A/E_1)}{\mathrm{P(A)}}$$

$$= \frac{0.6 \times 0.03}{0.034}$$

$$= \frac{0.018}{0.034}$$

$$= \frac{9}{17}.$$

$$\mathrm{P}\,(E_2/A) = \frac{\mathrm{P(E_2)P}(A/E_2)}{\mathrm{P(A)}}$$

$$= \frac{0.2 \times 0.05}{0.034}$$

$$= \frac{0.010}{0.034}$$

$$= \frac{5}{17}.$$

$$\mathrm{P\,(E_3/A)} = \frac{\mathrm{P(E_3)P(A/E_3)}}{\mathrm{P(A)}}$$

$$= \frac{0.006}{0.034}$$

$$= \frac{3}{17}.$$

It implies that

$$\mathrm{P\,(E_3/A)} = 1 - [\mathrm{P}\,(E_1/A) + \mathrm{P}\,(E_2/A)\,]$$

$$= 1 - \left(\frac{9}{17} + \frac{5}{17}\right) = \frac{3}{17}.$$

Example 4.6.2. From a vessel containing 3 white and 5 black balls, 4 balls are transferred into an empty vessel. From this vessel a ball is drawn and is found to be white. What is the probability that out of four balls transferred3 are white and 1 is black?

Solution. Let us define the following events:

E_1: Transfer of 0 white and 4 black balls.
E_2: Transfer of 1 white and 3 black balls.
E_3: Transfer of 2 white and 2 black balls.
E_4: Transfer of 3 white and 1 black balls

(Since the urn contains 3 white balls, more than 3 white balls cannot be transferred from the vessel)

E: Drawing of a white ball from the second vessel.

Then $P(E_1) = \frac{\binom{5}{4}}{\binom{8}{4}}$

$= \frac{1}{14}$,

$$P(E_2) = \frac{\binom{3}{1} \times \binom{5}{3}}{\binom{8}{4}}$$

$= \frac{3}{7}$,

$$P(E_3) = \frac{\binom{3}{2} \times \binom{5}{3}}{\binom{8}{4}}$$

$= \frac{3}{7}$,

$$P(E_3) = \frac{\binom{3}{3} \times \binom{5}{1}}{\binom{8}{4}}$$

$= \frac{1}{14}$,

Also

$\mathrm{P}(E/E_1) = 0$,
$\mathrm{P}(E/E_2) = \frac{1}{4}$,
$\mathrm{P}(E/E_3) = \frac{2}{4}$, and
$\mathrm{P}(E/E_4) = \frac{3}{4}$.

Hence, by Bayes Theorem, the probability that out of four balls transferred, 3 are white and is black is:

$$\mathrm{P}\,(\mathrm{E}_4/\mathrm{E}) = \frac{\frac{1}{14} \times \frac{3}{4}}{\frac{1}{14} \times 0 + \frac{3}{7} \times 4 + \frac{3}{7} \times \frac{1}{2} + \frac{1}{14} \times \frac{3}{4}}$$

$$= \frac{3}{6 + 12 + 3}$$

$$= \frac{1}{7}$$

$$= 0.14.$$

CHECK YOUR PROGRESS

CYP5. In 2002 there will be three candidates for the position of principal –Mr.Chatterji, Mr.Ayangar and Dr.Singh-whose chances of getting the appointment are in the proportion $4:2:3$ respectively. The probability that Mr.Chatterji if selected would introduce co-education in the college is 0.3 The probabilities of Mr.Ayangar and Dr. Singh doing the same are respectively 0.5 and 0.8.

- **i.** What is probability that there will be co-education in the college in 2003?
- **ii.** If there is coeducation in the college in 2003, what is the probability that Dr.Singh is the principal.

4.7. SUMMARY

This unit basically based on conditional probability. We are starting with the Introduction and Objectives. In this unit we are giving the definition and proof of Multiplication Theorem of Probability. We also explain the Definition of Independent Events and Theorems, Pairwise Independent Events, Mutually Independent Events. Our main focus in this unit is state and proof of Bayes' Theorem and the problems related with Bayes' Theorem.

4.8. GLOSSARY

- **i.** **Set:** In mathematics, a set is a well-defined collection of distinct objects, called elements, that can be numbers, letters, symbols, or even other sets.

ii. **Event:** In probability, an event is a set of outcomes of a random experiment to which a probability is assigned.

iii. **Trial:** In probability, a trial refers to a single performance of a random experiment. It's the individual execution of an experiment, where the outcome is uncertain and can be one of several possibilities.

iv. **Sample space:** In probability, the sample space is the set of all possible outcomes of a random experiment.

v. **Probability:** In statistics, probability is a numerical measure that represents the likelihood of an event occurring.

vi. **Mutually exclusive**: In statistics, mutually exclusive events are events that cannot occur simultaneously. If two events are mutually exclusive, the probability of both happening at the same time is zero.

4.9. REFRENCES

1. S. C. Gupta and V. K. Kapoor, (2020), *Fundamentals of mathematical statistics*, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), *Schaum's Outline: Introduction to Probability and Statistics*, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold , (2003), *Introduction to Probability and Statistics (4th Edition)*, Tata McGraw-Hill.
4. https://www.wikipedia.org.

4.10. REFRENCES

1. A.M. Goon,(1998), *Fundamental of Statistics (7th Edition)*, 1998.
2. R.V. Hogg and A.T. Craig, (2002), *Introduction to Mathematical Statistics*, MacMacMillan, 2002.
3. Jim Pitman, (1993), *Probability*, Springer-Verlag.
4. https://archive.nptel.ac.in/courses/111/105/111105090

4.11. TERMINAL QUESTIONS

TQ1 State and prove Baye's Theorem?

....

………………………………………………………………………………………………

………………………………………………………………………………………………

……………………………………………………………………………………………..

TQ2 What are the criticisms against the use of Baye's Theorem in probability theory?

……………………………………………………………………………………………..

……………………………………………………………………………………………..

TQ3. One shot is fired from each of the three guns. E_1, E_2, E_3denote the events that the target is hit by the first, second and third guns respectively. If $P(E_1) = 0.5,\ P(E_2) = 0.6$ and $P(E_3) = 0.8$ and E_1, E_2, E_3, are independent events, find the probability that

(a) exactly one hit is registered, and

(b) at least two hits are registered.

TQ4. Two computers A and B are to be marketed. A salesman who is assigned the job of finding customers for them has 60% and 40% chances respectively of succeeding in case of computer A and B. The two computers can be sold independently. Given that he was able to sell at least one computer, what is the probability that computer A has been sold?

TQ5. The probabilities of X, Y and Z becoming managers are $\frac{4}{9}, \frac{2}{9}$ and $\frac{1}{3}$ respectively. The probabilities that the Bonus Scheme will be introduced if $, Y$ and Z becomes managers are$\frac{3}{10}, \frac{1}{2}$ and $\frac{4}{5}$ respectively

i. What is the probability that Bonus Scheme will be introduced, and
ii. if the Bonus Scheme has been introduced, what is the probability that the manager appointed was X?

4.12. ANSWER

Answer of Check your progress Questions:-

CYQ 1: $(i)\frac{5}{12}(ii)\frac{1}{8}(iii)\frac{5}{24}$.

CYQ 2:$(a)\frac{3}{5}(ii)\frac{1}{2}(iii)\frac{4}{5}$.

CYQ3: Independent.

CYQ 4:.0.2when$A$ and B are mutually disjoint.

CYQ 5: $(i)\frac{23}{45}(ii)\frac{12}{23}$.

Answer of Terminal Questions:-

TQ 3: $(i)0.26\ \ (ii)0.70$.

TQ4: 0.79.

TQ5: $(i)\frac{23}{45}(ii)\frac{6}{23}$.

Unit-5: RANDOM VARIABLE

CONTENTS:

5.1 *INTRODUCTION*

A random variable is a numerical quantity whose value is determined by the outcome of a random experiment. It acts as a bridge between probability theory and statistical analysis, assigning numbers to events in a sample space. Random variables are broadly classified into discrete random variables, which take countable values (such as number of heads in coin tosses), and continuous random variables, which take values from an uncountable range (such as height or temperature). Each random variable is associated with a probability distribution that describes the likelihood of different outcomes, enabling the calculation of measures like expected value, variance, and standard deviation. This concept is fundamental for modeling uncertainty, making predictions, and solving real-world problems involving randomness.

5.2 *OBJECTIVE*

The main objectives of this unit are:

1. Define and explain the concept of a random variable and its role in probability theory.
2. Differentiate between discrete and continuous random variables.
3. Understand and interpret probability mass function (PMF), probability density function (PDF), and cumulative distribution function (CDF).
4. Calculate measures of central tendency and dispersion such as expected value, variance, and standard deviation for random variables.
5. Model real-life situations using appropriate types of random variables.

5.3 *RANDOM VARIABLE*

Definition: A random variable is a function that assigns a real number to each possible outcome of a random experiment, representing uncertain events in numerical form. It links the outcomes in the sample space to numerical values, enabling the application of probability and statistical methods.

Mathematically, a random variable is defined as: $X: S \to R$

where:

- S = sample space of a random experiment
- R = set of real numbers
- $X(s) =$ Real number assigned to the outcomes $s \in S$.

Example 1: If we toss a fair die, $S = \{1,2,3,4,5,6\}$

and a random variable X could be defined as $X(s) = s^2$, so X assigns each die face to its square: X={1,4,9,16,25,36}

Remarks 1: Despite being a function, random variable X is referred to as a (random) variable. Although this language doesn't seem very satisfactory, we will nonetheless use it because it is widely recognised.

2: A random variable has more to its definition than just being a function from S to R. Not every possible function between these sets qualifies as a random variable. For it to be valid, the event $\{X(s) = x\}$ for any real number x and the event $\{X(s) \in I\}$ for any interval III must have well-defined probabilities that follow the basic axioms of probability. These probabilities are determined in terms of "equivalent events" (explained later) within the original sample space SSS. This means that the sets $X^{-1}(x)$ and $X^{-1}(I)$ must be events in S. The challenge arises because, in general, not every subset of S is necessarily an event, as the relevant σ-field may not contain all subsets of S. However, since this issue does not typically appear in most practical applications, we will set it aside and continue to treat X simply as a function from S to R.

3: Since X is defined as a function from S to R, the definition of a function ensures that each $s \in S$ is associated with exactly one value $X(s)$ in R. However, it is possible for different elements of S to be assigned the same value of *X*.

4: For practical purposes, specifying the exact functional form of *X* is not necessary. What matters most is the set of possible values *X* can take, rather than the specific origin of these values. We will represent random variables using capital letters such as *X*, *Y*, *Z* and their corresponding values using lowercase letters such as *x*, *y*, *z*.

We now present the definition of "equivalent events," which provides a basis for determining the probability of an event associated with the random variable *X*.

Definition: Let E be a random experiment with sample space S, and let X be a random variable with range space R_X. Consider an event B for the random variable X, where $B \subset R_X$, and let $A = \{s \in S : X(s) \in B\}$ be its inverse image, denoted as $X^{-1}(B)$, under the mapping defined by X. The sets A and B are referred to as **equivalent events**. It is important to note that B belongs to the range space R_X while A belongs to the sample space S. However, the occurrence of B is considered "equivalent" to the occurrence of A, which makes it possible to determine the probability of B using the well-defined probability of A (since A is an event in S and X is a random variable). This leads us to the following definition:

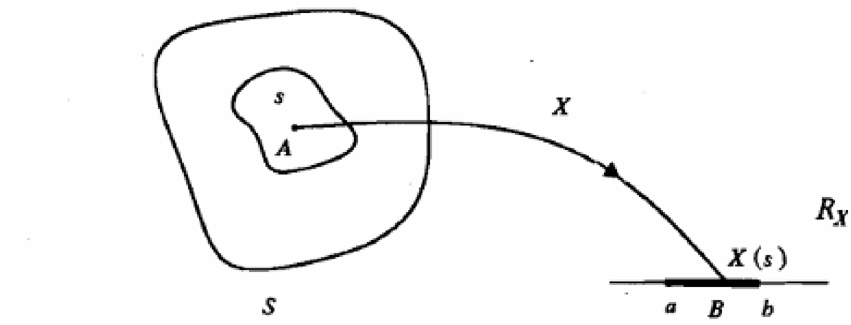
Definition: Let B be an event in the range space R_X, and define

$$A = X^{-1}(B) = \{s \in S : X(s) \in B\}$$

The probability of B is then given as the probability of A, *i.e.*,

$$P(B) = P(A).$$

As noted earlier, A and B belong to different sample spaces $B \subset R_X$ and $A \subset S$. Ideally, we should use a distinct notation for the probability of B, such as $P_X(B)$; however, for simplicity, we will use the same notation for both and define $P(B) = P(A)$.



The above definition provides a method for assigning probabilities to different events within the range space R_X of a random variable X. After these probabilities are established, the original sample space S is no longer needed and can often be disregarded. We will now demonstrate these ideas using the following simple examples:

Example 2: A fair coin is tossed twice, and let X represent the number of heads obtained. Determine the range space R_X of X and the probabilities corresponding to the events $\{X = 0\}$, $\{X = 1\}$ and $\{X = 2\}$.

Solution: The sample space S for this random experiment is:

$S = \{(H, H), (H, T), (T, H), (T, T)\}$.

Since the random variable X counts the number of heads obtained, we have:

$X(H, H) = 2$, $X(H, T) = 1$, $X(T, H) = 1$, $X(T, T) = 0$.

Thus, the range space R_X is {0, 1, 2}. To compute $P(X = 0)$, we first identify the equivalent event in the original sample space S and then use it to find the probability. For the event {X=0} in R_X, the corresponding equivalent event in S is {(T,T)}. Therefore

$$P(X = 0) = P(T, T) = \frac{1}{4}.$$

Similarly, $P(X = 1) = P((H, H), (T, H)) = P(H, T) + P(T, H) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

So, $P(X = 2) = P(T, T) = \frac{1}{4}$.

Example 3: Consider a production line where each item is inspected and classified as either defective (D) or non-defective (N). The inspection process continues until the first non-defective item is found. Let X represents the number of items inspected before the experiment ends. Determine the range space R_X for the random variable X. If the probability of an item being defective is 0.2 calculate P(X=3).

Solution: Since the probability of defective item is 0.2 then the probability of non-defective item will be 0.8 (=1-0.2).

Since X represents the number of items tested until the experiment ends, and the process continues until the first non-defective item appears, the sample space is

$S = \{N, DN, DDN, DDDN, ...\}$

By the definition of the random variable:

$X(N) = 1, X(DN) = 2, X(DDN) = 3, X(DDDN) = 4$, and so on.

Thus, the range space $R_X = \{1, 2, 3\}$.

To find P(X=3), we identify the equivalent event for {X=3} in the original sample space S and then use it to calculate the probability. Hence,

$P(X = 3) = P(DDN) = (0.2)^2(0.8) = 0.032$

Example 4: A light bulb is produced and tested for its lifespan T by placing it in a socket and recording the time (in hours) until it burns out. Is T considered a random variable? If so, determine its range space.

Solution: The sample space S for the associated random experiment E is the set $\{t : t \geq 0\}$. The lifespan T qualifies as a random variable since we can define $T : S \to R$ by $T(t) = t$. Hence, the range space of the random variable T is $\{t : t \geq 0\} \subset R$.

Example 5: Consider the random experiment E described in Example 4. Suppose the manufactured light bulbs are sold in the market, and past data shows that the profit per bulb is Rs. 1.00 if the lifespan is under 50 hours, Rs. 2.00 if the lifespan is between 50 and 150 hours, and Rs. 4.00 if the lifespan exceeds 190 hours. Is the profit function Y a random variable? Also, determine the range space of Y.

Solution: Referring to the solution of Example 4, we have $S = \{t : t \geq 0\}$. The profit function $Y : S \to R$ is defined as follows:

$$Y(t) = \begin{cases} 1, & t < 50 \\ 2, & 50 \leq t \leq 150 \\ 4, & t > 150 \end{cases}$$

Thus, Y is a random variable, and its range space is {1, 2, 4}. The probabilities of events within this range space are determined based on the probabilities of the corresponding equivalent events in the original sample space. Therefore, P(Y = 1, P(Y = 2), and P(Y = 4) can be obtained if the probabilities of the events $\{t : t < 50\}, \{t : 50 \leq t \leq 150\}$ and $\{t : t > 150\}$ are known in the original sample space $S = \{t : t \geq 0\}$.

For further discussion, it is helpful to classify random variables based on their range space. If the range space R_X is finite or countably infinite, then X is called a **discrete random variable**. If R_X is an interval, then X is called a **continuous random variable**. From Examples 1 to 5 above, we see that the random variables in Examples 2, 3, and 5 are discrete, while the one in Example 4 is continuous. A random variable defined over a discrete sample space must be discrete, but a discrete random variable can also be defined over a continuous sample space (as in Y from Example 5). In some cases, a random variable X may take certain distinct values $x_1, x_2, ..., x_n$ with positive probability and also take all values within an interval $a \leq x \leq b$; such random variables are called **mixed random variables**.

5.4 *TYPES OF RANDOM VARIABLE*

Random variables are classified based on the nature of their range space into these 3 catogories.

(1) Discrete random variable

(2) Continuous random variable

(3) Mixed-type random variable

Note: Here, mixed type random variable is less common but mostly important random variable.

5.5 *DISCRETE RANDOM VARIABLE*

You have already encountered some examples of discrete random variables in the previous section, and you will come across many more practical cases in the upcoming section of this unit. This section focuses on exploring the probability distribution and cumulative distribution function of a discrete random variable.

Definition: A random variable X is termed a discrete random variable if its range space R_X is either finite or countably infinite. In other words, the possible values of X can be listed as $x_1, x_2, ..., x_n$, where the list may end after a finite number of terms or extend infinitely.

5.5.1 *PROBABILITY DISTRBUTION FUNCTION OF DISCRETE RANDOM VARIABLE*

From the definition of a random variable and the concept of equivalent events discussed earlier, we can determine the probability of the event $[X = x_i]$, i.e., $P[X = x_i]$, for each $i = 1, 2, 3, ...$ using the relation:

$$P[X = x_i] = P\{s \in S : X(s) = x_i].$$

Therefore, for a discrete random variable *X*, we have the following data:

$$X : x_1, x_2, ..., x_i, ..$$

$$P(X) : P(x_1), P(x_2), ..., P(x_i), .. \qquad ...(1)$$

where $p(x_i) = P[X = x_i]$ and $\sum P[X = x_i] = 1$

because the probabilities are summed over all possible values of X, which together represent P(S) in the original sample space.

The function P given in equation (1) is known as the **probability function** (or **point probability function**) of the random variable X. The set of ordered pairs $\{(x_i, p(x_i)), i = 1, 2, ..\}$ is called the **probability distribution** of the random variable X. Based on this, we can state the following definition:

Definition: Let X be a discrete random variable with range space $R_X = \{x_1, x_2, ..., x_i, ..\}$. If $p(x_i) = P[X = x_i]$, then the set $\{(x_i, p(x_i)), i = 1, 2, ..\}$ is referred to as the **probability distribution** of the discrete random variable X.

It follows by definition that,

(1) $p(x_i) \geq 0$ for all i and

(2) $\sum_i p(x_i) = 1$

The main benefit of a probability distribution is that it allows us to determine the probability of any event related to a discrete random variable X. Specifically, let $B \subset R_X$ be an event, and let $I = \{i : x_i \in B\}$. Then, the probability of B is given by:

$$P(B) = P[s \in S : X(s) \in B] = \sum_{i:x_i \in B} P[X = x_i]$$

In other words, the probability of an event $B \subset R_X$ is found by summing the probabilities of each individual outcome that makes up B. Furthermore, within any finite interval [a, b], there will be only a finite number of possible values of X, say $x_1, x_2, ..., x_k$. Therefore,

$$P[a \le X \le b] = \sum_{i=1}^{k} P[X = x_i] = \sum_{i=1}^{k} p(x_i).$$

If the interval [a, b] contains none of the possible values x_i, we assign $P[a \le X \le b] = 0$

Example 6: A fair coin is tossed twice. Let *X* represents the number of heads obtained. Determine the probability distribution corresponding to the random variable *X*.

Solution: Referring to the solution of Example 2, observe that the possible values of *X* are 0, 1 and 2. Additionally, $P[X = 0] = \frac{1}{4}$, $P[X = 1] = \frac{1}{2}$ and $P[X = 2] = \frac{1}{4}$. Therefore,

$X:$	0	1	2
$P(x):$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

represents the probability distribution of the random variable *X*.

Example 7: Find the probability distribution of the total scores obtained when two fair dice are rolled. Here, the score refers to *x* + *y* for the outcome (*x*, *y*).

Solution: The most effective way to approach this problem is by listing and tabulating all possible dice outcomes to determine how each possible score can be obtained. For this purpose, we proceed as follows:

Score	Combination
2	(1, 1)
3	(1, 2) (2, 1)
4	(1, 3) (2, 2) (3, 1)
5	(1, 4) (2, 3) (3, 2) (4, 1)
6	(1, 5) (2, 4) (3, 3) (4, 2) (5, 1)
7	(1, 6) (2, 5) (3, 4) (4, 3) (5, 2) (6, 1)
8	(2, 6) (3, 5) (4, 4) (5, 3) (6, 2)
9	(3, 6) (4, 5) (5, 4) (6, 3)

10	(4, 6) (5, 5) (6, 4)
11	(5, 6) (6, 5)
12	(6, 6)

Let X represents the dice score. Based on the above discussion, we have:

$$P[X=2]=P((1,1))=\frac{1}{36}=0.028$$

$$P[X=3]=P((1,2))+P((2,1))=\frac{2}{36}=0.056$$

$$P[X=4]=P((1,3))+P((2,2))+P((3,1))=\frac{3}{36}=0.083$$

…

$$P[X=12]=P((6,6))=\frac{1}{36}=0.028$$

Thus, the probability distribution of *X* is:

X	2	3	4	5	6	7	8	9	10	11	12
P(*x*)	0.028	0.056	0.083	0.111	0.139	0.167	0.139	0.111	0.083	0.056	0.028

It can be observed that the probability P(x) reaches its highest value when the score *X*=7. Hence, in a betting game, the most favorable choice would be to place your bet on the score 7.

5.5.2 *CUMULATIVE DISTRBUTION FUNCTION OF DISCRETE RANDOM VARIABLE*

Definition: Let *X* be a random variable. The function $F_X(x)$, defined for a $F_X(x)=P[X\le x]$

is called the **Cumulative Distribution Function (c.d.f.)** of the random variable *X*.

This definition applies to both discrete and continuous random variables, although the method for calculating $P[X\le x]$ will differ depending on the type.

For a discrete random variable, the c.d.f. is given by:

$$(X, F_X(x))=P[X\le x]=\sum_{i\in I}P[X=x_i]=\sum_{i\in I}p(x_i)$$

Where, $I = \{i : x_i \leq x\}$. Thus,

$F_X(x)) = \sum_{i \in I} p(x_i)$ where the summation is taken over all indices *I*, satisfying $x_i \leq x$.

Example 8: A fair coin is tossed twice. Let *X* represent the number of heads obtained. Determine and illustrate both the probability distribution and the cumulative distribution function (c. d. f.) of the random variable X.

Solution: Referring to the solutions of Examples 2 and 6 above, the probability distribution *P*(*x*) of the random variable *X* can be expressed as follows.

$X:$	0	1	2
$P(x):$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

To determine the cumulative distribution function $F_X(x)$ of the random variable *X*, we examine the various intervals or ranges of *x* in $-\infty < x < \infty$. Thus for $x < 0$, $P[X \leq x] = 0$; for $0 \leq x < 1$, $P[X \leq x] = \frac{1}{4}$; for $1 \leq x < 2$, $P[X \leq x] = P[X = 0] + P[X = 1] = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$ and for for $x \geq 2$, $P[X \leq x] = P[X = 0] + P[X = 1] + P[X = 2] = 1$.

Therefore,

for $1 \leq x < 2$, $F_X(x) = \begin{cases} 0, & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$

Now, the plot of $p(x)$ and $F(x)$ are as follows:

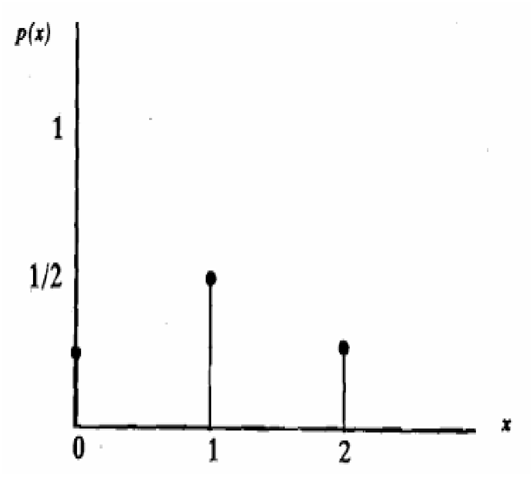


Fig 2: Probability distribution of *X*

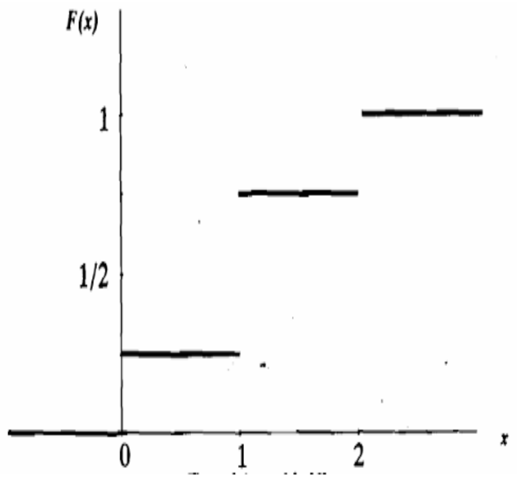


Fig 3: Cumulative Probability distribution of *X*

Properties of the Cumulative Distribution Function (c.d.f.).

1. From the definition of $F_X(x)$, for all $-\infty<x<\infty$, we have $0 \le F(x) \le 1$, since $F(x)$ represents a probability.

2. $F(x)$ is a monotone non-decreasing function of *x*. That is, if $x_1 \le x_2$, then $F(x_1) \le F(x_2)$. This holds because the interval $(-\infty, x_1]$ is contained within $(-\infty, x_2]$ whenever $x_1 \le x_2$.

3. The limits are given by:

$$\lim_{x\to\infty} F(x) = 1 \text{ and } \lim_{x\to\infty} F(x) = 0$$

4. At each value x_i in the range of *X*, $F(x)$ has a positive jump of size p_i (where $p_i = P[X = x_i]$)). Between x_i and x_{i+1}, $F(x)$ remains constant. In other words:

$$F(x_{i+1}) = F(x_i) + P[X = x_{i+1}]$$

Because of property 4, the c.d.f. of a discrete random variable is always a step function.

5.6 *CONTINUOUS RANDOM VARIABLE*

A **continuous random variable** is a type of random variable that can take **any value** within a given range (which could be finite or infinite).

Instead of assigning probabilities to individual values (like in a discrete random variable), we describe its behavior using a **probability density function (PDF)**.

Key points about a continuous random variable:

1. **Infinite possible values:** Between any two numbers, there are infinitely many possible outcomes.
2. **Probability at a single point is zero:** For example, $P(X = 5) = 0$; probabilities are only meaningful over intervals.
3. **Probabilities are found using integrals**:

   $P(a \le x \le b) = \int_a^b f(x)\,dx$, where $f(x)$ is the PDF.

4. **Total probability is 1:** $\int_{-\infty}^{\infty} f(x)\,dx = 1$, where $f(x)$ is the PDF.

5.6.1 *PROBABILITY DENSITY FUNCTION OF CONTINUOUS RANDOM VARIABLE*

Definition: A random variable *X* is said to follow a continuous distribution if there exists a non-negative function $f = f_X$ such that

$$P(a \le x \le b) = \int_a^b f(x)\,dx$$

for every pair of real numbers a and b. This function *f* is known as the probability density function (PDF) or simply the density function of *X*.

More specifically, such a function *f* satisfies certain properties that characterize continuous distributions.

Example 9: Find whether the following function is a probability density function:

$$f(x) = \begin{cases} x; & 0 \le x \le 1 \\ 2x; & 1 \le x \le 1 \end{cases}$$

Solution: $\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 x\,dx + \int_1^2 2x\,dx$ $\quad \{\because \int_{-\infty}^{0} f(x)\,dx = 0 = \int_2^{\infty} f(x)\,dx\}$

$$= \left[\frac{x^2}{2}\right]_0^1 + \left[\frac{2x^2}{2}\right]_1^2 = \left(\frac{1}{2} - 0\right) + \left(2^2 - 1^2\right) = \frac{1}{2} + 3 = \frac{7}{2} \ne 1$$

$\frac{1}{3} < X < \frac{1}{2}$ **Example 10:** If the function $f(x)$ is defined by

$$f(x) = ce^{-x}; 0 \le x \le \infty,$$

Find the value of c which changes $f(x)$ to a probability density function.

Solution: In order that $f(x)$ may be probability density function, we should have

(1) $f(x) \geq 0$ for every x.

(2) $\int_{-\infty}^{\infty} f(x) = 1$

Since e^{-x} is always positive for the values of x lying between 0 and ∞, the condition will be satisfied if

$$\int_{-\infty}^{\infty} ce^{-x}\, dx = 1$$

i.e., if $\left[-ce^{-x}\right]_0^{\infty} = 1$

i.e., if $c = 1$

Example 11: If $f(x)$ has probability density function $Cx^2, 0 < x < 1$, evaluate the value of c and find the probability that $\frac{1}{3} < X < \frac{1}{2}$ *i.e.*, $P\left(\frac{1}{3} < X < \frac{1}{2}\right)$.

Solution: $f(x) = Cx^2$ will be probability density function if,

$$\int_0^1 Cx^2 dx = 1 \text{ } i.e., \text{ if } \left[\frac{1}{3}Cx^3\right]_0^1 = 1 \text{ } i.e., \text{ } C = 3$$

Hence, $f(x) = 3x^2, 0 < x < 1$

Now, $P\left(\frac{1}{3} < X < \frac{1}{2}\right) = \int_{1/3}^{1/2} 3x^2 dx = [x^3]_{1/3}^{1/2} = \left(\frac{1}{8} - \frac{1}{27}\right) = \frac{19}{216}$

$$f(x) = ce^{-x}; 0 \leq x \leq \infty,$$

Find the value of c which changes $f(x)$ to a probability density function.

5.6.2 *CUMULATIVE DISTRBUTION FUNCTION OF CONTINUOUS RANDOM VARIABLE*

The **Cumulative Distribution Function (CDF)** of a continuous random variable describes the probability that the variable takes a value less than or equal to a given number.

Definition: Let X be a continuous random variable with probability density function (PDF) $f(x)$.

The cumulative distribution function $F_X(x)$ is defined as:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

Here:

- $f(t)$ is the PDF of X
- The CDF accumulates the total probability from $-\infty$ up to x.

Properties of CDF for a continuous random variable:

1: **Range:** $0 \le F_X(x) \le 1$ for all x.

2: **Non-decreasing:** If $x_1 \le x_2$, then $F_X(x_1) \le F_X(x_2)$

3: **Limits:** $\lim_{x \to \infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$

4: **Relationship with PDF:** $f(x) = \dfrac{d}{dx} F_X(x)$

5: **Continuity:** The CDF of a continuous random variable is a continuous function.

Example 12: A random variable X has PDF:

$$f(x) = \begin{cases} 2x; & 0 \le x \le \\ 0; & otherwise \end{cases},$$

then find the CDF.

Solution: $F_X(x) = \int_{-\infty}^{x} f(t)\,dt$

Case I: If $x < 0$, then $F_X(x) = 0$

Case II: If $0 \le x \le 1$, then $F_X(x) = \int_0^x 2t\,dt = \left[t^2\right]_0^x = x^2$

Case III: If $x > 1$, then $F_X(x) = 1$

Then final CDF is $F_X(x) = \begin{cases} 0; & x < 0 \\ x^2; & 0 \le x \le 1 \\ 1; & x > 1 \end{cases}$

Check your progress

Problem 1: Find the constant *C* for the probability density function

$$f(x) = \begin{cases} Cx^2; & 0 \le x \le 3 \\ 0; & elsewhere \end{cases}$$

and compute $P(1 \le X \le 2)$. Find also the distribution function

Problem 2: Find the value y_0 so that the function $f(x)$ defined as follows be a density function $f(x) = y_0 e^{-x/\sigma}$ $: 0 \le x \le \infty$

5.7 *SUMMARY*

A random variable is a numerical quantity whose value depends on the outcome of a random experiment. It can be discrete, taking countable values (like the number of heads in coin tosses), or continuous, taking values from an interval (like the height of students). The behavior of a random variable is described by its probability distribution—for discrete variables through a probability mass function (PMF) and for continuous variables through a probability density function (PDF). The cumulative distribution function (CDF) gives the probability that the variable is less than or equal to a certain value, and it applies to both discrete and continuous cases (with step functions for discrete and smooth curves for continuous variables). These tools allow us to compute event probabilities, analyze statistical behavior, and form the basis for further concepts like expectation, variance, and hypothesis testing.

5.8 *GLOSSARY*

- Random Variable
- Discrete random variable
- Continuous random variable
- Probability density function
- Cumulative distribution function

5.9 *REFERENCES*

- S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
- Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
- J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.

5.10 *SUGGESTED READINGS*

- https://www.wikipedia.org.
- A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
- R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
- Jim Pitman, (1993), Probability, Springer-Verlag.
- OpenAI. ChatGPT (GPT-4 model) [Large language model]. https://chat.openai.com/
- https://archive.nptel.ac.in/courses/111/105/111105090

5.11 *TERMINAL QUESTIONS*

Long answer type questions

1. The probability function of a discrete random variable is as follows:

$X = x$	0	1	2	3	4	5	6	7
$P(x) = P(X = x)$	0	k	$2k$	$2k$	$3k$	K^2	$2K^2$	$K^2 + k$

Then find **(*i*)** k ; **(*ii*)** $P(X < 6)$ and $P(X \geq 6)$; **(*iii*)** $P(0 < X < 5)$; **(*iv*)** Distribution function of X

2. Explain the term random variable by suitable example

3. The probability function of a random variable X is given by,

$$P(x) = \begin{cases} k; & if\ x = 0 \\ 2k; & if\ x = 1 \\ 3k; & if\ x = 2 \\ 0; & otherwise \end{cases}$$

Then find **(*i*)** k ; **(*ii*)** $P(X < 2)$, $P(X \le 2)$, $P(0 < X < 2)$ and $P(X \ge 6)$

(*iii*) Distribution function of X

4. If the random variable X has the values 0, 1, 2, … then for what value of t, $p(x) = (1-t)t^x; x = 0,1,2,...$ is probability function of X.

5. If a random variable X has the probability density function as follows:

$$P(x) = \begin{cases} 1/4; & if\ -2 < x < 2 \\ 0; & otherwise \end{cases}$$

Then obtain the values of **(*i*)** $P(X < 1)$ **(*ii*)** $P(|X| > 1)$

(*iii*) $P[(2X + 3) > 5]$ and $P(X \ge 6)$

Short answer type questions

1. Let X be a discrete random variable with $P(X = j) = C.\left(\frac{1}{2}\right)^j$; $j = 1, 2, ...$

Then find the value of C.

2. A continuous random variables X has the probability density function $f(x) = 3x^2, 0 \le x \le 1$, find a and b when

(*i*) $P(X \le a) = P(X > a)$

(*ii*) $P(X > b) = 0.05$

Objective type questions:

1. A random variable is

a) Always discrete

b) Always continuous

c) A numerical function of a sample space

d) Independent of the sample space

2. Which of the following is an example of a discrete random variable?

a) Height of students in a class
b) Time taken to finish a race
c) Number of defective items in a batch
d) Distance between two cities

3. For a discrete random variable X, the sum of all $P(X = x_i)$ over all possible x_i is equal to:

a) 0
b) 1
c) Between 0 and 1
d) Infinity

4. The probability density function (PDF) of a continuous random variable must satisfy:

a) $f(x) \geq 0$ for all x
b) $\int_{-\infty}^{\infty} f(x)dx = 1$
c) Both (a) and (b)
d) None of the above

5. The Cumulative Distribution Function (CDF), $F(x)$, is:

a) Always decreasing
b) Always non-decreasing
c) Constant for all x
d) Equal to the PDF

6. If X is a continuous random variable, then $P(X = a)$ is:

a) Always positive
b) Always zero
c) May be positive or zero
d) Depends on a

7. For a fair die, if X is the outcome, the probability distribution is:

a) Uniform
b) Normal
c) Binomial
d) Poisson

8. Which of the following is **not** a property of the CDF?

a) $0 \leq F(x) \leq 1$

b) $F(x)$ is a step function for discrete random variables

c) $F(x)$ decreases when x increases

d) $\lim_{x \to -\infty} F(x) = 0$

9. A random variable that can take only a finite or countably infinite set of values is called:

a) Continuous random variable

b) Discrete random variable

c) Uniform variable

d) Gaussian variable

10. The probability distribution of a continuous random variable is represented by:

a) Probability Mass Function (PMF)

b) Probability Density Function (PDF)

c) Cumulative Probability Function (CPF)

d) None of the above

11. For a discrete random variable, the probability of each outcome must be:

a) Between 0 and 1 (inclusive)

b) Greater than 1

c) Equal for all outcomes

d) Zero for at least one outcome

12. If $F_X(x)$ is the CDF of X, then the PDF $F_X(x)$ for a continuous random variable is obtained by:

a) Integrating $F_X(x)$

b) Differentiating $F_X(x)$

c) Squaring $F_X(x)$

d) None of the above

13. Which of the following is **true** for a continuous random variable?

a) $P(a \le X \le b) = P(a < x < b)$

b) $P(X = a) > 0$

c) Its distribution is always normal

d) It cannot take infinite values

14. For a discrete random variable X taking values {0, 1, 2}, with $P(X = 0) = 0.2$, $P(X = 1) = 0.5, P(X = 2) = 0.3$, the CDF at $x = 1$ is:

a) 0.2

b) 0.5

c) 0.7
d) 1.0

15. Which of the following is an example of a continuous random variable?
a) Number of books in a library
b) Time taken to boil water
c) Number of cars in a parking lot
d) Roll of a die:

Fill in the blanks questions:

1. A variable whose values are determined by the outcomes of a random experiment is called a __________.

2. A random variable that takes on only a finite or countably infinite number of possible values is called a __________ random variable.

3. A random variable that can take any value within a given range is called a __________ random variable.

4. The probability distribution of a discrete random variable is described by a __________.

5. The probability distribution of a continuous random variable is described by a __________.

6. The function $F_X(x) = P(X \le x)$ is called the __________ of *X*.

7. For a discrete random variable *X*, $\sum p(x_i) =$ __________.

8. For a continuous random variable *X*, the probability that *X* takes an exact value is __________.

9. For a continuous random variable, the total area under the PDF curve is __________.

10. If *X* follows a uniform distribution on [*a*, *b*], then the PDF is given by $f(x) =$ _______ for $a \le X \le b$.

11. The CDF of a continuous random variable is obtained by __________ the PDF.

12. The PDF of a discrete random variable is always __________.

13. The CDF of a discrete random variable is always a __________ function.

True and False questions

1. A random variable is always numerical in nature.

2. A discrete random variable can take an infinite number of values within a finite interval.

3. The probability mass function (PMF) is used for continuous random variables.

4. For a discrete random variable, the sum of all probabilities is always 1.

5. The probability density function (PDF) can take negative values.

6. For a continuous random variable, $P(X = a)$ is always zero.

7. The CDF of a random variable is always a decreasing function.

8. The area under the PDF of a continuous random variable equals 1.

9. The CDF of a discrete random variable is a step function.

10. The variance of a random variable is always non-negative.

11. The expected value is always equal to the most probable value of the random variable.

12. The PDF for a uniform distribution is constant over its range.

13. The CDF of a continuous random variable is obtained by integrating its PDF.

14. Every random variable must be either discrete or continuous.

15. A PMF for a discrete random variable can have probabilities greater than 1.

5.12 *ANSWERS*

Answer of check your progress

Problem 1: $C = 1;\ P(1 \le X \le 2) = \frac{7}{27};\ F(x) = \begin{cases} 0; & x < 0 \\ \frac{x^3}{27}; & 0 \le x \le 3 \\ 1; & x \ge 3 \end{cases}$

Problem 2: $y_0 = \frac{1}{\sigma}$

Answer of long answer type question:

1: (*i*) $k = \frac{1}{10}$ **(*ii*)** $P(X < 6) = 0.81$ and $P(X \ge 6) = 0.19$

(*iii*) $P(0 < X < 5) = 0.8$ **(*iv*)** $F(x) = P(X \le x); F(0) = 0; F(1) = \frac{1}{10}; F(2) = \frac{3}{10};$

$F(3) = \frac{5}{10};\ F(4) = \frac{8}{10};\ F(5) = \frac{81}{100};\ F(6) = \frac{83}{100};\ F(7) = 1$

3: **(*i*)** $k = \frac{1}{6}$ **(*ii*)** $\frac{1}{2};1;\frac{1}{3}$

(*iii*) $$F(x) = \begin{cases} 0; & if\ x < 0 \\ \frac{1}{6}; & if\ x = 0\, or\, 0 \le x < 1 \\ \frac{1}{2}; & if\ x = 1 or\, 0 \le x < 2 \\ 1; & if\ x \ge 2 \end{cases}$$

4: $0 \le t \le 1$

5: **(*i*)** $P(x < 1) = \frac{3}{4}$ **(*ii*)** $P(|x| > 1) = \frac{1}{2}$

(*iii*) $P[(2X + 3) > 5] = \frac{1}{4}$

Answer of short answer type question:

1. $C = 1$ **2:** **(*i*)** $a = \left(\frac{1}{2}\right)^{1/3}$ **(*ii*)** $b = (0.95)^{1/3}$

Answer of objective questions

1:	c	**2:**	c	**3:**	b	**4:**	c
5:	b	**6:**	b	**7:**	a	**8:**	c
9:	b	**10:**	b	**11:**	a	**12:**	b
13:	a	**14:**	c	**15:**	b		

Answer of fill in the blanks

1: random variable **2:** Discrete **3:** Continuous

4: Probability mass function **5:** Probability density function

6: cumulative distribution function **7:** 1

8: 0 **9:** 1 **10:** $\frac{1}{b-a}$

11:	Integratting	**12:**	undefined (or "not applicable")
13:	Step		

Answer of True and False:

1:	True	**2:**	False	**3:**	False
4:	True	**5:**	False	**6:**	True
7:	False	**8:**	True	**9:**	True
10:	True	**11:**	False	**12:**	True
13:	True	**14:**	False	**15:**	False

Unit-6: EXPECTED VALUE AND VARIANCE

CONTENTS:

6.1 *INTRODUCTION*

The **Expected Value and Variance** unit in statistics introduces two fundamental concepts used to describe the behavior of random variables. The expected value, also known as the mean, represents the long-run average or central tendency of a random variable's outcomes, providing a single value summary of its distribution. In contrast, the variance measures the degree of spread or dispersion of these outcomes around the mean, indicating how much the values deviate on average. Together, expected value and variance help in understanding, summarizing, and comparing probability distributions, making them essential tools in statistical analysis, risk assessment, and decision-making under uncertainty.

6.2 *OBJECTIVE*

After reading this unit you will be able to:

1. To introduce the expected value as a measure of the central tendency of a random variable, both for discrete and continuous distributions.
2. To develop skills for calculating the expected value of a random variable using summation or integration methods.
3. To explain how variance measures the spread or dispersion of a random variable around its mean, and how standard deviation is its square root.
4. To practice computing variance and standard deviation for both discrete and continuous random variables.
5. To examine important properties such as linearity of expectation and the behavior of variance under transformations of random variables.
6. To solve applied problems involving expected values and variances in areas such as economics, engineering, and risk assessment.
7. To show how expected value and variance help summarize and interpret the behavior of probability distributions.

6.3 *EXPECTED VALUE*

The expected value of a random variable X is the weighted average of the values that X can assume with probabilities of its various values as weights. Thus, the expected value of a random variable X, denoted as $E(X)$ is defined as

$$E(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n) = \sum_{i=1}^{n} x_i p(x_i)$$

Thus, the expected value of a random variable is obtained by considering the various values that the variable can take, multiplying these by their corresponding probabilities, and then summing these products.

Expectation is a very basic concept and is employed widely in decision theory, management science. systems analysis, theory of games, and many other fields. Some of these applications will be discussed in the chapter on Decision Theory.

Remark: Expected value is called by many other names also, such as the mathematical expectation, the arithmetic mean, or simply the expectation or the mean.

6.4 *RANDOM VALUE*

A variable whose value is determined by the outcome of a random experiment is called a random variable. To be more rigorous, we say that a random variable is a function defined over the sample space of an experiment, and generally assumes different values with a definite probability associated with each value.

A random variable is denoted by the capital letters X, Y, Z, etc., of the English alphabet and particular values which the random variable takes are denoted by the corresponding small letters of the English alphabet.

It may be noted that the actual value that the event assumes is not a random variable. For example, in three tosses of a coin, the number of heads obtained is a random variable which can take any one of the three values $0,1,2$ or 3, as long as the coin is not tossed. But after it is tossed and we get one head, then 1 is not a random variable.

A random variable may be discrete or continuous. If the random variable takes on the integral values such as $0,1,2, \ldots$, then it is called a discrete random variable. For example, the number of defective items in a sample, the number of printing mistakes in each page of a book, the number of telephone calls received by the telephone switchboard of a firm, etc. If the random variable takes on all values within a certain interval, then the random variable is called a continuous random variable. Any variable involving measurements of height, weight, time, volume, etc., is essentially a continuous random variable.

6.5 *PROBABILITY DISTRIBUTION*

The probability distribution of a random variable is a listing of the various values of the random variable with the corresponding probability associated with each value of the random variable.
The two tables shown below give some of the probabilities associated with two simple experiments, one consisting of the roll of a die and the other three flips of a balanced coin.

Number of points rolled with a die	Probability	Number of heads obtained in three flips of a coin	Probability
1	$1/6$	0	$1/8$
2	$1/6$	1	$3/8$
3	$1/6$	2	$3/8$
4	$1/6$	3	$1/8$
5	$1/6$		
6	$1/6$		

The first of these tables is easily obtained on the basis of the assumption that each face of the die has a probability of $1/6$, the second is obtained by considering as equally likely the eight possible outcomes TTT, HTT, THT, TTH, HHT, HTH, THH, and HHH, where H stands for 'head' and T for 'tail'.

Wherever possible, we try to express probability distribution by means of a formula which enables us to calculate the probabilities associated with the various numerical descriptions of the outcomes with the usual functional notations.

We can write

(*i*) $p(x) = 1/6; x = 1,2,3,4,5,6$

(*ii*) $p(x) = {}^3C_x \left(\frac{1}{2}\right)^3; x = 0,1,2,3$

In general, if a random variable X assumes the values $x_1, x_2, \dots, x_n$ with probabilities $p(x_1), p(x_2), \dots, p(x_n)$ respectively, the probability distribution of X is the specification of the set of values x_i together with their probabilities $p(x_\text{i}), (i = 1,2, \dots, n)$.

Values	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$ … …	… . $\boldsymbol{x_n}$	**Total**
Probability	$p(x_1)$	$p(x_2)$	$p(x_3)$ …	… $p(x_n)$	1

where $p(x_\text{i}) =$ Probability that X assumes the value $x_\text{i} = P(X = x_\text{i})$

Remarks:

1. It may be noted that since a random variable has to assume one of its values, the sum of the probabilities in a probability distribution must always be 1.

2. The concept of probability distribution is analogous to that of frequency distribution. Just as frequency distribution tells us how the total frequency is distributed among different values (or classes) of the variable, similarly, a probability distribution tells us how the total probability of 1 is distributed among the various values that the random variable can take.

3. Probability Density Function (*Continuous r.v.*). In case of a continuous random variable, we do not talk of probability at a particular point (which is always zero), but we always talk of probability in an interval. If $p(x)dx$ is the probability that the random variable X takes the value in a small interval of magnitude dx, e.g.,

$(x, x + dx)$ or $[\ x - (dx/2), x + (dx/2)\ ]$, then $p(x)$ is called the probability density function (*p.d.f.*) of the r.v. X.

Illustration. An unbiased coin is thrown repeatedly until a head appears. If X denotes the 'number of tails preceding the first head', then its probability distribution is given below :

No. of tails (x):·	0	1	2	3	...	x	...	Total
Probability (p):	$\frac{1}{2}$	$\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^3$	$\left(\frac{1}{2}\right)^4$	...	$\left(\frac{1}{2}\right)^{x+1}$	...	1

This distribution can be written as

$$p(x) = \frac{1}{2}x + 1;\ (x = 0,1,2, \ldots \infty)$$

6.6 *PROPERTIES OF EXPECTED VALUE*

(*i*) The expected value of a constant is the constant itself, *i.e.*,

$$E(k) = k, \text{ for every constant } k.$$

(*ii*) The expected value of the product of a constant and a random variable is equal to the product of the constant with the expected values of the individual random variables, *i.e.*,

$$E(kX) = kE(X)$$

(*iii*) The expected value of the sum or difference of the random variables is equal to the sum or difference of the expected values of the individual random variables, *i.e.*,

$$E(X \pm Y) = E(X) \pm E(Y)$$

(*iv*) The expected value of the product of two independent random variables is equal to the product of their individual expected values, *i.e.*,

$$E(XY) = E(X)E(Y)$$

(*v*)

$$E[X - E(X)] = 0$$

(*vi*)

$$E[\phi(X)] = \sum_{i=1}^{n} \phi(x_i)p(x_i)$$

In particular,

$$E(X^2) = \sum_{i=1}^{n} {x_i}^2 p(x_i)$$

Illustration 1: A die is thrown at random. What is the expectation of the number on it?
Let X denote the number on the die. Then X is a random variable which can take any one of the values $1,2,3,\ldots,6$ each with equal probability $1/6$ as given below :

x	1	2	3	4	5	6
p	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

$$\therefore E(X) = \sum x \cdot p(x)$$
$$=1 \times (1/6) + 2 \times (1/6) + 3 \times (1/6) + 4 \times (1/6) + 5 \times (1/6) + 6 \times (1/6)$$
$$= (1/6)(1+2+3+4+5+6)$$
$$=\frac{21}{6} = \frac{7}{2}.$$

Illustration 2: Two unbiased dice are thrown together at random. Find the expected value of the total number of points shown.

If a random variable X denotes the sum of the number of points obtained when two unbiased dice are thrown, the probability distribution is

Sum of Numbers (x)	**Favorable sample points**	**No. of favorable cases**	**Probability $p(x)$**
2	(1,1)	1	1/36
3	(1, 2), (2, 1)	2	2/36
4	(1, 3), (3, 1), (2, 2)	3	3/36
5	(1, 4), (4, 1), (2, 3), (3, 2)	4	4/36
6	(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)	5	5/36
7	(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)	6	6/36
8	(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)	5	5/36
9	(3, 6), (6, 3), (4, 5), (5, 4)	4	4/36
10	(4,6), (6,4), (5,5)	3	3/36
11	(5,6), (6,5)	2	2/36

12	(6,6)	1	1/36

$$E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{4}{36} + 6 \times \frac{5}{36} + 7 \times \frac{6}{36} + 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36}$$

$$= \frac{1}{36}[2 + 6 + 12 + 20 + 30 + 42 + 40 + 36 + 30 + 22 + 12] = \frac{252}{36} = 7.$$

Illustration 3: Under an employment promotion program, it is proposed to allow the sale of newspapers on the buses during off-peak hours. The vendor can purchase the newspaper at a special concessional rate of 25 paise per copy against the selling price of 40 paise. Any unsold copies are, however, a dead loss. A vendor has estimated the following probability distribution for the number of copies demanded.

Number of copies :	15	16	17	18	19	20
Probability :	0.04	0.19	0.33	0.26	0.11	0.07

How many copies should he order so that his expected profit will be a maximum? [**Hint:** Calculate expected profits for the purchase of 15 copies, 16 copies, and compare.]

The vendor can purchase the newspapers at a rate of 25 paise per copy against the selling price of 40 paise. Hence, he makes a profit of 15 paise per copy.

Expected profit = No. of copies × Probability × 15

No. of copies	:	15	16	17	18	19	20
Profits	:	2.25	2.40	2.55	2.70	2.85	3.00
Probability	:	0.04	0.19	0.33	0.26	0.11	0.07
Expected Profit in price	:	9	46	84	70	31	21

Thus, it is evident that 17 copies give the maximum expected profit. Hence, he should order 17 copies to earn a maximum profit.]

Example 1: *A lottery sells 10,000 tickets at Rs. 1 per ticket, and a prize of Rs. 5,000 will be given to the winner of the first draw. Suppose you have bought a ticket. How much should you expect to win?*

Solution: Here, the random variable 'win', W has two possible values: Rs. 1 and Rs. 4,999. Their respective probabilities are

$$\frac{9{,}999}{10{,}000} \text{ and } \frac{1}{10{,}000}$$

Thus $$E(W) = -1 \times \frac{9{,}999}{10{,}000} + 4{,}999 \times \frac{1}{10{,}000} = -\text{Rs.}\,0.50$$

Thus, a minus 50 paise is the amount we expect to win on average if we play this game over and over again.

Example 2: *From a life table, it is observed that the probability is 0.98 that a 35-year-old man will live for one more year. An insurance company offers to sell such a man a Rs. 10,000 one-year term life insurance policy at a premium of Rs. 220. What is the company's expected gain?*

Solution: Let X be the company's gain.

Then X :	220	$-(10{,}000 - 220)$
Probability :	0.98	0.02

Hence expected gain of the company

$$= 0.98 \times 220 + 0.02(-10{,}000 + 220)$$
$$= 215.6 - 195.6 = \text{ Rs. } 20$$

Example 3: *A box contains 6 tickets. Two of the tickets carry a prize of Rs. 5 each, the other four a prize of Rs. 1. (a) If one ticket is drawn, what is the expected value of the prize? (b) If two tickets are drawn, what is the expected value of the game?*

Solution: (*a*) The sample space consists of $^6C_1 = 6$ sample points. Let X be the random variable associated with the experiment and let it denote the amount of prize associated with the sample point. Here X assumes values Rs. 5 and Rs. 1 respectively for 2 and 4 sample points.

Also, $$p(5) = \frac{2}{6} = \frac{1}{3} \text{ and } p(1) = \frac{4}{6} = \frac{2}{3}$$

∴ $$E(X) = Expected \text{ value of the prize}$$
$$= x_1.p(x_1) + x_2.p(x_2)$$
$$= 5 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{5}{3} + \frac{2}{3} = \frac{7}{3} = \text{ Rs. } 2.33$$

The expected amount of prize is Rs. 2.33.

(b) The sample space consists of $^6C_2 = 15$ sample points. Let X be a random variable associated with the experiment and let it denote the amount of prize associated with sample points. Then X assumes the following values.

(i) Rs. 10 (when both these tickets carry a prize Rs. 5 each - Number of sample points $^2C_2 = 1$).

(ii) Rs. 6 (when one ticket carries a prize of Rs. 5 and the other Rs. 1 -Number of sample points $= {}^2C_1 \times {}^4C_1 = 8$).

(iii) Rs. 2 (when both the tickets carry a prize Rs. 1 each- No. of sample points $= {}^4C_2 = 6$).

$$\text{also} \qquad p(10) = \frac{1}{15}, p(6) = \frac{8}{15}, p(2) = \frac{6}{15}$$

$$\therefore\ E(X) = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + x_3 \cdot p(x_3)$$

$$= 10 \times \frac{1}{15} + 6 \times \frac{8}{15} + 2 \times \frac{6}{15}$$

$$= \frac{2}{3} + \frac{16}{5} + \frac{4}{5} = \frac{10 + 48 + 12}{15} = \frac{70}{15} = \frac{14}{3} = 4.67$$

Hence expected amount of prize is Rs. 4.67.

Example 4: The probability that there is at least one error in an accounts statement prepared by A is 0.2 and for B and C they are 0.25 and 0.4, respectively. A, B and C prepared 10,16, and 20 statements respectively. Find the expected number of correct statements in all.

Solution: We are given that

$$P(A) = 0.2,\ \ P(B) = 0.25 \text{ and } P(C) = 0.4,$$

where A, B, C stand for the error in accounts prepared respectively by A, B, and C.

$$\therefore\ P(\text{ no error by } A) = P(\bar{A}) = 1 - 0.2 = 0.80$$
$$P(\text{ no error by } B) = P(\bar{B}) = 1 - 0.25 = 0.75$$
$$P(\text{ no error by } C) = P(\bar{C}) = 1 - 0.40 = 0.60$$

Expected number of correct statements in all

$$= n_1 P(\bar{A}) + n_2 P(\bar{B}) + n_3 P(\bar{C})$$

(where n_1, n_2 and n_3 are the number of statements prepared respectively by A, B and C)

$$= 10 \times 0.80 + 16 \times 0.75 + 20 \times 0.60$$
$$= 8 + 12 + 12 = 32$$

Advantages of expected value:

1. **Summarizes Central Tendency:** It provides a single number that represents the average outcome of a random variable, making complex distributions easier to understand.
2. **Supports Decision-Making:** In fields like economics, finance, and insurance, expected value helps make rational decisions under uncertainty by comparing average outcomes.
3. **Useful in Risk Assessment:** It helps evaluate the long-term performance or profitability of uncertain processes or investments.
4. **Mathematically Convenient:** The expected value is easy to manipulate algebraically and has useful linearity properties, *i.e.*,

$$E(aX+b)=aE(X)+b$$

5. **Foundational in Probability and Statistics:** Expected value is a key component in calculating other statistical measures like variance, standard deviation, and moments.
6. **Applicable to Both Discrete and Continuous Variables:** The concept is flexible and applies across different types of probability distributions.
7. **Basis for Law of Large Numbers:** It plays a central role in this law, which states that sample averages converge to the expected value as the sample size increases.

Limitations of expected value:

1. **Ignores Variability:** Expected value does not provide any information about the spread or variability of outcomes. Two distributions can have the same expected value but very different risks.
2. **Not Always Realistic or Achievable:** The expected value may represent an average outcome that is not actually possible in real-world scenarios (e.g., an expected value of 2.5 heads in coin tosses).
3. **Misleading in Small Samples:** For small sample sizes or short-term decisions, the expected value might not reflect the actual likely outcomes.
4. **Sensitive to Extreme Values:** If a distribution includes very large or small outliers, the expected value can be heavily skewed and may not represent a "typical" value.
5. **Assumes Known Probabilities:** Calculating the expected value requires accurate knowledge of all possible outcomes and their probabilities, which may not always be available.

6. **Not Suitable for All Decision Contexts:** In situations involving high risk or where losses must be minimized, relying solely on expected value can lead to poor or risky decisions.
7. **Does Not Reflect Preference or Utility:** It does not account for individual risk preferences or utility, which are important in economics and behavioral decision-making.

6.7 *VARIANCE*

The variance of the probability distribution of a random variable X is the average squared deviation measured from the expected value of a probability distribution of X. Symbolically

$$V(X) = \sigma^2 = \sum_{i=1}^{n} [x_i - E(X)]^2 p(x_i).$$

Illustration: A dealer in television sets estimates from his past experience the probabilities of his selling television sets in a day. These are given below:

No. of television sets sold in a day	0	1	2	3	4	5	6
Probability	0.02	0.10	0.21	0.32	0.20	0.09	0.06

We observe that the number of television sets sold in a day is a random variable that can assume the values 0,1,2,3,4,5,6 with the respective probabilities given in the table. We may also note that the dealer has estimated the probability of selling seven or more sets in a day to be zero. Now,

Mean number of TV sets sold in a day

$$= 0 \times .02 + 1 \times .10 + 2 \times .21 + 3 \times .32 + 4 \times .20 + 5 \times .09 + 6 \times .06$$
$$= 3.09$$

Variance of the number of TV sets sold in a day

$$= (0-3.09)^2 \times .02 + (1-3.09)^2 \times .10$$
$$+(2-3.09)^2 \times .21 + (3-3.09)^2 \times .32$$
$$+(4-3.09)^2 \times .20 + (5-3.09)^2 \times .09$$
$$+(6-3.09)^2 \times .06$$
$$=1.871060$$

S.D. of the number of TV sets sold $= \sqrt{1.871060} = 1.36$.

Example 5: Anil's company estimates the net profit on a new product it is launching to be Rs. 3,000,000 during the first year, if it is successful: Rs. 10,00,000 if it is 'moderately successful' and a loss of Rs. 10,00,000, if it is 'unsuccessful'. The firm assigns the following probabilities to the first year for the product. Successful-0.15, moderately successful-0.25. What are the expected value and standard deviation of the first-year net profit for this product?

Solution: Regarding loss as negative profit, the probability distribution of net profit (in million Rs.) X on the new product in the first year is:

COMPUTATION OF EXPECTED VALUE AND VARIANCE

Value of the random variable (X) (in million Rs.)	Probability $\boldsymbol{p(x)}$	Product $\boldsymbol{p(x)}$	$[\boldsymbol{X - E(X)}]$	$[\boldsymbol{X - E(X)}]^2\boldsymbol{p(x)}$
(1)	(2)	(3)	(4)	(5)
3	0.15	0.45	2.90	$(2.90)^2 \times 0.15$
1	0.25	0.25	0.90	$(0.90)^2 \times 0.25$
-1	$1 - 0.15 - 0.25 = 0.60$	-0.60	-1.10	$(-1.10)^2 \times 0.60$
Total		$0 \cdot 10$		$2 \cdot 19$

$$E(X) = \sum xp(x) = 0.10 \text{ million Rs.} = \text{Rs. } 1,00,000$$
$$V(X) = \sum [x - E(X)]^2 p(x) = 2.19 \text{ (million Rs.)}$$
$$\therefore \quad \sigma_x = \sqrt{2.19} = 1.48 \text{ (million Rs.)}$$

Advantages of variance:

1. **Measures Data Dispersion:** Variance quantifies how much the values of a dataset differ from the mean, helping to understand the spread or variability in the data.
2. **Mathematically Useful:** It has desirable mathematical properties that make it suitable for algebraic manipulation and theoretical analysis in probability and statistics.
3. **Foundation for Standard Deviation and Other Metrics:** Variance is the basis for calculating the standard deviation, which is widely used to interpret variability in the same units as the data.
4. **Essential in Statistical Modeling:** Variance plays a crucial role in regression, hypothesis testing, ANOVA, and other inferential statistical techniques.
5. **Highlights Risk and Uncertainty:** In fields like finance and insurance, variance is used to measure the risk associated with uncertain outcomes.
6. **Applicable to Both Discrete and Continuous Variables:** Variance can be computed for any type of random variable, making it a versatile measure in statistics.
7. **Supports the Law of Large Numbers and Central Limit Theorem:** Variance is critical in establishing the behavior of sample means and distributions in large samples.

Limitations of variance:

1. **Uses Squared Units:** Variance is expressed in the square of the original data units (e.g., cm^2, kg^2), making it less intuitive and harder to interpret directly.
2. **Affected by Extreme Values (Outliers):** Since variance involves squaring deviations, it gives disproportionate weight to large differences, making it sensitive to outliers.
3. **Not Easily Comparable Across Datasets:** Because it's in squared units, variance is not easily comparable between datasets with different units or scales.
4. **Less Useful for Direct Interpretation:** Unlike standard deviation, which is in the same units as the data, variance is abstract and less meaningful without context.
5. **Assumes Mean is the Best Measure of Center:** Variance is based on deviations from the mean, which may not be suitable if the data is skewed or has a non-normal distribution.
6. **May Be Misleading in Small Samples:** In small samples, variance can be unstable and not accurately reflect the true variability in the population.
7. **Ignores Direction of Deviation:** Variance treats all deviations from the mean equally (squares them), so it doesn't indicate whether data tends to be above or below the mean.

Check your progress

Problem 1: A box contains 4 white and 6 black balls. A person drawn 2 balls and is given Rs. 14 for every white ball and Rs. 7 for every black ball. What is his expectation?

Problem 2: What are the limitations of expected value?

Problem 3: What are the characteristic of variance?

6.8 *SUMMARY*

The unit on expected value and variance focuses on two essential statistical measures used to describe the behavior of random variables. The expected value gives the average or mean outcome of a probability distribution, while the variance measures how much the values spread out from the mean. These measures help in understanding the distribution's shape, consistency, and reliability. The chapter includes methods for calculating expected value and variance for both discrete and continuous random variables, explores their properties, and demonstrates their use in practical applications like risk analysis, prediction, and decision-making. These concepts are fundamental to probability theory and serve as a base for further statistical methods and models.

6.9 *GLOSSARY*

- Statistics
- Statistics as a data
- Statistics as a method
- Limitation of statistics
- Distrust of statistics
- Fallacies in statistics

6.10 *REFERENCES*

- S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
- Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
- J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.

6.11 *SUGGESTED READINGS*

- https://www.wikipedia.org.
- A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
- R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
- Jim Pitman, (1993), Probability, Springer-Verlag.
- OpenAI. ChatGPT (GPT-4 model) [Large language model]. https://chat.openai.com/
- https://archive.nptel.ac.in/courses/111/105/111105090

6.12 *TERMINAL QUESTIONS*

Long answer type questions

1: A company introduces a new product in the market and expects to make a profit of Rs. 2.5 lakhs during the first year if the demand is ‘good’, Rs. 1.5 lakhs if the demand is ‘moderate’ and a loss of Rs. 1 lakh if the demand is ‘poor’. Market research studies indicate that the probabilities for the demand to be good and moderate are 0.2 and 0.5 respectively. Find the company’s expected profit and the standard deviation.

2: A and B play for a prize of Rs. 99. The prize is to be won by a player who first throws a ‘2’, with one die. A first throws, and if he fails B throws and if he fails A again throws, and so on. Find their respective expectations.

Short answer type questions

1: A balanced coin is tossed 4 times. Find probability distribution of the number of heads and its expectation.

2: In a business venture a man can make a profit of Rs. 2000 with a probability of 0.4 or have a loss of Rs. 1000 with a probability of 0.6. What is his expected profit?

3: If it rains a dealer in raincoats can earn Rs.400 per day. If it is fair he can lose Rs. 50 per day. What is his expectation if the probability of a fair day is 0.6?

4: If the probability that the value of a certain stock will remain the same is 0.46, the probabilities that its value will increase by Re. 0.50 or Re. 1.00 per share are respectively 0.17 and 0.23, and the probability that its will value will decrease by Re. 0.25 per share is 0.14, what is the expected gain per share?

5: A box contains 8 items of which 2 are defective. A man selects 3 items at random. Find the expected number of defective items he has drawn.

6: A player tosses 3 fair coins. He wins Rs. 5 if 3 heads appear, Rs. 3 if 2 heads appear, Re. 1 if 1 head occurs. On the other hand, he loss Re. 15 if 3 tails occur. Find expected gain of the player.

Objective type questions:

1: The expected value of a random variable is also known as:

A) Mode

B) Median

C) Mean

D) Range

2: Which of the following is not true about variance?

A) Variance is always non-negative

B) Variance measures the spread around the mean

C) Variance is expressed in the same units as the data

D) Variance is the square of standard deviation

3: If a constant is added to every value of a random variable, its variance:

A) Increases

B) Decreases

C) Remains unchanged

D) Becomes zero

4: The expected value of a constant is:

A) 0

B) 1

C) The constant itself

D) Undefined

5: Which of the following is true for expected value $E(X)$?

A) $E(aX) = a + E(X)$

B) $E(aX + b) = aE(X) + b$

C) $E(X + Y) = E(X).E(Y)$

D) $E(X^2) = [E(X)]^2$

6: If the variance of a random variable is 0, then:

A) The mean is 0

B) The random variable is constant

C) All values are different
D) Data has maximum spread

7: Which of the following is used to measure variability?
A) Mean
B) Variance
C) Median
D) Mode

8: Standard deviation is:
A) The square of variance
B) The square root of variance
C) The same as variance
D) Variance minus mean

9: For a discrete random variable ***X***, the expected value is calculated as:
A) $\sum x_i$
B) $\sum P(x_i)$
C) $\sum x_i.P(x_i)$
D) $\sum x_i^{\ 2}.P(x_i)$

10: Which of the following affects the value of variance?
A) Addition of a constant to all values
B) Multiplying all values by a constant
C) Taking square root of all values
D) None of the above

Fill in the blanks questions:

1: The expected value is also called the ___________ of a random variable.

2: The formula for expected value of a discrete random variable X is $E(X) = \sum x_i.$ ___________

3: The variance of a random variable is defined as $Var(X) = \ldots\ldots\ldots\ldots\ldots$

4: Variance is always ___________ or zero.

5: The square root of variance is called ___________.

6: If a constant is added to all values of a random variable, the variance ___________.

7: If all outcomes of a random variable are the same, then the variance is ___________.

8: The expected value is a measure of ___________ tendency.

9: Variance helps in understanding the ___________ of data around the mean.

10: The expected value of a constant *c* is ___________.

True and False questions

1. The expected value is always equal to one of the possible outcomes of a random variable.

2. Variance can be negative.

3. The expected value of a constant is the constant itself.

4. Adding a constant to all values of a random variable changes its variance.

5. Standard deviation is the square of the variance.

6. Expected value is a measure of central tendency.

7. Variance measures the spread of data around the mean.

8. The expected value is not defined for continuous random variables.

9. Multiplying all values of a random variable by a constant multiplies the variance by the square of that constant.

10. If the variance of a random variable is zero, then all its values must be the same.

6.13 *ANSWERS*

Check your progress answer:

Answer of problem 1: 19.60

Answer of long questions:

1:	0.95 and 1.331	**2:**	54 and 45

Answer of short questions:

2:	200	**3:**	130	**5:**	3/4
6:	0.25				

Answer of objective questions

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

Answer of fill in the blanks

1:	mean	**2:**	$P(x_i)$	**3:**	$E[X - E(X)]^2$
4:	non-negative	**5:**	Standard Deviation	**6:**	Remains unchanged
7:	Zero	**8:**	Central	**9:**	Spread Or Dispersion

10: *c*

Answer of True and False:

1:	False	**2:**	False	**3:**	True
4:	False	**5:**	False	**6:**	True
7:	True	**8:**	False	**9:**	True
10:	True				

COURSE NAME: BASIC STATISTICS

COURSE CODE: MT(N)-222

BLOCK - III:

MOMENT GENERATING FUNCTION

UNIT 7: *MOMENTS*

CONTENTS:

7.1 *INTRODUCTION*

In statistics, moments are quantitative measures that describe the characteristics of a probability distribution. They are essentially expected values of different powers of a random variable, providing information about the distribution's shape, center, and spread.

In mathematics, the **moments** of a function are certain quantitative measures related to the shape of the function's graph. If the function represents mass density, then the zeroth moment is the total mass, the first moment (normalized by total mass) is the center of mass, and the second moment is the moment of inertia. If the function is a probability distribution, then the first moment is the expected value, the second central moment is the variance, the third standardized moment is the skewness, and the fourth standardized moment is the kurtosis.

7.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Moments

(ii) Kurtosis

7.3 *MOMENTS*

Moment is a familiar mechanical term for the measure of a force with reference to its tendency to produce rotation. The strength of this tendency depends, obviously, upon the amount of the force and the distance from the origin of the point at which the force is exerted."

The term 'moment' in mechanics refers to the turning or rotating effect of a force. In statistics, it is used to describe the peculiarities of a frequency distribution. Using moments, one can measure the central tendency of a set of observations, their scatter, i.e., dispersion, their asymmetry and the peaked Ness of the curve.

- **Definition:**

Moments are calculated as the expected value of a random variable raised to a power (e.g., E(X), E(X^2), E(X^3), etc.).

Common Moments and their Significance:

1. First Moment (Mean):

The expected value of the random variable, representing the "center" or average of the distribution.

2. Second Moment (Variance/Standard Deviation):

Measures the spread or dispersion of the data around the mean.

3. Third Moment (Skewness):

Indicates the asymmetry of the distribution. A positive skew means a longer tail to the right, and a negative skew means a longer tail to the left.

4. Fourth Moment (Kurtosis):

Measures the "tailedness" or peakedness of the distribution. It indicates whether the distribution has heavy tails (high kurtosis) or light tails (low kurtosis).

7.4 *MOMENTS ABOUT THE MEAN*

The moments about the mean are generally represented by μ (read as mu). Thus, the various moments about the mean would be:

In a series of Individual Observations:

$$\mu_1 = \frac{1}{N}\Sigma(X - \bar{X}), \quad \mu_2 = \frac{1}{N}\Sigma(X - \bar{X})^2$$
$$\mu_3 = \frac{1}{N}\Sigma(X - \bar{X})^3, \quad \mu_4 = \frac{1}{N}\Sigma(X - \bar{X})^4$$

Where $\mu_1, \mu_2, \mu_3, \mu_4, \dots\dots$ are called first, second, third fourth,…….moments respectively.

For a frequency distribution,

$$\mu_1 = \frac{1}{N}\Sigma f_i(X_i - \bar{X}), \quad \mu_2 = \frac{1}{N}\Sigma f_i(Xi - \bar{X})^2$$

$$\mu_3 = \frac{1}{N}\Sigma f_i(X_i - \bar{X})^3 \text{ and } \mu_4 = \frac{1}{N}\Sigma f_i(X_i - \bar{X})^4$$

If the actual mean comes out to be in fractions, it becomes very tedious to compute moments by the direct method. In such cases, we first calculate moments about 'a working origin'. Symbolically

$$\mu_1' = \frac{1}{N}\Sigma(X_i - A), \qquad \mu_2' = \frac{1}{N}\Sigma(X_i - A)^2$$
$$\mu_3' = \frac{1}{N}\Sigma(X_i - A)^3, \qquad \mu_4' = \frac{1}{N}\Sigma(X_i - A)^4$$

For a frequency distribution.

$$\mu_1' = \frac{1}{N}\Sigma f_i(X_i - A), \qquad \mu_2' = \frac{1}{N}\Sigma f_i(X_i - A)^2,$$
$$\mu_3' = \frac{1}{N}\Sigma f_i(X_i - A)^3, \qquad \mu_4' = \frac{1}{N}\Sigma f_i(X_i - A)^4$$

- **Relation between Moments from Working Origin and Moments about actual Mean.**

These are important relations between central and non-central moments as follows:

$$\mu_2 = \mu_2' - (\mu_1')^2$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - (\mu_1')^4$$

In particular, using the first two moments μ_1' and μ_2' about an arbitrary origin A, the mean and the variance may be computed as follows:

$\bar{X} = \mu_1' + A, \sigma^2 = \mu_2' - \mu_1'^2$.

- ## Importance of Moments

1. The first central moment is always zero, i.e., $\mu_1 = 0$.
2. The second central moment about the mean indicates the variance, i.e., $\mu_2 = \sigma^2$. Thus, the standard deviation is the square root of the second central moment.
3. The third central moment μ_3 is used to measure skewness. Karl Pearson has suggested a different measure of skewness based on the third and second central moments, defined as follows:
4.

$$\beta_1 = \frac{\mu_3{}^2}{\mu_3{}^2}$$

R.A. Fisher has introduced the notation:

$$\gamma_1 = +\sqrt{\beta_1}$$

In a symmetrical distribution:

$\mu_3 = 0, \beta_1 = 0$ and $\gamma_1 = 0$.

7.5 *SHEPPARD'S CORRECTION FOR MOMENTS*

In the case of grouped frequency distribution, when calculating moments, we assume that the frequencies are concentrated at the center (or middle point) of the class intervals. However, this assumption is not always true in practice, and an error called the grouping error can creep into the calculation of the moments.

W.F. Sheppard has proved that if:

(*i*) the frequency distribution is continuous, and
(*ii*) the frequency tapers off to zero in both directions, the effect of grouping at the midpoints of the intervals can be corrected using the following formulas, known as Sheppard's corrections:

$$\mu_2(\text{ corrected }) = \mu_2 - \frac{h^2}{12}$$
$$\mu_3(\text{ corrected }) = \mu_3$$
$$\mu_4(\text{ corrected }) = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4$$

where h is the width of class interval.

- **COEFFICIENTS BASED ON MOMENTS**

On the basis of moments some very useful coefficients known as alpha (α), beta (β) and gamma (γ) coefficients have been devised.

Alpha coefficients	Beta coefficients	Gamma coefficients
$\alpha_1 = \frac{\mu_1}{\sigma} = 0$ $\alpha_2 = \frac{\mu_2}{\sigma^2} = 1$ $\alpha_3 = \frac{\mu_3}{\sigma^8} = \frac{\mu_3}{{\mu_2}^{3/2}}$ $\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{{\mu_2}^2}$	$\beta_1 = {a_3}^2$ $= \frac{{\mu_3}^2}{{\mu_2}^3}$ $\sqrt{\beta_1} = -\frac{\mu_3}{{\mu_2}^{3/2}}$ $\beta_2 = \frac{\mu_4}{{\mu_2}^2} = a_4$	$\gamma_1 = \sqrt{\beta_1} = a_3$ $\gamma_2 = \beta_2 - 3$ $= \frac{\mu_4}{{\mu_2}^2} - 3$

β coefficients are used for calculating mode, skewness and kurtosis. γ_1 and γ_2 are used to measure skewness and kurtosis.

Remark. It may be noted that these coefficients are pure numbers independent of units of measurement and as such can be conveniently used for comparative studies.

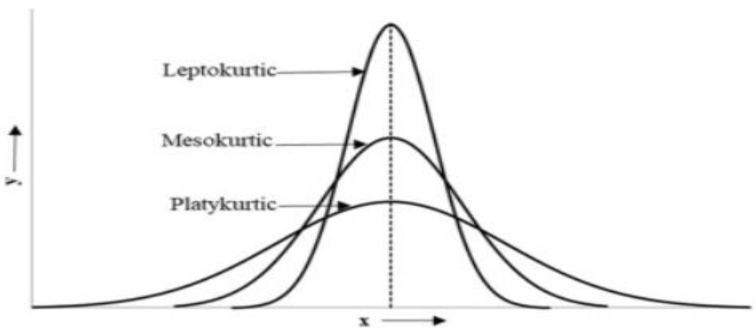
7.6 *KURTOSIS*

Kurtosis is yet another measure that tells us about the form of a distribution. It tells us whether the distribution, if plotted on a graph, would give us a normal curve, a curve more flat than the normal curve or a curve more peaked than the normal curve. The word kurtosis in the Greek language means 'bulginess'.

In the words of Simpson and Kafka 'The degree of Kurtosis of a distribution is measured relative to the peakedness of a normal curve'.

Croxton and Cowden state that 'A measure of Kurtosis indicates the degree to which a curve of a frequency distribution is peaked or flat topped."

Karl Pearson, in 1905, introduced three broad patterns of peakedness, which are illustrated in the following diagram.



https://www.sciencedirect.com/topics/social-sciences/kurtosis

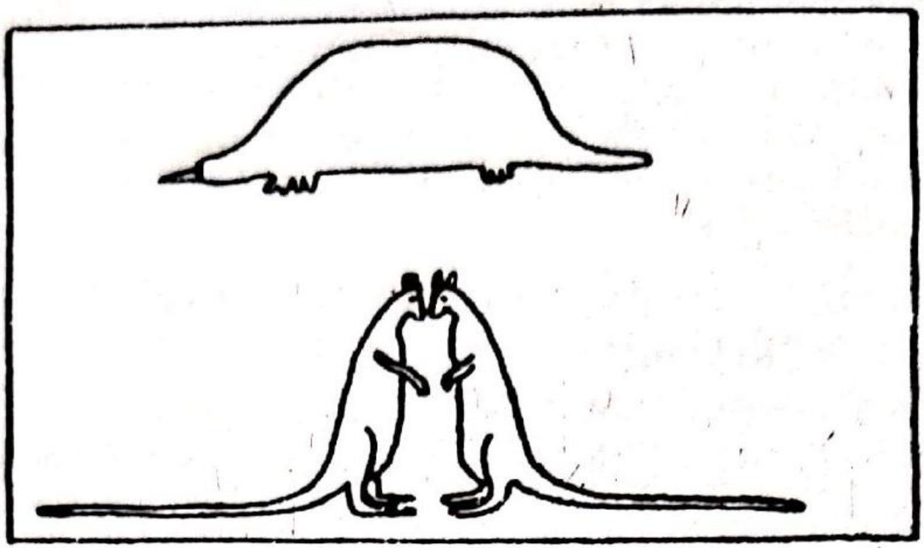
(i) A peaked curve is called 'Leptokurtic', and is said to lack kurtosis or to have negative kurtosis.

(ii) An intermediate peaked curve which is neither flat-topped nor peaked is known as normal or 'Mesokurtic' curve,

(iii) A flat-topped curve is termed 'Platykurtic' and possesses kurtosis In excess or have positive kurtosis.

It is interesting to quote here the words of a British statistician W.S. Gosset, (who wrote under the pen name of Student), who very humorously explains the use of the terms platykurtic and leptokurtic in the following sentence: "Platykurtic curves, like the platypus, are squat with short tails; leptokurtic curves are high with long tails like the kangaroos noted for leaping."

Gosset also gave a sketch to illustrate his definition. It is represented below:



In the above figure the top curve is Platykurtic and the bottom one is Leptokurtic.

- **MEASURES OF KURTOSIS**

As a measure of kurtosis, Karl Pearson gave the coefficient Beta two (β_2) or its derivative Gamma two (γ_2) defined as follows :

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} = \frac{\mu_4}{\sigma^4} = \alpha_4$$

$$r_2 = \beta_2 - 3 = a_4 - 3 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_1 - 3\sigma^4}{\sigma^4}$$

For a normal or meso-kurtic curve $\beta_2 = 3$ or $\gamma_3 = 0$. For a leptokurtic curve, $\beta_2 > 3$ or $\gamma_2 > 0$ and for a platykurtic curve, $\beta_2 < 3$ or $\gamma_2 < 0$.

The measure of kurtosis helps, among other things, in the choice of an average. For example, the mean is preferable for a normal distribution; for a leptokurtic distribution, the median is suitable, and for a platykurtic distribution, the quartile range is more appropriate.

Example 1. The calculation of the first, second, third, and fourth moments about the mean in the set of numbers (2,3,7,8,10) are shown below:

X	$(X - \bar{X})$	$(X - \bar{X})^2$	$(X - \bar{X})^3$	$(X - \bar{X})^4$
2	-4	16	-64	256
3	-3	9	-27	81
7	1	1	1	1
8	2	4	8	16

10	4	16	64	256
($\sum X = 30$)	0	46	-18	610

$$\bar{X} = \frac{\sum X}{n} = \frac{30}{5} = 6$$

$$\mu_1 = \frac{0}{5} = 0;\ \mu_2 = \frac{46}{5} = 9.2;\ \mu_3 = \frac{-18}{5} = -3.6;\ \mu_4 = \frac{610}{5} = 122$$

$$\therefore\ \text{Skewness } (\beta_1) = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{12.96}{778.688} = 0.0166 \text{ and}$$

$$\text{Kurtosis } (\beta_2) = \frac{\mu_4}{{\mu_2}^2} = \frac{122}{(9.2)^2} = 1.4$$

Therefore, the distribution is platykurtic, or it is flat at the top.

Example 2. From the following data, calculate the first four moments (i) about the value 15, (ii) about the mean, (iii) skewness based on moments, and (iv) kurtosis.

Class interval :	$0-10$	$10-20$	$20-30$	$30-40$
Frequency :	1	3	4	2

Solution: COMPUTATION OF MOMENTS

Class interval	Frequency (f)	Mid-point (m)	$d = \frac{m-15}{10}$	fd	fd^2	fd^3	fd^4
0-10	1	5	-1	-1	1	-1	1
10-20	3	15	0	0	0	0	0
20-30	4	25	1	4	4	4	4
30-40	2	35	2	4	8	16	32
Total	10			7	13	19	37

(i) $\mu_1' = \left(\frac{1}{N}\Sigma fd\right) \times h = \frac{7}{10} \times 10 = 7$

$\mu_2' = \left(\frac{1}{N}\Sigma fd^2\right) \times h^2 = \frac{13}{10} \times (10)^2 = 130$

$\mu_3' = \left(\frac{1}{N}\Sigma fd^3\right) \times h^3 = \frac{19}{10} \times (10)^3 = 1900$

$\mu_4' = \left(\frac{1}{N}\Sigma fd^4\right) \times h^4 = \frac{37}{10} \times (10)^4 = 37{,}000$

(ii) First moment about mean or $\mu_1 = 0$

Second moment or $\mu_2 = \mu_2{}' - \mu_1{}'^2$

$$= 130 - (7)^2 = 81$$

Third moment or $\mu_3 = \mu_3{}' - 3\mu_2{}'\mu_1{}' + 2\mu_1{}'^3$

$$= 1900 - 3 \times 130 \times 7 + 2(7)^3$$
$$= 1900 - 2730 + 686 = -144$$

Fourth Moment or $\mu_4 = \mu_4{}' - 4\mu_3{}'\mu_1{}' + 6\mu_2{}'\mu_1{}'^2 - 3\mu_1{}'^4$

$$= 37{,}000 - 4 \times 1900 \times 7 + 6 \times 130 \times (7)^2 - 3(7)^4$$
$$= 37{,}000 - 53200 + 38220 - 7203 = 14817$$

(iii) Skewness based on moments is studied by the calculation of β_1

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{(-144)^2}{(81)^3} = \frac{20736}{531441} = 0.039$$

Since β_1 is 0.038 , the distribution is skew. The skewness is negative as μ_3 has a negative value.

(iv) Kurtosis is studied by β_2

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} = \frac{14817}{(81)^2} = \frac{14817}{5561} = 2.26$$

Since the value of β_2 is less than 3, there is kurtosis in the curve. The curve is not normal. It is platykurtic,

Also

$$\gamma_2 = \beta_2 - 3 = 2.2 - 3 = -0 \cdot 8$$

If γ_2 is negative, it indicates that the curve is platykurtic. If γ_2 is a positive figure the curve is leptokurtic. If $\gamma_2 = 0$, the curve is mesokurtic.

Example 3. The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and the variance.

Solution. In the usual notations, we are given that

$$A = 5, \mu_1' = 2, \text{ and } \mu_2' = 20$$

We know that

$$\text{Mean } = A + \mu_1' = 5 + 2 = 7$$

and

$$\text{variance } = \mu_2 = \mu_2' - \mu_1'^2 = 20 - 4 = 16$$

Example 4. The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

Solution. Skewness is given by the formula:

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3} = \frac{(0.7)^2}{(2.5)^3} = +0.031$$

The distribution is not perfectly symmetrical as $\beta_1 = +0.03$. Kurtosis is given by the formula

$$\beta_2 = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

The distribution is mesokurtic or normal as $\beta_2 = 3$.

Example 5. The first four moments of a distribution about the value 5 are 2, 20, 40 and 50. Obtain, as far as possible, the various characteristics of the distribution on the basis of information given. Comment upon the nature of the distribution.

Solution. Here, in usual notations, we are given

$$A = 5, \mu_1' = 2, \mu_2' = 20, \mu_3' = 40, \text{ and } \mu_4' = 50$$

(i) Arithmetic mean $(\bar{X}) = A + \mu_1' = 2 + 5 = 7$
(ii) Variance $(\sigma^2) = \mu_2{}' - (\mu_1')^2 = 20 - (2)^2 = 16$

(iii)

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$
$$= 40 - 3 \times 20 \times 2 + 2 \times (2)^3 = -64$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$
$$= 50 - 4 \times 40 \times 2 + 6 \times 20 \times (2)^2 - 3 \times (2)^4 = 162$$
$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3} = \frac{(-64)^2}{(16)^3} = 1; \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-64}{16^{3/2}} = -1$$

Since $\gamma_1 < 0$, the distribution is negatively skewed.

(iv)

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2} = \frac{162}{256} = 0.63 \ \therefore \gamma_2 = \beta_2 - 3 = 0.63 - 3 = -2.37$$

Since $\gamma_2 < 0$, The distribution is platykurtic.

Example 6. The first three moments of a distribution about the value 2 of the variable are 1,16 and - 40. Show that the mean is 3, the variance is 15, and μ_3 is -86 . Also, show that the first three moments about $x = 0$ are 3, 24, and 76.

Solution. Here $\mu_1{}' = 1,\ \mu_2{}' = 16,\ \mu_3{}' = -40,\ A = 2$
(i) Mean = $A + \mu_1{}' = 2 + 1 = 3$
(ii) Variance is the second moment about the mean is (μ_2).

$$\mu_2 = \mu_2' - \mu_1' = 16 - (1)^2 = 15$$

(iii) Third moment about the mean is μ_3.

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -40 - 3 \times 1 \times 16 + 2(1)^3$$
$$= -40 - 48 + 2 = -86$$

(iv) The moments about the origin are given by:

$$v_1 = A + \mu_1' = 2 + 1 = 3$$
$$v_2 = \mu_2 + v_1{}^2 = 15 + (3)^2 = 24$$
$$v_3 = \mu_3 + 3\mu_2 v_1 + v_1^3$$
$$= -86 + 3 \times 15 \times 3 + (3)^3 = 76.$$

CHECK YOUR PROGRESS

True or false Questions

Problem 1. The first central moment is always zero.

Problem 2. The third central moment μ_3 is used to measure skewness.

Problem 3. A peaked curve is called 'Leptokurtic'.

7.7 *SUMMARY*

1. The first central moment is always zero, i.e., $\mu_1 = 0$.
2. The second central moment about the mean indicates the variance, i.e., $\mu_2 = \sigma^2$. Thus, the standard deviation is the square root of the second central moment.

7.8 *GLOSSARY*

(i) Permutation

(ii) Combinations

(iii) Probability

7.9 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.

5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

7.10 *TEWRMINAL QUESTIONS*

1. What is Kurtosis.
2. What is Moment.
3. Define Sheppard's correction for moments.

7.11 *ANSWERS*

CYQ 1. True

CYQ 2. True

CYQ 3. True

UNIT 8: *MOMENT GENERATING FUNCTION*

CONTENTS:

8.1 *INTRODUCTION*

In probability theory and statistics, the **moment-generating function** of a real-valued random variable is an alternative specification of its probability distribution. Thus, it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. There are particularly simple results for the moment-generating functions of distributions defined by the weighted sums of random variables. However, not all random variables have moment-generating functions.

As its name implies, the moment-generating function can be used to compute a distribution's moments: the n-th moment about 0 is the n-th derivative of the moment-generating function, evaluated at 0.

In addition to univariate real-valued distributions, moment-generating functions can also be defined for vector- or matrix-valued random variables, and can even be extended to more general cases.

The moment-generating function of a real-valued distribution does not always exist, unlike the characteristic function. There are relations between the behavior of the moment-generating function of a distribution and properties of the distribution, such as the existence of moments.

8.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Moment Generating Function

(ii) Cumulant Generating Function

8.3 *MOMENT GENERATING FUNCTION*

Definition

Let X be a random variable. Then $E(e^{tX})$ denoted by $M(t)$ is called moment generating function of X, where t is a real constant. The function. $M(t)$ generates moments, that is why it is called the moment gene ating function (in brief m.g.f.). Thus

$$M(t) = E(e^{tX}) \qquad \ldots\ldots\ldots..(1)$$
$$= E\left(1 + tX + \frac{t^2X^2}{2!} + \cdots . + \frac{tX^r}{r!} + \cdots .\right)$$
$$= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \cdots . + \frac{t^2}{r!}E(X') + \cdots .$$
$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \cdots . + \frac{t'}{r!}\mu_r' + \cdots . \qquad \ldots\ldots\ldots (2)$$

The coefficient of $\frac{t'}{r!}$ in the expression (2) is μ_r', the r th moment of X about origin. This shows that m.g.f. generate moments.

m.g.f. for a discrete random variable: Let X be a discrete random variable with probability distribution

x	:	x_1	x_2	….	x_n
$p(x)$	:	p_1	p_2	….	p_n

then

$$M(t) = E(e^{tX}) = \sum_{i=1}^{n} e^{tX} p_i \qquad \ldots\ldots\ldots.(3)$$

m.g.f. for a continuous random variable: Let X be a continuous random variable with probability density function.

$$f(x), -\infty < x < \infty \qquad \ldots\ldots\ldots\ldots (4)$$

Then $M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{\alpha} f(x)dx$

8.4 *MOMENT GENERATING FUNCTION ABOUT VARIOUS POINTS*

Moment Generating Function about Various Points

(i) $M_0(t) = E(e^{tX})$, is called m.g.f. about origin

(ii) $M_a(t) = E\left[e^{t(X-a)}\right]$, is called m.g.f. about any point a.

(iii) $M_{\bar{x}}(t) = E\left[e^{t(X-\bar{x})}\right]$, is called m.g.f. about mean.

$$M_a(t) = E\left[1 + t(X-a) + \frac{t^2}{2!}(X-a)^2 + \cdots . + \frac{t^r}{r!}(X-a)^r + \cdots .\right]$$
$$= 1 + t\mu_1'(a) + \frac{t^2}{2!}\mu_2'(a) + \cdots . + \frac{t^r}{r!}\mu_r'(a) + \cdots .$$

where $\mu_r'(a)$ is r th moment about a.

$$M_{\bar{x}}(t) = E\left[1 + t(X-\bar{x}) + \frac{t^2}{2!}(X-\bar{x})^2 + \cdots . + \frac{t^r}{r!}(X-\bar{x})^r + \cdots .\right]$$

$$= 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \cdots . + \frac{t^r}{r!}\mu_r + \cdots .$$

where μ_r = r th moment about mean or r^{th} central moment.

8.5 *EFFECT OF CHANGE OF ORIGIN AND SCALE ON M.G.F.*

Let X be a random variable with $M_X(t) = E(e^{tX})$.

Let $U = \frac{X-a}{h}$, where a and h are constants.

$\Rightarrow X = a + hU$

Then $M_U(t) = E(e^{tU}) = E\left[e^{t(X-a)/h}\right]$

$$\Rightarrow M_U(t) = e^{-\frac{nt}{h}} E\left(e^{\frac{tX}{h}}\right)$$

$$\Rightarrow M_U(t) = e^{-\frac{at}{h}} M_X\left(\frac{t}{h}\right) \qquad \ldots\ldots (1)$$

In other form

$$M_X(t) = E(e^{tX})$$

$$\Rightarrow \quad M_X(t) = E\left[e^{t(a+hU)}\right]$$

$$\Rightarrow \quad M_X(t) = e^{at} E(e^{thU})$$

$$\Rightarrow \quad M_X(t) = e^{at} M_U(th) \qquad \ldots\ldots\ldots (2)$$

- **Properties of m.g.f.**

Property 1. Let $X = cu$. Then

$$M_X(t) = E(e^{tX}) = E(e^{tcU}) = M_U(ct).$$

Property 2. Let $X = u + c$. Then

$$M_X(t) = E(e^{tX}) = E\left(e^{r(U+c)}\right) = e^{tc}.E(e^{tU}) = e^{tc} M_U(t).$$

Property 3. The moment generating function of the sum of two independent random variables in the product of their moment generating functions.

Proof: Let X and Y be two independent random variables, then the m.g.f. of the sum $X+Y$ is given by

$$M_{X+Y}(t) = Ee^{t(X+Y)}$$
$$\Rightarrow \quad M_{X+Y}(t) = E(e^{tX} \cdot e^{tY})$$
$$\Rightarrow \quad M_{X+Y}(t) = E(e^{tX}) \cdot E(e^{tY}), [\because X \text{ and } Y \text{ are independent }]$$
$$\Rightarrow \quad M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$$

Corollary: If $X_1, X_2, X_3, \ldots, X_n$ be n mutually indepedent random variables then $M \sum_{i=1}^{n} X_i(t) = M_{X_.}(t) M_{X_.}(t) \ldots . M_{X_.}(t)$.

8.6 *GENERATION OF MOMENTS BY M.G.F.*

There are two methods to generate moments by m.g.f. -

(i) **By expansion:**

$$M_X(t) = 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \cdots . + \frac{t'}{r!}\mu_r' + \cdots .$$

where $\mu_r' =$ coefficient of $\frac{t^r}{r!}$ in $M_X(t)$.

(ii) **By differentiation:**

$$\mu_r' = \frac{d^r M_X(t)}{dt^r}\bigg|_{t=0}$$

That is to find μ_r', differentiate $M_X(t)$ with respect to t, r times and put $= 0$.

Illustrative Examples

Example 1. Find the m.g.f. of the random variable X having the probability density function

$$f(x) = \begin{cases} x, & 0 \le x < 1 \\ 2 - x, & 1 \le x < 2 \\ 0, & \text{otherwise} \end{cases}$$

Also find the mean and variance of X using m.g.f.

Solution: $M_X(t) = E(e^{tX})$

$$= \int_{-\infty}^{\infty} e^{tX} f(x) dx$$

$$= \int_0^1 e^{tx} \cdot x dx + \int_1^2 e^{tx}(2-x) dx$$

$$= \left[\frac{e^{tx}}{t} \cdot x\right]_0^1 - \int_0^1 \frac{e^{tx}}{t} \cdot 1 dx + \left[\frac{e^{tx}}{t}(2-x)\right]_1^2 - \int_1^2 \frac{e^{tx}}{t}(-1) dx$$

$$= \frac{e^t}{t} - \left[\frac{e^{tx}}{t^2}\right]_0^1 + \frac{-e^t}{t} + \left[\frac{e^{tx}}{t^2}\right]_1^2$$

$$= \frac{e^t}{t} - \left(\frac{e^t}{t^2} - \frac{1}{t^2}\right) - \frac{e^t}{t} + \left(\frac{e^{2t}}{t^2} - \frac{e^t}{t^2}\right)$$

$$= \frac{e^{2t}}{t^2} - \frac{2e^t}{t^2} + \frac{1}{t^2}$$

$$= \frac{1}{t^2}(e^{2t} - 2e^t + 1) \qquad \ldots..(1)$$

$$= \frac{(e^t - 1)^2}{t^2} \qquad \ldots..(2)$$

Now expanding $M_X(t)$ given in (1), we have

$$M_X(t) = \frac{1}{t^2}\left[\left(1 + 2t + \frac{(2t)^2}{2!} + \frac{(2t)^3}{3!} + \frac{(2t)^4}{4!} + \cdots.\right)\right.$$

$$\left. -2\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots.\right) + 1\right]$$

$$\Rightarrow M_X(t) = \frac{1}{t^2}\left[t^2 + t^3 + \frac{7}{12}t^4 + \cdots.\right]$$

$$\Rightarrow M_X(t) = 1 + t + \frac{7}{12}t^2 + \cdots \qquad \ldots..(3)$$

Hence [from (3)]

$$\mu_1{}' = \text{ coefficient of t in } M_X(t) = 1 = \text{ Mean}$$

$$\mu_2{}' = \text{ coefficient of } \frac{t^2}{2!} \text{ in } M_X(t) = 2!\frac{7}{12} = \frac{7}{6}$$

$$\therefore \text{ Variance } = \mu_2 = \mu_2{}' - (\mu_1{}')^2 = \frac{7}{6} - 1 = \frac{7-6}{6} = \frac{1}{6}.$$

Example 2. Show that the m.g.f. of the random variable X having the probability density function

$$f(x) = \frac{1}{3}, \qquad -1 < x < 2$$
$$= 0, \qquad \text{elsewhere}$$

is $$M(t) = \frac{e^{2t} - e^{-t}}{3t}, \qquad t \neq 0$$
$$= 1, \qquad \text{when } t = 0.$$

Solution: $M(t) = E(e^{tX})$

$$= \int_{-1}^{2} f(x) \cdot e^{tx} dx = \int_{-1}^{2} \frac{1}{3} e^{tx} dx$$
$$= \frac{1}{3t}[e^{tx}]_{-1}^{2} = \frac{1}{3t}(e^{2t} - e^{-t}), t \neq 0$$

However, when $t = 0$, then

$$M(t) = \int_{-1}^{2} \frac{1}{3} e^{0} dx$$
$$= \int_{-1}^{2} \frac{1}{3} dx = \left[\frac{x}{3}\right]_{-1}^{2} = \frac{2-(-1)}{3} = \frac{3}{3} = 1$$

Thus $M(t) = \frac{e^{2t} - e^{-t}}{3t}, t \neq 0$

$$= 1, \text{ when } t = 0.$$

Example 3. Find the m.g.f. of a random variable X, whose probability function is given by

$$P(X = x) = p(1-p)^{x-1}, x = 1,2, \ldots$$

Solution:

$$M_X(t) = E(e^{tX})$$
$$= \sum_{x=1}^{\infty} e^{tx} \cdot p(1-p)^{x-1}$$
$$= \frac{p}{1-p} \sum_{x=1}^{\infty} e^{tx}(1-p)^{x}$$
$$= \frac{p}{1-p} \sum_{x=1}^{\infty} [(1-p)e^{t}]^{x}$$
$$= \frac{p}{1-p}\left[\sum_{x=0}^{\infty} [(1-p)e^{t}]^{x} - 1\right]$$
$$= \frac{p}{1-p}[\{1-(1-p)e^{t}\}^{-1} - 1]$$
$$= \frac{p}{q}[(1-qe^{t})^{-1} - 1], \text{ where } q = 1-p, \text{ say.}$$

Example 4. If $\mu_r{}' = r!\,2^r$, then find m.g.f.

Solution: $M_X(t) = 1 + \mu_1{}'\,t + \mu_2{}'\,\frac{t^2}{2!} + \mu_3{}'\,\frac{t^3}{3!} + \cdots$

$$= 1 + (1!\cdot 2)t + (2!\cdot 2^2)\cdot\frac{t^2}{2!} + (3!\cdot 2^3)\frac{t^3}{3!} + \cdots.$$
$$= 1 + 2t + (2t)^2 + (2t)^3 + \cdots.$$
$$= (1-2t)^{-1}.$$

8.7 *CUMULANT GENERATING FUNCTION*

Logarithm of the m.g.f. of a distribution is called cumulant generating function and is denoted by $K(t)$

Thus, $K(t) = \log_e M_0(t)$

$$= \kappa_1 t + \kappa_2\frac{t^2}{2!} + \kappa_3\frac{t^3}{3!} + \cdots \qquad \ldots\ldots\ (1)$$

where it is assumed that $\log_e M_0(t)$ can be expanded in a convergent series in powers of t.

The coefficients $\kappa_1, \kappa_2, \kappa_3$, ..., are called the first, second, third, ... **cumulants** of the distribution.

If we differentiate (1) r times with respect to t and put $t = 0$, we get

$$\kappa_r = \left[\frac{d^r}{dt^r}K(t)\right]_{t=0}$$

8.8 *CUMULANT IN TERMS OF MOMENTS*

We have

$$K(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \cdots . + \kappa_r \frac{t^r}{r!} + \cdots = \log M(t)$$
$$= \log_e \left[1 + \mu_1{}' t + \mu_2{}' \frac{t^2}{2!} + \mu_3{}' \frac{t^3}{3!} + \mu_4{}' \frac{t^4}{4!} + \cdots . + \mu_r{}' \frac{t^r}{r!} \ldots \right]$$
$$= \left(\mu_1{}' t + \mu_2{}' \frac{t^2}{2!} + \mu_3{}' \frac{t^3}{3!} + \mu_4{}' \frac{t^4}{4!} + \cdots \right)$$
$$- \frac{1}{2}\left(\mu_1{}' t + \mu_2{}' \frac{t^2}{2!} + \mu_3{}' \frac{t^3}{3!} + \cdots . \right)^2 + \frac{1}{3}\left(\mu_1{}' t + \mu_2{}' \frac{t^2}{2!} + \cdots . \right)^3$$
$$- \frac{1}{4}\left(\mu_1{}' t + \mu_2{}' \frac{t^2}{2!} + \cdots . \right)^4 + \cdots .$$

Hence from (1) and (2), we have

$$\kappa_1 t + \kappa_2 \frac{t^2}{2!} + \cdots . + \kappa_r \frac{t^r}{r!} + \cdots .$$
$$= (a + \kappa_1{}' h)t + \kappa_2{}' h^2 \frac{t^2}{2!} + \cdots . + \kappa_r{}' h' \frac{t^r}{r!} + \cdots .$$

Comparing the coefficients of $t, t^2, \ldots., t^r$ on both sides, we get

$$\kappa_1 = a + \kappa_i{}' h \qquad (3)$$
$$\kappa_r = \kappa_r{}' h \text{ for } r > 2 \qquad (4)$$

Thus, except the first cumulant all cumulants are unaltered by a change of origin.

But r th cumulant of x is h^r times the rth cumulant of Q.

Remark: Let $u = \frac{x}{h}$, then $\kappa_r = h'\kappa_r{}', r = 1,2, \ldots.$

where κ_r is the r th cumulant of x and $\kappa_r{}'$ is the rth cumulant of u.

Theorem 1. The cumulant generating function of the sum of two independent random variables is the sum of their cumulant generating functions.

Proof: Let X and Y be two independent random variables; then we have

$$M_{x+y}(t) = M_x(t) M_y(t)$$
$$\log M_{x+y}(t) = \log M_x(t) + \log M_y(t)$$
$$K_{x+y}(t) = K_x(t) + K_y(t)$$

Theorem 2. The r th cumulant of the sum of two independent random variables is the sum of their r th cumulants.

Proof: Let X and Y be two independent random variables we have

$$M_{x+y}(t) = M_x(t)M_y(t)$$
$$\text{or } \log M_{x+y}(t) = \log M_x(t) + \log M_y(t)$$
$$K_{x+y}(t) = K_x(t) + K_y(t)$$
$$\text{or } \kappa_{1(x+y)}t + \kappa_{2(x+y)}\frac{t^2}{2!} + \cdots.. + \kappa_r(x+y)\frac{t^r}{r!} + \cdots.$$
$$= \kappa_{1(x)}t + \kappa_{2(x)}\frac{t^2}{2!} + \cdots. + \kappa_{r(x)}\frac{t^r}{r!} + \cdots + \kappa_{1(y)}t + \kappa_{2(y)}\frac{t^2}{2!} + \cdots. + \kappa_{r(y)}\frac{t^r}{r!} + \cdots$$

Equating the coefficients of $\frac{t^r}{r!}$ we have

$$\kappa_{r(x+y)} = \kappa_{r(x)} + \kappa_{r(y)}$$

where $\kappa_{r(x+y)}$ = rth cumulant of X + Y.

Example 1. Find the moment generating function of x ~ f(x) = 1, where 0 <x< 1, and thereby confirm that E(x) = 1 / 2 and V (x) = 1 / 12.

Solution. The moment generating function is

$$M(x,t) = E(e^{xt}) = \int_0^1 e^{xt}dx$$
$$= \left[\frac{e^{xt}}{t}\right]_0^1 = \frac{e^t}{t} - \frac{1}{t}.$$

But

$$e^t = \frac{t^0}{0!} + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots,$$

so

$$M(x,t) = \left[\frac{1}{t} + 1 + \frac{t}{2!} + \frac{t^2}{3!} + \frac{t^3}{4!} + \cdots\right] - \frac{1}{t}$$
$$= 1 + \frac{t}{2} + \frac{t^2}{6} + \frac{t^3}{24} + \cdots.$$

By the process of differentiating M(x, t) with respect to t and the setting t = 0, we get

$$E(x) = \left.\frac{\partial M(x,t)}{\partial t}\right|_{t=0} = \left[\frac{1}{2} + \frac{2t}{3!} + \frac{3t^2}{4!} + \cdots\right]_{t=0} = \frac{1}{2},$$
$$E(x^2) = \left.\frac{\partial^2 M(x,t)}{\partial t^2}\right|_{t=0} = \left[\frac{2}{3!} + \frac{6t}{4!} + \cdots\right]_{t=0} = \frac{1}{3}.$$

Combining these results gives

$$V(x) = E(x^2) - \{E(x)\}^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

CHECK YOUR PROGRESS

True or false or MCQ Questions

Problem 1. If the moment generating function of a random variable X is $M_X(t) = e^{2t+8t^2}$, then what is the mean and variance of X?

(a) Mean = 2, Variance = 4

(b) Mean = 2, Variance = 8

(c) Mean = 4, Variance = 2

(d) Mean = 8, Variance = 2

Problem 2. Which of the following statements is true about moment generating functions?

(a) They always exist for all probability distributions.

(b) They can be used to find the probability density function.

(c) They are unique for each probability distribution.

(d) They are only defined for continuous random variables.

Problem 3. Let X be a random variable. Then $E(e^{tX})$ denoted by $M(t)$ is called moment generating function of X, where t is a real constant.

Problem 4. The cumulant generating function of the sum of two independent random variables is the sum of their cumulant generating functions.

Problem 5. Logarithm of the m.g.f. of a distribution is called cumulant generating function

8.9 *SUMMARY*

(i) Properties of m.g.f.

Property 1. Let $X = cu$. Then

$$M_X(t) = E(e^{tX}) = E(e^{tcU}) = M_U(ct).$$

Property 2. Let $X = u + c$. Then

$$M_X(t) = E(e^{tX}) = E\left(e^{r(U+c)}\right) = e^{tc}.E(e^{tU}) = e^{tc}M_U(t).$$

Property 3. The moment generating function of the sum of two independent random variables in the product of their moment generating functions.

(ii) The cumulant generating function of the sum of two independent random variables is the sum of their cumulant generating functions.

8.10 *GLOSSARY*

(i) Moment

(ii) Exponential function

8.11 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

8.12 *TEWRMINAL QUESTIONS*

1. Define moment generating function. Why is it called moment generating function?
2. What do you understand by moment generating function? How moments are deduced from it?

3. Prove that moment generating function of the sum of two independent random variables is the product of their moment generating functions.

4. If $P(X = x) = \frac{1}{2^x}, x = 1, 2, 3, \dots$ find the m.g.f. of x. Hence obtain the variance.

5. Define cumulant generating function and cumulant. Obtain also the first five cumulants in term of central moments.

6. Find the m.g.f. for $f(x) = ce^{-cx}, c > 0, 0 \leq x \leq \infty$.

8.13 *ANSWERS*

CYQ 1. b

CYQ 2. c

CYQ 3. True

CYQ 4. True

CYQ 5. True

TQ 4. $M_n(t) = \frac{e^t}{2 - e^t}$

TQ 6. $\left(1 - \frac{t}{c}\right)^{-1}$

UNIT 9: *CHARACTERISTIC FUNCTION*

CONTENTS:

9.1 *INTRODUCTION*

In probability theory and statistics, the characteristic function of any real-valued random variable completely defines its probability distribution. If a random variable admits a probability density function, then the characteristic function is the Fourier transform (with sign reversal) of the probability density function. Thus it provides an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. There are particularly simple results for the characteristic functions of distributions defined by the weighted sums of random variables. In addition to univariate distributions, characteristic functions can be defined for vector- or matrix-valued random variables, and can also be extended to more generic cases.

The characteristic function always exists when treated as a function of a real-valued argument, unlike the moment-generating function. There are relations between the behavior of the characteristic function of a distribution and properties of the distribution, such as the existence of moments and the existence of a density function.

9.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Characteristic Function

(ii) Fourier's Inversion Theorem

9.3 *CHARACTERISTIC FUNCTION*

The function $\varphi(t) = E\left(e^{itX}\right)$ is called the Characteristic function of the random variable X. here t is real number and i is the imaginary quantity.

If X has the discrete values $x_k(k = 1, 2, 3, \ldots.), P(x = x_k) = p_k$, then the characteristic function of the variable X is given by

$$\varphi(t) = E\left(e^{itX}\right) = p_1 e^{itx_1} + p_2 e^{itx_2} + \cdots + p_k e^{itx_k}$$

$$= \sum_k p_k e^{itx_k}$$

If X is a continuous random variate with p.d.f. f(x), then the characteristic function is given by

$$\varphi(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} f(x) e^{itX} dx$$

- **Properties of Characteristic Function.**

i. $\varphi(t)$ is a complex valued function of real t.

ii. $\varphi(t)$ always exists.

iii. $\varphi(t)$ is continuous in t.

iv. $\varphi(t)$ is defined in every finite 't' interval.

v. $\varphi(t)$ is uniquely determined by the distribution.

vi. $\varphi(t)$ completely determines the distribution of X.

vii. $\varphi(0) = 1$.

viii. $|\varphi(t)| \leq 1$.

ix. $\varphi(t)$ and $\varphi(-t)$ are conjugate functions.

x. $\varphi(t)$ generates moments.

xi. $\varphi_{cx}(t) = \varphi_x(ct)$ where c is a constant.

xii. $\varphi_{x_1+x_2+\dots\dots+x_n}(t) = \varphi_{x_1}(t)\varphi_{x_2}(t) \dots \dots .. \varphi_{x_n}(t)$

- **Necessary Conditions for a Function to be a Characteristic Function.**

The characteristic function $\varphi(t)$ has the following necessary but not sufficient conditions for any function to be a characteristic function. That is a function satisfying these conditions may or may not be a characteristic function of a well-defined probability distribution but a characteristic function will necessary have these properties:

i. $\varphi(t)$ is continuous in t.

ii. $\varphi(t)$ is defined in every finite 't' interval.

iii. $\varphi(0) = 1$.

iv. $|\varphi(t)| \leq 1$.

v. $\varphi(t)$ and $\varphi(-t)$ are conjugate functions.

- **Necessary and Sufficient Condition for a Function to be a Characteristic Function.**

The necessary and sufficient conditions for a function to be a characteristic function is that

i. $\varphi(0) = 1$.

ii. $\varphi(x,c) = \int_0^c \int_{-\infty}^{\infty} \varphi(t-z) e^{ix(t-z)} dt\, dz$ in real and non-negative for all real x and all $c > 0$.

Proof: (i) $\varphi(t) = E\left(e^{itX}\right)$

Therefore, $\varphi(0) = E\left(e^{i\times 0\times X}\right) = E\left(e^{i\times 0\times X}\right) = E(e^0) = E(1) = 1$

(ii) $\varphi(x,c) = 2\int_{-\infty}^{\infty} \frac{1-\cos c(x+y)}{(x+y)^2} dF(y)$

Where $F(x) =$ distribution function of x, which is evidently real and non-negative. Hence the condition.

Example 1. Find the characteristic function of binomial distribution $B(n,p)$.

Solution. $X \sim B(n,p)$

$$f(x) = n_{C_x} p^x q^{n-x}, x = 0, 1, 2, \dots., n$$

$$\varphi(t) = E\left(e^{itX}\right)$$

$$= \sum_{x=0}^{n} e^{itx} f(x)$$

$$= \sum_{x=0}^{n} e^{itx} n_{C_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} n_{C_x} (pe^{it})^x q^{n-x}$$

$$= (q + pe^{it})^n$$

Example 2. Find the characteristic function of the random variable x where x takes values $x_1 = -1$ and $x_2 = +1$ with probability

$P(x = -1) = P(x = +1) = \frac{1}{2}$.

Solution. $\varphi(t) = E\left(e^{itX}\right) = p_1 e^{ix_1} + p_2 e^{ix_2}$

$$= \frac{1}{2} e^{-it} + \frac{1}{2} e^{it} = \frac{e^{-it} + e^{it}}{2} = \cos t.$$

Example 3. The random variable x has a rectangular distribution

$f(x) = \frac{1}{2h}$, for $a - h \le x \le a + h$

= otherwise

Find the characteristic function of the random.

Solution. characteristic function

$$\varphi(t) = E\left(e^{itX}\right) = \frac{1}{2h}\int_{a-h}^{a+h} e^{itx}dx = \frac{e^{it(a+h)} - e^{it(a+h)}}{2ith}$$

$$= \frac{1}{ht}e^{ita}\sin th.$$

Example 4. Show that $\varphi(-t) =$ the complex number conjugate to $\varphi(t)$.

Solution. $\varphi(-t) = E\left(e^{-itX}\right) = E(\cos tx - i\sin tx)$

$$= E(\cos tx) - iE(\sin tx)$$

And $\varphi(t) = E\left(e^{itX}\right) = E(\cos tx + i\sin tx)$

$$= E(\cos tx) + iE(\sin tx).$$

Example 5. Show that $\varphi(t)$ is complex value function of real t.

Solution. $|\varphi(t)| = \left|E\left(e^{itX}\right)\right|$

$$= |\cos tX + i\sin tX| = Cos^2tX + Sin^2tX = 1$$

$\Longrightarrow \varphi(t) = E\left(e^{itX}\right)$ always exists.

9.4 *GENERATION OF MOMENTS AND CUMULANTS BY CHARACTERISTIC FUNCTION*

Let $\varphi(t) = E\left(e^{itX}\right)$

$$= E\left[1 + itX + \frac{(itX)^2}{2!} + \cdots + \frac{(itX)^r}{r!} + \cdots\right]$$

$$= 1 + itE(X) + \frac{(it)^2}{2!}EX^2 + \cdots + \frac{(it)^r}{r!}EX^r + \cdots$$

$$= 1 + it\mu_{1'} + \frac{(it)^2}{2!}\mu_{2'} + \cdots + \frac{(it)^r}{r!}\mu_{r'} + \cdots$$

where, $\mu_{1'}$ = coefficient of $\frac{(it)^l}{l!}$ in $\varphi(t)$.

Example 1. For a distribution the cumulants are given by $k_r = n(r-1)!, n > 0.$ Using cumulant generating function as the logarithm of characteristic function, find the characteristic function.

Solution. Let $k_r(t) = \log \varphi_x(t)$ ……. (1)

$$= \sum_{r=1}^{\infty} \frac{(it)^r}{r!} \mathrm{k_r}$$

Where k_r = coefficient of $\frac{(it)^r}{r!}$ Is the rth cumulant.

We are given $k_r = n(r-1)!$

$k_r(t) = \sum_{r=1}^{\infty} \frac{(it)^r}{r!} n(r-1)!$

$= n \sum_{r=1}^{\infty} \frac{(it)^r}{r!} = \text{n}[\text{it} + \frac{(it)^2}{2!} + \frac{(it)^3}{3!} + \cdots]$

$= n[-log_e(1-it)]$

$= (1-it)^{-n}$ ……….. (2)

$\log \varphi_x(t) = log_e(1-it)^{-n}$ {from (1) and (2)}

$\Rightarrow \varphi_x(t) = (1-it)^{-n}.$

9.5 *FOURIER'S INVERSION THEOREM*

Statement: if $F(x)$ and $\varphi(t)$ be the distribution function and characteristic function of a random variable X then $f(x) = F'(x)$ the p.d.f. is given by

$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itX} \varphi(t)\ dt$

Provided $\varphi(t)$ is integrable and $\int_{-\infty}^{\infty} |\varphi(t)| dt < \infty$.

Moreover, the density function $f(x)$ is everywhere continuous, but if tends to be a positive number f_i, the distribution is discontinuous at that point.

Proof: Suppose $(a-h, a+h)$ be a continuity interval of the distribution function $F(x)$, then we have

$F(a+h) - F(a-h) = \log_{T\to\infty} \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-ita} \varphi_x(t)\ dt$

In order to prove the inversion theorem, first we shall prove the expression as follows. Let us write

$$J = \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-ita} \varphi_x(t)\ dt$$

$$= \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-ita} \left[\int_{-\infty}^{\infty} e^{-itx} dF(x)\right]\ dt$$

Since $\left|\frac{\sin ht}{t} e^{-it(x-a)}\right| \le h$, the conditions for the reversion of the order of integration are satisfied. Hence

$$J = \frac{1}{\pi} \int_{-\infty}^{\infty} dF(x) \int_{-T}^{T} \frac{\sin ht}{t} e^{it(x-a)}\ dt$$

$$= \frac{2}{\pi} \int_{-\infty}^{\infty} \text{dF(x)} \int_{-\text{T}}^{\text{T}} \frac{\sin ht}{t} \cos(x-a)\ \text{t dt}$$

$$= \int_{-\infty}^{\infty} g(x,T)dF(x)$$

where, $g(x,T) = \frac{2}{\pi}\int_0^T \frac{\sin ht}{t}\cos(x-a)tdt$

$$= \frac{1}{\pi}\int_0^T \frac{\sin(x-a+h)t}{t}dt - = \frac{1}{\pi}\int_0^T \frac{\sin(x-a-h)t}{t}dt$$

$$= \frac{1}{2}s(x - a + h, T) - \frac{1}{2}s(x - a - h, T)$$

Thus, $|g(x,T)|$ is less than an absolute constant, and we have

$$\log_{T\to\infty} g(x,T) = \begin{cases} 0, & for\ x < a-h \\ \frac{1}{2}, & for\ x = a-h \\ 1, & for\ a-h < x < a+h \\ \frac{1}{2}, & for\ x = a+h \\ 0, & for\ x > a+h \end{cases}$$

We may thus apply theorem 'if $g_n(x) = g(x)$ exists for almost everywhere, then $\log_{n\to\infty}\int g_n(x)dF = \int g(x)dF$ we obtain

$\log_{n\to\infty} J = \int_{a-b}^{a+b} dF(x) = F(a+h) - F(a-h)$

Since $F(x)$ is continuous for $x = a \pm h$.

Now, in particular, when $\varphi_n(t)$ is integrable over $(-\infty, \infty)$, it follows that

$$\frac{F(x+h)-F(x-h)}{2h} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\sin ht}{ht} e^{itx}\varphi_n(t)dt$$

As soon as $F(x)$ is continuous for $x = x \pm h$.

When h tends to zero, the function under the 'integral sign tends to $e^{-itx}\varphi_x(t)$, $\left[\sin \log_{h\to\infty} \frac{\sin ht}{t} = 1\right]$ while its modulus is dominated by the integrable function $|\varphi_n(t)|$.

Thus, $\log_{h\to\infty} \frac{F(x+h)-F(x-h)}{2h} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_x(t)dt$ which gives

$$F'(x) = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_x(t)dt$$

This important theorem shows that a distribution is uniquely determined by its characteristics function.

Example 1. Find the density function of the variate for which characteristic function is

$$e^{-t^2}/_2.$$

Solution. $\varphi_x(t) = e^{-t^2/2}$

$$\therefore \quad f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx} e^{-t^2/2}\, dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(t+ix)^2/2}\, e^{(ix)^2/2} dt$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} . \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(t+ix)^2/2}\, dt$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Example 2. Show that the distribution for which characteristic function is $e^{-|t|}$ has the density function $f(x) = \frac{1}{\pi}.\frac{1}{(1+x^2)}, -\infty < x < \infty$.

Solution. $\varphi(t) = e^{-|t|}$

$$\therefore \quad f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-|t|}\, e^{-itx}\, dt$$

$$= \frac{1}{2\pi}\left[\int_{-\infty}^{0} e^{t} e^{-itx}\, dt + \int_{0}^{\infty} e^{-t} e^{-itx}\, dt\right]$$

$$= \frac{1}{2\pi}\int_{0}^{\infty} e^{-t}(\, e^{-itx} + e^{-itx})\, dx = \int_{0}^{\infty} e^{-t} costx\, dt \quad \ldots\ldots (1)$$

$$= \frac{1}{\pi}[-e^{-t} costx]_0^{\infty} - \frac{x}{\pi}\int_{0}^{\infty} e^{-t} sintx\, dt$$

$$= \frac{1}{\pi} + \frac{x}{\pi}[-e^{-t} sintx]_0^{\infty} - \frac{x^2}{\pi}\int_{0}^{\infty} e^{-t} costx\, dt$$

$$= \frac{1}{\pi} - x^2 f(x) \qquad \text{[from (1)]}$$

$$\therefore \quad f(x) = \frac{1}{\pi}.\frac{1}{(1+x^2)}, -\infty < x < \infty.$$

$$\therefore \quad f_w(w) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itw}(1-it)^{-n}\, dt \qquad \text{[inversion theorem]}$$

Solving the integral by contour integration in the complex plane by making transformation

$z = -w(1-it)$ or $z = -w + wit$ or $z + w = wit$

Which gives $f_w(w) = w^{n-1} e^{-w}.H$

Where H is the Hankel integration, given by

$$H = \frac{1}{2\pi}\int_{-w-i\infty}^{-w+i\infty} e^{-z}(-i)^{-n}\, dz = \frac{1}{\Gamma(n)}$$

$$\therefore \quad f_w(w) = \begin{cases} \frac{1}{\Gamma(n)} w^{n-1} e^{-w}, & w \geq 0 \\ 0, & w < 0 \end{cases}.$$

CHECK YOUR PROGRESS

True or false or MCQ Questions

Problem 1. What does the characteristic function uniquely determine?

a) The mean of a distribution
b) The variance of a distribution
c) The probability distribution itself
d) The skewness of a distribution

Problem 2. What is the value of the characteristic function at t = 0?

a) 0 b) 1 c) -1 d) Varies depending on the distribution

Problem 3. If the characteristic function of a random variable is real-valued and an even function of t, what can be inferred about the distribution?

a) It is skewed to the right
b) It is skewed to the left
c) It is symmetric about zero
d) It is not a valid distribution

Problem 4. What is the relationship between the characteristic function and the probability distribution?

a) The characteristic function is the Fourier transform of the probability density function
b) The probability density function is the Fourier transform of the characteristic function
c) They are unrelated
d) The characteristic function is a moment generating function

Problem 5. What is the significance of the "uniqueness theorem" related to characteristic functions?

a) It states that a characteristic function is always unique for a given distribution.
b) It means that if two distributions have the same characteristic function, they must be identical.
c) It implies that the characteristic function can be used to find the moments of a distribution.
d) It suggests that the characteristic function is a complex-valued function.

9.6 *SUMMARY*

(i) Characteristic function:

Characteristic function of the variable X is given by

$$\varphi(t) = E(e^{itX}) = p_1 e^{itx_1} + p_2 e^{itx_2} + \cdots + p_k e^{itx_k}$$

$$= \sum_k p_k e^{itx_k}$$

(ii) Fourier's Inversion Theorem: if $F(x)$ and $\varphi(t)$ be the distribution function and characteristic function of a random variable X then $f(x) = F'(x)$ the p.d.f. is given by $f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itX}\varphi(t)\,dt$ Provided $\varphi(t)$ is integrable and $\int_{-\infty}^{\infty}|\varphi(t)|dt < \infty$.

9.7 *GLOSSARY*

(i) Moment

(ii) Exponential function

9.8 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

9.9 *TEWRMINAL QUESTIONS*

1. Define characteristic function. States a set of necessary condition for a function to be a characteristic function.
2. Show that $\varphi_{cx}(t) = \varphi_x(ct)$, c being a constant.
3. Find the characteristic function of $f(x) = \frac{1}{b-a}, a < x < b$.
4. Find the distribution if $\varphi_X(t) = \frac{1}{1+t^2}$.
5. Prove that $\varphi_u(t) = e^{-iat/h}\varphi_x(t/h)$, where, $u = \frac{x-a}{h}$.

9.10 *ANSWERS*

CYQ 1. c

CYQ 2. b

CYQ 3. c

CYQ 4. a

CYQ 5. b

TQ 3. $\frac{e^{ibt} - e^{iat}}{it(b-a)}$

TQ 4. $f(x) = \frac{1}{2}e^{|x|}, -\infty < x < \infty$

COURSE NAME: BASIC STATISTICS

COURSE CODE: MT(N)-222

BLOCK - IV:

DISCRETE DISTRIBUTION

UNIT 10: DISCRETE DISTRIBUTIONS

CONTENTS:

10.1 *INTRODUCTION*

A discrete distribution is a probability distribution that depicts the occurrence of discrete (individually countable) outcomes, such as 1, 2, 3, yes, no, true, or false. The binomial distribution, for example, is a discrete distribution that evaluates the probability of a "yes" or "no" outcome occurring over a given number of trials, given the event's probability in each trial—such as flipping a coin one hundred times and having the outcome be "heads."

Statistical distributions can be either discrete or continuous. A continuous distribution is built from outcomes that fall on a continuum, such as all numbers greater than 0 (including numbers whose decimals continue indefinitely, such as pi = 3.14159265...). Overall, the concepts of discrete and continuous probability distributions and the random variables they describe are the underpinnings of probability theory and statistical analysis.

10.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) discrete distribution

(ii) probability distributions

10.3 *DISCRETE PROBABILITY DISTRIBUTION*

A discrete probability distribution assigns probabilities to each possible outcome of a discrete random variable, meaning a variable that can only take on specific, separate values (like integers). These distributions are characterized by the fact that the sum of all probabilities must equal one.

A discrete probability distribution describes the likelihood of each possible outcome for a discrete random variable. A discrete random variable is a variable that can take on a countable number of distinct values, typically whole numbers. In simpler terms, it's a way to assign a probability to each value that the variable can have, showing how likely each outcome is.

The common examples of discrete probability distributions include Bernoulli, Binomial, Poisson, and Geometric distributions.

Conditions for the discrete probability distribution are:

1. The probability of a discrete random variable lies between 0 and 1:
   $0 \leq P\ (X = x) \leq 1$
2. Sum of Probabilities is always equal to 1:
   $\sum P\ (X = x) = 1$

Example 1. Let two coins be tossed; then the probability of getting a tail is an example of a discrete probability distribution. The sample space for the given event is {HH, HT, TH, TT}, and let X be the number of tails. Then, the discrete probability distribution table is given by:

X	0{HH}	1{HT, TH}	2{TT}
P(X = x)	1/4	1/2	1/4

Example 2. A pair of fair dice is rolled. Let XX denote the sum of the number of dots on the top faces.

a. Construct the probability distribution of XX for a paid of fair dice.
b. Find P $(X \geq 9)$.

Solution.

The sample space of equally likely outcomes is

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

where the first digit is die 1 and the second number is die 2.

(a) The possible values for X are the numbers 2 through 12. X = 2 is the event {11}, so P (2) = 1/36. X=3 is the event {12, 21}, so P(3) =2/36. Continuing this way we obtain the following table

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This table is the probability distribution of X.

(b) The event X ≥ 9 is the union of the mutually exclusive events X = 9, X = 10, X = 11, and X = 12. Thus

$$P(X \geq 9) = \mathrm{P}(9) + \mathrm{P}(10) + \mathrm{P}(11) + \mathrm{P}(12)$$
$$= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$$
$$= \frac{10}{36} = 0.2\overline{7}$$

Remark 1. The different formulas for the discrete probability distribution, like the probability mass function, the cumulative distribution function, and the mean and variance, are given below.

Probability Mass Function (PMF) of Discrete Probability Distribution:

PMF of a discrete random variable X is the value completely equal to x. The PMF i.e., probability mass function of a discrete probability distribution is given by:

f(x) = P (X = x)

Example:

A discrete random variable X as the number of heads obtained when tossing two fair coins.

Possible Outcomes:

HH → 2 heads and HT, TH → 1 head and TT → 0 heads

Then PMF are given as:

- P (X = 0) = ¼, → Only **TT**
- P (X = 1) = 2/4 → **HT** or **TH**
- P (X = 2) = ¼ → Only **HH**

Total Probability: ¼ + ½ + ¼ =1

Properties of PMF:

- The sum of all probabilities must always equal 1. i.e. ∑P (X = x) = 1
- P (X = x) ≥ 0

Remark 2. CDF of Discrete Probability Distribution

CDF of a discrete random variable X is less than or equal to value x. The CDF i.e., cumulative distribution function of discrete probability distribution is given by: **f(x) = P (X ≤ x)**

Example. A random variable X as the outcome of rolling a fair 4-sided die. To find P (1 < X ≤ 3).

X	PMF	CDF
1	0.2	0.2
2	0.3	0.2+0.3 = 0.5
3	0.2	0.5+0.2 = 0.7
4	0.3	0.7+0.3 = 1

$P(1 < X \le 3) = F(3) - F(1) = 0.7 - 0.2 = 0.5$

Remark 3. Discrete Probability Distribution Mean

Mean of discrete probability distribution is the average of all the values that a discrete variable can obtain. It is also called as the expected value of the discrete probability distribution. The mean of discrete probability distribution is given by:

$$E[X] = \sum x\, P(X = x)$$

Remark 4. Discrete Probability Distribution Variance

Variance of discrete probability distribution is defined as the product of squared difference of distribution and mean with PMF. The variance of the discrete probability distribution is given by:

$$Var[X] = \sum (x - \mu)^2\, P(X = x)$$

Example:

A random variable X as the outcome of rolling a fair 4-sided die.

To find $P(1 < X \le 3)$

X	PMF	CDF
1	0.2	0.2
2	0.3	0.2 + 0.3 = 0.5
3	0.2	0.5 + 0.2 = 0.7
4	0.3	0.7 + 0.3 = 1

$P(1 < X \leq 3) = F(3) - F(1) = 0.7 - 0.2 = 0.5$

- **How to Find Discrete Probability Function**

Steps to find the discrete probability function are given below:

Step 1: First determine the sample space of the given event.

Step 2: Define random variable X as the event for which the probability has to be found.

Step 3: Consider the possible values of x and find the probabilities for each value.

Step 4: Write all the values of x and their respective probabilities in tabular form to get the discrete probability distribution.

- **Calculation of Discrete Probability Distribution**

How you calculate a discrete probability distribution depends on your test, what you're trying to measure, and how you measure it. For instance, if you're flipping a coin twice, the possible combinations are:

Tails/tails (TT), Heads/tails (HT), Tails/heads (TH), Heads/heads (HH)

Because you're flipping the coin twice and there are two possible outcomes, there are four possibilities. Each of the results represents one-quarter of the possibilities. The HT and TH combinations are each one-quarter (and essentially the same thing), representing one-half of the results. Therefore, the probability is that one-quarter of the time, you'll get a TT or HH, and one-half of the time, you'll get HT or TH.

This works similarly for rolling two dice because the results of a dice roll are discrete. There are 36 possibilities because each die has six faces, but there cannot be a result of one since the lowest number on each die is one. So the lowest result you can get is two, and the highest is 12. Many of the combinations will repeat, just as in the coin example—so the more possibilities that repeat, the more instances will be graphed.

As seen in the table below, if you add the figures for dice roll results together, you have one instance where the result is two and one where it is 12—creating odds of one in 36 for the numbers two and 12.

$$\begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

The probability (P) that X (the outcome) will equal x (the chosen number) would be:

- P (X = 2) = 1 / 36
- P (X = 3) = 2 / 36
- P(X = 4) = 3 /36
- P(X = 5) = 4 / 36
- P(X = 6) = 5 /36
- P(X = 7) = 6 / 36
- P(X = 8) = 5 / 36
- P(X = 9) = 4 / 36
- P(X = 10) = 3 / 36
- P(X = 11) = 2 / 36
- P(X = 12) = 1 / 36

The probability that the roll equals two is one in 36; the probability of it equalling three is two in 36, and so on.

10.4 *TYPE OF DISCRETE PROBABILITY DISTRIBUTIONS*

The most common discrete probability distributions include binomial, Bernoulli, multinomial, and Poisson.

i. Binomial

A binomial probability distribution is one in which there are only two possible outcomes. In this distribution, data are collected in one of two forms after repetitive trials and classified into either success or failure. It generally has a finite set of just two possible outcomes, such as zero or one. For instance, flipping a coin gives you the list (Heads, Tails).

Binomial Distribution Formula

The Binomial Distribution Formula, which is used to calculate the probability, for a random variable X = 0, 1, 2, 3,....,n is given as

$$P(X = r) = {}^{n}C_{r}\, p^{r}\, (1-p)^{n-r},\ r = 0, 1, 2, 3....$$

Where, **n** = Total number of trials, **r** = Number of successes, **p** = Probability of success

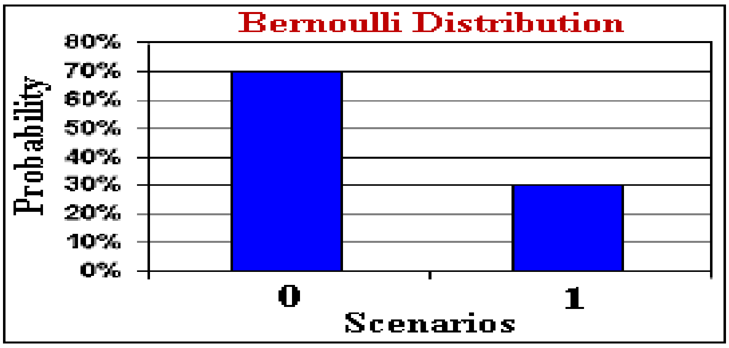
The binomial distribution is used in options pricing models that rely on binomial trees. In a binomial tree model, the underlying asset can only be worth exactly one of two possible values—with the model, there are just two probable outcomes with each iteration—a move up or a move down with defined values.

ii. Bernoulli

Bernoulli distributions are similar to binomial distributions because there are two possible outcomes. One trial is conducted, so the outcomes in a Bernoulli distribution are labeled as either a zero or a one. A one indicates success, and a zero means failure—one trial is called a Bernoulli trial.

The Bernoulli distribution formula, which describes the probability of success or failure in a single trial, is: $\mathbf{P\ (X = x) = p^{x}(1-p)^{1-x}}$ **for x = 0 or 1**. Here, 'p' represents the probability of success, and '1-p' is the probability of failure.

So, if you used one green marble (for success) and one red marble (for failure) in a covered bowl and chose without looking, you would record each result as a zero or a one rather than success or failure for your sample. Bernoulli distributions are used to view the probability that an investment will succeed or fail.



https://d18l82el6cdm1i.cloudfront.net/uploads/IFPlPEydqU-bernoulli.gif

iii. Multinomial

Multinomial distributions occur when there is a probability of more than two outcomes with multiple counts. For instance, say you have a covered bowl with one green, one red, and one yellow marble. For your test, you record the number of times you randomly choose each of the marbles for your sample.

iv. Poisson Distribution

The Poisson distribution expresses the probability that a given number of events will occur over a fixed period.

The Poisson distribution is a discrete distribution that counts the frequency of occurrences as integers, whose list (0, 1, 2, ...) can be infinite. For instance, say you have a covered bowl with one red and one green marble, and your chosen period is two minutes. Your test is to record whether you pick the green or red marble, with the green indicating success. After each test, you place the marble back in the bowl and record the results.

In this model, the distribution would plot the results over a period of time, indicating how often green is chosen.

Poisson distribution is commonly used to model financial data where the tally is small and often zero. For example, it can be used to model the number of trades a typical investor will make in a given day, which can be 0 (often), 1, 2, and so on.

CHECK YOUR PROGRESS

True or false Questions

Problem 1. P (X = x) = $p^x(1-p)^{1-x}$ for x = 0 or 1 is the Bernoulli distribution formula.

Problem 2. Var[X] = $\sum (x - \mu)^2$ P (X = x).

Problem 3. The probability of a discrete random variable lies between 0 and 1.

10.5 *SUMMARY*

(i) A discrete probability distribution describes the likelihood of each possible outcome for a discrete random variable. A discrete random variable is a variable that can take on a countable number of distinct values, typically whole numbers.

(ii) The binomial distribution is used in options pricing models that rely on binomial trees In a binomial tree model, the underlying asset can only be worth exactly one of two possible values with the model, there are just two probable outcomes with each iteration a move up or a move down with defined values.

10.6 *GLOSSARY*

(i) Permutation

(ii) Combinations

(iii) Probability

10.7 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.

2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.

3. J. S. Milton and J. C. Arnold, (2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.

4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

10.8 *TEWRMINAL QUESTIONS*

1. What do you mean by Discrete Probability Distributions.
2. What do you mean by Types of Discrete Probability Distributions Define Pascal's triangle.
3. Define probability mass function of a discrete probability distribution.

10.9 ANSWERS

CYQ 1. True

CYQ 2. True

CYQ 3. True

UNIT 11: *BINOMIAL DISTRIBUTION*

CONTENTS:

11.1 *INTRODUCTION*

The binomial distribution has a rich history, with key developments occurring over time. While the binomial theorem itself dates back to ancient times, the formalization of the binomial distribution, as we know it today, is largely attributed to Jacob Bernoulli in the late 17th and early 18th centuries.

The binomial theorem, which forms the foundation for the binomial distribution, has roots in the work of Greek mathematicians like Diophantus.

Isaac Newton, around 1665, discovered the general form of the binomial theorem, which was later stated (without proof) in 1676.

11.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Probability Function of Binomial Distribution

(ii) Pascal's Triangle

(iii) Constants of Binomial Distribution

11.3 *BINOMIAL DISTRIBUTION*

Binomial distribution is also known as the "Bernoulli distribution" after the Swiss mathematician James Bernoulli (1654-1705) who discovered it in 1700 and was first published in 1713, eight years after his death. This distribution can be used under the following conditions:

1. The random experiment is performed repeatedly a finite and fixed number of times. In other words, the n, referring to the number of trials, is finite and fixed.

2. The result of any trial can be classified only under two mutually exclusive categories called success (the occurrence of the event) and failure (the non-occurrence of the event). For example, if a coin is tossed, the result of each throw can be either a head or a tail; when one card is drawn, it may be a black or a red card; an item when examined from a lot can be either defective or non-defective. Similarly, a firm may be classified as large or not large depending on its total output or the level of employment. The proportion of outcomes falling in the "success" category is generally denoted by p and the proportion of items falling in the category of "failure" by $q = 1 - p$.

3. The probability of success in each trial remains constant and does not change from one trial to another. For example, the probability of obtaining a head in successive throws

of a coin is always $\frac{1}{2}$; The probability of obtaining a defective article from a batch does not change in successive drawings with replacement, and practically remains constant even in drawings without replacement, when the batch is large.

4. The trials are independent, so that the result of any trial is unaffected by the results of previous trials. For example, in successive throws of a coin the occurrence of a head at any trial will in no way affect the probability of a head or a tail in any subsequent trial: or if several coins are thrown together, the occurrence of a head or a tail in any particular coin does not alter the probability of occurrence of a head in any other coin.

11.4 *PROBABILITY FUNCTION OF BINOMIAL DISTRIBUTION*

If a trial of an experiment can results in success with probability p and failure with probability $q = 1 - p$, the probability of exactly r successes. in n trials is given by

$$P(X = r) = p(r) = {}^nCrp^r q^{n-r}; r = 0,1,2, \dots, n$$

The quantities n and p are called parameters of the binomial distribution and the notation ($r: n, p$) reads "the binomial probability of given n and p.

The entire probability distribution of $r = 0,1,2, \dots, n$

successes with two types expressions, where p stands for success and q for non-success can be written as follows:

BINOMIAL PROBABILITY DISTRIBUTION

Number of Successes r	$(p+q)^n$ Probability p(r)	Number of Successes r	$(g+p)^n$ Probability p(r)
n	${}^nCnp^nq^0$	0	${}^nC_0q^np_0$
$n-1$	${}^nC_{n-1}p^{n-1}q^0$	1	${}^nC_1q^{n-1}p$
$n-2$	${}^nC_{n-2}p^{n-2}q^2$	2	${}^nC_2q^{n-2}p^2$
⋮	${}^nC_{n-3}p^{n-3}q^3$	3	${}^nC_3q^{n-3}p^3$
⋮		⋮	⋮
0	${}^nCop0q^n$	n	${}^nC_nq^0p^n$
Total	1		1

The probability distribution is called binomial since the probabilities specified above, viz.,

$$q^n,\ {}^nC_1 q^{n-1}p,\ {}^nC_2 q^{n-2}p^2, \dots, p^n$$

are successive terms in the binomial expansion of $(q+p)^{\mathrm{n}}$.

Remarks:

1. The Binomial distribution is completely determined, i.e., all the probabilities can be obtained, if n and p are known. Obviously q is known when p is given ($\because q = 1 - p$).
2. Since the random variable X takes only integral values, the binomial distribution is a *discrete* probability distribution.
3. The values of the Binomial coefficient for different values of n can be obtained conveniently from the Pascal's triangle given below:

11.5 *PASCAL'S TRIANGLE*

Pascal's Triangle is a triangular array of numbers where each number is the sum of the two numbers directly above it. It's a fundamental concept in mathematics, particularly in algebra, probability, and combinatorics. The triangle starts with a "1" at the top, and each subsequent row begins and ends with 1, with the numbers in between calculated by adding the two numbers above.

[Showing coefficients of terms $(a+b)^n$

Value of n	*Binomial coefficients*	*Sum* (2^n)
1	1 1	2
2	1 2 1	4
3	1 3 3 1	8
4	1 4 6 4 1	16
5	1 5 10 10 5 1	32
6	1 6 15 20 15 6 1	64
7	1 7 21 35 35 21 7 1	128
8	1 8 28 56 70 56 28 8 1	256
9	1 9 36 84 126 126 84 36 9 1	512
10	1 10 45 120 210 252 210 120 45 10 1	1024

It can be easily seen that, taking the first and last terms as 1, each term in the above table can be obtained by adding the two terms on either side of it in the preceding line (i.e., the line above it). It can be seen that the binomial coefficients are symmetric.

(i). To investigate how this distribution can arise, consider a particular sequence of outcomes of n In independent trials of the experiment, each outcome is either a success (S) or a failure (F) and we suppose there exactly r successes and (n − r) failures.

Let ' S ' denote a success and ' F 'denotes a failure. Let us first find the probability that the first r trials are successes, and the remaining ($n - r$) trials are failures, ie., the probability of

$$\underbrace{S\,S\ldots\ldots S}_{\text{1st } r \text{ trials}} \quad \underbrace{F\,F\ldots\ldots F}_{(n-r)\text{ trials}}.$$

Since n trials are independent, by the multiplication rule of probability, the probability of r successes is $\underbrace{p \times p \times \ldots \times p}_{r\text{ times}} = p^r$ and the probability of $(n-r)$ failures is $\underbrace{q \times q \times \ldots \times q}_{(n-r)\text{ times}} = q^{n-r}$

∴ the probability that the first r trials are successes and the remaining $(n-r)$ trials are failures is $p^r . q^{n-r}$ where $q = 1 - p$. Similarly, the probability of obtaining r successes and $(n-r)$ failures in any other definite specified order is also $p^r . q^{n-r}$.

But we are interested in finding the probability that in n independent trials any r trials are successes [and, therefore, remaining $(n-r)$ trials are failures]. r trials can be selected out of n in ${}^nC_r$ mutually exclusive ways in each of which the probability of r successes is $p^r q^{n-r}$.

Hence the probability of getting r successes and consequently $(n-r)$ failures in n trials in any order (whatsoever) is given by

$$\begin{aligned} P(X = r) = p(r) &= p^r q^{n-r} + p^r q^{n-r} + \cdots \; {}^nC_r \text{ times} \\ &= {}^nC_r p^r q^{n-r}; r = 0,1,2,\ldots,n \end{aligned}$$

(ii). Binomial distribution is symmetrical if $q = p = \frac{1}{2}$. Even if $p \neq q$ the distribution tends to be symmetrical if n is increased sufficiently.

(iii). Graphical Representation of the Binomial Distribution. The binomial distribution is commonly represented graphically by constructing a variable scale along the horizontal axis to represent the possible number of successes and another scale along the vertical axis to represent the probability of occurrence. The probability for each possible value is represented by the height of a vertical line erected above the value.

11.6 *CONSTANTS OF BINOMIAL DISTRIBUTION*

- **Constants of Binomial Distribution**

r	$P(X=r)=p(r)$	$r\cdot p(r)$	$r^2p(r)$
0	q^n	0	0
1	${}^nC_1q^{n-1}p$	${}^nC_1q^{n-1}p$	$1^2\ {}^nC_1q^{n-1}p$
2	${}^nC_2q^{n-2}p^2$	$3^nC_3q^{n-2}q^n$	$2^2\ {}^nC_2q^{n-2}p^2$
3	${}^nC_3q^{n-3}p^3$	$3^nC_3q^{n-3}p^3$	$3^2\ {}^nC_3q^{n-3}p^3$
$\vdots$	$\vdots$		$\vdots$
n	p^n	np^n	n^2p^n

By definition:

$$(\boldsymbol{i})\ \textbf{Mean}\ =\frac{\Sigma rp(r)}{\Sigma p(r)}=\sum_{r=0}^{n} r\cdot\ {}^nC_rp^rq^{n-\mathrm{r}}$$

$$=0\times({}^nC_0p^0q^n)+1\times({}^nC_1pq^{n-1})+2\times({}^nC_2p^2q^{n-2})$$
$$+3\times({}^nc_3p^3q^{n-3})+\cdots+n\times({}^nC_np^nq^{n-\mathrm{n}})$$
$$=0+1\times npq^{n-1}+2\times\frac{n(n-1)}{1\times 2}p^2q^{n-2}$$
$$+3\times\frac{n(n-1)(n-2)}{1\times 2\times 3}p^3q^{n-3}+\cdots+np^n$$
$$=npq^{n-1}+n(n-1)p^2q_1^{n-2}+\frac{n(n-1)(n-2)}{1\times 2}p^3q^{n-3}+\cdots+np^n$$
$$=np\left[q^{n-1}+(n-1)pq^{n-3}+\frac{n(n-1)(n-2)}{1\times 2}p^2q^{n-3}+\cdots+p^{n-1}\right]$$

$$=np[q^{n-1}+\ {}^{n-1}C_1pq^{n-2}+\ {}^{n-1}C_2p^2q^{n-8}+\cdots+\ {}^{n-1}C_{n-1}p^{n-1}]=np(q+p)^{n-1}$$

$$=np(1)=np$$

$\therefore$ **Mean of B.D.** $=\boldsymbol{np}$

(ii) **Variance** $=\sum_{r=0}^{n} r^2p(r)-\left[\sum_{r=0}^{n} rp(r)\right]^2$

$$=\sum_{r=0}^{n}\{r(r-1)+r\}p(r)-\left[\sum_{r=0}^{n} rp(r)\right]^2$$
$$=\sum_{r=0}^{n} r(r-1)p(r)+\sum_{x=2}^{n} rp(r)-\left[\sum_{x=0}^{n} rp(r)\right]^2$$

Now $\sum_{r=0}^{n} r(r-1)p(r)=\sum_{r=0}^{n} r(r-1)\cdot\ {}^nC_rp^rq^{n-r}$

$$= 0(-1)\{{}^nC_0p^0q^n\} + 1\times 0\{nC_1pq^{n-1}) + 2\times 1\{{}^nC_2p^2q^{n-2}$$
$$+3\times 2\{{}^nC_3p^3q^{n-3}\} + \cdots + n(n-1)\{{}^nC_np^nq^0)\}$$
$$= 0 + 0 + 2\times 1\left\{\frac{n(n-1)}{1\times 2}p^2q^{n-2}\right\}$$
$$+3\times 2\left\{\frac{n(n-1)(n-2)}{1\times 2\times 3}p^3q^{n-3}\right\}$$
$$+\cdots + n(n-1)\{1.p^n\}$$
$$= n(n-1)p^2q^{n-2} + n(n-1)(n-2)p^3q^{n-3} + \cdots + n(n-1)p^n$$
$$= n(n-1)p^2[q^{n-2} + (n-2)pq^{n-3} + \cdots + p^{n-2}]$$
$$= n(n-1)p^2[{}^{n-2}C_0q^{n-2} + n^{n-2}C_1pq^{n-3} + \cdots + {}^{n-2}C_{n-2}p^{n-2}]$$
$$= n(n-1)p^2(q+p)^{n-2} = n(n-1)p^2(1) = n(n-1)p^2$$
$$\therefore \text{ Variance } = n(n-1)p^2 + np - (np)^2 = n^2p^2 - np^2 + np - n^2p^2$$
$$= np(1-p) = npq$$

∴ **Standard deviation =** $\sqrt{npq}$

(iii) **Moments**:

1st moment	$\mu_1 = 0$
2nd moment	$\mu_2 = \sigma^2 = npq$
3rd moment	$\mu_3 = npq(q-p)$
4th moment	$\mu_4 = 3n^2p^3q^3 + npq(1-6pq)$

Hence the moment coefficient of skewness is:

$$\beta_1 - \frac{\mu_3^2}{\mu_2^3} = \frac{n^2p^2q^2(q-p)^2}{n^3p^3q^3} = \frac{(q-p)^2}{npq}$$

and

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}}$$

Coefficient of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2p^2q^2 + npq(1-6pq)}{n^2p^2q^2} = \frac{3n^2p^2q^2}{n^2p^2q^2} + \frac{npq(1-6pq)}{n^2p^2q^2}$$
$$= 3 + \frac{1-6pq}{npq}$$

and

$$\gamma_2 = \beta_2 - 3 = \frac{1 - 6pq}{npq}.$$

Remarks:

1. For Binomial distribution, variance < mean.
2. Binomial distribution is symmetrical if $p = q = \frac{1}{2}$. It is positively skewed if $p < \frac{1}{2}$ and negatively skewed if $p > \frac{1}{2}$.
3. Binomial distribution is unimodal if np, is a whole number (i.e., integer) and the mean and mode are equal, each being np.

Example 1. Four coins are tossed simultaneously. What is the probability of getting (i) 2 heads and 2 tails, (ii) at least two heads, and (iii) at least one head.

Here the 'random experiment' consists in tossing 4 coins and observing the number of heads. Let us call the occurrence of heads as 'success', then

$$p = P(\text{ head with single coin }) = \frac{1}{2}, q = 1 - \frac{1}{2} = \frac{1}{2}, n = 4$$

Since the value of p is constant for each coin and the trials are in independent, using formula for binomial distribution, the probability of r successes is

$p(r) = {}^4C_r(\frac{1}{2})^r(\frac{1}{2})^{4-r} = {}^4C_r(\frac{1}{2})^4$ ………… (1)

(i) Putting $r = 2$ in (1), we get

$P(2 \text{ heads and } 2 \text{ tails }) = p(2) = {}^4C_2\left(\frac{1}{2}\right)^4 = 3/8.$

(ii) Since 'at least 2 heads' implies '2 or 3 or 4' heads, the probability of at least 2 successes is given by the sum of probabilities $p(2) + p(3) + p(4)$.

$$\therefore P(\text{ at least 2 heads }) = {}^4C_2\left(\frac{1}{2}\right)^4 + {}^4C_3\left(\frac{1}{2}\right)^4 + {}^4C_4\left(\frac{1}{2}\right)^4$$
$$= \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{11}{16}$$

(iii) $P(\text{ at least 1 head }) = 1 - P(\text{ no head }) = 1 - {}^4C_0\left(\frac{1}{2}\right)^4 = \frac{15}{16}.$

Example 2. Three percent of a given lot of manufactured parts are defective. What is the probability that in a sample of four items, none will be defective?

Here $p = P(\text{ defective item }) = 3/100 = 0.03$

$$q = 1 - p = 1 - 0.03 = 0.97; n = 4$$

Using the formula for the binomial distribution, the probability of getting defectives in a sample of size 4 is given by

$$P(X = r) = p(r) = {}^4C_r(0.03)^r(0.97)^{4-r}$$

The required probability that none of the four items will be defective is

$$p(0) = {}^4C_0(0.03)^0(0.97)^{4-0} = (0.97)^4 = .885$$

Example 3. 8 coins are tossed at a time, 256 times. Find the expected frequencies of success (getting a head) and tabulate the result obtained. Find the mean and S.D. of the fitted distribution.

Probability of success or $p = \frac{1}{2}, n = 8$ and $N = 256$. Using the formula for the binomial distribution, the probability of getting r successes are:

$$p(r) = {}^nC_r p^r q^{n-r} = {}^8C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{8-r}$$
$$= {}^8C_r \left(\frac{1}{2}\right)^8 = \frac{1}{256} \times {}^8C_r$$

Hence in 256 throws of 8 coins, the frequency of r successes are:

$$f(r) = 256 \times \frac{1}{256}\, {}^8\text{Cr} = {}^8\text{Cr}$$

COMPUTATION OF EXPECTED FREQUENCIES

Success	Expected Frequency
0	${}^8C_0 = 1$
1	${}^8C_1 = 8$
2	${}^8C_2 = 28$
3	${}^8C_3 = 56$
4	${}^8C_4 = 70$
5	${}^8C_5 = 56$
6	${}^8C_6 = 28$
7	${}^8C_7 = 8$

8	${}^{8}C_{8} = 1$

Mean $= np$ or $8 \times \frac{1}{2} = 4$

Standard deviation $= \sqrt{npq}$ or $\sqrt{8 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{2} = 1.4142.$

Example 4. A box contains 100 transistors, 20 of which are defective, 10 are selected for inspection. Indicate what is the probability that

(i) all 10 are defective, and

(ii) all 10 are good,

(iii) at least one is defective, and

(iv) at the most 3 are defective?

Solution. Let X represent the number of defective transistors selected. Then the possible values of X are $0,1,2,\ldots,10$. Now

$\mathrm{p} = \mathrm{P}(\text{ transistor is defective }) = \frac{20}{100} = \frac{1}{5}; \mathrm{q} = 1 - \frac{1}{5} = \frac{4}{5}$

Using formula for binomial distribution, the probability of r defective transistors is

$$\mathrm{P}(\mathrm{X} = \mathrm{r}) = \mathrm{p}(\mathrm{r}) = {}^{10}\mathrm{Cr}(1/5)^{\mathrm{r}}(4/5)^{10-\mathrm{r}}$$

(i) Probability that all 10 are defective is

$$\mathrm{P}(\mathrm{X} = 10) = \mathrm{p}(10) = {}^{10}\mathrm{C}_{10}(1/5)^{10}(4/5)^{0} = \frac{1}{5^{10}}$$

(ii) Probability that all 10 are good $= \mathrm{P}(\mathrm{X} = 0) = \mathrm{p}(0)$

$$= {}^{10}\mathrm{C}_{0}\left(\frac{1}{5}\right)^{0}\left(\frac{4}{5}\right)^{10} = \left(\frac{4}{5}\right)^{10}$$

(iii) Probability that at least one is defective is given by the sum of probabilities, viz.,

$$\mathrm{p}(1) + \mathrm{p}(2) + \mathrm{p}(3) + \cdots + (10)$$

or

$$1 - \mathrm{p}(0) = \mathrm{I} - {}^{10}\mathrm{C}_{0}(1/5)^{0}(4/5)^{10} = 1 - (4/5)^{10}$$

(iv) Required probability of at the most three defective items is

$$\begin{aligned} P(X \leqslant 3) &= P(X=0)+P(X=1)+P(X=2)+P(X=3) \\ &= p(0)+p(1)+p(2)+p(3) \\ &= {}^{10}C_0(1/5)^0(4/5)^{10} + {}^{10}C_1(1/5)^1(4/5)^9 \\ &\quad + {}^{10}C_2(1/5)^2(4/5)^8 + {}^{10}C_3(1/5)^8(4/5)^2 \\ &= 1(\cdot 017)+10(\cdot 026)+45(\cdot 0067)+120(\cdot 0017) \\ &= 107+\cdot 26+\cdot 30+\cdot 204 = 871 \end{aligned}$$

Example 5. The incidence of occupational disease in an industry is such that the workmen have a 20% chance of suffering from it. What is the probability that out of six workmen, 4 or more will contract the disease?

Solution. Let the random variable X denote the number of workers suffering from the disease. Then the possible values of X are 0,1,2, ... ,6.
$p = P(\text{ worker suffer from a disease }) = 20/100 = 1/5;$

$$q = 1-(1/5) = 4/5, n = 6$$

Using the formula for binomial distribution, we have

$$P(X=r) = p(r) = {}^6C_r(1/5)^r(4/5)^{6-r}$$

The probability that 4 or more workers contract the disease is -

$$\begin{aligned} P(X \leqslant 4) &= P(X=4)+P(X=5)+P(X=6) \\ &= {}^6C_4(1/5)^4(4/5)^2 + {}^6C_5(1/5)^5(4/5) + {}^6C_6(1/5)^6 \\ &= \frac{15\times 16}{15625}+\frac{6\times 4}{15625}+\frac{1}{15625} = \frac{265}{15625} = 0.0169 \end{aligned}$$

Example 6. Assuming that half of the population is vegetarian so that chance of an individual being a vegetarian is $\frac{1}{2}$. Assuming that 100 investigators can take sample of 10 individuals each to see whether they are vegetarians, how many investigators would you expect to report that: three people or less were vegetarians?

Solution. In the usual notations, we have:

n = 10
p = probability that an individual is a vegetarian $= \frac{1}{2}$
q = 1 – p = $\frac{1}{2}$

Then, using the formula for binomial distribution, the probability that there are r vegetarians in a sample of 10 is given by

$$P(X = r) = p(r) = {}^{10}Crp^r q^{10-r} = {}^{10}Cr\left(\frac{1}{2}\right)^r\left(\frac{1}{2}\right)^{10-r}$$

$$= {}^{10}Cr\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}\,{}^{10}Cr$$

Thus, the probability that in a sample of 10, three or less people are vegetarian is:

$$p(0) + p(1) + p(2) + p(3) = \frac{1}{1024}[\,{}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3][\text{ From } (*)]$$

$$= \frac{1}{1024}[1 + 10 + 45 + 120] = \frac{176}{1024} = \frac{11}{64}$$

Hence, out of 100 investigators, the number of investigators who will report 3 or less vegetarians in a sample of 10 is

$$100 \times \frac{11}{64} = \frac{275}{16} = 17 \cdot 2 = 17,$$

since the number of investigators cannot be in fraction.

Example 7. Five coins are tossed 3,200 times: find the frequencies of the distribution of heads and tails and tabulate the results. Calculate the mean number of success and standard deviation.

Solution. $p = P$ (head with single coin) $= \frac{1}{2}; q = 1 - \frac{1}{2} = \frac{1}{2}, n = 5.$

Now, using the formula for binomial distribution, probability of r successes in a toss of 5 coins is given by

$$p(r) = {}^5C_r\left(\frac{1}{2}\right)^r\left(\frac{1}{2}\right)^{5-r} = {}^5C_r\left(\frac{1}{2}\right)^5$$

Hence in 3200 throws of 5 coins, the frequencies of r successes are:

$$f(r) = N.p(r) = 3{,}200 \times {}^5C_r\left(\frac{1}{2}\right)^5 = 100 \times {}^5Cr$$

No. of heads (r)	$f(r) = 100 \times {}^5Cr$
0	$f(0) = 100 \times {}^5C_0 = 100$
1	$f(1) = 100 \times {}^5C_1 = 500$
2	$f(2) = 100 \times {}^5C_2 = 1{,}000$
3	$f(3) = 100 \times {}^5C_3 = 1{,}000$

4	$f(4) = 100 \times {}^5C_4 = 500$
5	$f(5) = 100 \times {}^5C_5 = 100$
Total	3,200

Mean number of successes $= np = 5 \times \frac{1}{2} = 2.5$

Standard deviation $= \sqrt{npq} = \sqrt{5 \times \frac{1}{2} \times \frac{1}{2}} = 1 \cdot 118.$

Example 8. (a) Comment on the following:
For a binomial distribution, mean $= 7$, variance $= 11$.
(b) For a binomial distribution, the mean is 6 and the standard deviation is $\sqrt{2}$. Write out all the terms of the binomial distribution.

Solution. For a binomial distribution,
Mean $= np = 7$
Variance $= npq = 11$
Dividing Variance by Mean, we get

$$\frac{npq}{np} = \frac{11}{7} \quad \text{or} \quad q = 1.6$$

Since q cannot be more than 1, the given data are inconsistent.
(b) Here we are given:
Mean $= np = 6$
Standard deviation $= \sqrt{npq} = \sqrt{2}$ or $npq = 2$
Dividing Standard deviation by Mean, we get

$$\frac{npq}{np} = \frac{2}{6} \text{ or } q = \frac{1}{3} \Rightarrow p = 1 - \frac{1}{3} = \frac{2}{3}$$

Substituting the values of $p = \frac{2}{3}$ in Mean, we have

$$n \times \frac{2}{3} = 6 \text{ or } n = 9$$

Now $p(r) = {}^nC_r, p^r q^{n-r} = {}^9C_r \left(\frac{1}{3}\right)^{9-r} \left(\frac{2}{3}\right)^r ; r = 0,1,2, \dots, 9$
∴ The terms of the binomial distribution would be:
$\left(\frac{1}{3}\right)^9 \cdot 9\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^8, 36\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^7, \dots, \left(\frac{2}{3}\right)^9.$

CHECK YOUR PROGRESS

True or false or MCQ Questions

Problem 1. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by
a) np
b) n
c) p
d) np(1-p)

Problem 2. In a Binomial Distribution, if p, q and n are probability of success, failure and number of trials respectively then variance is given by
a) np
b) npq
c) np^2q
d) npq^2

Problem 3. In a Binomial Distribution, the mean and variance are equal.
a) True
b) False

Problem 4. It is suitable to use Binomial Distribution only for
a) Large values of ‘n’
b) Fractional values of ‘n’
c) Small values of ‘n’
d) Any value of ‘n’

Problem 5. Binomial Distribution is a
a) Continuous distribution
b) Discrete distribution
c) Irregular distribution
d) Not a Probability distribution

11.7 *SUMMARY*

(i) For Binomial Distribution $P(X=r)=p(r)={}^nCrp^rq^{n-r}; \quad r=0,1,2,\ldots,n.$

(ii) Pascal's Triangle is a triangular array of numbers where each number is the sum of the two numbers directly above it. It's a fundamental concept in mathematics, particularly in algebra, probability, and combinatorics. The triangle starts with a "1" at the top, and each subsequent row begins and ends with 1, with the numbers in between calculated by adding the two numbers above.

11.8 *GLOSSARY*

(i) Permutation

(ii) Combinations

(iii) Probability

11.9 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

11.10 *TEWRMINAL QUESTIONS*

1. What do you mean by Binomial Distribution.
2. What do you mean by Probability function of Binomial Distribution.
3. Define Pascal’s triangle.
4. Define constants of Binomial Distribution.

11.11 ANSWERS

CYQ 1. a

CYQ 2. b

CYQ 3. b

CYQ 4. c

CYQ 5. b

UNIT 12: *POISSON DISTRIBUTION*

CONTENTS:

12.1 *INTRODUCTION*

A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space. The Poisson distribution has only one parameter, λ (lambda), which is the mean number of events.

12.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Poisson Distribution

(ii) Characteristics of Poisson Distribution

12.3 *POISSON DISTRIBUTION*

In the Binomial distribution, it was found that there is a sample of a definite size so that it is possible to count the number of times an event is observed; in other words, n was precisely known. But there are certain situations where this may not be possible. The basic reason for this is that the event is rare and casual. We can also say that the successful events in the total event space are few, i.e., the events like accidents on a road, defects in a product, goals scored at a football match, etc. In these, we know the number of times an event occurs but not how many times it does not occur. Obviously, the toral number of trials in regard to a given experiment is not precisely known. The Poisson distribution is very suitable in case of such rare events. Only the average chance of occurrence based on past experience or a small sample extracted for the purpose will enable us to construct the whole distribution. Now, whereas the Binomial distribution requires two parameters p and n .This distribution requires only the value of m which is the mean of the occurrences of an event (np) based on existing knowledge on the matter.

Poisson distribution was derived in 1837 by a French mathematician, Simeon D. Poisson (1781-1840). Poisson distribution may be obtained as a limiting case of the Binomial probability distribution under the following conditions:

(i) n, the number of trials is indefinitely large, i.e., $n \rightarrow \infty$

(ii) p, the constant probability of success for each trial is indefinitely small, i.e., $p \rightarrow 0$.

(iii) $np = m$, (say), is finite.

The probability function of random variable X following Poisson distribution is

$$P(X = r) = p(r) = \frac{m^r}{r!} \cdot e^{-m}; r = 0,1,2, \ldots$$

where $X =$ the number of successes (occurrences of the event)

$e = 2 \cdot 71828$ [The base of the system of natural logarithms]

and $r! = r(r-1)(r-2)..3.2.1.$

12.4 *BINOMIAL APPROXIMATION TO POISSON DISTRIBUTION*

In case of binomial distribution, the probability of successes is given by

$$p(r)\sigma^n Crp^r q^{n-r} = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} p^r \cdot q^{n-r}$$

Put now $p = \frac{m}{n}$ and $q = 1 - p = 1 - \frac{m}{n}$. We get

$$p(r) = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!}\left(\frac{m}{n}\right)^r \cdot \left(1 - \frac{m}{n}\right)^{n-r}$$

$$= \frac{1.[1-(1/n)][1-(2/n)]\ldots[1-[(r-1/n)]m^r}{r!}\left[\frac{[1-(m/n)]^n}{(1-(m/n)^r)}\right]$$

For fixed r, as $n \rightarrow \infty$

$$\left[1 - \frac{1}{n}\right], \left[1 - \frac{2}{n}\right], \ldots \ldots, \left[1 - \frac{r-1}{n}\right] \text{ all tend to 1 and}$$

$$\left[1 - \frac{m}{n}\right]^{-n} \text{ to } e^{-m}.$$

Hence in the limiting case: $p(r) = \frac{e^{-m} m^r}{r!}; r = 0,1,2, \ldots$

Remarks. 1. The distribution based on the above function is derived from the function of the constant ' e ' which is derived from the natural law of exponential growth indicated below :

$$e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots$$

$$= 1 + 1 + \cdot 5 + 0 \cdot 1667 + 0 \cdot 00833 + \cdots = 2 \cdot 71828$$

Now, for e^m, it is

$$e^m = \frac{m^0}{0!} + \frac{m^1}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots + \frac{m^n}{n!} + \cdots$$

$$= 1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots + \frac{m^n}{n!} + \cdots$$

In order to have the above in the form of a probability distribution based on the following law of indices

$$e^{-m} \times e^{m} = e^{0} = 1$$

we write the expression as

$$e^{-m}\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots \frac{m^n}{n!} + \cdots\right]$$

The successive terms after opening brackets are

$$e^{-m}, me^{-m}, \frac{m^2}{2!}e^{-m}, \frac{m^3}{3!}e^{-m}, \ldots, \frac{m^n}{n!}e^{-m}, \ldots$$

which gives the probabilities of that variate X which can assume infinite values of $0,1,2,3,4,5,\ldots$ as shown below:

Values of variable (X):	0	1	2	3		Total
Probability $P(X = r) = p(r)$	e^{-m}	$e^{-m}m$	$\frac{e^{-m}m^2}{2!}$	$\frac{e^{-m}m^3}{3!}$	$\cdots$	1

12.5 *CHARACTERISTICS OF POISSON DISTRIBUTION*

The Poisson distribution possesses the following characteristics:

1. Discrete distribution. Like binomial distribution, Poisson distribution is also a discrete probability distribution, i.e., it is concerned with occurrences that can take integral values, $0,1,2,\ldots$
2. Values of p and q. It is applied in situations where the probability of the success of an event (p) is very small and that of failure (q) is very high, almost equal to 1, and n is also very large.
3. Main parameter. If we know m, all the probabilities of the Poisson distribution can be obtained. m is, therefore, called the parameter of the Poisson distribution.
4. Form of distribution. The Poisson distribution is a positively skewed distribution, i.e., to the left. With an increase in the value of m, the distribution shifts to the right and the skewness is reduced.
5. Assumptions. The Poisson distribution is based on the following assumptions :
   (i) the occurrence or non-occurrence of an event does not influence the occurrence or non-occurrence of any other ;
   (ii) the probability of success for a short time interval or a small space is proportional to the length of the time interval or space, as the case may be;
   (iii) the probability of the happening of more than one event in a very small interval is negligible.

- **Constants of Poisson Distribution**

Mean and S.D. for the Poisson distribution are calculated in the same manner as for the Binomial Distribution. Obviously,

r	p(r)	rp(r)	$r^2p(r)$
0	e^{-m}	0	0
1	me^{-m}	$1. me^{-m}$	me^{-m}
2	$\frac{m^2e^{-m}}{2!}$	$2. \frac{m^2e^{-m}}{2!}$	$2^2. \frac{m^2e^{-m}}{2!}$
4	$\frac{m^8e^{-m}}{3!}$	$3. \frac{m^8e^{-m}}{3!}$	$3^2. \frac{m^3e^{-m}}{3!}$
	$\frac{m^4e^{-m}}{4!}$	$4. \frac{m^4e^{-m}}{4!}$	$4^2. \frac{m^4e^{-m}}{4!}$
:	⋮	⋮	⋮

$$\mu_1{}' = \sum_{r=0}^{\infty} rp(r) = \sum_{r=0}^{\infty} r \cdot \frac{e^{-m}m^r}{r!}$$
$$= 0 \times e^{-m} + 1 \times me^{-m} + 2 \times \frac{m^2}{2!}e^{-m} + \cdots \text{to} \infty$$
$$= e^{-m} \cdot m \left[1 + m + \frac{m^2}{2!} + \cdots \text{to } \infty\right]$$
$$= e^{-m} \cdot m \cdot e^m = m$$

since the infinite series in the rectangular brackets is the expansion of e^m. Similarly,

$$\mu'_2 = \sum_{r=0}^{\infty} r^2 p(r) = \sum_{x=0}^{\infty} r^2 \cdot \frac{e^{-m} m^r}{r!}$$
$$= 0^2 \times e^{-m} + 1^2 \times \frac{m}{1!} e^{-m} + 2^2 \times \frac{m^2}{2!} e^{-m} + \cdots \text{to } \infty$$
$$= e^{-m} m \left\{1 + 2m + \frac{3m^2}{2!} + \cdots\right\}$$
$$= e^{-m} m \left\{1 + m + \frac{m^2}{2!} + \cdots\right\} + e^{-m} \cdot m^2 \left\{1 + m + \frac{m^2}{2!} + \cdots\right\}$$
$$= e^{-m} \cdot m \cdot e^m + e^{-m} \cdot m^2 \cdot e^m$$
$$= m + m^2$$

Hence $\mu_2 = \mu_2{}' - \mu_1{}'^2 = m + m^2 - (m)^2 = m$ and $\sigma = \sqrt{\mu_2} = \sqrt{m}$.
Thus, the mean of the Poisson distribution is m and the standard deviation is $\sqrt{m}$.

Other constants. The moments about the mean of a Poisson distribution are :

First moment, $\mu_1 = 0$
Second moment, $\mu_2 = m$
Third moment, $\mu_3 = m$
Fourth moment, $\mu_4 = m + 3m^2$

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3} = \frac{m^2}{m^3} = \frac{1}{m} \Rightarrow \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$$
$$\beta_2 = \frac{\mu_4}{\mu_2{}^2} = \frac{3m^2 + m}{m^2} = 3 + \frac{1}{m} \Rightarrow \gamma_2 = \beta_2 - 3 = \frac{1}{m}$$

- ## Importance of Poisson Distribution

Poisson distribution has found application in a variety of fields such as waiting time problems (Queueing Theory), Insurance, Business, Economics, Industry, Physics and Biology.

Most of the temporal and spatial distributions follow the Poisson law. Distributions dealing with events which are supposed to occur in equal intervals of time are called temporal distributions. Spatial distributions deal with events which are supposed to occur in intervals of equal lengths along a straight line.

(a) Examples of temporal distributions:

1. Number of telephone calls received at a particular switchboard per minute during a certain hour of the day.
2. The number of customers arriving per hour at the supermarket.

3. Number of deaths per day due to a specific disease (not in epidemic form), such as heart attack or cancer, in a certain city for a number of days, or the number of suicides reported on a particular day.

4. Number of cars passing a busy corner per minute during a certain part of the day for a number of minutes.

5. Number of α-particles emitted by a radioactive substance and received at a certain portion of a plate during intervals, say, of 7.5 seconds for a number of these intervals.

(b) Examples of spatial distribution:

6. Number of flying bombs that hit London during World War II. The city was divided into 576 units of $\frac{1}{4}$ sq. kilometres each, and the number of units receiving $0,1,2,$ ...Bombs were recorded, and the distribution was found to follow the Poisson law remarkably well.

7. Number of typographical errors per page in a typed material or the number of printing mistakes per page in a book.

8. Number of bacteria colonies on a slide per 01 sq. mm. for a number of such units.

9. The number of defective materials, say, blades, pins, optical lenses, etc., in a packing manufactured by a good concern.

10. Number of defects or scratches on a sheet of glass or a piece of furniture.

- **TABLE OF VALUES OF e^{-m}**
(m lying between 0 and +1)

m	0	1	2	3	4	5	6	7	8	9
0.0	1.000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9321	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8207
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

For example, $e^{-.5} = .6065$. Similarly, $e^{-.57} = 0.5655$.

- **TABLE OF e^{-m}**

(m lying between +1 and +10)

m	1	2	3	4	5
e^{-m}	0.36788	0.13534	0.04979	0.01832	0.006738
m	6	7	8	9	10
e^{-m}	0.002479	0.000912	0.000337	0.000123	0.00045

For example, $e^{-1.5} = e^{-1} \times e^{-.5} = .36788 \times .6065 = .22312922$ or $\cdot\, 2231$..

Remark: Expectation and variance of the Poisson distribution:

The expectation and variance of the Poisson distribution can be derived directly from the definitions which apply to any discrete probability distribution. However, the algebra involved is a little lengthy. Instead we derive them from the binomial distribution from which the Poisson distribution is derived. Intuitive Explanation One way of deriving the mean and variance of the Poisson distribution is to consider the behavior of the binomial distribution under the following conditions:

1. n is large

2. p is small

3. np = λ (a constant) Recalling that the expectation and variance of the binomial distribution are given by the results E(X) = np and V(X) = np(1 − p) = npq it is reasonable to assert that condition (2) implies, since q = 1 − p, that q is approximately 1 and so the expectation and variance are given by E(X) = np and V(X) = npq ≈ np

In fact the algebraic derivation of the expectation and variance of the Poisson distribution shows that these results are in fact exact. Note that the expectation and the variance are equal.

Example 1. A company knows on the basis of past experience that 3% of its bulbs are found to be defective. We have to find out the probability of 0,1,2,3,4, and 5 defectives in a sample of 100 bulbs.

Solution: Here $m = np = 100 \times .03 = 3$

Using Poisson distribution, the probability of r defectives is

$$p'(r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-3.} 3^r}{r!}$$

The probability of 0 defective:

$$p(0) = e^{-3} = 0.04979 \text{ (from tables)}$$

The probability of 1 defective:

$$p(1) = \frac{e^{-m}m^1}{1!} = p(0) \times m = 3 \times 0.04979 = 0.1494$$

The probability of 2 defectives:

$$p(2) = \frac{e^{-m}m^2}{2!} = p(1) \times \frac{m}{2} = 0.1494 \times \frac{3}{2} = 0.2241$$

The probability of 3 defectives:

$$p(3) = \frac{e^{-m}m^3}{3!} = p(2) \times \frac{m}{3} = 0.2241 \times \frac{3}{3} = 0.2241$$

The probability of 4 defectives:

$$p(4) = \frac{e^{-m}m^4}{4!} = p(3) \times \frac{m}{4} = 0.2241 \times \frac{3}{4} = 0.1681$$

The probability of 5 defectives:

$$p(5) = \frac{e^{-m}m^5}{5!} = p(4) \times \frac{m}{5} = 0.1681 \times \frac{3}{5} = 0.1009$$

Example 2. A manufacturing concern employing a large number of workers finds that, over a period of time, the average absentee rate is three workers per shift. Calculate the probability that, on a given shift,
(i) exactly two workers will be absent,
(ii) more than four workers will be absent.

Solution: Average number of absentees per shift $(m) = 3$
Using Poisson probability distribution:

$$P(X = r) = p(r) = \frac{e^{-m}m^r}{r!} = \frac{e^{-3}3^r}{r!}$$

(i) Probability that exactly two workers will be absent is

$$p(2) = \frac{e^{-m}3^2}{2!} = \frac{0.0498 \times 9}{2} = 0.2241$$

(ii)

$$p(0) = \frac{e^{-m}m^0}{0!} = e^{-3} = 0.0498$$

$$p(1) = \frac{e^{-m}m^1}{1!} = e^{-m}m = mp(0) = 3 \times 0.0498 = 0.1494$$

$$p(2) = \frac{e^{-m}m^2}{2!} = \frac{m}{2}(me^{-m}) = \frac{m}{2}p(1) = \frac{3}{2} \times 0.1494 = 0.2241$$

$$p(3) = \frac{m}{3}p(2) = 1 \times 0.2241 = 0.2241$$

$$p(4) = \frac{m}{4}p(3) = \frac{3}{4} \times 0.2241 = 0.1681$$

$$\therefore\ p(0) + p(1) + p(2) + p(3) + p(4) = 0.8155$$

Hence the probability that more than four workers will be absent = 1 – 0.8555 = 0.1845.

Example 3. In a town, 10 accidents take place in a span of 50 days. Assuming that the number of accidents follows the Poisson distribution, find the probability that there will be 3 or more accidents in a day.

Solution: Average number of accidents per day $= \frac{10}{50} = 0.2$

Using Poisson probability law:

$$P(x = r) = p(r) = \frac{e^{-m}m^r}{r!} = \frac{e^{-0.2}(0.2)^r}{r!}$$

Now $e^{-0.2} = \frac{1}{\text{Antilog }(0.4343 \times 0.2)} = \frac{1}{1.2214} = 0.8187$

Probability of 3 or more accidents

$= 1 - P(2 \text{ or less accidents })$
$= 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$
$= 1 - [e^{-0.2}(1 + 0.2 + 0.02)]$
$= 1 - 0.8187 \times 1.22 = 1 - 0.9988 = 0.0012.$

Example 4. Fit a Poisson distribution to the following data and calculate the theoretical frequencies:

Number of mistakes per page:	0	1	2	3	4
Number of pages on which mistake occurred:	109	65	22	3	1

Solution:

CALCULATIONS FOR MEAN

X	f	fX
0	109	0
1	65	65
2	22	44
3	3	9
4	1	4
	$\Sigma f = 200$	$\Sigma. fX = 122$

$$\text{Mean} = \frac{\Sigma fX}{\Sigma f} = \frac{122}{200} = 0.61$$

Probability of 0 mistake or

$$p(0) = e^{-0.61} = (2.71828)^{-0.61} = \frac{1}{\text{Antilog}\,\{\log 2.71828 \times (-0.61)\}}$$

$$= \frac{1}{\text{Antilog}\,\{0.4343 \times (0.61)\}}$$

$$= \frac{1}{\text{Antilog}\,(0.26492)} = \frac{1}{1.841} = 0.5432$$

CALCULATIONS FOR EXPECTED FREQUENCIES

X	Probability $p(r) = e^{-m}m^r/r!$	Frequency (in round figures)
0	$p(0) = e^{-0.61} = 0.5432$	$200 \times 0.5432 = 108.64 \approx 109$
1	$p(1) = 0.5432 \times 0.61 = 0.3313$	$108.64 \times 0.61 = 66.27 \approx 66$
2	$p(2) = 0.5432 \times \frac{(0.61)^2}{2} = 0.101$	$66.27 \times \frac{0.61}{2} = 20.21 \approx 20$
3	$p(3) = 0.5432 \times \frac{(0.61)^3}{6} = 0.021$	$2021 \times \frac{0.61}{3} = 4.11 \approx 4$
4	$p(4) = 0.5432 \times \frac{(0.61)^4}{24} = 0.003$	$4.11 \times \frac{0.61}{4} = 0.63 = 1$

Example 5. Between 2 and 4 p.m. the average number of phone calls per minute coming into the switchboard of a company is 2.5. Find the probability that during one particular minute there will be (i) no phone call at all, (ii) exactly 3 calls, (iii) at least 7 calls. (Given $e^{-2} = 0.13534$ and $e^{-0.5} = 0.60650$).

Solution. Let X represent the number of telephone calls received during the period 2 to 4 p.m. Then the random variable X follows the Poisson distribution with mean $m = 2 \cdot 5$. Using the formula for the Poisson distribution, we have

$$P(X = r) = p(r) = \frac{e^{-2\cdot 5}(2.5)^r}{r!}$$

(i) Probability of no phone call during the period (i.e., $X = 0$) is

$$P(X = 0) = \frac{e^{-2\cdot 5}(2.5)^0}{0!} = e^{-2\cdot 5} = e^{-2} \cdot e^{-\cdot 5} = 0.1353 \times 0.6065 = 0.0821$$

(ii) Probability of exactly 3 calls during the period (i.e., **X** $= 3$) is

$$P(X = 3) = \frac{e^{-2\cdot 5}(2 \cdot 5)^3}{3!} = 0.2138$$

(iii) Required probability is

$$P(X > 7) = 1 - P(X \leqslant 6) = 1 - \sum_{r=0}^{6} P(X = r)$$

$$= 1 - e^{-2.5} \sum_{r=0}^{6} \frac{(2.5)^r}{r!}$$

Example 6. Using Poisson approximation to the Binomial distribution, solve the following problem: If the probability that an individual suffers a bad reaction from a particular injection is 0.001, determine the probability that out of 2,000 individuals (i) exactly three, (ii) more than two individuals will suffer a bad reaction (Given $e^2 = 7.4$)

Solution. The random variable X, 'number of individuals out of 2,000 who suffer from bad reaction, follows Binomial distribution with $n = 2{,}000$ and $p = 0.001$. However, since p is small and n is large, and

$$np = 2000 \times 0.001 = 2$$

is finite, the binomial distribution can be approximated by Poisson distribution with $m = 2$. Using the formula for the Poisson distribution,

$$P(X = r) = p(r) = \frac{e^{-2} 2^r}{r!}, \text{ we have}$$

(i) The required probability of exactly 3 individuals suffering from bad reaction is

$$P(X = 3) = \frac{e^{-2} 2^3}{3!} = \frac{e^{-2} \times 4}{3} = 0.1351 \times \frac{4}{3} = 0.1801.$$

(ii) The required probability that more than 2 individuals will suffer from bad reaction is

$$\begin{aligned} P(Y > 2) &= P(X = 3) + P(X = 4) + \cdots \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - [p(0) + p(1) + p(2)] \\ &= 1 - \left\{\frac{e^{-2} \cdot 2^0}{0!} + \frac{e^{-2} \cdot 2^1}{1!} + \frac{e^{-2} \cdot 2^1}{2!}\right\} \\ &= 1 - \left\{\frac{1}{e^2} + \frac{2}{e^2} + \frac{2}{e^2}\right\} = 1 - \left\{\frac{1}{7 \cdot 4} + \frac{2}{7 \cdot 4} + \frac{2}{7 \cdot 4}\right\} \\ &= 1 - \{0.1351 + 0.2702 + 0.2702\} = 0.3245. \end{aligned}$$

Example 7. If a random variable X follows Poisson distribution such that $P(X = 1) = P(X = 2)$, find
(i) the mean of the distribution, (ii) $p(X = 0)$.

Solution. The probability function of the distribution is:
$P(X = r) = p(r) = \frac{e^{-m} m^r}{r!}; r = 0,1,2, \ldots$
Since, $P(X = 1) = P(X = 2)$,

$$me^{-m} = \frac{e^{-m} m^2}{2}$$

Cancelling me^{-m} on both sides, we get

$$1 = \frac{m}{2} \text{ or } m = 2 = \text{mean}$$

Now, $P(X = 0) = e^{-m} = e^{-2} = 0.1353.$

Example 8. In a certain factory turning out blades, there is a 0.2% probability for any blade to be defective. Blades are supplied in packets of 10. Using the Poisson distribution, calculate the approximate number of packets containing no defective, one defective, two defective, three defective blades, respectively, in a consignment of 20,000 packets. You are given that $e^{-0.02} = 0.9802$.

Solution. In the usual notations, we are given:

$$N = 20{,}000, n = 10 \text{ and}$$

$$p = \text{ Probability of a defective blade } = .2\% = \frac{2}{1000} = \frac{1}{500}$$

$$m = np = 10 \times \frac{1}{500} = 0.02$$

Let the random variable X denote the number of defective blades in a packet of 10. Then by Poisson probability law, the probability of r defective blades in a packet is given by

$$p(r) = \frac{e^{-0.02}(0.02)^r}{r!} = \frac{0.9802 \times (0.02)^r}{r!}$$

Hence in a consignment of 20,000 packets, the frequency (number) of packets containing r defective blades is:

$$f(r) = N.p(r) = \frac{20000 \times 0.9802 \times (0.02)^r}{r!} \qquad \dots (*)$$

Putting $r = 0,1,2,3$ and 4 in (*), we get
No. of packets containing no defective blades, i.e.,

$$f(0) = 20000 \times 0.9802 = 19604$$

No. of packets containing 1 defective blade, i.e.,

$$f(1) = \frac{20{,}000 \times 0.9802 \times (0.02)}{1}$$
$$= 19604 \times 0.02 = 392.08 \approx 392$$

No. of packets containing 2 defective blades, i.e.,

$$f(2) = \frac{20000 \times 0.9802 \times (0.02)^2}{2!}$$
$$= \frac{392.08 \times 0.02}{2} = 3.9208 \simeq 4$$

No. of packets containing 3 defective blades, i.e.,

$$f(3) = \frac{20000 \times 0.9802 \times (0.02)^3}{3!}$$
$$= \frac{3.9208 \times 0.02}{3} = 0.026138 \simeq 0,$$

since the number of packets cannot be in fraction.
Obviously $f(4)$ is also zero.

Hence, the number of packets containing $0,1,2,3$ and 4 defective blades are respectively 19604, 392, 4, 0, 0.

12.6 *FITTING OF POISSON DISTRIBUTION*

To fit the Poisson distribution, the probabilities of 0,1,2,3,4 ..., successes are obtained, as discussed below:

(i) We compute the mean $\bar{X}$ of the given distribution and take it equal to the mean of the fitted (Poisson) distribution, i e., we take $m = \bar{X}$.

(ii) The value of e^{-m} is obtained as follows:

$$e^{-m} = \frac{1}{e^m} = \frac{1}{(2 \cdot 7183)^m}$$

$$= \frac{1}{\text{Antilog}\ (\log 2.7183 \times m)} = \frac{1}{\text{Antilog}\ (.4343 \times m)}$$

$$= \text{Reciprocal of}\ [\text{Antilog}\ (.4343 \times m)]$$

The value of e^{-m} can also be obtained from tables.

(iii) The various probabilities of the Poisson distribution can be obtained using the formula:

$$P(X = r) = p(r) = \frac{e^{-m} m^r}{r!}; r = 0,1,2, \ldots$$

(iv) By using the following formula, the expected or theoretical frequencies according to Poisson distribution are given by

$$f(r) = N \cdot p(r)$$

where N is the total observed frequency.

(v) We compare the observed and expected frequencies class by class and determine the measure of agreement between the two kinds of frequencies by means of what is called the χ^2-test.

Example 1. The distribution of typing mistakes committed by a typist is given below. Assuming a Poisson Model, find out the expected frequencies.

Mistakes per page	:	0	1	2	3	4	5
No. of pages	:	142	156	69	27	5	1

Solution. Mean $= \frac{\Sigma fX}{N} = \bar{X}$

$$= \frac{0 \times 142 + 1 \times 156 + 2 \times 69 + 3 \times 27 + 4 \times 5 + 5 \times 1}{100} = \frac{400}{400} = 1$$

The above distribution is approximated by a Poisson distribution, then the parameter of Poisson distribution (number) of $m = \bar{X} = 1$ and by Poisson probability law, the frequency (number) of pages containing r mistakes is given by

$$f(r) = N \cdot p(r) = 400 \times \frac{e^{-1}}{r!}$$

Putting $x = 0,1,2, \ldots ,5$, we get the expected frequencies of Poisson distribution,

Also,

$$p(0) = e^{-m} = e^{-1} = 0.3679 \qquad \text{(From table)}$$

COMPUTATION OF EXPECTED FREQUENCIES

x	Expected frequency $f(r) = N.p(r)$	Observed frequency
0	$400 \times 0.3679 = 147.16 \simeq 147$	142
1	$400 \times 0.3679 \times 1 = 147.16 \simeq 147$	156
2	$\frac{\Delta 01 \times 0.3679 \times (1)^2}{2!} = 73.58 \simeq 74$	69
3	$\frac{400 \times 0.3679 \times (1)^3}{3!} = 24.52 \simeq 25$	27
4	$\frac{400 \times 0.3679 \times (1)^4}{4!} = 6.13 \simeq 6$	5
5	$\frac{400 \times 0.3679 \times (1)^5}{5!} = 1.21 \simeq 1$	1

Remark. Expected frequencies can also be very conveniently computed as explained below:

Value of Variable (X)	Probability $p(r)$	Expected or theoretical Frequencies $f(r) = Np(r)$
0	$p(0) = e^{-m}$	$f(0) = Np(0) = Ne^{-m}$

1	$p(1) = me^{-m}$ $= mp(0)$	$f(1) = mNp(0) = mf(0)$
2	$p(2) = \frac{m^2e^{-m}}{2!}$ $= \frac{m}{2}me^{-m}$ $= \frac{m}{2}p(1)$	$f(2) = \frac{m}{2} \cdot Np(1) = \frac{m}{2}f(1)$
3	$p(3) = \frac{m^3e^{-m}}{3!}$ $= \frac{m}{3}\frac{m^2e^{-m}}{2!}$ $= \frac{m}{3}p(2)$	$f(3) = \frac{m}{3}Np(2) = \frac{m}{3}f(2)$
4 . . .	$p(4) = \frac{m^4e^{-m}}{4!}$ $= \frac{m}{4}\frac{m^3e^{-m}}{3!}$ $= \frac{m}{4}p(3)$ . . .	$f(4) = \frac{m}{4} \cdot Np(3) = \frac{i.1}{4}f(3)$ . . .

CHECK YOUR PROGRESS

True or false or MCQ Questions

Problem 1. In a Poisson Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by?

a) m = np

b) m = $(np)^2$

c) m = np(1-p)

d) m = p

Problem 2. f 'm' is the mean of a Poisson Distribution, then variance is given by
a) m^2
b) $m^{1/2}$
c) m
d) $m/2$

Problem 3. In a Poisson Distribution, the mean and variance are equal.
a) True
b) False

Problem 4. Poisson distribution is applied for
a) Continuous Random Variable
b) Discrete Random Variable
c) Irregular Random Variable
d) Uncertain Random Variable

Problem 5. For a Poisson Distribution, if mean(m) = 1, then P(1) is?
a) 1/e
b) e
c) e/2
d) Indeterminate

12.7 SUMMARY

1. The probability function of random variable X following Poisson distribution is

$$P(X = r) = p(r) = \frac{m^r}{r!} \cdot e^{-m}; r = 0,1,2, \dots$$

2. Poisson distribution has found application in a variety of fields such as waiting time problems (Queueing Theory), Insurance, Business, Economics, Industry, Physics and Biology.

   Most of the temporal and spatial distributions follow the Poisson law. Distributions dealing with events which are supposed to occur in equal intervals of time are called temporal distributions. Spatial distributions deal with events which are supposed to occur in intervals of equal lengths along a straight line.

12.8 *GLOSSARY*

(i) Permutation

(ii) Combinations

(iii) Probability

12.9 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold, (2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

12.10 *TEWRMINAL QUESTIONS*

1. Define Poisson distribution.
2. A man can send an email on average 2 emails per hour. What is the probability that the man sends no email in a given hour?
3. Calculate the mean of the Poisson Distribution given that the number of trials is 20 and probability of success is 0.6.
4. Find the mean of the Poisson distribution given that the variance of the Poisson distribution is 4.

12.11 *ANSWERS*

CYQ 1. (a)

CYQ 2. (c)

CYQ 3. **(**a)

CYQ 4 (b)

CYQ 5 (a)

COURSE NAME: BASIC STATISTICS

COURSE CODE: MT(N)-222

BLOCK - V: CONTINUOUS DISTRIBUTION

UNIT 13: *UNIFORM DISTRIBUTION*

CONTENTS:

13.1 *INTRODUCTION*

A uniform distribution is a probability distribution where all outcomes within a specified range are equally likely. In other words, the probability density function (PDF) is constant over that range. This means there's an equal chance of any value within the interval occurring. Uniform distribution is simplest continuous distribution

- It has prime importance in the field of simulations
- The distribution function F(x) of continuous random variable isa random variable follows uniform distribution over the interval [0,1], irrespective of the distribution of X

13.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Uniform Distribution

(ii) Characteristics of Poisson Distribution

13.3 *UNIFORM DISTRIBUTION*

In statistics, probability distributions can help you decide the probability of a future event that is, the likelihood of something happening. A uniform distribution is a type of probability distribution where the probability remains constant across the entire range of possible values (i.e., all outcomes are equally likely).

A deck of cards is expected to have a uniform distribution because the likelihood of drawing a heart, a club, a diamond, or a spade is equally likely. A coin also has a uniform distribution expectation because the odds of getting either heads or tails are the same.

A uniform distribution can be visualized as a straight horizontal line. For a coin flip, heads or tails each has a probability of occurring 50% of the time (p = 0.50), so it would be plotted on a chart with a line from the y-axis at 0.50.

Types of uniform distribution

1) **Discrete Uniform Distribution:** he possible results of rolling a die provide an example of a discrete uniform distribution. It is possible to roll a 1, 2, 3, 4, 5, or 6, but it is not possible to roll, for example, a 2.3, 4.7, or 5.5. Therefore, the roll of a die generates a discrete distribution with the probability of 1/6 for each outcome. There are only six possible values to return and nothing in between. The possibilities are finite. This applies to a finite number of outcomes, each with the same probability. For instance, when rolling a fair six-sided die, each face (1-6) has an equal chance of appearing, specifically 1/6 probability.
2) **Continuous Uniform Distribution:** This is used when outcomes can be any value within a continuous interval. Because there are infinitely many values, a probability density function (PDF) is used instead of direct probabilities for individual points.

Definition: A continuous random variable X is said to follow uniform distribution over the interval [a, b], if it's probability density function (p. d. f.) is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & otherwise \end{cases}$$

Remark 1. The p. d. f. of uniform distribution is constant over its support [a, b].

2. Nature of p. d. f. curve is flat over its range.

Example: Imagine a fair spinner with numbers 1 to 10. Each number has an equal probability of being landed on (1/10), making it a uniform distribution.

13.4 *DISTRIBUTION FUNCTION*

Distribution function: If $X \sim U(a, b)$ then the distribution function is

$$F(x) = P(X \le x) = \int_a^x f(t)dt = \frac{x-a}{b-a}$$

Therefore, $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x \ge b \end{cases}$

13.5 *MOMENT GENERATING FUNCTION*

Moment Generating Function: If $X \sim U(a, b)$ then the distribution function of X is

$$M_x(t) = E(e^{tx}) = \int_a^b e^{tx} f(x)dx = \int_a^b e^{tx}\left(\frac{1}{b-a}\right)dx = \frac{e^{bt} - e^{at}}{t(b-a)}$$

13.6 *MEAN AND VARIANCE*

Mean and Variance:

$$E(X) = \int_a^b x\, f(x)dx = \int_a^b x\left(\frac{1}{b-a}\right)dx = \frac{b^2 - a^2}{2(b-a)} = \left(\frac{a+b}{2}\right)$$

Therefore,

Mean = $E(X^2) = \int_a^b x^2\, f(x)dx = \int_a^b x^2\left(\frac{1}{b-a}\right)dx = \frac{b^3 - a^3}{3(b-a)} = \left(\frac{a^2 + ab + b^2}{3}\right)$

And $\quad \mathrm{Var}(X) = E(X^2) - \left(E(X)\right)^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2$

Key Points

- In a uniform distribution, all possible outcomes are equally probable.
- In a discrete uniform distribution, outcomes are discrete and have the same likelihood.
- In a continuous uniform distribution, outcomes are continuous and infinite.
- In a normal distribution, data around the mean (average) occur more frequently than occurrences farther from it.
- Uniform distributions can be plotted on charts.

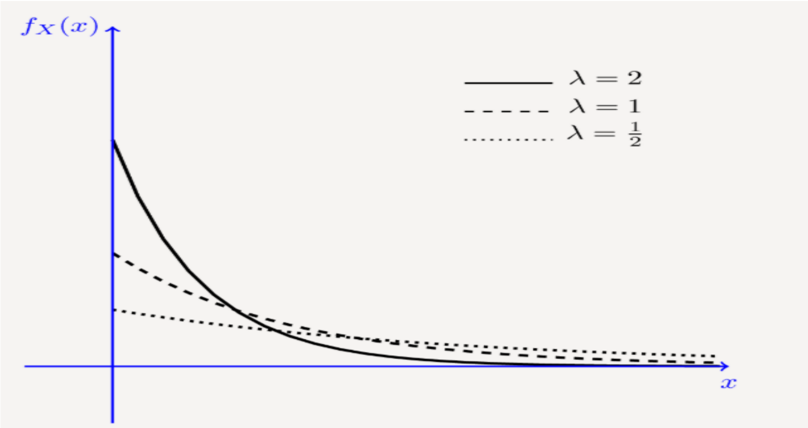
13.7 *EXPONENTIAL DISTRIBUTION*

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. We will now mathematically define the exponential distribution, and derive its mean and expected value. Then we will develop the intuition for the distribution and discuss several interesting properties that it has.

A continuous random variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim$ Exponential (λ), if its P.D.F. is given by

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & otherwise \end{cases}$$

P.D.F. of exponential distribution for several values of λ.



https://www.probabilitycourse.com/chapter4/4_2_2_exponential.php

It is convenient to use the unit step function defined as

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

So, we can write the PDF of an Exponential(λ) random variable as

$$f_X(x) = \lambda e^{-\lambda x} u(x)$$

Let us find its CDF, mean and variance. For $x > 0$, we have

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$

So, we can express the CDF as

$$F_X(x) = (1 - e^{-\lambda x}) u(x)$$

Let $X \sim$ Exponential(λ). We can find its expected value as follows, using integration by parts:

$$\begin{aligned} EX &= \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \qquad \text{choosing } y = \lambda x \\ &= \frac{1}{\lambda} \Big[-e^{-y} - y e^{-y} \Big]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Now let's find Var(X). We have

$$\begin{aligned} EX^2 &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy \\ &= \frac{1}{\lambda^2} \Big[-2e^{-y} - 2ye^{-y} - y^2 e^{-y} \Big]_0^\infty \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

13.8 *GAMMA DISTRIBUTION*

The gamma distribution is another widely used distribution. Its importance is largely due to its relation to exponential and normal distributions. Here, we will provide an introduction to the gamma distribution. Before introducing the gamma random variable, we need to introduce the gamma function.

Gamma function: The gamma function, shown by Γ(x), is an extension of the factorial function to real (and complex) numbers. Specifically, if n ∈ {1,2,3, ...} then

$$\Gamma(n) = (n-1)!$$

More generally, for any positive real number α, Γ(α) is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \text{ for } \alpha > 0.$$

Gamma Distribution:

A continuous random variable XX is said to have a gamma distribution with parameters α > 0 and λ > 0, shown as X ∼ Gamma (α, λ), if its PDF is given by

$$f_X(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda}}{\Gamma(\alpha)}, & x > 0 \\ 0, & otherwise \end{cases}$$

If we let α = 1, we obtain

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & otherwise \end{cases}$$

Thus, we conclude Gamma (1, λ) = Exponential(λ). More generally, if you sum n independent Exponential(λ) random variables, then you will get a Gamma (n, λ) random variable. We will prove this later on using the moment generating function. The gamma distribution is also related to the normal distribution as will be discussed later.

Example 1. If a random variable X is uniformly distributed between 32 and 42, what is the probability that X will be between 32 and 40?

Solution. The probability density function is f(x) = $\frac{1}{42-32} = \frac{1}{10}$. The probability is (40 - 32) /10 = 8 / 10 = 0.8 or 80%.

Example 2. Finding the mean and variance, If X is uniformly distributed in (-1, 4), find its mean and variance.

Solution. Mean = $\frac{-1+4}{2}$ = 1.5 and Variance = $\frac{(4-(-1))^2}{12} = \frac{25}{12}$.

Example 3. Let X be a random variable with PDF given by

$$f_X(x) = \begin{cases} cx^2 & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a. Find the constant c.

b. Find EX and Var(X).

c. Find P (X ≥ 12)

a. To find c, we can use $\int_{-\infty}^{\infty} f_X(u)du = 1$:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(u)du \\ &= \int_{-1}^{1} cu^2 du \\ &= \frac{2}{3}c. \end{aligned}$$

Thus, we must have $c = \frac{3}{2}$.

b. To find EX, we can write

$$\begin{aligned} EX &= \int_{-1}^{1} u f_X(u)du \\ &= \frac{3}{2}\int_{-1}^{1} u^3 du \\ &= 0. \end{aligned}$$

In fact, we could have guessed $EX = 0$ because the PDF is symmetric around $x = 0$. To find Var(X), we have

$$\begin{aligned} \text{Var}(X) &= EX^2 - (EX)^2 = EX^2 \\ &= \int_{-1}^{1} u^2 f_X(u)du \\ &= \frac{3}{2}\int_{-1}^{1} u^4 du \\ &= \frac{3}{5}. \end{aligned}$$

c. To find $P(X \ge \frac{1}{2})$, we can write

$$P(X \ge \frac{1}{2}) = \frac{3}{2}\int_{\frac{1}{2}}^{1} x^2 dx = \frac{7}{16}.$$

Example 4.

Let X be a continuous random variable with PDF given by

$$f_X(x) = \frac{1}{2}e^{-|x|}, \qquad \text{for all } x \in \mathbb{R}.$$

If $Y = X^2$, find the CDF of Y.

Solution.

First, we note that $R_Y = [0, \infty)$. For $y \in [0, \infty)$, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}e^{-|x|}dx \\ &= \int_0^{\sqrt{y}} e^{-x}dx \\ &= 1 - e^{-\sqrt{y}}. \end{aligned}$$

Thus,

$$F_Y(y) = \begin{cases} 1 - e^{-\sqrt{y}} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

13.9 *OCCURRENCE AND APPLICATIONS*

The probabilities for uniform distribution function are simple to calculate due to the simplicity of the function form. Therefore, there are various applications that this distribution can be used for as shown below: hypothesis testing situations, random sampling cases, finance, etc. Furthermore, generally, experiments of physical origin follow a uniform distribution (e.g. emission of radioactive particles). However, it is important to note that in any application, there is the unchanging assumption that the probability of falling in an interval of fixed length is constant.

a) Economics example for uniform distribution

In the field of economics, usually demand and replenishment may not follow the expected normal distribution. As a result, other distribution models are used to better predict probabilities and trends such as Bernoulli process. But according to Wanke (2008), in the particular case of investigating lead-time for inventory management at the beginning of the life cycle when a completely new product is being analyzed, the uniform distribution proves to be more useful. In this situation, other distribution may not be viable since there is no existing data on the new product or that the demand history is unavailable so there isn't really an appropriate or known distribution. The uniform distribution would be ideal in this situation since the random variable of lead-time (related to demand) is unknown for the new product but the results are likely to range between a plausible range of two values. The lead-time would thus represent the random variable. From the uniform distribution model, other factors related to lead-time were able to be calculated such as cycle service level and shortage per cycle. It was also noted that the uniform distribution was also used due to the simplicity of the calculations.

b) Sampling from an arbitrary distribution

The uniform distribution is useful for sampling from arbitrary distributions. A general method is the inverse transform sampling method, which uses the cumulative distribution function (CDF) of the target random variable. This method is very useful in theoretical work. Since simulations using this method require inverting the CDF of the target variable, alternative methods have been devised for the cases where the CDF is not known in closed form. One such method is rejection sampling.

The normal distribution is an important example where the inverse transform method is not efficient. However, there is an exact method, the Box–Muller transformation, which uses the inverse transform to convert two independent uniform random variables into two independent normally distributed random variables.

CHECK YOUR PROGRESS

True or false Questions

Problem 1. What is the mean of a Gamma distribution with shape parameter 'α' and scale parameter 'β' ?

(a) αβ

(b) α/β

(c) β/α

(d) α + β

Problem 2. If α = 1 in the Gamma distribution, it simplifies to which other distribution?

(a) Normal distribution

(b) Poisson distribution

(c) Exponential distribution

(d) Binomial distribution

Problem 3. What does the scale parameter (β) in the Gamma distribution represent?

(a) The number of events

(b) The rate of events

(c) The average time between events

(d) The probability of an event

Problem 4. The Gamma distribution is often used to model:

(a) Discrete data

(b) Normally distributed data

(c) Positive, skewed data

(d) Negative data

Problem 5. Γ(1/2) = √(π)

13.10 *SUMMARY*

(i) Gamma Distribution:

A continuous random variable XX is said to have a gamma distribution with parameters α > 0 and λ > 0, shown as X ~ Gamma (α, λ), if its PDF is given by

$$f_X(x) = \begin{cases} \dfrac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda}}{\Gamma(\alpha)}, & x > 0 \\ 0, & otherwise \end{cases}$$

If we let α = 1, we obtain

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & otherwise \end{cases}$$

Thus, we conclude Gamma (1, λ) = Exponential(λ).

(ii) Distribution function: If $X \sim U(a, b)$ then the distribution function is

$$F(x) = P(X \leq x) = \int_a^x f(t)dt = \frac{x-a}{b-a}$$

Therefore, $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x \geq b \end{cases}$

13.11 *GLOSSARY*

(i) integrations

(ii) Combinations

(iii) Probability

13.12 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold, (2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

13.13 *TEWRMINAL QUESTIONS*

1. What do you mean by uniform Distribution.
2. Explain types of Uniform distributions.
3. A rectangle has a perimeter of 20 cm. The length, x cm, of one side of this rectangle is uniformly distributed between 1 cm and 7 cm. Find the probability that the length of the longer side of the rectangle is more than 6 cm long.
4. The continuous random variable X is uniformly distributed over the interval [–2, 7].

   (a) Write down fully the probability density function f(x) of X.

   (b) Sketch the probability density function f(x) of X.

   (c) E (X^2)

   (d) P (–0.2 < X < 0.6)

13.14 *ANSWERS*

CYQ 1. a

CYQ 2. c

CYQ 3. c

CYQ 4. c

CYQ 5. True

UNIT 14: *NORMAL DISTRIBUTION*

CONTENTS:

14.1 *INTRODUCTION*

The normal distribution, also known as the Gaussian distribution, has a rich history rooted in the study of probability and statistics. While its origins can be traced back to the 18th century, it wasn't until the 19th century that it became a cornerstone of statistical analysis

14.2 *OBJECTIVES*

After completion of this unit learners will be able to:

(i) Normal Distribution

(ii) Standard Normal Distribution

14.3 *NORMAL DISTRIBUTION*

The most important continuous probability distribution used in the entire field of Statistics is the normal distribution. Its graph, called the normal curve, is a bell-shaped curve that extends indefinitely in both directions, coming closer and closer to the horizontal axis without ever reaching it. The mathematical equation of the normal curve was developed by De Moivre in 1733. It was later rediscovered and applied in sciences, both natural and social, by the French mathematician Laplace (1749-1827). The normal distribution is often referred to as the Gaussian distribution in honor of Karl Friedrich Gauss (1777-1855), who also derived its equation from the study of errors in repeated measurements of the same quantity.

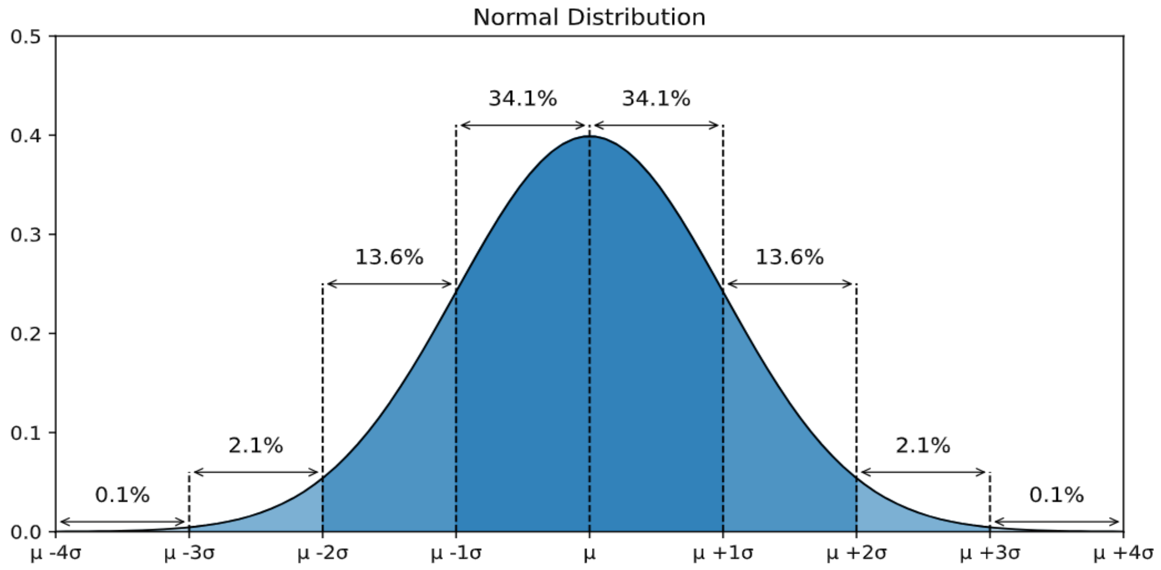
Definition. A continuous random variable X is said to be normally distributed if it has the probability density function represented

by the equation:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty \leqslant x \leqslant \infty \qquad \ldots (*)$$

where μ and σ, the mean and the standard deviation are known as two parameters and $\pi = 3 \cdot 14159, e = 2 \cdot 7183$ are the two constants.

GRAPH OF NORMAL DISTRIBUTION



https://www.w3schools.com/statistics/statistics_normal_distribution.php

Remarks:

1. It may be noted that normal distributions can have different shapes depending on different values of μ and σ, but there is one and only one normal distribution for any given pair of values of μ and σ.

2. The normal distribution with a mean μ and variance σ^2 may be denoted by the symbol $N(\mu, \sigma^2)$.

3. (a) Normal distribution is a limiting form of the Binomial distribution when (i) n, the number of trials is very large, (ii) neither p nor q is very small.
(b) Normal distribution is a limiting case of Poisson distribution when its mean m is large.

4. The probability that a normally distributed random variable with mean μ and variance σ^2 lies between two specified values a and b is $P(a < X < b) =$ Area under the curve $p(x)$ between the specified values $X = a$ and $X = b$.
Since $p(x)$ is a probability distribution, the total area under the curve
$p(x)$ is equal to 1.

14.4 *PROPERTIES OF NORMAL DISTRIBUTION*

1. It is perfectly symmetrical about the mean (μ) and is bell-shaped. This means that if we fold the curve along the vertical line at the centre, the two halves of the curve would coincide.
2. Mean = Median = Mode.
3. It has only one mode, i.e., it is unimodal.
4. The ordinate at the mean of the distribution divides the total area under the normal curve into two equal parts. Further, since the total area under the normal probability curve is 1, the area to the right of the ordinate as well as to the left of the ordinate is 0.5.
5. The following are the descriptive measures of the normal. distribution :

   Mean $= \bar{X}$ or μ (Standard form : $\bar{X} = 0$)

   Standard deviation $= \sigma$ (Standard form : $\sigma = 1$).

   Variance or $\mu_2 = \sigma^2$

   Third central moment, $\mu_3 = 0$

   Fourth central moment, $\mu_4 = 3\sigma^4 = 3{\mu_2}^2$

   Moment coefficient of skewness, $\sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = 0$

   Moment coefficient of kurtosis, $\beta_2 = \frac{\mu_4}{{\mu_2}^2} = 3$

   Hence, it is a mesokurtic curve.
6. The normal curve is concave near the mean value, while near $\pm 3\sigma$ it is convex to the horizontal axis. The points of inflexion, i.e., the points where the change in curvature occurs are $\pm\sigma$.

7. The quartiles Q_1 and Q_3 are equidistant from the median and are given (in terms of μ and σ) by

$$Q_1 = \mu - 0.6745\sigma, \qquad Q_3 = \mu + 4.6745\sigma$$

$\therefore$ Quartile deviation $= \frac{Q_3 - Q_1}{2} = 0.6745\sigma = \frac{5}{6} \times \frac{4}{5}\sigma = \frac{5}{6}$ M.D.

8. The mean deviation about mean is $\frac{4}{5}\sigma$ or 0.7979σ.

9. It is asymptotic to the base line on its either side. As the distance of the curve from the mean increases, the curve comes closer and closer to the axis but never touches it.

10. As x increases numerically [i.e., on either side of $X = \mu$], the value of $p(x)$ decreases rapidly, the maximum probability occurring at $x = \mu$ and is given by [Put $x = \mu$ in (*)]

$$[p(x)]_{max} = \frac{1}{\sigma\sqrt{2\pi}}$$

Thus, maximum value of $p(x)$ is inversely proportional to the standard deviation. For larger values of $\sigma, p(x)$ decreases, i.e., the curve tends to flatten out and for small values of $\sigma, p(x)$ increases, i.e., the curve has a sharp peak.

11. A linear combination of independent normal variates is also a normal variate. If $X_1, X_2, \ldots, X_n$ are independent normal variates with means $\mu_1, \mu_2, \ldots, \mu_n$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively the $_n$ their linear combination

$$a_1X_1 + a_2X_2 + \cdots \ldots + a_nX_n \ldots \qquad (**)$$

where $a_1, a_2, \ldots, a_n$, are constants is also a normal variate with

$$\left.\begin{aligned} \text{Mean } &= a_1\mu_1 + a_2\mu_2 + \cdots + a_{n\mu n} \\ \text{Variance } &= a_1{}^2\sigma_1{}^2 + a_2{}^2\sigma_2{}^2 + \cdots + a_n{}^2\sigma_n{}^2 \end{aligned}\right\} \qquad (**)$$

In particular, if we take $a_1 = a_2 = \cdots a_n = 1$ in (**), then we get " $X_1 + X_2 + \cdots + X_n$ is a normal variate with mean $\mu_1 + \mu_2 + \cdots + \mu_n$ and variance $\sigma_1{}^2 + \sigma_2{}^2 + \cdots + \sigma_n{}^2$."

Thus, the sum of independent normal variates is also a normal variate. This is known as the Additive or 'Re-productive Property' of the Normal distribution.

12. Area Property. One of the most fundamental properties of the normal probability curve is the area property. The area under the normal probability curve between the ordinates at $X = \mu - \sigma$ and $X = \mu + a$ is 0.6826 . In other words, the range $\mu \pm \sigma$ covers 68.26% of the observations.

The area under the normal probability curve between the ordinates at $X = \mu - 2\sigma$ and $X = \mu + 2\sigma$ is 0.9544 , i.e., the range $\mu \pm 2\sigma$ covers more than 95% of the observations.

The area under the normal probability curve between the ordinates at $X = \mu - 3\sigma$ and $X = \mu + 3\sigma$ is 0.9973 , i.e., the range $\mu \pm 3\sigma$ covers 99.73% of the observations. Hence, for practical purposes, the range $\mu \pm 3\sigma$ covers the entire area, which is 1 [or all the observations].

It may also be noted that :

$\mu \pm 1.96\sigma$ covers 95% area, 47.5% on each side,
$\mu \pm 2.5758\sigma$ covers 99% area, 49.5% on each side,
$\mu \pm 0.6745\sigma$ covers 50% area, 25% on each side.

13. The limiting form of the binomial distribution as $n \to \infty$ for fixed p is the normal distribution. To be more precise, if x be the binomial variate, $\frac{x-np}{\sqrt{npq}}$ is a standard normal variate for large n.

- ## IMPORTANCE OF NORMAL DISTRIBUTION

It is said "...if a statistician could select one distribution to work with during his life time, he would almost surely select the normal distribution......Although modern statisticians and applied economists could perhaps get along without computer, it would be exceedingly difficult to do without normal distribution." The following points will further highlight the importance of normal distribution:

1. Most of the discrete probability distributions (e.g., Binomial distribution, Poisson distribution) tend to the normal distribution as n, the number of trials increases. For large values of n, computations of probabilities for discrete distributions becomes quite tedious and time-consuming. In such cases, normal approximation can be used with great ease and convenience.

2. Almost all the exact sampling distributions, e.g., t-distribution, F-distribution, Z-distribution and the chi-square distribution conform to normal distribution for large degrees of freedom (i.e., as $n \rightarrow \infty$).

3. The whole theory of exact sample (small sample) tests, viz., t, F, χ^2 tests, etc., is based on the fundamental assumption that the parent population from which the samples are drawn follows normal distribution.

4. Perhaps, one of the most important applications of the normal distribution is inherent in one of the most fundamental theorems in the theory of statistics, viz., the Central Limit Theorem which may be stated as follows :

"If $X_1, X_2, \ldots, X_n$ are n independent random variables following normal distribution, then under certain very general conditions, their sum $\mathbf{\Sigma X = X_1 + X_2 + \cdots + X_n}$ is asymptotically normally distributed, i.e., $\mathbf{\Sigma X}$ follows normal distribution as $n \rightarrow \infty$ ".

An immediate consequence of this theorem is the following result :

"If $X_1, X_2, \ldots, X_n$ is a random sample of size n from any population with mean μ and variance σ^2, then the sample mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\Sigma X,$$

is asymptotically normal (as $n \rightarrow \infty$) with mean μ and variance σ^2/n''.

5. It is extensively used in large sampling theory, to find estimates of parameters from statistics and confidence limits, etc.

6. Many of the distributions which are not normal can be made normal by simple transformation.

Theory of normal curves can be applied to the graduation of the curves which are not normal.

7. It finds many applications in statistical quality control and industrial experiments for setting control limits.

8. It has numerous mathematical properties which make it popular and comparatively easy to manipulate for use in social and natural sciences. The following quotation, due to W.J. Youlden, reveals the popularity and importance of normal distribution.

THE
NORMAL
LAW OF ERROR
STANDS OUT IN
THE EXPERIENCE OF
MANKIND AS ONE OF THE
BROADEST GENERALISATION OF
NATURAL PHILOSOPHY. IT SERVES AS THE
GUIDING INSTRUMENT TO RESEARCHES IN THE
PHYSICAL AND SOCIAL SCIENCES AND IN MEDICINE,
AGRICULTURE AND ENGINEERING. IT IS AN INDISPENSABLE
TOOL FOR THE ANALYSIS AND THE INTERPRETATION OF
THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Artistically, it gives also the shape of a normal curve.

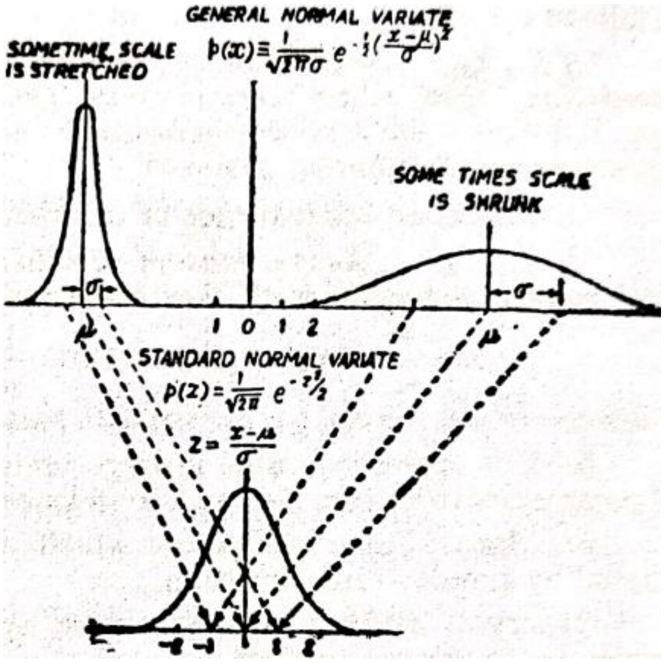
9. The importance of normal distribution is further revealed in the following quotation due to Lipman.

"Everybody believes in the law of errors (the normal curve), the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

14.5 *STANDARD NORMAL DISTRIBUTION*

It was explained earlier that a normal distribution is completely specified by its mean and standard deviation. There will then be different bell-shaped curves with different means and standard deviations. But it is possible to convert these different normal distributions into one standardized form as shown in the right.

Once this is done, the areas under normal distribution represent probabilities and can be obtained from the use of tables prepared with reference to such standardized normal distribution.



We know

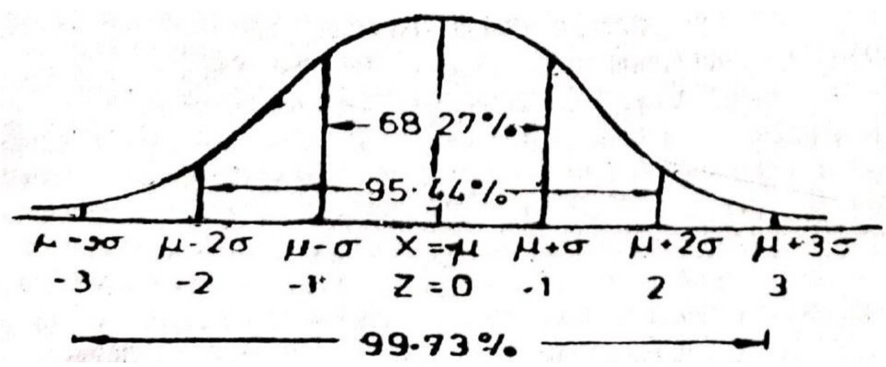
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2}$$

Now, *Standard Normal Distribution* or Z-Distribution is

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \mathrm{Z} = \frac{X-\mu}{\sigma}$$

The standard normal distribution has mean 0 and standard deviation 1. The 99.73% area under the curve is covered by $\pm 3\sigma$.

This has been depicted in the following diagram :



The normal distribution with mean 0 and variance 1 may be denoted by the symbol $N(0,1)$. Since $p(z)$ is a probability distribution,

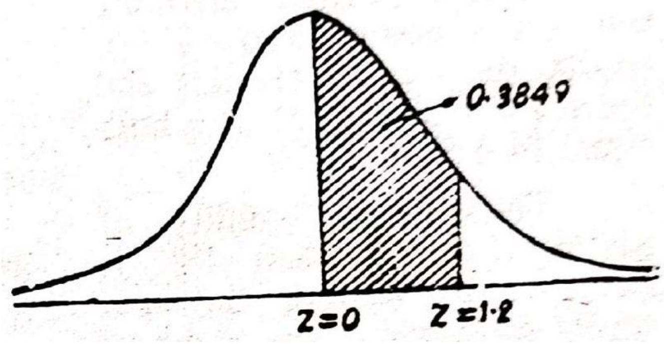
the total area under the curve is 1.

The probability that a standard normal variable (S.N.V.) lies between two specified values c and d is

$P(c < Z \leq d) =$ Area under the curve between two specified values $Z = c$ and $Z = d$.

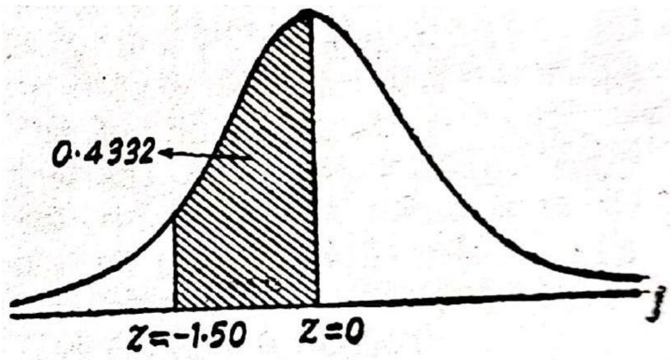
The table of areas (or probabilities) under the standard normal distribution is given in the Appendix at the end of the book.

Example 1. Find the probability that the S.N.V. lies between 0 and 1.2 .



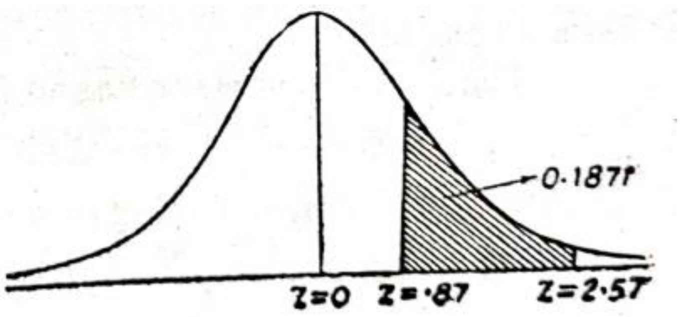
From the tables of standard normal distribution, the area between $Z=0$ and $Z=1.2$ is 0.3849 which is the required probability.

Example 2. Find the probability that the standard normal variate lies between $-1\cdot 50$ to 0 .



By virtue of the symmetry of the standard normal distribution about its mean, namely $Z=0$, to find the area between $Z=-1.50$ and $Z=0$, we simply look up the area between $Z=0$ and $Z=1.50$. From the table this is found to be 0.4332 , which is the required probability.

Example 3. Find the probability that the S.N.V. lies between 0.87 and $2:57$. The required probability is the area between the ordinates at 0.87 and 2.57 . This area cannot be looked up directly, but we can look up the area under the curve between $Z=0$ and $Z=2\cdot 57$, the area between $Z=0$ and $Z=0.87$ and then take the difference between the two. Thus, the required probability is

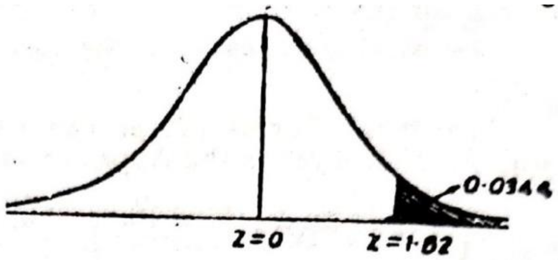


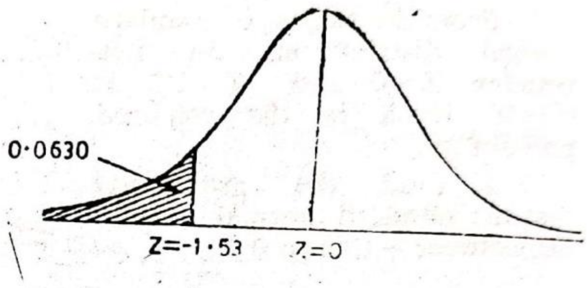
$$0.4949-0.3078=0.1871$$

Example 4. Find the probability that a S.N.V. is

(i) greater than 1.82 , (ii) less than -1.53 , (iii) greater than -2.47 .

Questions concerning area under normal distribution arise in various ways, and the ability to find any desired area quickly can be a big help. Although the table gives only area between the mean $.Z = 0$ and selected positive values of Z, we often have to find areas to the left or to the right of given positive or negative values of Z, or areas between two given values of Z.
We make use of the fact that the standard normal distribution is symmetrical about $Z = 0$, so that the area to the left and the area to the right are both equal to 0.5000

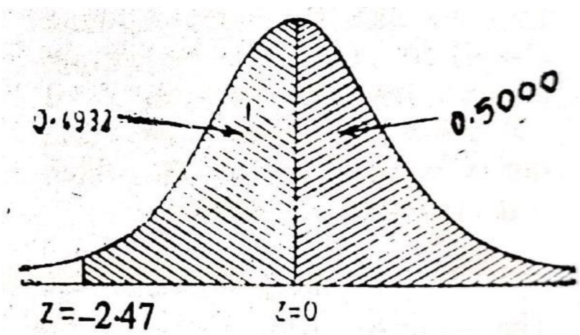




Thus, the probability of getting Z greater than $1 \cdot 82$, is obtained by finding the area under the curve to the right of $Z = 1.82$ which is $0.5000 - 0.4656 = 0.0344$. Probability of getting Z less than -1.53 is obtained by finding the area under the curve to the left which is $0.5000 - 0.4370 = 0.0630$.

Similarly, the probability of getting a Z greaterthan -2.47 (namely the area under the curve to the right of $Z = -2{:}47$ shaded in the diagram) is

$$0.5000 + 0.4932 = 0.9932$$



Example 5. *Given a normal distribution with $\mu = 50$ and $\sigma = 8$, find the probability that X assumes a value between 34 and 62.*

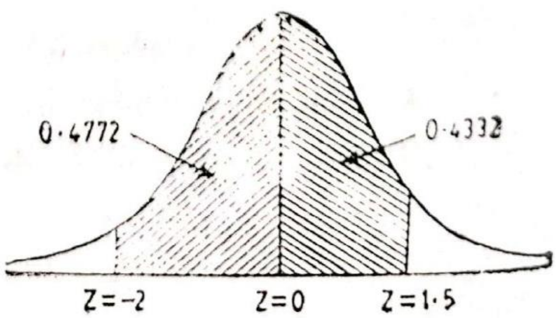
The S.N.V. corresponding to $X = 34$ is

$$Z_1 = \frac{34 - 50}{8} = -2$$

and the S.N.V. corresponding to $X = 62$

$$Z_2 = \frac{62-50}{8} = 1.5$$

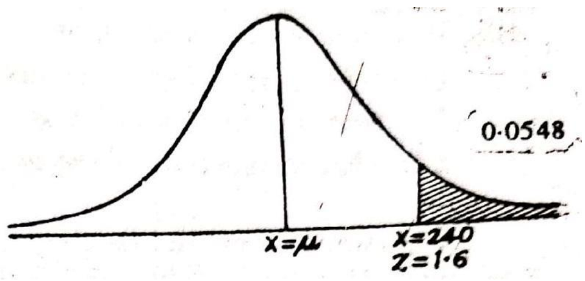
$$\therefore \quad P(34 < X < 62) = P(-2 < Z < 1.5)$$



The $P(-2 < Z < 1.5)$ is given by the area of the shaded region is the adjoining figure. This area may be found out by adding the area between $Z = -2$ and $Z = 0$ as well as the area between $Z = 0$ and $\boldsymbol{Z} = 1.5$

$$\therefore P(34 < X < 62) = P(-2 < Z < 1.5)$$
$$= P(-2 < Z < 0) + P(0 < Z < 1.5)$$
$$= 0.4772 + 0.4332 = 0.9105$$

Example 6. Given a normal distribution with $\mu = 200$ and $\sigma = 25$, find the probability that X assumes a value greater than 240.



The normal probability distribution showing the desired area is given in the adjoining figure. To find the $P(X > 240)$, we need to evaluate the area under the normal curve to the right of $X = 240$. This can be done by transforming $X = 240$ to the corresponding Z value, obtaining the area between $Z = 0$ and $Z = 1.6$ from tables, and then subtracting this area from $0 \cdot 5$. We find

$$Z = \frac{240 - 200}{25} = 1.6$$

Hence

$$P(X > 240) = PZ > 1.6)$$
$$= 0.5 - P(0 < Z < 1.6) = 0.5 - 0.4452 = 0.0548$$

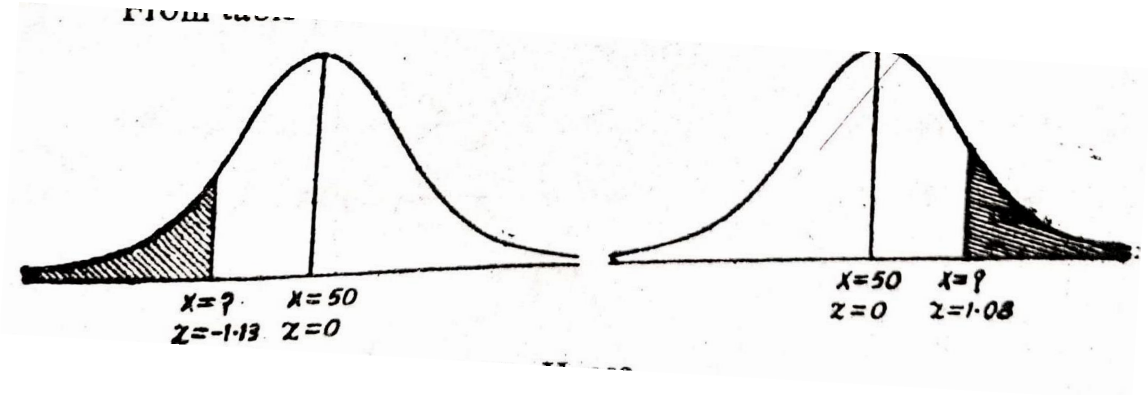
Example 7. Given a normal distribution with $\mu = 50$ and $\sigma = 10$, find the values of X that has (a) 13% of the area to its left and (b) 14% of the area to its right.

The preceding illustrations were solved by first going from a value of X to a Z value and then computing the desired area. In this illustration we reverse the process. Here we begin with a known area or probability, find Z value, and then determine X by rearranging the formula

$$Z = \frac{X - \mu}{\sigma} \text{ to give } X = \sigma Z + \mu$$

(a) An area of 0.13, to the left of the desired X value is shaded in the following figure. We required a Z value that leaves an area of 0.13

From the table we find $P(Z < -1.13) = 0.13$ to the left.



so that desired Z value is -1.13 . Hence

$$X = 10(-1 \cdot 13) + 50 = 38 \cdot 7$$

(b) In this case we require a Z value that leaves 0.14 of the area. to the right and hence an area of 0.36 is between $Z = 0$ and $Z = 1.08$. Sor. the desired Z value is 1.08 and

$$X = 10(1.08) + 50 = 60.8$$

Example 8. The customer accounts of a certain departmental store thave an average balance of Rs. 120 and a standard deviation of Rs. 40. Assuming that the account balances are normally distributed: -
(i) What proportion of the accounts is over Rs. 150?
(ii) What proportion of the accounts are between Rs. 100 and Rs. 150?
(iii) What proportion of the accounts are between Rs. 60 and Rs. 90?

Solution. Let the random variable X denote the balance of the customer accounts. Then we are given that X is normally distributed with mean $\mu = 120$ and s.d. $\sigma = 40$.

The standard normal variable Z is given by

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 120}{40} \quad \ldots (*)$$

(i) Probability that the customer accounts is over Rs. 150 is given by

$$P(X > 150).$$

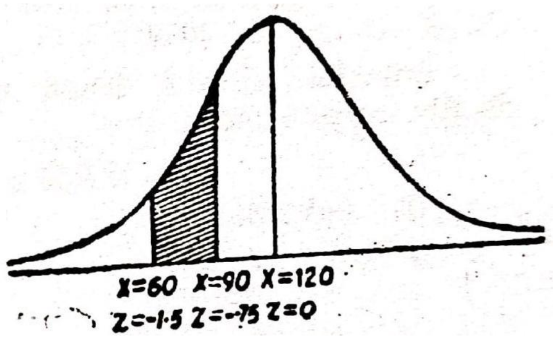
When $X = 150, Z = \frac{150-120}{4} = 0.75$
$\therefore\ P(X > 150) = P(Z > 0.75)$
= Area to the right of

$$Z = 0.75$$

$$= 0.5000 - 0.2734$$

$= 0.2266$

Hence, 22.66% of the accounts have a balance in excess of Rs. 150.

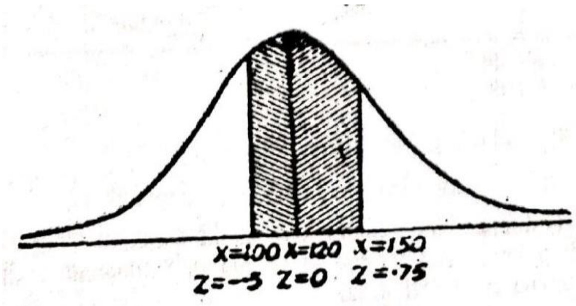


(ii) Probability that the accounts lie between Rs. 100 and Rs. 150 as given by

$$P(100 \leqslant X \leqslant 150)$$

When $X = 100$, $Z = \frac{100-120}{40} = -0.5$

When $X = 150$ $Z = \frac{150-120}{40} = 0.75$



$\therefore\ P(100 \leq X \leq 150) = P(-0.5 \leq Z \leq 0.75)$
$= \text{Area between } Z = -0.5 \text{ and } Z = 0.75$
$= (\text{ Area between } Z = -0.5 \text{ and } Z = 0)$
$+(\text{ Area between } Z = 0 \text{ and } Z = 0.75)$
$=P(0 \leqslant Z \leqslant 0.5) + P(0 \leqslant Z \leqslant 0.75)$ (By symmetry)
$=0.1915 + 0.2734 = 0.4649.$

Hence, 46.49% of the accounts have an average balance between Rs. 100 and Rs. 150.

(iii) Here we want $P(60 \leqslant X \leqslant 90)$.

When $X = 60$, $Z = \frac{60-120}{40} = -1.5$

When $X = 90$, $Z = \frac{90-120}{40} = -0.75$

$\therefore P(60 \leq X \leq 90)$
$= P(-1.5 \leq Z \leq -0.75)$
$= P(0.75 < Z \leqslant 1.5)$ (By symmetry)

$= \text{Area between } Z = +0.75$
$\text{and } Z = +1 \cdot 5$

= (Area between $Z = 0$ and $Z = 1.5$) - (Area between $Z = 0$ and $Z = 0.75$) $= 0.4332 - 0.2734 = 0.1598$

Hence, 15.98 per cent of the accounts lie between Rs. 60 and Rs. 90.

CHECK YOUR PROGRESS

True or false Questions

Problem 1. The shape of the Normal Curve is
a) Bell Shaped
b) Flat
c) Circular
d) Spiked

Problem 2. Normal Distribution is symmetric is about
a) Variance
b) Mean
c) Standard deviation
d) Covariance

Problem 3. The area under a standard normal curve is?
a) 0
b) 1
c) ∞
d) not defined

Problem 4. The standard normal curve is symmetric about the value
a) 0.5
b) 1
c) ∞
d) 0

Problem 5. Normal Distribution is also known as
a) Cauchy's Distribution
b) Laplacian Distribution
c) Gaussian Distribution
d) Lagrangian Distribution

14.6 *SUMMARY*

(i) A continuous random variable X is said to be normally distributed if it has the probability density function represented by the equation:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty \leqslant x \leqslant \infty$$

(ii) The ordinate at the mean of the distribution divides the total area under the normal curve into two equal parts. Further, since the total area under the normal probability curve is 1, the area to the right of the ordinate as well as to the left of the ordinate is 0.5.

14.7 *GLOSSARY*

(i) Permutation

(ii) Combinations

(iii) Probability

14.8 *REFERENCES AND SUGGESTED READINGS*

1. S. C. Gupta and V. K. Kapoor, (2020), Fundamentals of mathematical statistics, Sultan Chand & Sons.
2. Seymour Lipschutz and John J. Schiller, (2017), Schaum's Outline: Introduction to Probability and Statistics, McGraw Hill Professional.
3. J. S. Milton and J. C. Arnold,(2003), Introduction to Probability and Statistics (4th Edition), Tata McGraw-Hill.
4. https://www.wikipedia.org.
5. A.M. Goon, (1998), Fundamental of Statistics (7th Edition), 1998.
6. R.V. Hogg and A.T. Craig, (2002), Introduction to Mathematical Statistics, MacMacMillan, 2002.
7. Jim Pitman, (1993), Probability, Springer-Verlag.
8. https://archive.nptel.ac.in/courses/111/105/111105090

14.9 *TEWRMINAL QUESTIONS*

1. What do you mean by Normal Distribution.
2. The weights of walnut cakes, W grams, are Normally distributed with a mean of 460 and a standard deviation of 10. Find the probability the weight of such cake will be between 465 and 475 grams.
3. The lifetimes, T hours, of a certain brand of battery are assumed to be Normally distributed with a mean of 825 and a variance of 900. Find the probability that one such battery will last between 801 and 870 hours.

14.10 *ANSWERS*

CYQ 1. a

CYQ 2. b

CYQ 3. b

CYQ 4. d

CYQ 5. c

Teen Pani Bypass Road, Transport Nagar
Uttarakhand Open University,
Haldwani, Nainital-263139
Phone No. 05946-261122, 261123
Toll free No. 18001804025
Fax No. 05946-264232,

E-mail:info@uou.ac.in
Website: https://www.uou.ac.in/