Creating By-
rh_sohan

STAT-1211: Statistics & Probability

Department of Computer Science and Engineering

BSc Honors 1st Year 2nd Semester Exam-2023

Exam Session: 2022-23

Admission Session: 2022-23



Lecture Note 01

Md. Ayub Ali
Assistant Professor
Department of Statistics
University of Barishal, Barishal-8254
maali@bu.ac.bd

Scanned with

Syllabus and Suggested Materials

Department of Computer Science Engineering

2nd Semester

Course Descriptions

Course Title: Statistics and probability

Course Code: STAT-1211

STAT-1211: Statistics and Probability [3.0 Credits, 45 Hours Lectures]

(i) Statistics – definition and scope: definitions of statistics – past and present, its nature and charateristics, population and sample, descriptive and inferential statistics, scope and applications of statistics, abuse of statistics, sources of statistical data, primary and secondary sources. Data collection tools, types, etc. Construction of questionnaire and other field problems of data collection. Types of data, cross sectional, longitudinal, follow-up and panel data.

(ii) Presentation of data: graphical presentation of data, details of different types of graphs and charts with their relative merits and dements, concept of explorative data analysis, stem-and-leaf plot, schematic plots, extremes and median, hinges, outlieres and 5 number summaries, measures of location, dispersion, skewness, kurtosis and their properties, moments box-and-whiskers plots, trimean, trimmed mean, interpretation of dat with these measures.

(iii) Correlation analysis: bivariate data, scatter diagram, simple correlation, rank correlation, correlation ratio, multiple and partial correlations, intraclass and biserial correlation.

(iv) Regression analysis: basic concept of regression, regression model, estimation of parameters (OLS method) in regression model, properties of estimators, interpreting the constants, some ideas of polynomial regression, 3-variable regression, estimation of parameters, standard error and other properties.

(v) Elements of probability: experiment, random experiment, sample spacem events, event space, union and intersection of events, different types of events. (vi) Basic concepts of probability: different approaches of defining probability- classical, axiomatic, empirical and subjective, laws and theorems of probability, conditional probability, Bayes' theorem and its importance in statistics. (vii) Random variable and its probability distribution: discrete and continuous random variables, probability mass function, probability density function, distribution function, function of random variable and its distribution, joint distribution, marginal and conditional distributions, independence of random variables. (viii) Mathematical expectation: concept, expectations of sums and products of random variables, conditional expectation and conditional variance, moments and moment generating functions, cumulants and cumulant generating functions, relation between moments and cumulants, probability generating functions, characteristic function. (ix) Some basic distributions: detailed study of bionomial, Poisson, normai, uniform, geometric, neigative binomial, hypergeometric, exponential, gamma, beta distributions

Scanned with

Statistics

❖ **Statistics and its chief characteristics.**

Answer:

Statistics is a branch of scientific knowledge refers to the body of techniques and methodology developed for the collection, classification, organization, presentation and analysis of statistical data and for the use of such data in decision-making in the face of uncertainty in any field of enquiry.

or

Statistics in the science of data collection, analysis, interpretation, presentation and organization. It is a branch of mathematics that deals with all aspects of data, from its collection to its use in decision-making.

or

According to Professor R.A. Fisher, ***"The science of statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data".***

Main characteristics of statistics:

1. **Data collection:** The first step in any statistical analysis is to collect data. This can be done through a variety of methods such as surveys, experiments and observational studies.
2. **Data organization:** Data can be organized in a variety of ways such as databases and spreadsheets. The best way to organize data will depend on the purpose of data collection and analysis.
3. **Data presentation:** Data can be presented in a variety of ways such as tables, graphs and charts. The best way to present data will depend on the audience and the purpose of the presentation.
4. **Data analysis:** Once data has been collected it must be analyzed to identify patterns and trends. This involves using a variety of statistical techniques such as descriptive statistics and inferential statistics.
5. **Data interpretation:** Once data has been analyzed it must be interpreted to draw meaningful conclusions. This involves understanding the limitation of the data and context which it was collected.

❖ **Two basic types of statistics and its description with examples.**

Answer:

The two basic types of statistics are descriptive and inferential statistics.

1. **Descriptive statistics is a branch of statistics that summarize the characteristics of a dataset. It is used to describe the center, spread, and shape of the data. Descriptive statistics can be calculated using a variety of measures such as:**

***Mean:** The average of a set of numbers.

***Median:** The middle number in a set of number when the numbers are arranged.

***Mode:** The most frequent number in a set of number.

***Range:** The difference of how spread out a set of numbers is from the mean.

***Standard deviation:** The square root of variance.

2. **Inferential statistics is a branch of statistics that uses sample data to draw conclusions of a population. It is uses a variety of techniques to measure inferential statistics.**

***Hypothesis testing:** A method if determining whether there is enough evidence to reject a null hypothesis.

***Confidence intervals:** A range of values that is likely to contain the true population parameter.
***Regression analysis:** A method of determining the relationship between two or more variables.

❖ **Functions and limitations of statistics.**

Answer:

Functions of statistics:

1. **Data description and summarization:** Statistics can provides techniques to summarize and describe the characteristics of data set including central tendency, measure of dispersion and inference about a population based on sample data.
2. **Hypothesis testing and inference:** Statistics allows us to draw inference about a population based on sample data.
3. **Estimation and precision:** If enables us to estimate population parameters such as the mean or proportion, based on sample data.
4. **Relationship analysis:** Statistics provides methods to identify and quantify relationships between variables.
5. **Decision making and risk assessment:** It provides insights into the probabilities and uncertainties associated with various options and scenarios.
6. **Quality control and improvement:** Statistics is widely used in quality control and improvement processes.

Limitations of statistics:

1. **Data quality and representativeness:** Statistical results are heavily depends on the quality and representativeness of the data. Biased or inaccurate data can lead to misleading conclusion.
2. **Assumption violations and robustness:** Statistical method often rely on certain assumption. Violation of these assumptions can affect the validity and reliability if the statistical inferences.
3. **Uncertainty and probability:** Statistical inferences are probabilistic in nature and there is always some degree of uncertainty associated with them.
4. **Misuse and manipulation:** Statistics can be misused or manipulates to support biased conclusions or agendas.
5. **Context and interpretation:** Misinterpretations can arise from ignoring relevant contextual factors or misapplying statistical concepts.

❖ **The importance/uses of statistics.**

Answer:

HW

- **"Statistics is liable to be misused"-explain.**

Answer:

First, statistics can be used to miss-present data by selectively regulations only the information that supports a particular conclusion while ignoring or downplaying contradictory evidence. This can be done through various techniques like using misleading graph and charts, cherry-picking data points or misinterpreting statistical results.

Second, statistics can be used to draw false conclusions due to a lack of understanding of statistical concept or due to intentional manipulation. For example, someone might claim that a correlation which is a logical fallacy.

Third, it can be used to support biased or agenda-driven claims. This is often seen in political debates or marketing campaigns where statistics are used to persuade people to support a particular viewpoint or purchase a particular product, regardless of the validity if the underlying data.

- **"Statistical results are not accurate"-explain.**

Answer:

The statement "statistics are aggregates of facts" is a concise and accurate description of the nature of statistics. It highlights two keys

Aspects of statistics:

1. Aggregation: Statistics involves collecting and combining individual fact into a larger body of data. This aggregation allows us to see patterns and trends that would not be visible if we only examine individual facts in isolation.

2. Quantification: Statistics is concerned with numerical representations of facts. This mean that we quantify facts using measures such as averages, percentages, standard deviation which allows us to make comparisons and draw inferences about the data.

By aggregating and quantifying facts, statistics provides a powerful fool for understanding and interpreting data. It allows us to summarize complex data sets, identify relationships between variables and make predictions about future events.

- **"Statistics is the main tool for the decision making in CSE"-explain.**

Answer:

Statistics helps a student in decision making in computer science and engineering (CSE) due to its ability to analyze and interpret data effectively. Here is some example how statistics is relevant in CSE for decision making:

1. **Data Analysis:** In CSE, vast amount of data is generated through various processes, including software development, system, performance monitoring and network activities. Statistics helps to analyze and interpret this data.
2. **Software testing and quality assurance:** Statistics is used in software testing and quality assurance to access the reliability and of performance for software applications.
3. **Performance evaluation:** Through statistical analysis, engineers can identify bottlenecks , optimize performance and make decisions

4. **Resource Allocation and optimization:** In the design and management of computer system, statistics helps in optimizing resources allocation.
5. **User behavior Analysis:** Statistics facilitates the analysis if user interactions, preferences and feedback to inform decisions related to interface design and system improvements.
6. **Project management:** Statistics aids in project management by providing tools for estimating project timelines, monitoring progress and identifying areas that may require adjustments or additional resources.

Scanned with

Some Important Concepts

❖ **Some key elements like, population, sample, parameter, statistic, population size and sample size.**

Answer:

Population: Population consists of all subjects which are being studied.

Sample: In statistics, a sample is a subset of a larger population. The purpose of taking a sample is to make inferences about the population, based on the characteristics of the sample. Sample is the representative part of a population.

Parameter: Parameter is the measure, quantity or numerical value obtained from population.

Statistic: Statistic is the measure, quantity, or numerical value obtained from population.

Population size: Population size refers to the total number of individuals, objects, events, or other units in a particular group or population. Generally, denoted by N.

Sample size: Sample size refers to the number of participants or observations included in a study. Generally, denoted by n.

❖ **Variable and different types of variables.**

Answer:

Variable is a changeable characteristic of the experimental units under consideration. Actually, it is the characteristic of experiment units which varies from experimental unit to experimental unit.

For example,

1. *Age of a worker*
2. *Religion of a student*
3. *Wage of a worker*
4. *Gross profits of a company*
5. *Height of a student etc.*

❖ **Various kinds of variables with examples.**

Answer:

➢ **Kinds of Variables: According to whether a variable takes numerical or non-numerical values.**

- **Qualitative variable:** A variable that cannot assume a numerical value but can be classified into two or more non-numerical categories is called a qualitative or categorical variable. Or a variable is called qualitative when it measures a qualitative characteristic on each experimental unit.

Examples: **Religion, Gender, Economic status, Teaching performance, Color** etc.

- **Quantitative variable:** A variable that can be measured numerically is called a quantitative variable. Or a variable is called quantitative variable when it measures a numerical quantity or amount on each experimental unit.

Examples: **Number of children, Price, Production in kg, Daily rainfall, Systolic BP etc.**

➢ **Kinds of Variables:** According to whether a variable is countable or measurable.

- Discrete Variable: A variable, which can take, only isolated or countable finite or infinite number of values is called a discrete variable.

***Examples*: Size of shoes, Size of nails, Coins etc.**

- **Continuous Variable:** A variable that can take infinitely many values over a certain interval or intervals is called a continues variable.

***Examples*: Age, Systolic blood pressure, Weight, Height etc.**

❖ **Distinguish between variable and constant.**

Answer:

HW

❖ **Change of origin and scale with an example.**

Answer:

- **Change of Origin and Scale:**
  - **Change of Origin: Shifting all values in a variable by a constant amount.**

Example: If temperatures in Celsius are converted to Fahrenheit (adding 32), it's a change of origin.

- **Change of Scale: Multiplying all values in a variable by a constant factor.**

Example: Converting lengths from meters to centimeters (multiplying by 100) is a change of scale.

❖ **The necessity of shifting origin and scale.**

Answer:

- *Simplify analysis or interpretation.*
- *Standardize units for comparison.*
- *Enhance visualization.*

❖ **The measurement of scale and different measurement scales with examples**

Answer:

The measurement of scale refers to the unit of measurement used for a variable.

Examples: Inches, centimeters, kilograms, degrees Celsius.

➢ **Measurement Scales:**

- **Nominal Scale: The scale of measurement by which we can classify and identify a qualitative variable according to different categories is called nominal scale.**

Example: Gender, Religion, Marital status etc.

Scanned with

- **Ordinal Scale: The scale of measurement by which we can classify, identify and rank a qualitative variable according to different categories is called ordinal scale.**

Example: Grading of a student, Size of a worker, Rating of an executive etc.

- **Interval Scale: The scale of measurement by which we can measure a qualitative variable numerically on experimental unit with arbitrary zero as origin is called interval scale.**

Quantitative data with equal intervals, but no true zero point.

Example: Body temperature, Marks obtained by students, Calendar time etc.

- **Ratio Scale: The scale of measurement is called ratio scale when a quantitative variable is measured numerically on experimental unit with absolute zero as origin.**

Quantitative data with equal intervals and a true zero point.

Example: Age, Weight, Height, Number of children etc.

❖ **Identify the types of variables and their scales of measurement on which the following variables are measured:**

I. **Religion of an employee of a factory (Muslim, Hindu, Buddhist, Christen etc.), Official status of an employee, Monthly salary of an employee, Number of family members of an employee, Age of an employee, IQ score of an employee, Name of the students of BBA and Smoking status of an employee.**

Answer:

Characteristics	Variable Type1	Variable Type2	Measurement scale
Religion	Qualitative	Discrete	Nominal
Official status of an employee	Qualitative	Discrete	Ordinal
Monthly salary of an employee	Quantitative	Continuous	Ratio
Number of family members of an employee	Quantitative	Discrete	Ratio
Age of an employee	Quantitative	Continuous	Ratio
IQ score of an employee	Quantitative	Discrete/Continuous	Interval
Name of the students of BBA	Qualitative	Discrete	Nominal
Smoking status of an employee	Qualitative	Discrete	Nominal/Ordinal

❖ **Practicing Ingredients**

II. Classify each of the followings as different measurement of scales: Sex, Marital status, Religious affiliation, Class room number, Political affiliation, Level of education, Opinion, Satisfaction score, Mental health status, Disease risk, Telephone number and shoe size.

III. State whether each of the followings are qualitative or quantitative: Prime interest rate, Bacteria count in drinking water, Percentage of arsenic in tubewell water, Divorce rate, Investment in business, Bank interest rate, Response to a four-point Likert scale, Car number and Prevalence.

IV. A building contractor has a chance to buy an old lot of 5000 used bricks at an auction. She is interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for her current

project, but she does not have enough time to inspect all. Instead, she checks 100 bricks to determine whether each is cracked. Find population, sample, experimental unit, measurable variable and nature of the variable.

V. Suppose a medical researcher wants to find the average systolic blood pressure of the employees of a big firm. For this purpose, a sample of 50 employees has been selected randomly from that firm. Indicate population, sample, experimental unit, measurable variables and nature of the variable.

VI. University authority needs the following information about the students before making decision whether the students are eligible for any financial aid. Classify each of the statement as qualitative or quantitative and corresponding scale of measurements: GPA of HSC, GPA of SSC, Sex of applicant, Parents' monthly income and Age of applicant.

Scanned with

Methods of Data Collection and All

❖ **Data and dataset.**

Answer:

Data: Data refers to individual pieces of information such as numbers, text, or symbols. It represents the raw material from which insights and knowledge are derived. Data can take various forms, including quantitative (measurable values like temperature), qualitative (descriptive characteristics like color), and categorical (grouped into specific categories).

Dataset: A dataset is a structured collection of data organized into rows and columns. Each row represents an individual observation, while each column represents a specific variable or attribute. Datasets are formatted for analysis and can vary in size and complexity. They serve as the foundation for tasks such as statistical analysis, machine learning, and research, providing a coherent framework for extracting meaningful insights.

❖ **Primary data and secondary data.**

Answer:

Primary Data: Primary data refers to information collected directly from original sources for a specific purpose or research study. This data is firsthand and is not previously gathered or analyzed. Methods of collecting primary data include surveys, interviews, observations, experiments, and direct interactions with individuals or entities relevant to the research. Primary data is often considered more accurate and tailored to the specific research objectives, but it can be more time-consuming and costly to collect compared to secondary data.

Secondary Data: Secondary data, on the other hand, refers to information that has already been collected and analyzed by someone else for a purpose other than the current research. It is pre-existing data that is repurposed for a different study. Sources of secondary data include books, articles, government reports, databases, and other published materials. While secondary data is generally quicker and more cost-effective to obtain, researchers must be cautious about its relevance, accuracy, and how well it aligns with the specific needs of their study.

❖ **Discuss various methods of collecting primary data and secondary data with its relative advantages and disadvantages.**

Answer:

Methods of Collecting Primary Data:

- **Surveys**
- **Interviews**
- **Observations**
- **Experiments**

Methods of Collecting Secondary Data:

- **Literature Review:**
- **Government Reports:**
- **Published Research Papers:**
- **Databases and Online Sources:**

❖ **Questionnaire and schedule for data collection.**

Answer:

Questionnaire: A questionnaire is a structured set of questions designed to collect information from respondents. It serves as a standardized instrument for gathering data, commonly used in surveys and quantitative research. Questionnaires can be administered through various means, including interviews, mailed forms, or online surveys. They offer efficiency in collecting data from a large sample size, but their reliability can be influenced by the wording of questions and the potential for respondent bias.

Schedule for Data Collection: A schedule for data collection is a plan outlining the timeline and sequence of activities for gathering research data. It includes details such as when, how, and for how long data will be collected. This organized plan ensures systematic and timely data collection, helping researchers coordinate activities and allocate resources effectively. While a schedule provides a roadmap, it needs periodic review and adjustment to accommodate unforeseen challenges, and external factors like participant availability can impact the planned timeline.

❖ **Discuss how you can select an appropriate method for collecting data.**

Answer:

- **Define Research Objectives:** Clearly articulate the research goals and objectives. Understand the type of information needed, the research questions to be answered, and the level of detail required. This clarity guides the selection of a data collection method that best suits the study.
- **Understand the Data Type:** Determine the nature of the data—whether it is qualitative or quantitative, categorical or numerical. Different data types may require different methods of collection. For example, surveys and questionnaires are often used for quantitative data, while interviews and observations may be more suitable for qualitative data.
- **Consider Resource Constraints:** Assess the available resources, including time, budget, and personnel. Some data collection methods may be more time-consuming or expensive than others. Choose a method that fits within the constraints of the research project.
- **Evaluate Feasibility:** Evaluate the practicality and feasibility of each data collection method. Consider the accessibility of the target population, the ease of implementation, and the potential for obtaining high-quality data. Practical considerations such as the availability of technology or the need for specialized skills should also be taken into account.
- **Assess Participant Comfort:** Consider the comfort level of participants with the chosen data collection method. Some individuals may prefer anonymity in surveys, while others

may be more comfortable expressing themselves in interviews. Ensuring participant comfort enhances the quality and reliability of the data.

- **Balance Reliability and Validity:** Strive for a balance between the reliability and validity of the data. Reliability ensures consistency and reliability, while validity ensures that the data accurately measures what it intends to. Different methods may have varying levels of reliability and validity, so choose one that best suits the research goals.
- **Review Ethical Considerations:** Review and address ethical considerations associated with each data collection method. Ensure that the chosen method respects participant rights, maintains confidentiality, and complies with ethical standards in research.
- **Pilot Testing:** Before full-scale implementation, conduct pilot tests of the chosen data collection method. This helps identify potential issues, refine questions or procedures, and ensure that the method is effective in capturing the intended data.

❖ **A local newspaper wants to ascertain the extent of satisfaction of the consumers of different brands of beverage of a company. Design a questionnaire for this purpose.**

Answer:

Consumer Satisfaction Survey: Beverages

Thank you for participating in our survey! Your feedback is valuable to us as we strive to enhance your experience with our beverage products.

Demographic Information:

- **Gender:**
  Male Female Other (please specify)
- **Age Group:**
  18-24 25-34 35-44 45-54 55 and above
- **Which of the following beverages do you regularly consume? (Select all that apply)**
  Brand A Brand B Brand C

❖ **Other (please specify)**

Product Satisfaction:

- **How satisfied are you with the overall quality of Brand A's beverages?**
  Very Satisfied Satisfied Neutral Dissatisfied Very Dissatisfied
- **Please rate your satisfaction with the taste of Brand B's beverages:**
  Excellent Good Average Poor Very Poor
- **How likely are you to recommend Brand C's beverages to a friend or family member?**
  Very Likely Likely Neutral Unlikely Very Unlikely
- **What factors influence your choice of beverage brand? (Select all that apply)**
  Taste Price Brand Reputation Packaging

❖ **Availability**

Brand Perception:

- **How would you describe the packaging of Brand A's beverages?**
  Attractive Average Unappealing
- **Do you associate any specific values or qualities with Brand B's beverages? (e.g., health-conscious, premium, eco-friendly)**
  Yes (please specify) No

Consumption Patterns:

- **How often do you purchase beverages from Brand C in a month?**
  Daily Weekly Monthly Rarely Never
- **On average, how much do you spend on beverages per month?**
  Less than $10 $10-$20 $20-$30 $30-$40 More than $40

Additional Feedback:

Is there anything specific you would like to highlight about your experience with any of the beverage brands surveyed? (Open-ended response).

Scanned with

Data Presentation

❖ **Data representation.**

Answer:

Data representation refers to the way data is presented and organized to make it understandable and interpretable. It involves choosing the appropriate methods, such as graphs, charts, tables, or other visual or numerical formats, to convey information effectively. The goal is to facilitate analysis and interpretation of the data, making it easier for individuals to draw insights and make informed decisions.

❖ **Necessity to represent statistical data?**

Answer:

Representing statistical data is crucial for several reasons:

Communication: It helps in conveying complex information in a clear and concise manner, making it easier for others to understand the patterns and trends present in the data.

Analysis: Visualization aids in the analysis of data by providing a visual context, enabling individuals to identify patterns, outliers, and relationships that may not be apparent in raw data.

Comparison: Representation allows for easy comparison between different data sets or variables, aiding in the identification of similarities, differences, and trends.

Decision-making: Well-presented data facilitates informed decision-making by providing a basis for understanding and evaluating various options or scenarios.

Storytelling: Visualization can turn data into a compelling narrative, making it more accessible and engaging for a broader audience.

In summary, representing statistical data enhances communication, analysis, comparison, decision-making, and the ability to tell a data-driven story.

❖ **Statistical methods for presenting raw data.**

Answer:

There are several statistical methods for presenting raw data. Here are some common ones:

Frequency Distribution: This method organizes raw data into intervals or classes and shows the number of observations falling into each interval.

Histograms: A graphical representation of the distribution of a dataset. It uses bars to represent the frequency of data within certain ranges or bins.

Stem-and-Leaf Plots: This is a visual representation of the distribution of a dataset, where each data point is split into a "stem" (the first digit or digits) and a "leaf" (the final digit).

Dot Plots: Simple plots where each data point is represented by a dot along a number line.

Box-and-Whisker Plots (Boxplots): These plots provide a visual summary of the distribution of a dataset, showing the median, quartiles, and potential outliers.

Scatterplots: Useful for displaying the relationship between two continuous variables. Each data point is represented by a dot on a two-dimensional graph.

Bar Charts: Useful for categorical data, where each category is represented by a bar. The height of the bar corresponds to the frequency or percentage of observations in that category.

Pie Charts: A circular statistical graphic divided into slices to illustrate numerical proportions. Each slice represents a proportionate part of the whole.

The choice of method depends on the nature of the data and the insights you want to convey.

- **What is meant by classification.**

Answer:

In statistics and data analysis, "classification" refers to the process of categorizing or organizing data into distinct groups or classes based on certain characteristics or criteria. The goal of classification is to assign each data point to a predefined category or class, making it easier to understand, analyze, and draw insights from the data.

Classification can be applied to both categorical and numerical data. In the context of machine learning, classification is a type of supervised learning where a model is trained on a labeled dataset to learn the patterns and relationships between input features and corresponding classes. Once trained, the model can then predict the class of new, unseen data based on its learned patterns.

In summary, classification is the act of assigning data into predefined categories or classes, and it plays a crucial role in various fields, including statistics, data analysis, and machine learning.

- **Describe the basics of classification and discuss its necessity.**

Answer: HW

- **Histogram and Bar diagram.**

Answer:

Histogram:

Definition: A histogram is a graphical representation of the distribution of a dataset. It displays the frequencies or relative frequencies of data points within specified intervals or bins.

Usage: Used to visualize the shape and spread of a continuous data distribution.

Appearance: Consists of bars, where the area of each bar corresponds to the frequency of data points in the corresponding interval. The bars are typically contiguous.

Bar Diagram (Bar Chart):

Definition: A bar diagram, or bar chart, is a graphical representation of categorical data. It uses rectangular bars of varying lengths to represent the values or frequencies of different categories.

Usage: Used for displaying and comparing the values of different categories or groups.

Appearance: Consists of bars that may or may not be contiguous, depending on the nature of the categorical data. The length of each bar represents the quantity or frequency associated with a particular category.

- **Dot plot.**

Scanned with

Answer:

A dot plot is a simple and straightforward graphical representation of individual data points along a number line. In a dot plot, each data point is represented by a dot. It's a useful tool for displaying the distribution and density of a dataset, especially when dealing with a relatively small amount of data.

Key features of a Dot plot:

Data Representation: Each data point is shown as a dot, and the position of the dot along the axis represents the value of the data point.

Density: The density of dots at a particular position indicates the frequency of data points at that value. When multiple dots overlap, it suggests higher density in that region.

Simplicity: Dot plots are simple and easy to create. They provide a quick visual summary of the distribution of the data.

Comparison: Dot plots are particularly effective for comparing the distribution of two or more groups or conditions.

Example:

Consider the data set: {2, 3, 3, 4, 4, 4, 5, 5, 6}

The dot plot would represent each of these values with dots along a number line. For instance, you might have a dot at 2, two dots at 3, three dots at 4, two dots at 5, and one dot at 6.

Dot plots are especially useful when you want to maintain the granularity of individual data points and provide a visual sense of their distribution.

❖ **Distinction between Pie chart and histogram plot.**

Answer:

Pie Chart	Histogram
Represents parts of a whole; shows the relationship of each part to the whole.	Displays the distribution and frequency of a continuous data set.
Typically used for categorical or nominal data.	Usually used for numerical and continuous data.
Circular chart divided into slices, each representing a category.	Bar chart with bars representing the frequency or count of data within predefined intervals (bins).
Does not have axes; categories are represented by slices.	Typically has two axes: x-axis for the variable being measured and y-axis for frequency or count.
Each slice is labeled with the corresponding category.	Bars may be labeled with the frequency or count within each interval.
Summarizes the whole data set by showing the percentage or proportion of each category.	Provides a visual representation of the distribution of the data, highlighting patterns and trends.

Scanned with

Suitable for comparing the proportions of different categories within a whole.	Suitable for comparing the frequency or count of data within different intervals.
Typically represents one variable.	Typically represents one variable.
Useful for displaying data with a clear overall pattern or for showing the composition of a whole.	Useful for visualizing the distribution and central tendency of numerical data.

❖ **How would you summarize categorical data by frequency distribution?**

Answer:

Summarizing categorical data using frequency distribution involves counting the number of observations (frequency) in each category. Here are the general steps to create a frequency distribution for categorical data:

List Categories: Identify and list all the distinct categories or groups present in your categorical data.
Tally Frequencies: Count the number of occurrences of each category. This is your frequency.
Organize Data: Arrange the categories in a table with two columns: one for the categories and another for their corresponding frequencies.
Calculate Relative Frequencies: Optionally, calculate relative frequencies, which represent the proportion of each category relative to the total number of observations. This is done by dividing the frequency of each category by the total number of observations and multiplying by 100 to get a percentage.
Present Results: Display the frequency distribution in a clear and organized manner. You can use a table or a bar chart to visually represent the frequencies.

❖ **Frequency and frequency distribution.**

Answer:

Frequency represents the count of occurrences of a specific value or category in a dataset. A frequency distribution is a tabular or graphical representation of these counts. Constructing a frequency distribution involves simplifying data, aiding analysis and interpretation, and facilitating comparison and communication.

❖ **Purpose of constructing a frequency distribution?**

Answer:

Purpose of Constructing a Frequency Distribution:

1. **Simplification:** It simplifies large datasets by condensing them into a manageable form.
2. **Analysis:** It facilitates the analysis of data distribution, allowing for the identification of patterns, central tendency, and variability.
3. **Interpretation:** It makes data more interpretable and aids in drawing conclusions about the dataset.
4. **Comparison:** It allows for easy comparison of different datasets or different aspects of the same dataset.
5. **Communication:** It provides a clear and organized way to communicate the distribution of data to others.

❖ **Steps of constructing a grouped frequency distribution.**

Answer:

Range Determination: Identify the range of the data, which is the difference between the maximum and minimum values. So, the range, $R = maximum - minimum$

Scanned with

Determine Number of Classes: Decide on the number of classes (intervals or bins) you want to use. Hence we can use $k = 1 + 3.32 \log_{10} n$ or $k = \sqrt{n}$, where n is the total number of observation.

Class Width Calculation: Calculate the class width by dividing the range by the number of classes. Hence, $c = R/k$.

Choosing Starting Point: Select a starting point for the first class within the range of the data.

Constructing Classes: Use the class width and starting point to construct non-overlapping intervals or classes covering the entire data range.

Tallying Frequencies: Count the number of data points falling into each class.

Constructing the Table: Create a table with columns for classes and their corresponding frequencies, and optionally include columns for relative and cumulative frequencies.

Visualization (Optional): Create a histogram or bar chart for a visual representation of the grouped frequency distribution.

- **Exclusive methods, inclusive methods, class intervals, equal class interval, unequal class interval, open-ended class interval, class midpoint, class boundaries, percentage frequency, frequency density, relative frequency and cumulative frequency distribution.**

Answer:

Exclusive method: when the class intervals are so fixed that the upper limit of one class is the lower limit of the next class. In this case, upper limit of each class is excluded from the count in that class, the observation equal to upper limit is counted in the next class interval.

For example, the audit time of 10 clients are: 11, 22, 14, 15, 16, 15, 20, 24, 18, 17

Audit time	Number of clients
10-15	2
15-20	5
20-25	3
Total	10

Inclusive method: Under the inclusive method of classification, the upper limit of one class is included in that class itself.

For example, the audit time of 10 clients are: 11, 22, 14, 15, 16, 15, 19, 24, 18, 17

Audit time	Number of clients
10-14	2
15-19	6
20-24	2
Total	10

Relative, percentage and cumulative frequency

For example, the audit time of 10 clients are: 11, 22, 14, 15, 16, 15, 19, 24, 18, 17

Audit time	Number of clients	Relative frequency	Percent frequency	Cumulative frequency
10-14	2	2/10	20	2
15-19	6	6/10	60	8

Scanned with

20-24	2	2/10	20	10
Total	10	1.0	100	

Example.

I. Construct a continuous frequency table (exclusive method or type) by calculating the suitable class interval from the following marks of 30 students. Also make a Histogram and Pie chart.

34	36	31	46	76	86	42	44	32	46
40	54	66	56	50	42	33	80	77	81
46	40	60	63	64	76	56	57	57	70

Answer:

Here, the maximum value=86 and minimum= 31

Range=86-31=55

According to the formula of H.G. Sturges

$$k = 1 + 3.32logN = 1 + 3.32\log(30) \cong 6$$

Now the class interval, c=Range/Number of classes=55/6=9.166

So let the class interval be 10. Then

Class	Tally marks	Frequency
30-40	~~IIII~~	5
40-50	~~IIII~~ III	8
50-60	~~IIII~~ I	6
60-70	IIII	4
70-80	IIII	4
80-90	III	3
		$N = 30$

Histogram

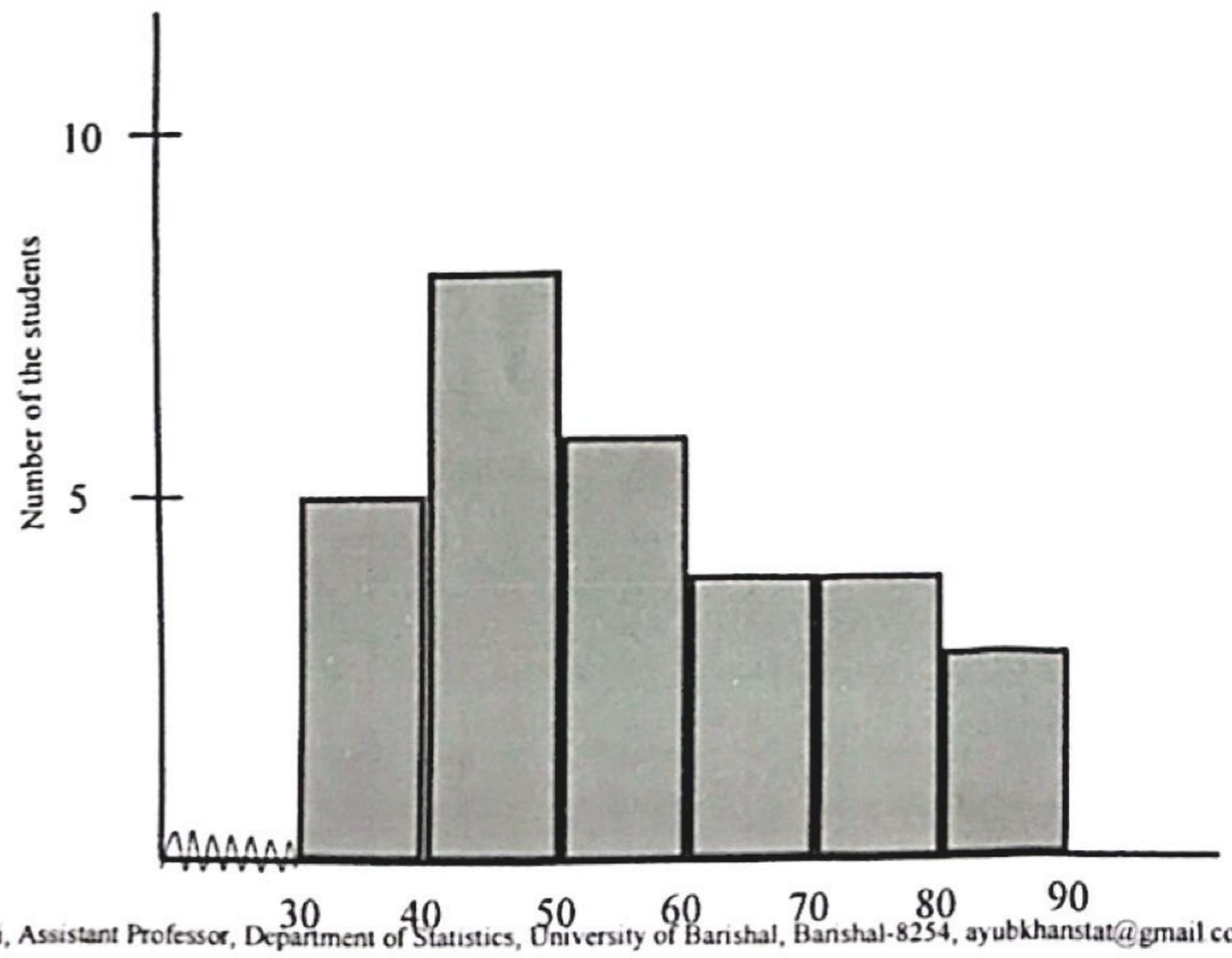


Figure 1: Histogram of the marks of 30 students

Pie chart

Class	Frequency	Angles
30-40	5	$\frac{5}{30} \times 360 = 60°$
40-50	8	$\frac{8}{30} \times 360 = 96°$
50-60	6	$\frac{6}{30} \times 360 = 72°$
60-70	4	$\frac{4}{30} \times 360 = 48°$
70-80	4	$\frac{4}{30} \times 360 = 48°$
80-90	3	$\frac{3}{30} \times 360 = 36°$
	$N = 30$	Total 360°

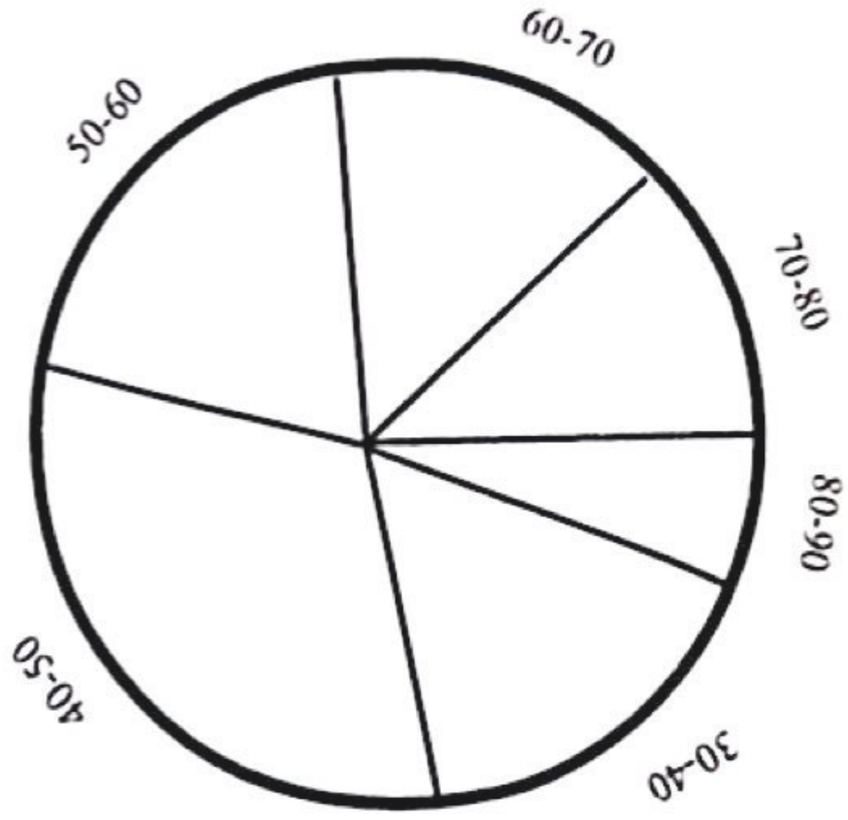


Figure 2: Pie chart of the marks of the students

Practicing Ingredients

HW

II. The following data represent the height of 50 students in centimeter in CSE. By calculating suitable class interval construct a frequency distribution and find the mid values, relative frequency, cumulative frequency, frequency density and percentage frequency of each class.

184.4	158.0	184.9	147.5	159.7	166.4	141.0	178.6	145.5	143.4
162.1	135.1	105.2	159.3	134.0	163.6	113.8	175.5	132.7	160.5
175.5	106.6	144.5	120.4	118.5	138.7	166.6	107.4	157.0	122.5
122.9	142.7	143.8	144.6	158.3	162.4	149.4	127.0	133.8	140.5
153.4	114.0	130.0	127.5	132.4	158.2	167.4	171.8	115.2	117.5

III. The following data represent the daily income (TK.) of 30 child workers. By taking class interval 5, construct a frequency distribution and draw more than Ogive and less than Ogive in the same graph. Also calculate median income from this graph. How much percentage of workers would have income more than TK. 50. Are these values same? Why or why not, explain.

65	80	55	60	70	72	85	50	52	60
57	87	75	68	65	56	78	80	86	62
70	60	50	56	74	66	62	88	80	75

IV. The data below represent the number of employees of an organization who were sick per day on 30 working days:

7	5	14	11	10	12	8	9	14	13
10	10	11	11	14	9	11	7	13	12
12	16	18	13	13	17	11	14	14	12

a) **Display the dataset with an stem-leaf display.**
b) **Construct a frequency distribution and draw histogram.**
c) **Plot Ogive curve and from that find out the number of sick days that 20 or more employees were sick. Compare the results obtained from Ogive with that from stem-leaf plot and comment.**

V. Draw a suitable diagram for the following data:

Expenditure	Food	Clothing	Education	Unforeseen	Savings or Deficit
Family A (Income in TK.)	500	150	125	25	10
Family B (Income in TK.)	300	150	60	50	-30

❖ **Five number summary to summarize numerical dataset with example with hypothetical dataset.**

Answer:

The five-number summary is a statistical summary of a dataset that provides a concise description of its distribution. It consists of five values: the minimum, the first quartile (Q1), the median (Q2), the third quartile (Q3), and the maximum. These values help to understand the spread, central tendency, and potential skewness of the data.

Here's how to calculate the five-number summary:
1. Minimum (Min): The smallest value in the dataset.
2. First Quartile (Q1): The median of the lower half of the dataset (excluding the overall median). It represents the point below which 25% of the data falls.
3. Median (Q2): The middle value of the dataset when it is ordered. It represents the point below which 50% of the data falls.
4. Third Quartile (Q3): The median of the upper half of the dataset (excluding the overall median). It represents the point below which 75% of the data falls.
5. Maximum (Max): The largest value in the dataset.

Example with a Hypothetical Dataset:

Consider the dataset: {10, 15, 18, 22, 25, 28, 30, 35, 40, 45, 50}

1. Minimum (Min): 10 (smallest value)
2. First Quartile (Q1): The median of the lower half {10, 15, 18, 22, 25} is 18.
3. Median (Q2): The middle value of the dataset {10, 15, 18, 22, 25, 28, 30, 35, 40, 45, 50} is 28.
4. Third Quartile (Q3): The median of the upper half {30, 35, 40, 45, 50} is 40.
5. Maximum (Max): 50 (largest value)

So, the five-number summary for this dataset is: 10, 18, 28, 40, 50.

Interpretation:

- The range (Max - Min) is 40.
- The interquartile range (IQR = Q3 - Q1) is 22, representing the middle 50% of the data.
- The median (Q2) is 28, indicating that 50% of the data falls below this value.
- Q1 and Q3 provide insights into the spread of the lower and upper halves of the dataset.

❖ **What do you mean by steam and leaf display? Explain with example with hypothetical dataset.**

Answer: HW

❖ **What do you mean by Box-whisker plot? Explain with example with hypothetical dataset. For the following dataset make Box and whisker plot to identify whether there is a outlier in the dataset: 16, 15, 18, 22, 25, 28, 30, 35, 36, 35, 50**

Answer: HW

❖ **What do you mean by Ogive curve? What is more than Ogive and less than Ogive? For the following dataset make more than Ogive and less than Ogive plot: 10, 15, 18, 22, 25, 28, 30, 35, 40, 45, 50**

Answer: HW

-The end-

STAT-1211: Statistics & Probability

Department of Computer Science and Engineering

BSc Honors 1st Year 2nd Semester Exam-2023

Exam Session: 2022-23

Admission Session: 2022-23



Lecture Note 02

Md. Ayub Ali
Assistant Professor
Department of Statistics
University of Barishal, Barishal-8254
maali@bu.ac.bd

Measures of Location, Dispersion, Skewness and Kurtosis

Measures of Location: Central Tendency and All

❖ **What do you mean by central tendency?**

Answer:

In a representative sample, the value of a series of data has always been a tendency to cluster around a certain point usually at the centre of the series is called central tendency. For example, we often talk of average income, average hourly product of a firm, average weight etc.

❖ **What do you mean by measures of central tendency?**

Answer:

Measures of central tendency are statistical values that provide insight into the center or average of a data set. The three main measures are:

1. **Mean (Average):** The sum of all values divided by the number of observations.
2. **Median:** The middle value in a sorted list of data. If there's an even number of observations, the median is the average of the two middle values.
3. **Mode:** The value that occurs most frequently in a data set.

❖ **Mention the principles or purposes of using measures of central tendency**

Answer:

Principle or Purposes of Using Measures of Central Tendency:

1. **Summary:** They provide a single representative value that summarizes the entire dataset.
2. **Comparison:** Facilitate comparisons between different datasets or subsets within a dataset.
3. **Decision-Making:** Aid in decision-making by highlighting the typical or central value around which data clusters.
4. **Interpretation**: Simplify complex data by offering a central reference point for interpretation.
5. **Estimation:** Serve as estimates for predicting future values or outcomes.
6. **Analysis:** Form a basis for further statistical analysis, such as inferential statistics.

❖ **What are the requisites of measures of central tendency?**

Answer:

Requisites of Measures of Central Tendency:

1. **Scale of Measurement:** The data should be at least on an interval scale for meaningful calculations.
2. **Representativeness:** The measure should truly represent the center of the data distribution.

3. **Robustness:** The measure should not be highly sensitive to extreme values (outliers).
4. **Applicability:** The choice of measure depends on the nature and distribution of the data.
5. **Objective:** The choice of measure should align with the specific objective or purpose of the analysis.
6. **Data Type:** Different measures may be appropriate for different types of data (e.g., nominal, ordinal, interval, ratio).

❖ **Define arithmetic mean (AM), geometric mean (GM), harmonic mean (HM), combined mean, median, and mode with examples.**

Answer:

1. **Arithmetic Mean (AM):**
   - **Definition:** The arithmetic mean, or simply the mean, is calculated by adding up all values in a dataset and dividing by the number of observations.
   - **Formula:** Arithmetic Mean, AM = $$\frac{\Sigma_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
   - **Example:** For the dataset [2, 5, 8, 12, 16], the mean is $\frac{2+5+8+12+16}{5}$ =8.6.
2. **Geometric Mean (GM):**
   - **Definition:** The geometric mean is the nth root of the product of n values in a dataset.
   - **Formula:** Geometric Mean, GM = $$\sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$
   - **Example:** For the dataset [2, 5, 8], the geometric mean is $\sqrt[3]{2 \times 5 \times 8}$ = 4.3.
3. **Harmonic Mean (HM):**
   - **Definition:** The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the values in a dataset.
   - **Formula:** Harmonic Mean, $$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots \ldots \ldots \ldots + \frac{1}{x_n}} = \frac{n}{\Sigma_{i=1}^{n} \frac{1}{x_i}}$$
   - **Example:** For the dataset [2, 4, 8], the harmonic mean is $\frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}} = \frac{24}{7}$
4. **Combined Mean:**
   - **Definition:** A combined mean is a measure that combines the characteristics of two or more different types of means.
   - **Example:** The weighted mean, which combines different subsets of data with different weights.
5. **Median:**
   - **Definition:** The median is the middle value of a dataset when it is ordered. If there's an even number of observations, the median is the average of the two middle values.
   - **Example**: For the dataset [3, 6, 8, 10, 15], the median is 8. For [2, 4, 6, 8], the median is $\frac{4+6}{2} = 5$
6. **Mode:**
   - Definition: The mode is the value that occurs most frequently in a dataset.

Example: For the dataset [2, 4, 4, 6, 8], the mode is 4.

❖ **What are the desirable characteristics of an ideal average should have?**

Answer:

An ideal average, or measure of central tendency, should possess certain desirable characteristics to effectively represent the central value of a dataset. Here are some key characteristics:

1. **Representativeness:** The average should accurately represent the center of the data distribution, reflecting the typical value.
2. **Uniqueness:** The average should be a unique value for a given dataset, providing a single-point summary.
3. **Sensitivity:** The measure should be sensitive to changes in the data values, capturing variations in the dataset.
4. **Robustness:** The average should not be unduly influenced by extreme values or outliers, ensuring stability in the presence of unusual data points.
5. **Additivity:** The average of combined datasets should be the sum of the averages of the individual datasets, facilitating the analysis of grouped data.
6. **Applicability:** The measure should be suitable for the type of data under consideration, whether it's nominal, ordinal, interval, or ratio data.
7. **Ease of Computation:** The average should be computationally feasible and easy to understand, promoting practicality in real-world applications.
8. **Comparability:** The average should allow for meaningful comparisons between different datasets or subsets within a dataset.
9. **Interpretability:** The average should be interpretable in the context of the data, providing insights that are easily understandable.
10. **Stability:** The average should not change significantly with minor variations in the dataset, ensuring stability in the face of small changes.

❖ **State the merits and demerits of using arithmetic mean (AM), geometric mean (GM), harmonic mean (HM), combined mean, median and mode with example.**

Answer:

Merits and Demerits of Central Tendency Measures:

1. **Arithmetic Mean (AM):**
   - Merits:
     - Widely used and easy to understand.
     - Utilizes all data points, providing a comprehensive summary.
   - Demerits:
     - Sensitive to extreme values (outliers), which can distort the mean.
     - Not suitable for highly skewed distributions.

Example: Consider the dataset [5, 8, 10, 12, 100]. The mean is influenced significantly by the outlier (100).

2. **Geometric Mean (GM):**

- Merits:
    - Less affected by extreme values compared to the arithmetic mean.
    - Suitable for data involving ratios, growth rates, or multiplicative relationships.

- Demerits:
    - Not applicable to datasets with negative values.
    - Requires positive values; undefined for zero.

Example: For investment returns of [5%, 10%, 20%], the geometric mean is appropriate.

3. **Harmonic Mean (HM):**
    - Merits:
        - Useful for rates and ratios, especially when dealing with reciprocals.
        - Resistant to extremely large values.
    - Demerits:
        - Highly sensitive to small values; affected by outliers on the lower end.
        - Less intuitive and less commonly used.

Example: If time taken to travel each of 3 segments is [2, 4, 8] hours, the harmonic mean represents the average speed.

4. **Combined Mean:**
    - Merits:
        - Allows for the incorporation of different weights for different subsets of data.
        - Useful when subsets have different levels of importance.
    - Demerits:
        - Requires clear justification for the choice of weights.
        - Interpretation can be complex.

Example: Weighted mean of exam scores where mid-term is weighted more heavily than quizzes.

5. **Median:**
    - Merits:
        - Not influenced by extreme values; robust to outliers.
        - Appropriate for skewed distributions.
    - Demerits:
        - May not use all data points, less comprehensive than the mean.
        - Computationally more complex, especially for large datasets.

Example: In the dataset [5, 8, 10, 12, 100], the median is 10.

6. **Mode:**
    - Merits:
        - Applicable to nominal data.
        - Easily understandable; suitable for categorical data.
    - Demerits:

- May not exist (no mode) or may be ambiguous (multimodal).
- Not sensitive to all values; may not represent the typical value.

Example: In the dataset [2, 4, 4, 6, 8], the mode is 4.

❖ **Determine general procedures for calculating median and mode from graphs with appropriate diagrams.**

Answer:

To calculate the median and mode from a graph, follow these general procedures:

1. Median Calculation:

- Locate the midpoint of the graph along the x-axis.
- If the graph in a histogram or bar chart finds the interval containing the median.
- If the graph is a line graph, identify the value-x value where the cumulative frequency is approximately half of the total.
- The corresponding y-value gives the median.

2. Mode calculation:

- Identify peaks or high on the graph, as these may represent modes.
- For histograms or high points on the graph, as these may represent modes.
- For histograms or bar charts, the mode is often the interval with the highest bar.

3. Appropriate diagram:

- Draw a sketch or use the graph itself to illustrate key points.
- Label the median and mode points on the diagram.
- Clearly indicate any interval or values associated with these measures.

❖ **Distinguish between AM and GM.**

Answer:

1. **Calculation:**
   - **Arithmetic mean:** Calculated by adding up all values in a dataset and dividing by the number of observations
   - **Geometric Mean:** Calculated as the nth root of the product of n values in a dataset.
2. **Sensitivity to values:**
   - **Arithmetic mean:** Sensitive to extreme values (outliers) in the dataset.
   - **Geometric mean:** Lens affected by extreme values if tends to moderate the impact of every height or very low value.
3. **Nature of values:**
   - **Arithmetic mean:** Utilize the actual valves in the dataset directly.

- **Geometric mean:** Involves the multiplication of values, making it suitable for situations where growth rates involved.

4. **Applicability to Data types:**
   - **Arithmetic Mean:** Applicable to all types of data, including nominal, ordinal, interval, and ratio data.
   - **Geometric Mean:** The Geometric mean of a combined dataset in the product of the geometric mean of the individual datasets.
5. **Interpretation:**
   - **Arithmetic mean:** Represents the balance point set or average value in the data.
   - **Geometric mean:** Represents the Central tendency in terms of the multiplicative relationship between values.
6. **Negative value:**
   - **Arithmetic Mean:** Can handle datasets with negative values.
   - **Geometric Mean:** Not applicable to data with negative values; undefined for zero values.
7. **Uniqueness.**
   - **Arithmetic mean:** The arithmetic mean is unique for a given dataset.

Geometric mean: The geometric mean is not unique, the value depends on the order of the data.

❖ **Distinguish between median and mode.**

Answer:

1. **Definition:**
   - **Median**: The median is the middle value in a dataset when it is ordered. If there's an even number of observations, the median is the average of the two middle values.
   - **Mode**: The mode is the value that occurs most frequently in a dataset. A dataset may have no mode (no value repeating), one mode (unimodal), or more than one mode (multimodal).
2. **Position in the Data:**
   - **Median:** It is related to the position of values in the dataset. It represents the middle point.
   - **Mode:** It is related to the frequency of values in the dataset. It represents the most frequently occurring value.
3. **Sensitivity to Extreme Values:**
   - **Median:** It is not sensitive to extreme values (outliers). The median is determined by the position of values and not their magnitudes.
   - **Mode:** It is not greatly influenced by extreme values, but it may be affected by slight variations in the frequency of values.
4. **Applicability:**
   - **Median:** Useful for skewed distributions or datasets with outliers, as it is resistant to extreme values.
   - **Mode:** Suitable for categorical data and distributions with clear peaks. It may not be suitable for continuous data with no clear mode.
5. **Uniqueness:**
   - **Median:** The median is unique for a given dataset.
   - **Mode:** A dataset can have no mode (no repeating values), one mode (unimodal), or more than one mode (multimodal).

6. **Type of Data:**
   - **Median:** Applicable to ordinal, interval, or ratio data.
   - **Mode:** Commonly used for nominal data, but it can be used for other types as well.

Theorem

❖ **Prove that AM depends on origin and scale.**

Answer:

We know that,

If the relation between two variables x and y is

$y = a + bx$ [Where a and b are constant]

Then, $\bar{y} = a + b\bar{x}$

So, $y = a + bx$

⇨ $\sum y = \sum a + \sum bx$ [Talking $\sum$ over both sides]

⇨ $\frac{\sum y}{n} = \frac{\sum a}{n} + \frac{\sum bx}{n}$ [Dividing both sides by n]

⇨ $\bar{y} = \frac{na}{n} + b\bar{x}$

⇨ $\bar{y} = a + b\bar{x}$

$\therefore \bar{y} = a + b\bar{x}$

This proves that AM depends on origin and scale.

❖ **If $\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$ are means of three series of sizes n1, n2 and respectively then the combined mean is $\bar{x}_c = \frac{n_1\bar{x}_1+n_2\bar{x}_2+n_3\bar{x}_3}{n_1+n_2+n_3}$.**

Answer:

Suppose, $x_{11}, x_{12}, \ldots . x_{1_{n_1}}$ be the n_1 observations of the first series,

$x_{21}, x_{22}, \ldots . x_{2_{n_2}}$ be the n_2 observations of the second series and

$x_{31}, x_{32}, \ldots ., x_{3_{n_3}}$ be the n_3 observations of the third series.

$\therefore$

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n1} x_{1i}$$

$\therefore$

$$\bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n2} x_{2j}$$

$\therefore$

$$\bar{x}_3 = \frac{1}{n_3}\sum_{k=1}^{n3} x_{3k}$$

The mean of the combined series,

$$\bar{x}_c = \frac{\left(x_{11} + x_{12} + x_{13} + \cdots + x_{1_{n_1}}\right) + \left(x_{21} + x_{22} + x_{23} + \cdots + x_{2_{n_2}}\right) + \left(x_{31} + x_{32} + x_{33} + \cdots + x_{3_{n_3}}\right)}{n_1 + n_2 + n_3}$$

$$= \frac{\frac{n_1}{n_1}\sum_{i=1}^{n1} x_1 + \frac{n_2}{n_2}\sum_{j=1}^{n2} x_2 + \frac{n_3}{n_3}\sum_{k=1}^{n3} x_3}{n_1 + n_2 + n_3}$$

$\therefore$

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

- **For two non-zero positive numbers, prove that, $AM \times HM = GM^2$**

Answer:

Suppose that the positive non-zero quantities are x_1 and x_2 then,

$AM = \frac{x_1 + x_2}{2}$; $GM = \sqrt{x_1 x_2}$; $HM = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$

Now, L.S.H = $AM \times HM$

$= \frac{x_1 + x_2}{2} \times \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$

$= (x_1 + x_2) \times \frac{x_1 x_2}{x_1 + x_2}$

$= x_1 x_2$

$= (\sqrt{x_1 x_2})^2$

$= (GM)^2$

= R.H.S (Proved)

[HW] For any set of non-zero positive numbers, prove that, $AM \times HM = GM^2$

- **For two non-zero positive numbers, prove that, $AM \geq GM \geq HM$.**

Answer:

Suppose that, the positive non-zero quantifies are x^1 and x^2 then,

$AM = \frac{x_1 + x_2}{2}$; $GM = \sqrt{x_1 x_2}$; $HM = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$

Consider,

$(\sqrt{x_1} + \sqrt{x_2})^2 \geq 0$

$\Rightarrow (x_1 + x_2 - 2)\sqrt{x_1 x_2} \geq 0$

$\Rightarrow \frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$

$\therefore$ AM $\geq$ GM ... (i)

Again, $(\frac{1}{\sqrt{x_1}} + \frac{1}{\sqrt{x_2}})^2 \geq 0$

$\Rightarrow (x_1 + x_2 - 2)\frac{1}{\sqrt{x_1 x_2}} \geq 0$

$\Rightarrow \frac{1}{x_1} + \frac{1}{x_2} \geq \frac{2}{\sqrt{x_1 x_2}}$

$\Rightarrow \sqrt{x_1 x_2} \geq \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$

$\therefore$ GM $\geq$ HM ... (ii)

So, $AM \geq GM \geq HM$. (Proved)

- **For the first natural numbers, prove that, $AM = \frac{n+1}{2}$ and $SD = \sqrt{\frac{n^2-1}{12}}$**

Answer:

For the first natural number, Arithmetic mean (AM) $= \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$

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

$$= \frac{n+1}{2}$$

$\therefore$ AM = $\frac{n+1}{2}$ (Proved)

We Know,

$$\text{Variance}(\sigma^2) = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$
$$= \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - \left(\frac{n+1}{2}\right)^2$$
$$= \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] - \left(\frac{n+1}{2}\right)^2$$
$$= \frac{(n+1)(n+1)}{6} - \frac{(n+1)^2}{4}$$
$$= \frac{(n+1)}{2}\left[\frac{2n+1}{3} - \frac{n+1}{2}\right]$$
$$= \frac{n+1}{2} \times \frac{n-1}{6}$$
$$= \frac{n^2-1}{12}$$

Standard deviation, $SD = \sqrt{\text{variance}} = \sqrt{\frac{n^2-1}{1n}}$ (Proved)

Mathematical Problems

- **The scones of 10 students of a college are 68, 70, 60, 64, 67, 50, 54, 60, 80, 74. Find the AM, GM, HM, median and mode of the given scores.**

Answer:

Arrange the given scores in ascending order: 50,54,60,60,64,67,68,70,74.80

$$AM = \frac{50+54+60+60+64+67+68+70+74+80}{10}$$
$$= 64.7.$$

$$GM = \sqrt[10]{50 \times 54 \times 60 \times 60 \times 64 \times 67 \times 68 \times 70 \times 74 \times 80}$$
$$= 64.12$$

$$HM = \frac{10}{\frac{1}{50}+\frac{1}{54}+\frac{1}{60}+\frac{1}{60}+\frac{1}{64}+\frac{1}{67}+\frac{1}{68}+\frac{1}{70}+\frac{1}{74}+\frac{1}{80}}$$
$$= 63.53$$

Here, $n = 10$, which is an even number

We know,

$$\text{median} = \frac{\frac{n}{2}\text{th number} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{ number}}{2}$$
$$= \frac{\left(\frac{10}{2}\text{ th} + (\frac{10}{2}+1)th\right) \text{number}}{2}$$
$$= \frac{\left(5^{\text{th}} + 6^{\text{th}}\right) \text{number}}{2}$$
$$= \frac{64+67}{2}$$
$$= 65.5$$

Mode is that number which occurs most frequently.

Here 60 is appears two times, more than any other score.

So, the mode is 60.

- **Calculate AM, GM, HM, Median and mode for the following dataset. Also verify that $AM > GM > HM$**

Marks:	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**
No. of Students:	**15**	**20**	**35**	**20**	**10**

Answer:

The given frequency distribution table:

class	**Mid value x_i**	**Frequency f_i**	**Common frequency**	$f_i x_i$
40-50	45	15	15	675
50-60	55	20	35	1100
60-70	65	35	70	2275
70-80	75	20	90	1500
80-90	85	10	100	850
		n = 100		$\sum f_i x_i = 6400$

We know,

$$AM = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$
$$= \frac{6400}{100}$$
$$= 64$$

$$GM = \left(45^{15} \cdot 55^{20} \cdot 65^{35} \cdot 75^{20} \cdot 85^{10}\right)^{\frac{1}{100}}$$
$$= 62 \cdot 88$$

$$H \cdot M = \frac{n}{\sum \frac{f_i}{x_i}}$$
$$= \frac{100}{\frac{15}{45}+\frac{20}{55}+\frac{35}{65}+\frac{20}{75}+\frac{10}{85}}$$
$$= 61.73$$

So, $AM > GM > HM$.

We know,

Here, $\frac{n}{2} = 50$.

so, $60 - 70$ is the median class

$$\text{Median} = L + \left(\frac{n}{2} - F_c\right) \times \frac{h}{f_m}$$
$$= 60 + (50 - 35) \times \frac{10}{35}$$
$$= 60 + 4.29$$
$$= 64.29.$$

$LL = 60$
$h = 10$
$F_c = 35$
$f_m = 35$

Here, $60 - 70$ is the modal class.

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$
$$= 60 + \left(\frac{35-20}{2\times35-20-20}\right) \times 10$$
$$= 60 + \frac{15}{30} \times 10$$
$$= 65.$$

$L = 60$
$f_1 = 35$
$f_2 = 20$
$f_0 = 20$
$h = 10$

❖ **Construct a frequency distribution by taking class interval 10 from the given dataset and find AM, median and mode. Also comment on the Skewness by graphical representation of the values.**

31	44	45	60	56	52	47	80	75	72
55	65	35	62	78	82	38	32	58	60
54	48	62	68	55	48	66	70	74	40

Answer:

From above data set.
Total frequency, $n = 30$

$$x_{\max} = 82$$
$$x_{\min} = 31$$
$$\text{Range} = 82 - 31 = 51$$

Given that,

Class interval, $h = 10$.

Total class $= \frac{51}{10} = 5.1 \approx 6$

Frequency distribution table:

Class	Tally	Mid value x_i	Frequency f_i	Cumulative frequency	$f_i x_i$
31-40		35.5	5	5	177.5
41-50		45.5	5	10	227.5
51-60		55.5	8	18	444
61-70		65.5	3	24	393
71-80		75.5	5	29	377.5
81-90		85.5	1	30	85.5
			N= 30		$\sum f_i x_i = 1705$

We know,

$$AM = \frac{\sum_1^n f_i x_i}{N} = \frac{1705}{30} = 56.83$$

Here, $51 - 60$ is median class.

$$\text{Median} = L + \left(\frac{N}{2} - F_c\right) \times \frac{h}{f_m} = 51 + (15 - 10) \times \frac{10}{6} = 51 + 5 \times \frac{10}{6} = 51 + 8.33 = 59.33$$

$F_c = 10$, $n = 30$, $f_m = 5$, $\therefore \frac{n}{2} = 15$, $h = 10$, $L = 51$

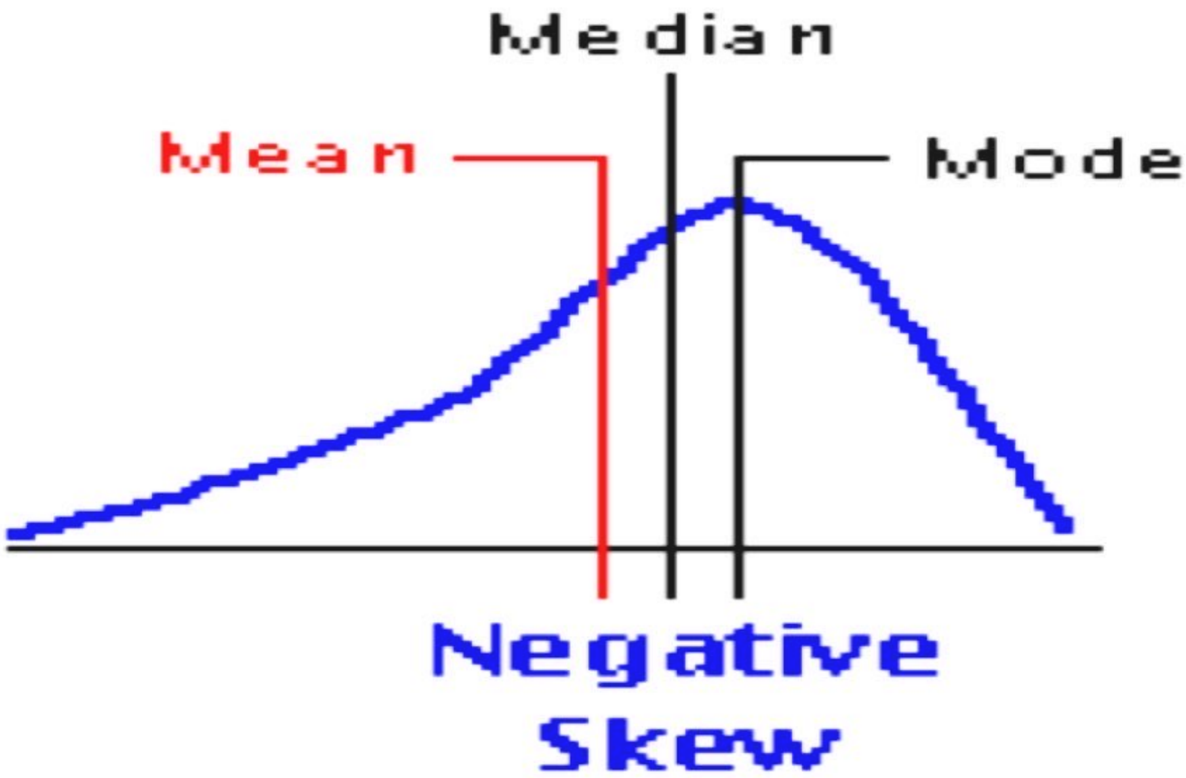
Here,

Modal class is $51 - 60$

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$
$$= 51 + \left(\frac{8-5}{2\times 8-5-6}\right) \times 10$$
$$= 51 + \frac{3}{5} \times 10$$
$$= 51 + 6$$
$$= 57$$

$f_1 = 8$
$f_0 = 5$
$f_2 = 6$
$L = 51$
$h = 10$

We know,



Since $mean < median < mode = 56.83 < 59.33 < 57$

So, the graph is negatively skewed distribution.

- **Calculate the mean, median and mode for the following dataset. Also, verify that A.M > G.M > H.M by graphical representation**

Marks	less than 50	50-60	60-70	70-80	greater than 80
No. of students	15	20	35	20	10

Answer:

The given data table is below

Class	less than 50	50-60	60-70	70-80	greater than 80	
Frequency, f	15	20	35	20	10	n=100
Cumulative frequency	15	35	70	90	100	

Here, $\frac{n}{2} = \frac{100}{2} = 50$ so, 60-70 is the median class.

We know,

$$\text{median} = L + (\frac{n}{2} - F_c) \times \frac{h}{f_m}$$

$$= 60 + (50 - 35) \times \frac{10}{35}$$

$$= 64.29$$

Here,

$L = 60$

$F_c = 35$

$f_m = 20$

$h = 10$

Here 35 is the largest frequency number.

So, 60-70 is the mode class.

We know,

$$mode = L + (\frac{f_1 - f_0}{2f_1 - f_0 - f_2}) \times h$$

$$= 60 + (\frac{35 - 20}{2 \times 35 - 20 - 20}) \times 10$$

$$= 65$$

Here,

$h = 10$

$L = 60$

$f_1 = 35$

$f_0 = 20$

$f_2 = 20$

We know,

$$mode = 3\, median - 2\, mean$$

$$\Rightarrow 65 = 3 \times 67.5 - 2 \times mean$$

$$\Rightarrow 2mean = 202.5 - 65$$

$$\Rightarrow mean = \frac{137.5}{2}$$

$$mean = 68.75$$

❖ **An incomplete frequency distribution is as follows, where median is 65**

Income (TK.)	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**	

No. of families	**10**	**15**	f_1	**20**	f_2	**5**	**Total=100**

Answer:

Given Table

Income (TK.)	40-50	50-60	60-70	70-80	80-90	90-100	
Frequency	10	15	f_1	20	f_2	5	Total=100
Cumulative frequency	10	20	$25 + f_1$	$45 + f_1$	$45 + f_1 + f_2$	$50 + f_1 + f_2$	

From this table,

$50 + f_1 + f_2 = 100$

$\Rightarrow f_1 + f_2 = 50$

Given that median is 65 so, the median class is 60-70

We know,

Median $= L + (\frac{n}{2} - F_c) \times \frac{h}{f_m}$

$\Rightarrow 65 = 60 + (\frac{100}{2} - (25 + f_1)) \times \frac{10}{20}$

$\Rightarrow 65 = 60 + \frac{50 - 25 - f_1}{2}$

$f_1 = 15$

So, $f_1 + f_2 = 50$

$f_2 = 35$

So, $f_1 = 50$ and $f_2 = 35$

Here,

n = 100

L = 60

h = 10

$F_c = 25 + f_1$

$f_1 = f_m$

- **The mean distance of 200 items was 50Km. Later on, it was found that two observations were wrongly read as 92 and 8 instead of 192 and 88 Km respectively. Find the corrected mean.**

Answer:

Mean of 200 items = 50

Sum of 200 items = (50×200)

$= 1000$

Correct sum of 200 items $= \{1000 - (sum\ of\ wrong\ items) + (sum\ of\ correct\ items)\}$

$= \{1000 - (92 + 8) + (192 + 88)\}$

$= 1000 - 100 + 280$

$= 10180$

So, correct mean $= \frac{10180}{200}$

$= 50.9$

❖ **The mean of marks in CSE of 100 students of a class was 72. The mean of marks of boys was 75, while their number was 70. Find out the mean marks of girls in the class.**

Answer:

The number of girl students $n = 100 - 70 = 30$

The total marks of all students $= 100 \times 72 = 7200$

The total marks of all boys $= 75 \times 70 = 5250$

So, the total marks of 30 girls $= 7200 - 5250 = 1950$

The mean marks of girl students $= \frac{1950}{30} = 65$

❖ **The average monthly income of male employees in a firm is Tk. 1000 and that of females is Tk. 800. But the mean income of all the employees is Tk. 920. Find the percentages of male and female employees.**

Answer:

Let the number of male employees be n_1

and the number of female employees be n_2

The average salary of male employees be $x_1 = 1000\ Tk$

The average salary of female employees be $x_2 = 800\ Tk$

The average salary of all employees be $x = 920\ Tk$

Now, $\underline{x} = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$

$$\Rightarrow 920 = \frac{n_1 \times 1000 + n_2 \times 800}{n_1 + n_2}$$

$$\Rightarrow 1000n_1 + 800n_2 = 920n_1 + 920n_2$$

$$\Rightarrow 80n_1 = 120n_2$$

$$\Rightarrow \frac{n_1}{n_2} = \frac{120}{80}$$

$$\Rightarrow n_1 = \frac{3}{2} \times n_2$$

Let,

The number of female employees be x

Then the number of male employees be $\frac{3}{2}x$

Total number of employees $= x + \frac{3}{2}x = \frac{5}{2}x$

Percentage of female employees$= \frac{x}{\frac{5}{2}x} \times 100\% = 40\%$

Percentage of male employees$= \frac{\frac{3}{2}x}{\frac{5}{2}x} \times 100\% = 60\%$

- **The geometric mean of scores in statistics of 15 male students is 75 and 10 female students is 80 respectively. Compute their combined geometric mean.**

Answer:

The geometric mean of 15 male students be $G.M_1 = 75$

The number of male students is $n_1 = 15$

The number of female students is $n_2 = 10$

The geometric mean of female students be $G.M_2 = 80$

So, combined geometric mean G.M $= log^{-1}[\frac{n_1 log(G.M_1) + n_2 log(G.M_2)}{n_1 + n_2}]$

$$= log^{-1}[\frac{15log(75) + 10log(80)}{15+10}]$$

$$= 76.$$

❖ **John spends Tk. 100 for apples costing Tk. 10 per kg, Tk. 100 for apples costing Tk. 12 per kg and another Tk. 100 for apples costing Tk. 14 per kg. Find the average price per kg.**

Answer:

The weight of apples costing Tk. 10 per Kg $x_1 = \frac{100}{10} = 10\ Kg$

The weight of apples costing Tk. 12 per Kg $x_2 = \frac{100}{12} = 8.33\ Kg$

The weight of apples costing Tk. 14 per Kg $x_3 = \frac{100}{14} = 7.14\ Kg$

Total weight of apples $x = x_1 + x_2 + x_3 = 25.47\ Kg$

Total costing in Tk. $= 100 + 100 + 100 = 300\ Tk$

So, The average price per Kg $\underline{x} = \frac{300}{25.47} = 11.77\ Tk\ .per\ Kg$

Measures of Dispersion: Variation and All

❖ **What do you mean by measures of dispersion?**

Answer:

Measures of dispersion are statistical measures that quantify the spread or variability of data. They are used to describe how different the values in a data set are from each other and from the central value of dataset. The most commonly used measures of dispersion are:

1. Range
2. Variance
3. Standard Deviation
4. Mean Deviation
5. Quartile Deviation

❖ **What are the purposes of measuring dispersion?**

Answer:

The purposes of measuring dispersion are:

1. To access the reliability of the central tendency: A central tendency measure, such as the mean or median is only representative of the data set, if the data is too dispersed. If the data is highly dispersed, then the central tendency measure may not be a good representation of data set.
2. To compare the variability of different datasets: Measures of dispersion can be used to compare the variability of different datasets. For example, a teacher might use measures of dispersion to compare the variability of test scores in different classes.
3. To identify outliers in a dataset: An outlier is a data point that is significantly different from the rest of the data. Measure of dispersion can be used to identify outliers in a data set.
4. To make prediction about the future data: Measures of dispersion can be used to make predictions about future data. For example, a business might use measures of dispersion to predict the variability of sales in the future.

❖ **What are the properties of a good measure of dispersion?**

Answer:

A good measure of dispersion should have the following properties:

1. Easy to calculate and understand: The measure should be easy to calculate and understand, even for people with limited statistical knowledge.
2. Based on all observations: The measure should be based on all of the observations in the data set, not just a few of them.
3. Unaffected by extreme values: The measure should not be unduly affected by extreme values in the data set.
4. Comparable across datasets: The measure should be comparable across different datasets, even if the data sets have different central tendencies.

5. Amenable to mathematical-treatment: The measure should be amenable to mathematical treatment, so that it can be used in statistical calculations and analysis.
6. Consistent: The measure should produce consistent results when applied to the same data set multiple times.
7. Efficient: The measure should use the information in the data set efficiently, so that, it provides the most accurate possible estimate of dispersion.
8. Robust: The measure should be robust to the violations of its assumptions, such as normality.

❖ **What are the classifications of measures of dispersion?**

Answer:

Measures of dispersion can be classified into two broad categories:

1. Absolute measures of dispersion.
2. Relative measures of dispersion.

Absolute measures of dispersion are expressed in the same units as the data. They measure the average distance of the data points from the central value of the data set. Common absolute measures of dispersion include:

1. Range
2. Variance.
3. Standard deviation
4. Mean deviation
5. Quartile deviation.

Relative measures of dispersion are dimensionless measures of dispersion. They are expressed as a percentage of the central value of the dada set common relative measures of dispersion include:

1. Coefficient of range.
2. Coefficient of mean deviation.
3. Coefficient of variation.
4. Coefficient of quartile deviation.

❖ **Define Range, Mean deviation, Standard Deviation, Quartile Deviation, Coefficient of Range (CR), Coefficient of Mean Deviation, Coefficient of Variation (CV), and Coefficient of Quartile Deviation.**

Answer:

The **Range** in statistics is the difference between the largest and smallest value in a data set. It is the simplest measure of dispersion, but it is the also informative. Simply, Range, $R = X_{Largest} - X_{Smallest}$.

The **mean deviation** is a measure of dispersion that calculates the average distance between each data point and the mean.

Mean deviation $= \frac{\sum|x_i - a|}{n}$

Where a may be arithmetic mean, median or any other value.

The **standard deviation** is a measure of the dispersion of a set of data from this mean. It is calculated as the square root of the variance, which is the average of the squared deviations of each data point from the mean.

Population standard deviation, $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$

Sample standard deviation, $\sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

Quartile deviation also known as the semi-inter quartile range is a measure of statistical Dispersion that describes the spread of a data set by focusing on the inter quartile range. The inter-quartile range is the range between first quartile (Q1) and the third quartile (Q3), representing the middle 50% of the data.

$$\text{Quartile Deviation (QD)} = \frac{Q_3-Q_1}{2}$$

Here, Q_3 is the third quartile.
Q_1 is the first quartile.
Q_D is the Quartile Deviation.

The **coefficient of range**, also known as the range ratio or relative range is a measure of statistical dispersion that express the spread or variability of a data set.

$$CR = \frac{Range}{Mean} \times 100$$

The **coefficient of mean deviation** is a relative measure of dispersion that is calculated by dividing the mean deviation by the mean. It is expressed as a percentage.

$$\text{Coefficient of Mean Deviation} = \frac{MD}{Mean} \times 100$$

The **coefficient of quartile deviation** is a relative measure of dispersion that is calculated by dividing the inter quartile range by the median.

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3-Q_1}{Q_3+Q_1} \times 100$$

- **Write down the advantages and disadvantages of range and SD.**

Answer:

Range:

Range is the easiest to understand of all the measures of dispersion. The difference between the largest and smallest item in a distribution is called range.

The main advantages of range con be expressed as follows:

- It is rigidly defined.
- It is easy to understand.
- It provides a quick measure of variability.
- Range provides an overview of the data at once.
- It takes the least possible time to be calculated.
- It is very easy to calculate as its formula rests only on two simple factors.

The main disadvantages of range can be expressed as follows:

- Range is not based on all of the observations. The range of distribution remains the same if every item is changed except if for the smallest and largest item.
- Fluctuations in sampling have a big impact on range. Its value differs widely between samples.
- It does not provide any insight into the pattern of distribution. It is possible for two distributions to have the range but different patterns of distribution.
- It is not capable for further algebraic treatment.
- It cannot reveal the different dispersal characters of different series in which two extreme values remain the same.

Standard Deviation (SD):

The standard deviation of a dataset is a way to measure the typical deviation of individual values from the mean value.

We now list out some of the advantages and disadvantages of standard deviation as a measure of dispersion.

Advantages of Standard Deviation:

- It is rigidly defined and free from any ambiguity.
- It is the most well-known measure of dispersion and is applicable in a wide variety of situation.
- The standard deviation takes all observations into consideration.
- The standard deviation is used to calculate the skewness and kurtosis of the data who is tells us about the symmetry and shape of the data.
- It is capable of further algebraic treatment. Since the formula involves squaring the error. The numerator can be further simplified by expanding the square.

Disadvantages of Standard Deviation:

- The standard deviation ignores the actual sign of the error term.
- The standard deviation does not exactly measure the actual distance of each observation from the mean but the square of the differences between the observations from the mean.
- When squaring the differences, outliers will contribute a large value to the numerator because squaring large values make them even larger. This means that the standard deviation gives more weightage to extreme values. Therefore, the standard deviation is not resistant to the effect of outliers.
- The calculation is very cumbersome to carry out by hand since it involves extracting square roots. Although this drawback is easily taken care by using calculator.

It cannot be used for comparing the dispersion of two, or more series given in different units.

❖ **Distinguish between CV and variance.**

Answer:

Coefficient of Variation (CV)	Variance
1. Coefficient of variation is a statistical measure of the dispersion of observations in a data set around the mean.	1. Variance is a measure of how spread out a set of data points from the mean.
2. It is calculated by divine the standard deviation by the mean. $\text{CV} = \frac{Standard\ Deviation}{Mean}$	2. It is calculated by taking the average of the squared deviation from the mean. $\text{Variance, } V_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$
3. Unit is percentage.	3. Same unit as original data.
4. Relative variability of a set of data points.	4. Spread of a set of data points from the mean.
5. Comparing the variability of different datasets, even if they are measured in different units.	5. Assassin the difference between population or groups in research.

❖ **Distinguish between SD and MD.**

Answer:

Standard Deviation (SD)	Mean Deviation (MD)
1. These are calculated from arithmetic mean only.	1. Mean deviations are calculated from mean median and mode.
2. In the calculation of standard deviation we make all the signs positive by squaring the deviations.	2. The algebraic signs have to be ignored, only values of deviation are taken.
3. It is mathematically sound because algebraic signs are not ignored.	3. It lacks mathematical properties since only absolute values are considered.
4. It is based on the square root of the average of the squared deviation.	4. It is based on the simple average of the sum of absolute deviations.
5. It is least influenced by Extreme values.	5. It tends to increase with the increase in the size of the sample.
6. It is somewhat complex but it is suitable in all cases whether the mean is whole number or a fraction.	6. It is simple to calculate when the mean is whole number. The shortcut method is somewhat cumbersome.

❖ **Distinguish between absolute measures of dispersion and relative measures of dispersion.**

Answer:

Absolute measures of Dispersion	Relative measures of Dispersion
1. Provide information in the same unit as data.	1. Express the spread in terms of a ratio or percentage.
2. Units are as same as the original data units.	2. Dimensionless (No units).
3. Directly interpretable in the context of the data.	3. Facilities comparison between data sets with different units or scales.
4. Can be sensitive to Extreme values.	4. Generally less sensitive to outliers.

5. Examples are range, variance, standard deviation etc.	5. Examples are the coefficient of variation, percentage deviation.
6. Range = X_{max} - X_{min}	6. CV = $\frac{SD}{Mean} \times 100\ \%$

❖ **"Variance is not always greater than Standard deviation"-explain.**

Answer:

The statement variance is not always greater than standard deviation might seem counterintuitive at first, but it is indeed accurate. The key to understanding this lies in the relationship between variance and standard deviation.

Variance is the average of the square difference between each data point and the mean. The standard deviation is the square root of the variance.

Variance, $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$

Standard deviation, $\sigma = \sqrt{Variance\ (\sigma^2)}$

Now consider a dataset with value close to the mean. In such a case the squared difference from the mean (used in the variance calculation) will be relatively small. When you take the square root of this small variance to get the standard deviation the result with also be small. On the other hand, if the dataset has the larger deviation from the mean, the squared difference will be larger, resulting in a larger variance. Taking the square root of this larger variance the standard deviation will yield of larger value.

This is why it is correct to say that variance is not always greater than the standard deviation, it depends on the characteristics of the data set.

❖ **"Coefficient of variance is an absolute number"-comment.**

Answer:

The statement "Coefficient of variance is an absolute number" is incorrect. The coefficient of variation is a relative measure of dispersion and it is impressed as a percentage. It is used to compare the relative variability of two or more sets of data that may have different units or scales. The formula for calculating the coefficient of variation is,

$CV = (\frac{Standard\ Deviation}{Mean}) \times 100\ \%$

Coefficient of variation is not an absolute number because it gives a relative measure of variability as a percentage. It depends on the characteristics of the specific data set being analyzed.

❖ **"Variance is not always non-negative"-explain.**

Answer:

The statement "Variance is not always non-negative" is not accurate. Variance is always non negative, it cannot be negative by definition. Variance is a measure of the spread or dispersion of a set of data points and it is calculated as the average of the squared difference between each data points and the mean of the data set.

The formula for variance is,

Variance $(\sigma^2) = \frac{\sum(x_i - \mu)^2}{N}$

Where,

x_i= Each individual data points.

μ = mean of the data.

N = the number of the data points.

Here the sum of squared differences is always non-negative and when divided by the number of data points(N), the variance is also non-negative.

- **"Mean deviation is least when calculated from median"-explain.**

Answer:

The statement "Mean deviation is least when calculated from the median" is accurate. Mean deviation is a measure of the average absolute difference between each data point and a central measure, which could be the mean or the median.

Mean deviation from the mean $= \frac{\sum|x_i - \bar{x}|}{N}$

Mean deviation from the Median $= \frac{\sum|x_i - Me|}{N}$

Now, consider the nature of the median. The median is the middle value of a dataset when it is arranged in ascending or descending order. Unlike the mean, the median is not affected by extreme value because it depends only on the middle values. Because the median is less sensitive to extreme values. The absolute difference between data point and the mean, especially in datasets with outliers. Therefore, you can calculate the mean deviation from the median, the values are generally smaller, resulting in a smaller mean deviation compared to when calculated from the mean.

Mathematical Problems

- **The run scores of two cricketers for five innings are given below:**

Cricketer A	**45**	**0**	**31**	**8**	**75**
Cricketer B	**25**	**18**	**0**	**23**	**100**

Who of the two is a more consistent batsman?

Answer:

For Cricketer A:

Given data,

Cricketer A 45 0 31 8 75

Number of data (n) =5

$$\text{Mean, } \bar{x} = \frac{45+0+31+8+75}{5} = 31.8$$

We know that, $SD_A = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$

$$= \sqrt{\frac{(45-31.8)^2+(-31.8)^2+(31-31.8)^2+(8-31.8)^2+(75-31.8)^2}{5}}$$

$$= 26.90$$

$\therefore$ Coefficient of variation $(CV)_A = \frac{SD_A}{\bar{x}} \times 100\%$

$$= \frac{26.90}{31.8} \times 100\%$$

$$= 8.46\%$$

For Cricketer B:

Given Data,

Cricketer B	25	18	0	23	100

Number of data (n) =5

$$\text{Mean, } \bar{x} = \frac{25+8+0+23+100}{5} = 33.2$$

We know that, $SD_B = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$

$$= \sqrt{\frac{(25-33.2)^2+(18-33.2)^2+(0-33.2)^2+(23-33.2)^2+(100-33.2)^2}{5}}$$

$$= 34.54$$

$\therefore$ Coefficient of variation $(CV)_A = \frac{SD_B}{\bar{x}} \times 100\%$

$$= \frac{34.54}{33.2} \times 100\%$$

$$= 104.04\%$$

$\therefore CV_B > CV_A$

So, **Cricketer A** is more consistent.

- **If the obtained marks of 5 students (out of 20) are 8, 7, 6, 3, 12 then prove that, $\bar{x}\sqrt{n-1} > \sigma$ and $\sqrt{n-1} > CV$.**

Answer:

Here the obtained marks of 5 students (out of 20) are 8, 7, 6, 3, 12

Number of data (n) = 5

Mean $\bar{x} = \frac{8+7+6+3+12}{5} = 7.2$

We know,

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$$

$$= \sqrt{\frac{(8-7.2)^2+(7-7.2)^2+(6-7.2)^2+(3-7.2)^2+(12-7.2)^2}{5}}$$

$$= 2.93$$

Now,

$\bar{x}\sqrt{n-1} = 7.2 \times \sqrt{5-1} = 14.4$

$\therefore \bar{x}\sqrt{n-1} > \sigma$

Again, CV = $\frac{SD}{\bar{x}} = \frac{2.93}{7.2} = 0.41$

And $\sqrt{n-1} = \sqrt{5-1} = 2$

$\therefore \sqrt{n-1} > CV$

- **Mean and standard deviation of the age of 100 students are 30 years and 5 years respectively. But the age of one student was written 30 years instead of 40 years at the time of processing the data. Find out the corrected mean and standard deviation.**

Answer:

Here,

Total students (N) = 100

Mean = 30 Years

Standard Deviation = 5 Years

The sum of the age for total student is = $100 \times 30 = 3000$ Years

If we correct 30 years by replacing with 40 years then,

The sum of the ages = (3000 – 30) + 40 = 3010 Years

Correct Mean $\bar{x} = \frac{3010}{100} = 30.1$

We know that,

Variance, $\sigma^2 = \frac{\sum x^2}{N} - \bar{x}^2 = 5^2 = 25$

$\Rightarrow \frac{\sum x^2}{N} - 30^2 = 25$

$\Rightarrow \sum x^2 = 2500 + 90000 = 92500$

If we correct 30 years by replacing with 40 years then,

$correct(\sum x^2) = 92500 - 30^2 + 40^2 = 93200$

So, correct variance$(\sigma^2) = \frac{correct(\sum x^2)}{N} - correct(\bar{x}^2)$

$= \frac{93200}{100} - 30.1^2 = 25.99$

$\therefore$ Correct standard deviation (SD) = $\sqrt{25.99} = 5.09$

- ***A* and *B* are two factories. The average weekly wages and standard deviation of wages of the workers are given below:**

Factory	Average weekly wages (TK.)	Standard deviation	Number of workers
A	1560	90	200
B	1580	70	160

Calculate the combined mean and standard deviation of wages of the whole workers in the two factories.

Answer:

Given, $N_1 = 200$

$N_2 = 160$

$\bar{x}_1 = 1560$

$\bar{x}_2 = 1580$

$\sigma_1 = 90$

$\sigma_2 = 70$

$\therefore$ Combined mean, $\bar{x}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1+N_2} = \frac{200\times1560+160\times1580}{200+160} = 1568.89$

Now, $d_1 = |\bar{x}_1 - \bar{x}_{12}| = |1560 - 1568.89| = 8.89$

$d_2 = |\bar{x}_2 - \bar{x}_{12}| = |1580 - 1568.89| = 11.11$

$\therefore$ Combined standard deviation

$$\delta_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1+N_2}} = \sqrt{\frac{200\times90^2+160\times70^2+200\times(8.89)^2+160\times(11.11)^2}{200+160}} = 82.32$$

So, combined standard deviation is 82.32.

❖ **The following table gives an average and variances of prices of shares of three companies *X*, *Y* and *Z*.**

Company	Average price (Tk.)	Variances (Tk.)
X	1470	80
Y	1580	70
Z	1200	50

Which company's share is more stable/consistent and which one would you like to dispose?

Answer:

For X- $\bar{x} = 1470, \sigma = \sqrt{80} = 8.94$

$$\text{C.V} = \frac{\sigma}{\bar{x}} \times 100\% = \frac{8.94}{1470} \times 100\% = 0.61\%$$

For Y- $\bar{x} = 1580, \sigma = \sqrt{70} = 8.37$

$$\text{C.V} = \frac{\sigma}{\bar{x}} \times 100\% = \frac{8.37}{1580} \times 100\% = 0.53\%$$

For Z- $\bar{x} = 1200, \sigma = \sqrt{50} = 7.07$

$$\text{C.V} = \frac{\sigma}{\bar{x}} \times 100\% = \frac{7.07}{1200} \times 100\% = 0.59\%$$

The **Y Company** is more consistent and **X Company** is less consistent than the others. So, I would dispose **X Company**.

Skewness and Kurtosis

- **What do you mean by moments in statistics? Distinguish between raw moments and Central moments. Mention the relation between raw moments and Central moments. What do you mean by Skewness? Graphically explain different types of Skewness. What do you mean by Kurtosis? Graphically explain different types of Kurtosis.**

Answer:

Moment in statistics:

Moments in statistics are quantitative measure that describes the specific characteristics of a distribution and hence can determine the distribution uniquely is called moments. It helps to understand the data set's shape, spread and Central tendency.

The difference between Raw moments and Central moments is given below:

Raw moments	Central moments
The n^{th} moment about zero of a probability density function f(x) is the expected value of x^n and is called a Raw moment.	A central moment is a moment of a probability distribution of a random variable, about a random variable's mean.
Raw moments provide information about the distribution's overall shape and spread.	Central moments offer insights into the distribution's shape and spread's relation to its mean.
Formula: $> \mu_1 = E(x) = \mu_x$ The mean of x. The r^{th} moment is sometimes written as function of Θ, where Θ is a vector of parameters that characterize the distribution of x.	Formula: The r^{th} central moment of x about a is defined as $E(x-a)^r$. If $a = \mu_x$, we have the r^{th} central moment of x about μ_x.

The relation between raw moments and central moments: -

Let $x_1, x_2, \ldots\ldots x_n$ be n values of variable x.

$\mu_1 = 0$

$\mu_2 = \mu_2^{'} - (\mu_1^{'})^2$

$\mu_3 = \mu_3^{'} - 3\mu_2^{'}\mu_1^{'} + 2(\mu_1^{'})\text{^}3$

$\mu_4 = \mu_4^{'} - 4\mu_3^{'}\mu_1^{'} + 6\mu_2^{'}(\mu_1^{'})^2 - 3(\mu_1^{'})^4$

Skewness:-

A measure of the degree of asymmetry of the distribution of a random variable, the standardized third moments of the distribution, is defined as: -

$\gamma = E[\frac{(x-\mu)^3}{\sigma^3}]$ simply, $\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}}$

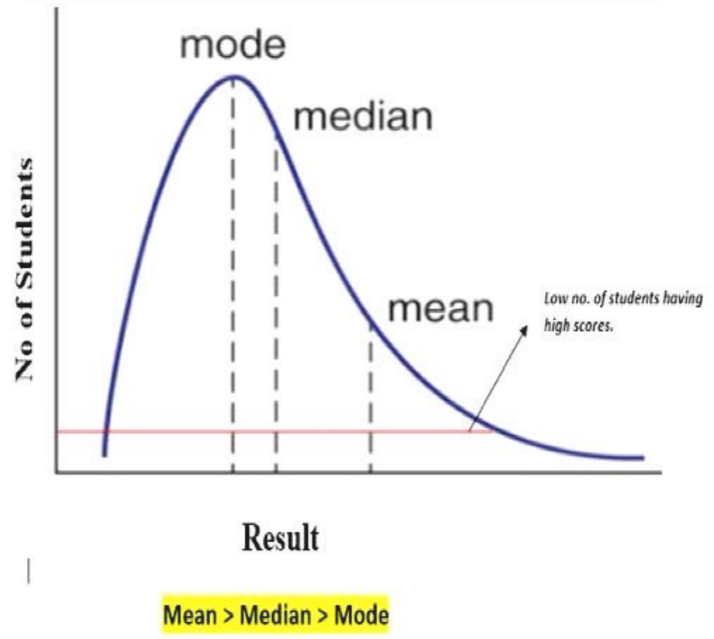
Where μ is the mean and σ is the standard deviation. For a symmetrical distribution γ = 0 and the median of the distribution equals to mean.

Graphically explanation of different types of Skewness: -

Skewness gives us the shape of the data. 'It is the Lack of Symmetry'. Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset.

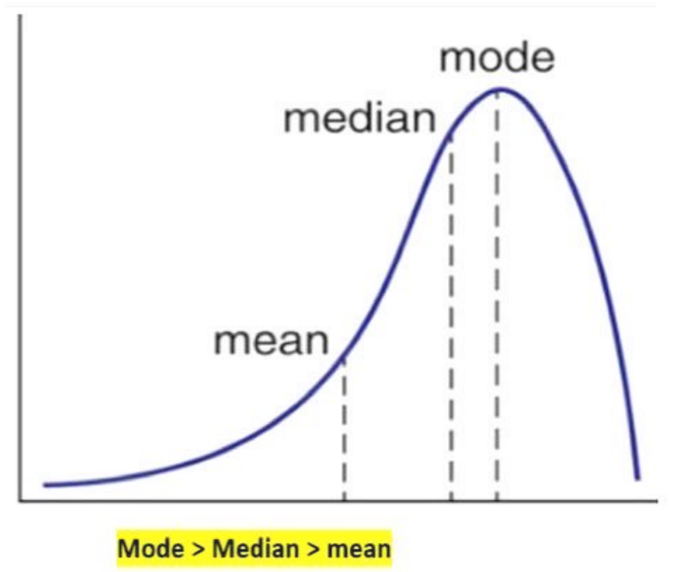
Positive Skewness:

1. The right tail is longer.
2. The mass of the distribution is concentration on the left.
3. Positive skewness is often associated with the presence of outliers on the right side of the distribution.
4. $Mode < Median < Mean$



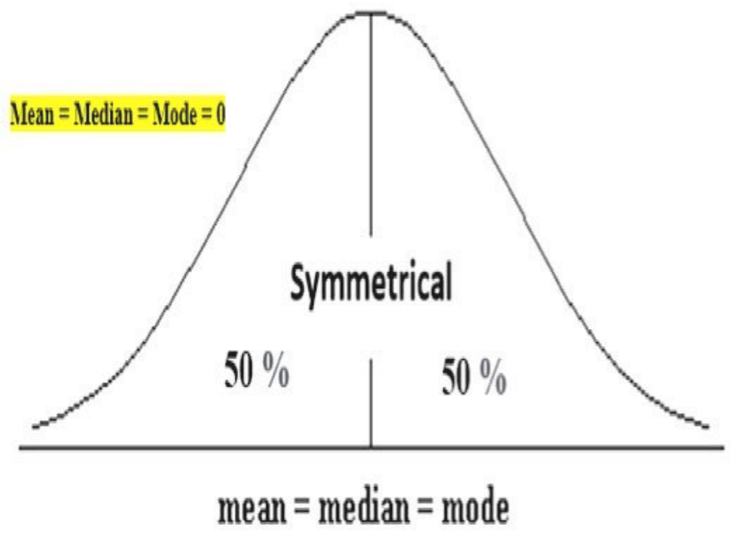
Negative Skewness:

1. The left tail is longer.
2. The mass of the distribution is concentrated on the might.
3. Negative skewness is often associated with the presence of outliers on the left side of the distribution.
4. $Mean < Median < Mode$



Symmetrical distribution:

1. Both tails are equal.
2. Mass of distribution is equally distributed.
3. Skewness is a measure of the asymmetry of a distribution.
4. $Mean = Median = Mode$



Kurtosis:- Kurtosis is a Statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. Mainly Kurtosis identifies whether the tails of a given distribution contain extreme values. Kurtosis helps in analyzing the characteristics and outliers of a dataset. The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur. It is denoted by γ_2 and defined as, $\gamma_2 = \frac{\mu_4}{\mu_2^2}$

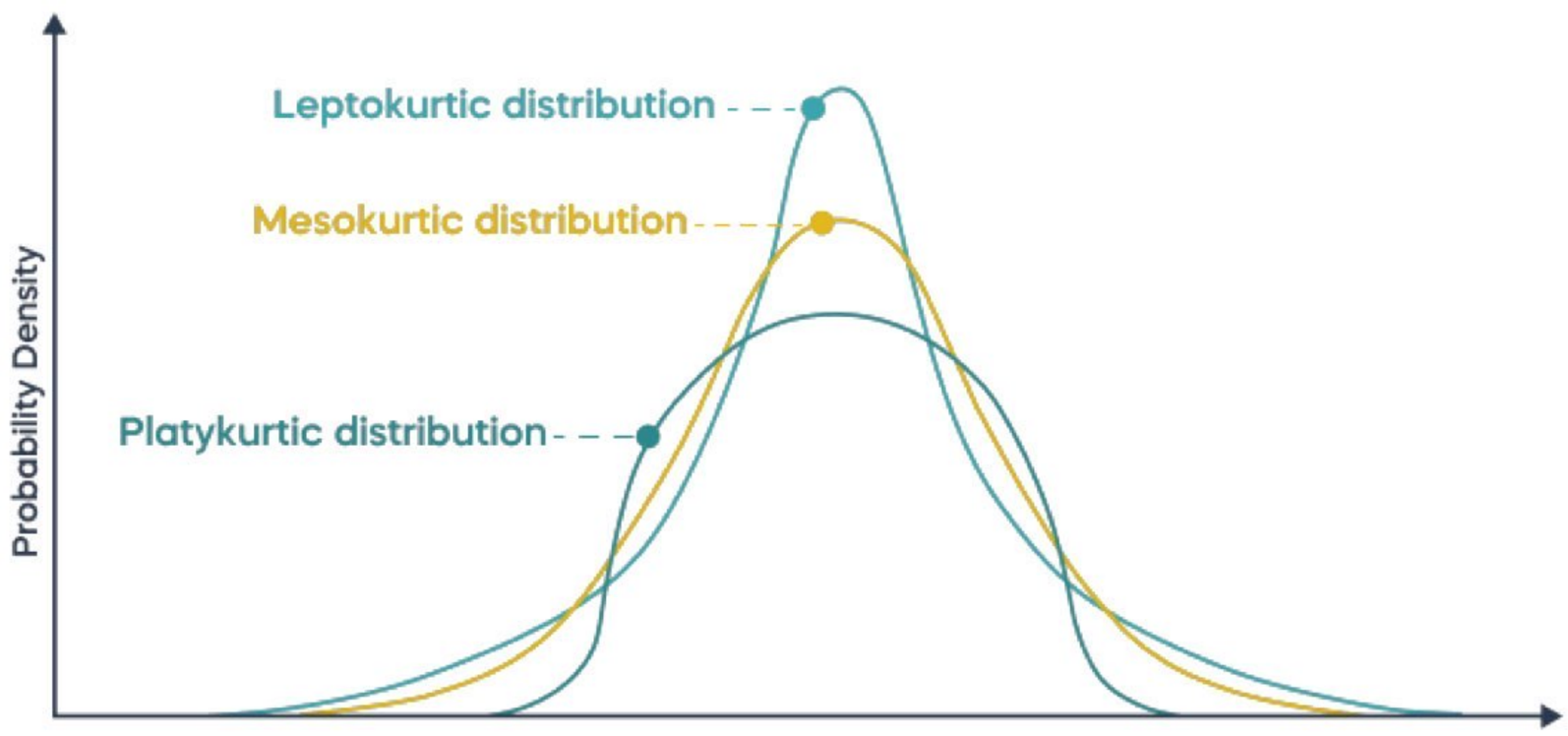
Graphically explanation of different types of Kurtosis: -

Mesokurtic ($Kurtosis = 3$):

1. The kurtosis of a mesokurtic distribution is approximately zero.
2. Mesokurtic is the same as the normal distribution.
3. In mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.
4. The data in a mesokurtic distribution is moderately spread around the mean.

Leptokurtic ($Kurtosis > 3$):

1. The key characteristic of a leptokurtic distribution is that its kurtosis is greater than zero.
2. Leptokurtic distributions have a peak that is taller and narrower than that of a normal distribution.
3. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.
4. Due to the heavier tails, leptokurtic distributions have an increased probability of outliers or extreme values.



Platykurtic($Kurtosis < 3$):

1. The key characteristic of a platykurtic distribution is that its kurtosis is less than zero.
2. Platykurtic distributions have a peak that is wider and shallower than that of a normal distribution.
3. Due to the lighter tails, platykurtic distributions have a decreased probability of outliers or extreme values.
4. The peak of a platykurtic distribution is lower, and the curvature of the distribution is broader compared to a normal distribution.

Mathematical Problems

❖ **Calculate the coefficient of Skewness and Kurtosis for the following dataset when mean is 2.86: -25, -10, -5, 0, 12, 18, 30**

Answer:

Given that, Mean, $\bar{X} = 2.86$

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^3$	$(X_i - \bar{X})^4$
-25	-27.86	776.18	-21624.36	602454.08
-10	-12.86	165.37	-2126.78	27350.41
-5	-7.86	61.78	-458.59	3816.72
0	-2.86	8.18	-23.40	66.91
12	9.14	83.54	763.55	6978.86
18	15.14	229.22	3470.38	52541.62
30	27.14	736.58	19990.17	542549.51

Now,

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$
$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 2060.9$$
$$\sum_{i=1}^{n}(X_i - \bar{X})^3 = -36.03$$
$$\sum_{i=1}^{n}(X_i - \bar{X})^4 = 1235758.11$$

Now calculate that,

$$\mu_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \frac{2060.90}{7} = 294.41$$
$$\mu_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n} = \frac{-36.03}{7} = -5.15$$
$$\mu_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n} = \frac{1235758.11}{7} = 176536.87$$

We have,

$$\text{Skewness, } \gamma_1 = \frac{\mu_3}{\sqrt{(\mu_2)^3}} = \frac{-5.15}{\sqrt{(294.41)^3}} = -0.001$$

$$\text{Kurtosis, } \gamma_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{176536.87}{(294.41)^2} = 2.04$$

Skewness is -0.001 indicates that dataset follows negatively skewed distribution and

Kurtosis is 2.02 indicates that dataset follows platykurtic distribution.

❖ **The first four moments of distribution about the value 5 are 2, 20, 40and 50 respectively obtain the first four central moments, skewness and kurtosis. Also comment on the shape of the distribution.**

Answer:

Central moment:

The first central moment is mean at the distribution which is 5 in this ease.

$$\mu_1 = 5$$

The second central moment is the variance

$$\mu_2 = \mu_2^{'} - \mu_1^2 = 20 - 2^2 = 16$$

The third central moment is the skewness which measures the asymmetry of the distribution.

$$\mu_3 = \mu_3^{'} - 3\mu_2^{'}\mu_1 + 3\mu_1^3 \quad = 40 - 3(20)(5) + 3(5)^3 \quad = -120$$

$$\gamma_1 = \mu_3 / \mu_2^{\frac{3}{2}}$$

$$= \frac{-120}{(16^2)^{\frac{3}{2}}}$$

$$= -0.875$$

The fourth central moment is kurtosis, which measures the "Peak" of the distribution.

$$\mu_4 = \mu_4^{'} - 4\mu_3\mu_1 + 6\mu_2^2 \quad = 50 - 4(-120)(5) + 6(16)^2 \quad = 15360$$

$$\gamma_2 = \mu_3 / \mu_2^2$$

$$= \frac{15360}{16^2}$$

Skewness: The skewness (-0.875) is negative. Indicating that the distribution is left skewed. This means that the tail of the distribution intends to the left.

Kurtosis: The kurtosis (60) is very high, indicating that the distribution is leptokurtic.

Shape of the distribution: Based on the calculated skewness and kurtosis, the distribution is left skewed and very peaked. This means that the tail intending to the left and a sharp peak around the mean.

- **For a distribution, Mean = 65, Median = 70 and coefficient of skewness = -0.5**

  Find the mode and coefficient of variation.

Answer:

Given, Mean $\bar{x}$ = 65; Median, M = 70; Skewness = - 0.5

We know that, $skewness = \frac{3(Mean-Madian)}{Standard\ davition}$

$$\Rightarrow \quad SD \quad = \frac{3(Mean-Median)}{skewness}$$

$$\Rightarrow \quad \text{SD} \quad = \frac{3(65-70)}{-0.5}$$

$$\Rightarrow \quad \text{SD} \quad = 30$$

Again, we know,

$$\text{Skewness} = \frac{(\text{Mean} - \text{Mode})}{\text{SD}}$$
$$\Rightarrow \quad Mode = Mean - (skewness \times SD)$$
$$\Rightarrow \quad Mode = 65 - (-0.5 \times 30)$$
$$Mode = 80$$

And,

$$CV = \frac{SD}{Mean} \times 100\%$$
$$\Rightarrow CV = \frac{30}{65} \times 100\%$$
$$CV = 46.15\%$$

Coefficient of variation = 46.15%, Mode = 80

❖ **For n moderately skewed data, $\bar{x} = 400, CV = 8\%$ and $\gamma_1 = 0.4$. Find mode and median.**

Answer:

Given, skewed data, $\bar{x} = 400, CV = 8\%, skewness = 0.40$

We know, $$CV = \frac{SD}{Mean} \times 100\%$$

$$\Rightarrow SD = \frac{CV \times SD}{100}$$
$$\Rightarrow SD = \frac{8 \times 400}{100}$$
$$= 32$$
$$skewness = \frac{Mean - Mode}{SD}$$

Here, $Mode = Mean - (skewness \times SD)$

$$\Rightarrow Mode = 400 \ (0.40 \times 32)$$
$$= 387.2$$
$$Mode = 387.2$$

Again know,

$$skewness = \frac{3(Mean - Median)}{SD}$$
$$\Rightarrow Median = \frac{3Mean - (SD \times skewness)}{3}$$
$$= \frac{3 \times 400 - (32 \times 0.40)}{3}$$
$$Median = 395.73$$

$Mode = 387.2, Median = 395.73$

-The end-

STAT-1211: Statistics & Probability

Department of Computer Science and Engineering

BSc Honors 1st Year 2nd Semester Exam-2023

Exam Session: 2022-23

Admission Session: 2022-23



Lecture Note 03

Md. Ayub Ali
Assistant Professor
Department of Statistics
University of Barishal, Barishal-8254
maali@bu.ac.bd

Correlation and Regression

Correlation Analysis

❖ **Correlation and correlation coefficient. Its types with graphical plot.**

Answer:

Correlation: Correlation is a statistical measures that expresses the extent to which two (more) quantitative variables are linearly related.

Correlation Coefficient (CC): Correlation Coefficient is a statistical measures that quantifies linear relationship between a pair (or more) of quantitative variables.

Correlation coefficient functional form:

$$r = \frac{Cov(x,y)}{\sqrt{V(x) \times V(y)}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}$$

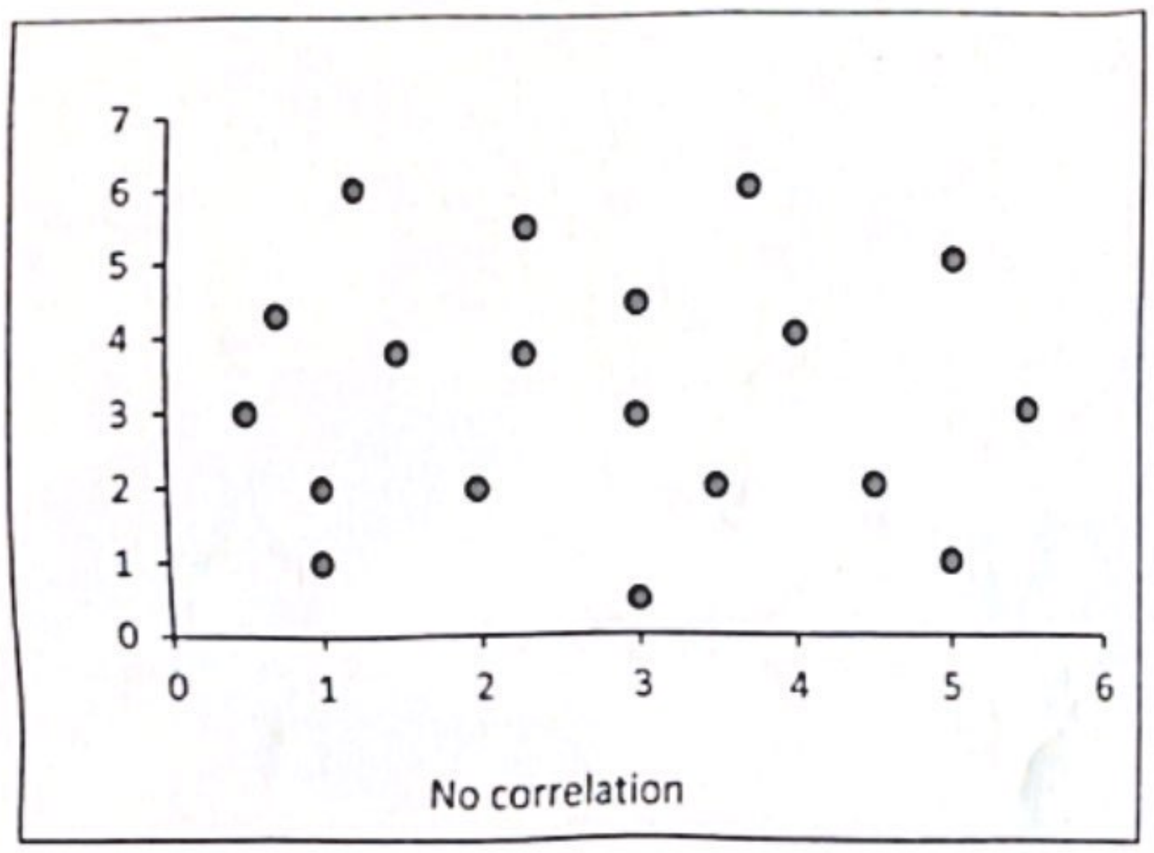
Types of Correlation: There can be three such situations to see the relation between the two variables.

Positive Correlation: When the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by increase/decrease of the other variable.

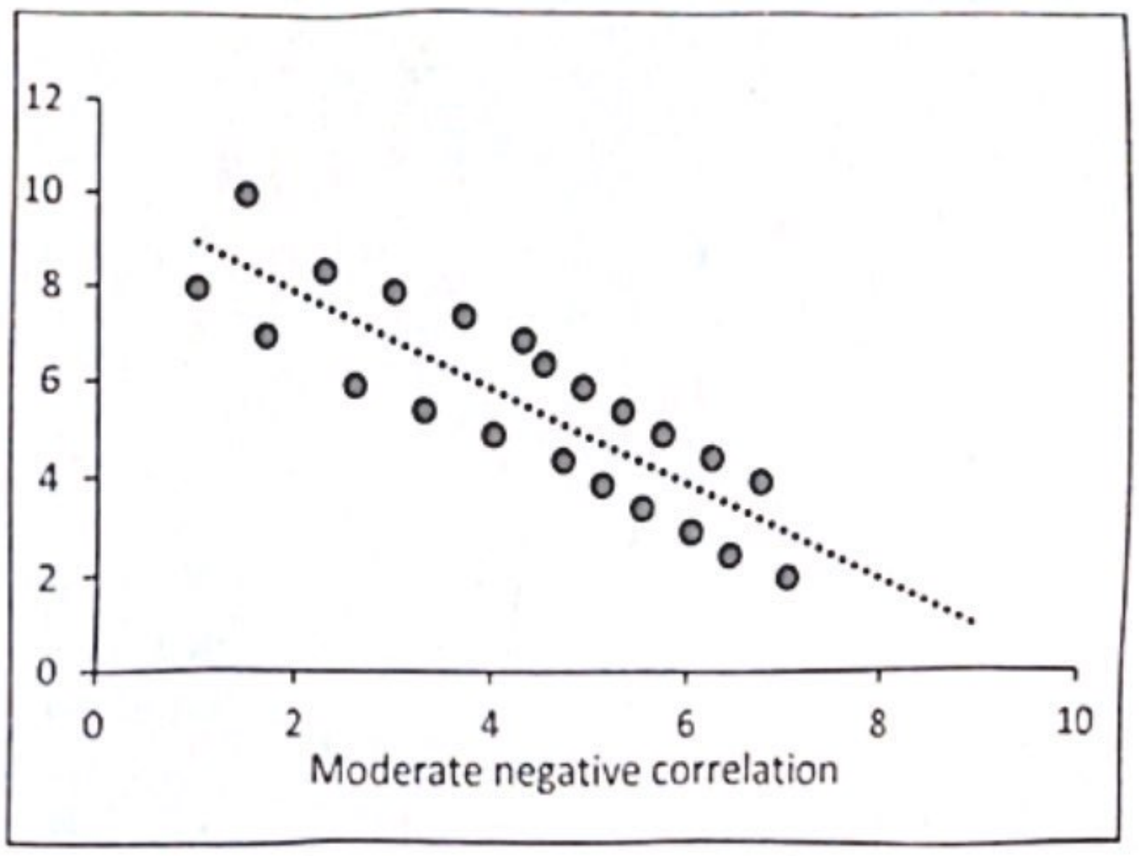
Negative Correlation: When the values of two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.

No Correlation: When there is no linear dependence or no relation between the two variables.

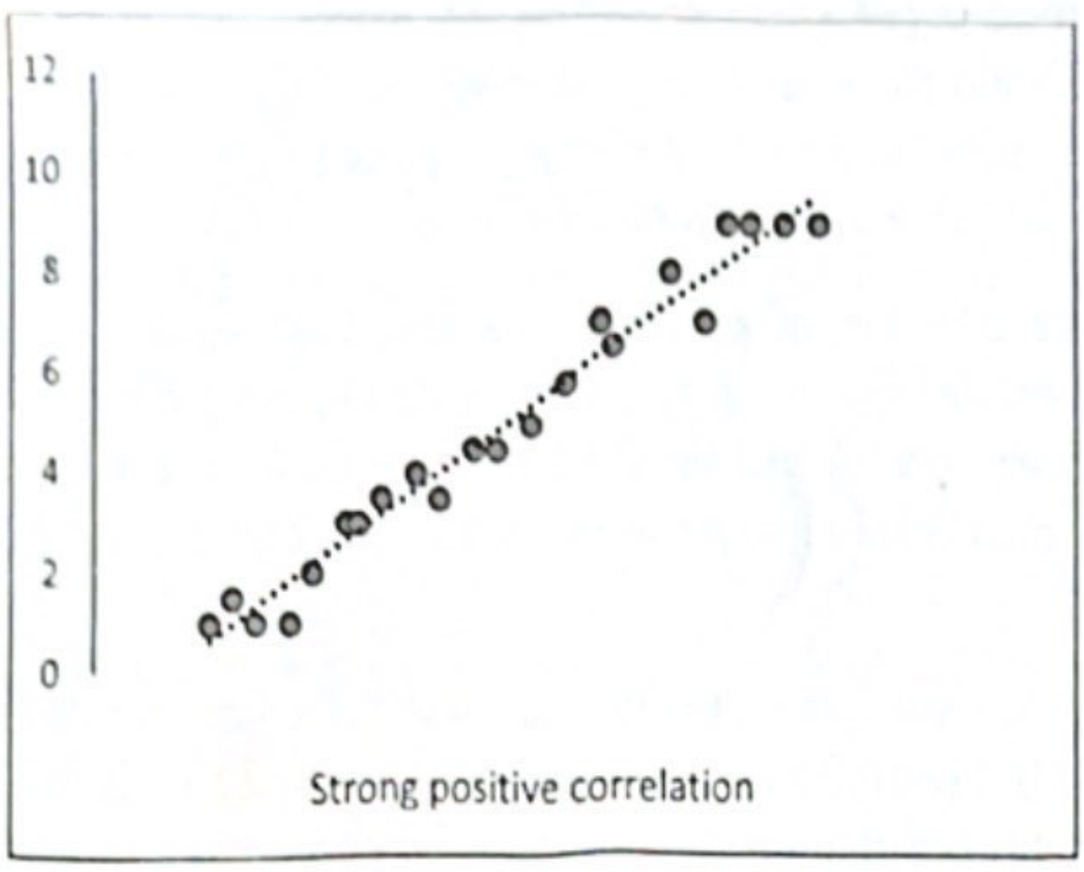
Five types of simple correlation:



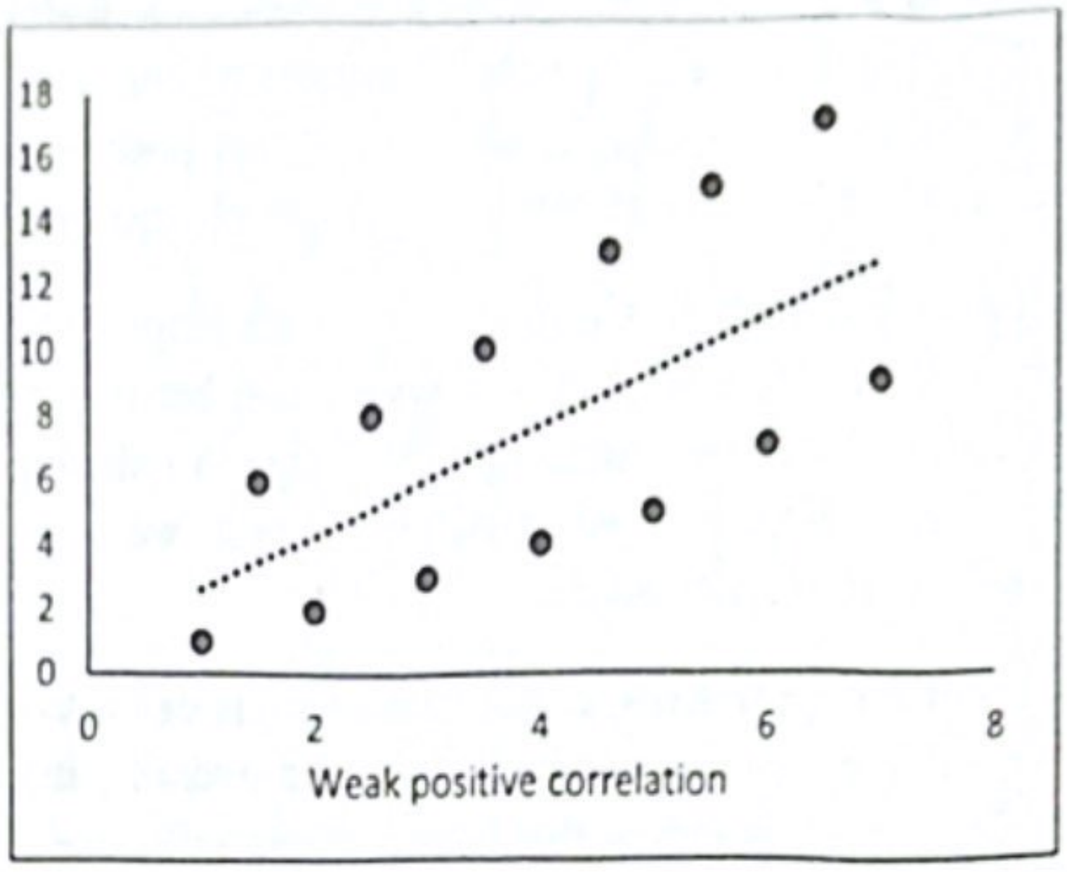
No correlation



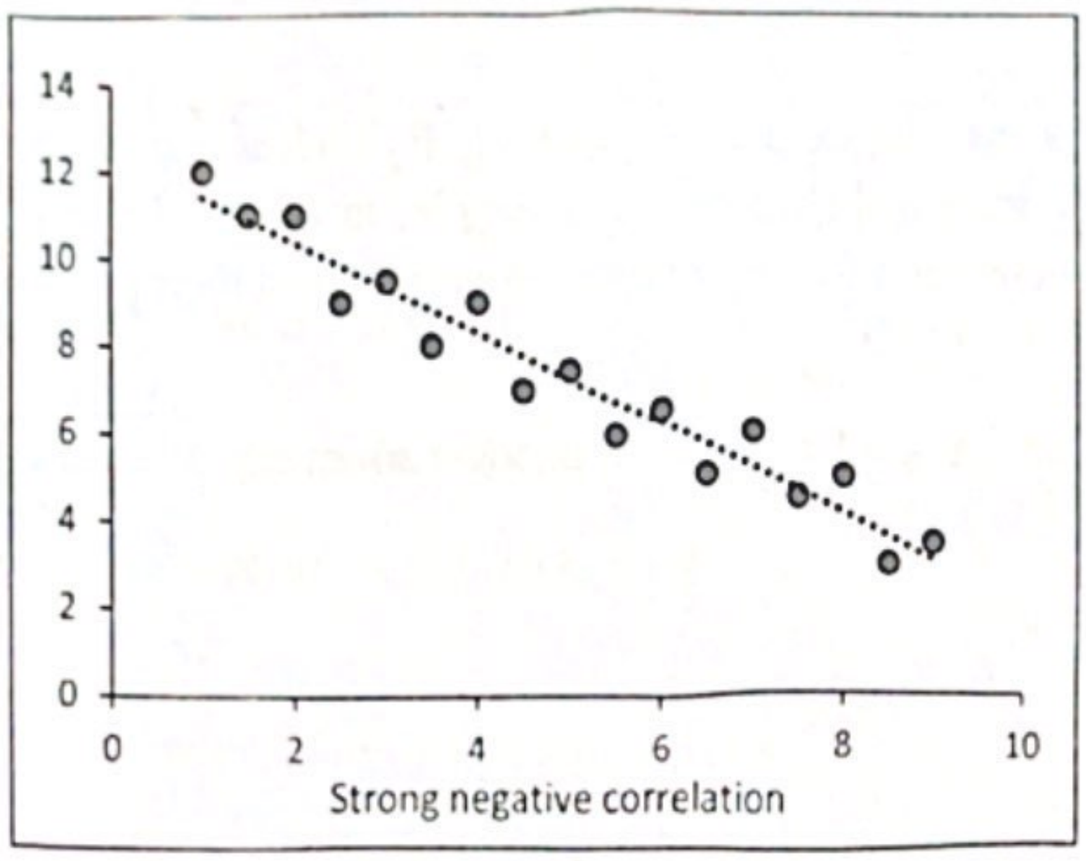
Moderate negative correlation



Strong positive correlation



Weak positive correlation



Strong negative correlation

Types of Correlation coefficient:

Correlation coefficient	Relationship	Levels of measurement	Distribution
Pearson's r	Linear	Two Quantitative variable interval or ratio	Normal distribution
Spearman's rho	Non-linear	Two Ordinal variable interval or ratio	Any distribution
Cramer's V	Non-linear	Two Nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal variable	Any distribution

❖ **Uses of correlation coefficient (CC) in statistics. Properties of correlation coefficient.**

Answer:

The uses of CC:

1. Understanding Relationships: Correlation analysis plays a crucial role in data analysis by helping us understand the relationships between variables. By examining the correlation coefficient, we can determine the strength and direction of the relationship between two variables. For example, in a study analyzing the relationship between hours of study and exam scores, a positive correlation indicates that as study hours increase, exam scores also tend to increase.

2. Predictive Modeling: Correlation analysis is widely used in predictive modeling. By identifying variables that are highly correlated with the target variable, we can build more accurate predictive models. For instance, in a marketing campaign, if we find a strong positive correlation between customer age and purchase behavior, we can use age as a predictor to target specific age groups for better campaign effectiveness.

3. Feature Selection: Correlation analysis helps in feature selection, especially in machine learning and data mining tasks. By examining the correlation between features and the target variable, we can identify the most relevant features for prediction. This helps in reducing dimensionality and improving model performance. For example, in a spam email classification task, we can use correlation analysis to select features that are highly correlated with spam emails.

4. Identifying Outliers: correlation analysis can also help in identifying outliers in a dataset. Outliers are data points that deviate significantly from the overall pattern. By examining the correlation between variables, we can identify observations that have a strong influence on the correlation coefficient. These outliers may indicate data quality issues or unusual patterns in the data.

5. Decision Making: Correlation analysis provides valuable insights for decision making. By understanding the relationships between variables, decision-makers can make informed choices. For example, in financial analysis, correlation analysis helps in assessing the relationship between different investment options and their potential risks and returns.

Remember, correlation analysis is a powerful tool in data analysis, allowing us to uncover meaningful insights and make informed decisions.

The properties of CC:

a) The correlation coefficient is a symmetric measures.

b) The correlation coefficient will be negative or positive.

c) The correlation coefficient lies between -1 to +1.

d) The correlation coefficient is a dimensionless quantity.

e) The coefficient of correlation is independent of origin and scale of measurement.

❖ **Coefficient of correlation between x and y lies between -1 to +1 , i.e. $-1 \leq r \leq +1$.**

Answer:

The coefficient of correlation between x and y lies between -1 to +1 , i.e. $-1 \leq r \leq +1$.

Let us consider the following expression

$$>\left(\frac{x_i-\bar{x}}{S_x} \pm \frac{y_i-\bar{y}}{S_y}\right) \ldots\ldots\ldots (i)$$

Which when is squared is always positive,

$$>\left(\frac{x_i-\bar{x}}{S_x} \pm \frac{y_i-\bar{y}}{S_y}\right)^2 \geq 0$$

$$>\frac{\Sigma(x_i-\bar{x})^2}{S_x^2} + \frac{\Sigma(y_i-\bar{y})^2}{S_y^2} \pm \frac{2\Sigma(x_i-\bar{x})(y_i-\bar{y})}{S_x S_y} \geq 0 \ldots\ldots\ldots (ii)$$

Now recall that,

$>\sum(x_i - \bar{x})(y_i - \bar{y}) = rnS_xS_y$

$>nS_x^2 = \sum(x_i - \bar{x})^2, nS_y^2 = \sum(y_i - \bar{y})^2$

Subtracting those values in (ii)

$>\frac{nS_x^2}{S_x^2} + \frac{nS_y^2}{S_y^2} \pm \frac{2nrS_xS_y}{S_xS_y} \geq 0$

$>2n + 2nr \geq 0$

$>1 \pm r \geq 0$

$>1 + r \geq 0,\ r \geq -1\ and\ 1 - r \geq 0,\ r \leq +1$

So $-1 \leq r \leq +1$ (proved)

❖ **The coefficient of correlation of independent of origin and scale of measurement**

Answer:

The coefficient of correlation of independent of origin and scale of measurement:

Let correlation coefficient between x and y

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

We now transform x and y to u and v

$u_i = \frac{x_i - a}{h}$ and $v_i = \frac{y_i - b}{k}$

$x_i = a + u_i h$ $y_i = b + kv_i$

Resulting in

$$\bar{x} = a + h\bar{u}^{\cdot} \quad and \quad \bar{y} = b + k\bar{v}$$

here, a and b are the origin factors and h and k are the scale factors. We shall demonstrate, that do not have any effect on the correlation coefficient that is, $r_{xy} = r_{uv}$.

$$r_{xy} = \frac{\sum(a + hu_i - a - h\bar{u})(b + kv_i - b - k\bar{v})}{\sqrt{\sum(a + hu_i - a - h\bar{u})^2 \sum(b + kv_i - b - k\bar{v})^2}}$$
$$= \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{(u_i - \bar{u})^2 (v_i - \bar{v})^2}}$$
$$= r_{uv} \qquad (Proved)$$

❖ **Correlation coefficient is the geometric mean coefficient of the regression coefficient**

Answer:

HW

❖ **Correlation coefficient in at most of the arithmetic mean of the regression coefficient**

Answer:

HW

❖ **Find out the CC for the following data of the sales and expenses:**

Sales (Tk.)	65	63	67	64	68	62	70	66	68	67
Expenses (Tk.)	68	66	68	65	69	66	68	65	71	67

Answer:

Funding out the correlation coefficient (cc) for the following data of scales and expenses:

Sales	Expanses	dx	dy	dx^2	dy^2	dxy
65	68	-2	1	4	1	-2
63	66	-4	-1	16	1	4
67(A)	68	0	1	0	1	0
64	65	-3	-2	9	4	6
68	69	1	2	1	4	2
62	66	-5	-1	25	1	5
70	68	3	1	9	1	3
66	65	-1	-2	1	4	2
68	71	1	4	1	16	4
67	67(A)	0	0	0	0	0
	N=10	$\sum dx = -10$	$\sum dy = 3$	$\sum dx^2 = 66$	$\sum dy^2 = 33$	$\sum dxy = 24$

$$r = \frac{N\sum dx\,dy - \sum dx \sum dy}{\sqrt{N\sum dx^2 - (dx)^2} \times \sqrt{N\sum dy^2 - (dy)^2}}$$

Let's substitute those values into the formula:

$$r = \frac{10\times 24 - (-10)\times 3}{\sqrt{10\times 66 - (10)^2} \times \sqrt{10\times 33 - (3)^2}}$$

$$= \frac{240+30}{\sqrt{560\times 321}}$$

$$= \frac{270}{423.98}$$

$$r \approx 0.6368$$

The value of r is approximately 0.6368 which is strong positive correlation.

- **For twelve observation in X and Y, following results are obtained, $\sum x = 30$, $\sum y = 5$, $\sum x^2 = 670$, $\sum y^2 = 285$ and $\sum xy = 334$. Find the correlation coefficient and comment.**

Answer:

For twelve observation X and Y, the correlation coefficient 'r' can be calculated using the formula:

$$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

Given,

N=12, $\sum x$ =30; $\sum y = 5$; $\sum x^2$=670; $\sum y^2 = 285$; $\sum xy = 334$

Let's substitute these value into the formula:

$$r = \frac{12\times 334 - 30\times 5}{\sqrt{12\times 670 - (30)^2} \times \sqrt{12\times 285 - (5)^2}}$$

$$= \frac{4008-150}{\sqrt{7140} \times \sqrt{3395}}$$

$$= \frac{3858}{\sqrt{24228900}}$$

$r \approx 0.7838$

The value of 'r' is approximately 0.7838 which is strong positive correlation coefficient.

- **For twelve observation in X and Y, following results are obtained, $\sum x = 30$, $\sum y = 5$, $\sum x^2 = 670$, $\sum y^2 = 285$ and $\sum xy = 334$. On subsequent verification it was found that pair $(x = 11, y = 4)$ was wrongly copied and the correct value being $((x = 10, y = 14))$ Find the corrected correlation coefficient.**

Answer:

The given initial values:

$\sum x = 30 \qquad \sum x^2 = 670$

$\sum y = 5 \qquad \sum y^2 = 285$

$\sum xy = 334$

Let's correct the values based on the given correction:

Initial value:(x=11, y=4)

Corrected value:(x=10, y=14)

This correction effects the summations. Let's update the values accordingly:

Corrected sum of x, $\sum x = 30 - 11 + 10 = 29$

Corrected sum of y, $\sum y = 5 - 4 + 14 = 15$

Corrected sum of x squared, $\sum x^2 = 670 - (11)^2 + (10)^2 = 649$

Corrected sum of y squared, $\sum y^2 = 285 - 4^2 + (14)^2 = 465$

Corrected sum of xy, $\sum xy = 334 - 11 \times 4 + 10 \times 14 = 430$

Substitute the corrected values into the formula:

$$r = \frac{N \times \sum xy - \sum x \sum y}{\sqrt{N \times (\sum x^2 - (\sum x)^2) \times (\sum y^2 - (\sum y)^2)}}$$

$$= \frac{12 \times 430 - 15 \times 29}{\sqrt{12 \times (649 - 29^2) \times (465 - 15^2)}}$$

$$= \frac{4725}{6494.98}$$

$$\approx 0.727$$

Therefore, after correcting the values, the updated correlation coefficient is approximately 0.727

Regression Analysis

❖ **Regression analysis. Functional form of it and some of its types.**

Answer:

Regression analysis(RA): Regression analysis is a statistical technique that serves as a basis of studying the dependence of one variable called dependent variable, on one or more other variables called independent or explanatory variable(s).

Functional formula of Regression analysis:

$$Y_i = f(x_i, \beta) + e_i .$$

Y_i= dependent variable.

f = function.

x_i = independent variable.

β = unknown parameters.

e_i = stochastic error term.

Different type of Regression Analysis:

1. Linear regression.	6. Poisson regression.	11. Ordinal regression.
2. Polynomial regression.	7. Elastic net regression.	12. Quasi Poisson regression.
3. Logical regression.	8. Principle components regression.	13. Negative Binomial regression.
4. Quantile regression.	9. Partial least regression.	14. Cox regression.
5. Ridge regression.	10. Support vector regression.	15. Tobit regression.

❖ **Plot linear regression and logistic regression (NA for CSE).**

Answer:

Graphically explanation of regression models that use in statistics:

1. **Linear regression:**

The equation is used to denote the linear regression model:

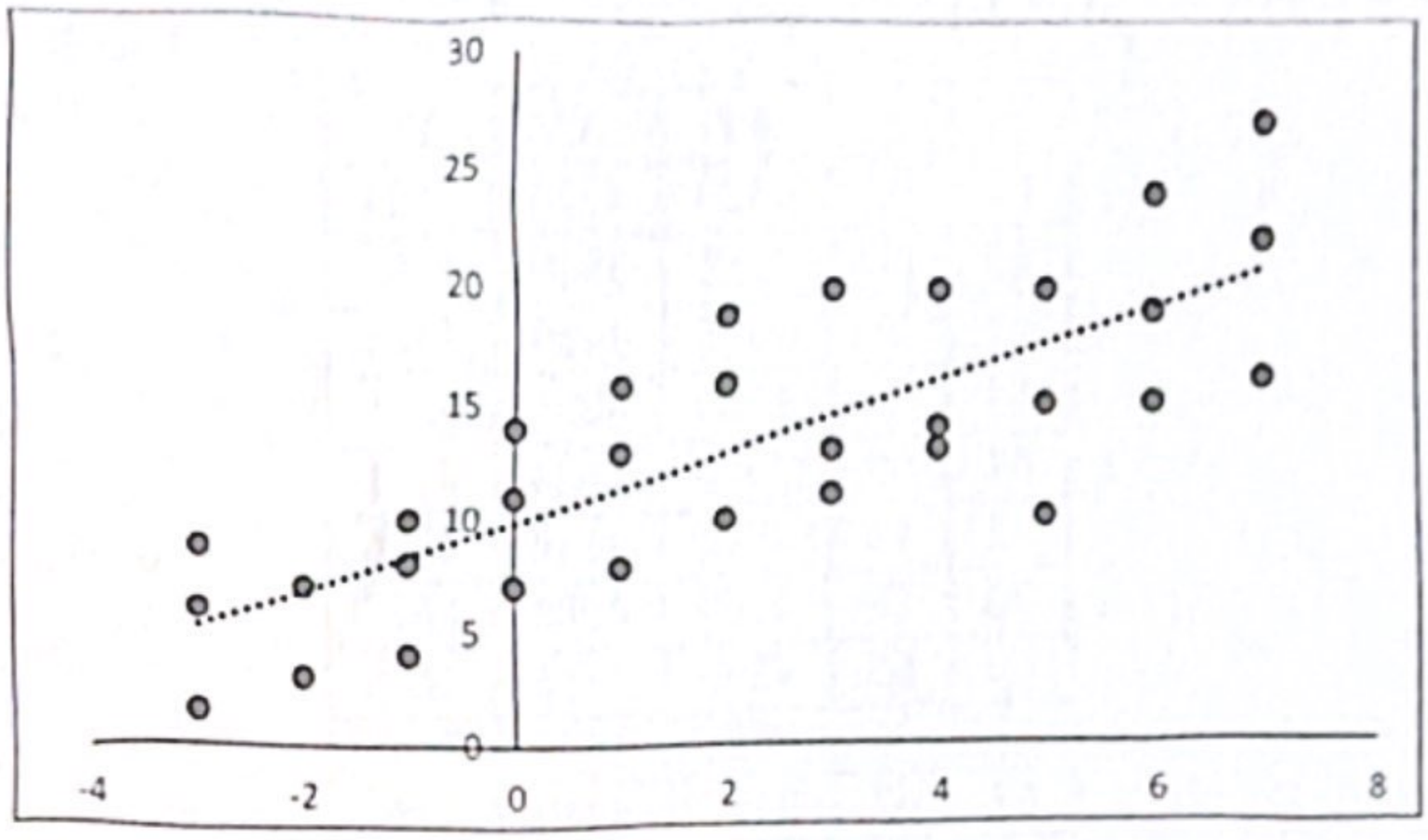
$$y = mx + c + e.$$

here,

m=slope of the line.

c = intercept.

e = error in the model.

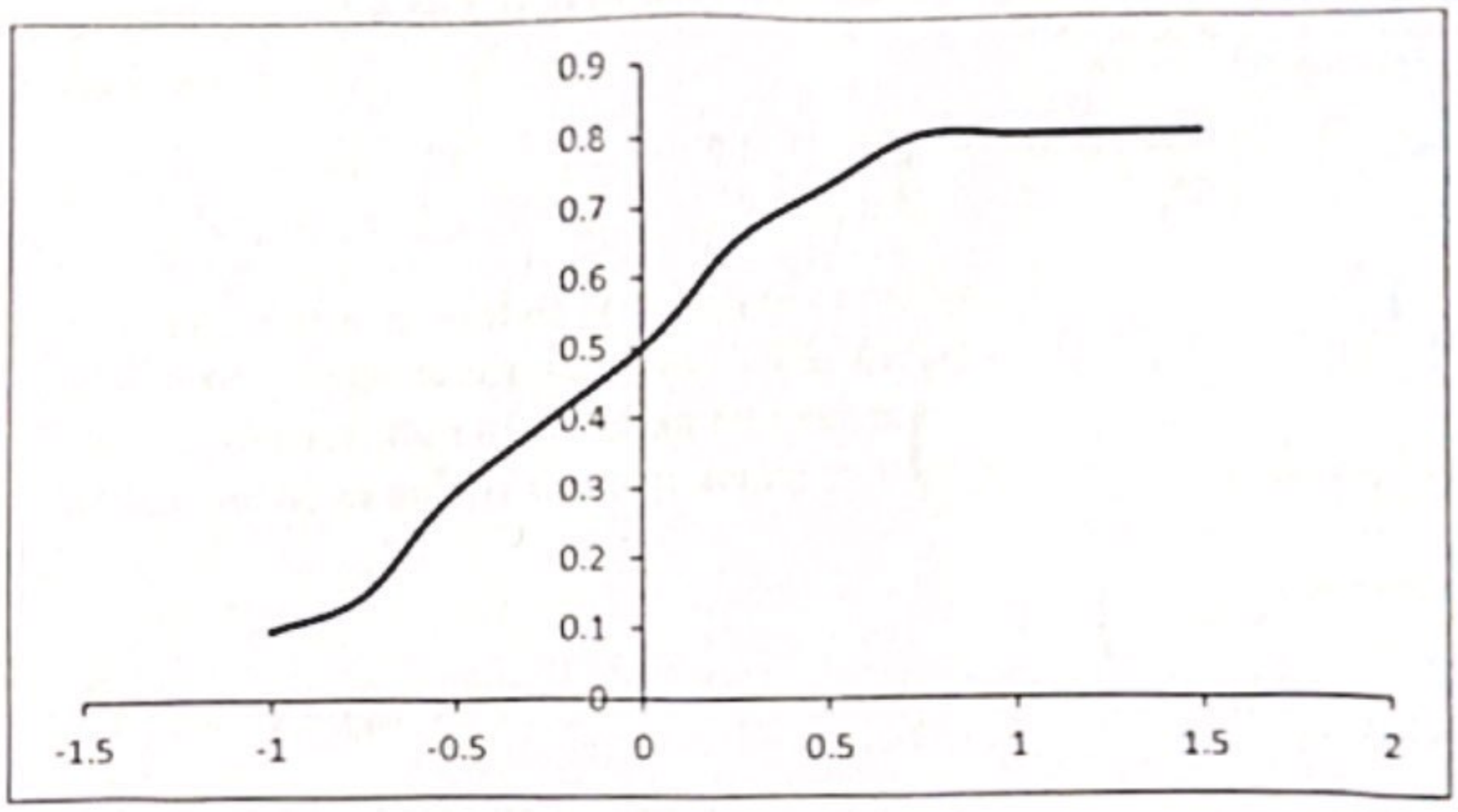


2. **Logistic Regression:**

The equation that denotes this,

$$\log(P) = \ln(P/(1-P)) = b_o + b_1x_1 + b_2x_2 + \cdots\cdots\cdots + b_kx_k$$

Where p is the probability of occurrence of the future.



Uses of Regression analysis in CSE:

HW

The properties of regression Analysis:

HW

- **By using the following data, estimate regression equation of y on x and x on y.**

x	65	63	67	64	68	62	70	66	68	67
y	68	66	68	65	69	66	68	65	71	67

Answer:

Regression equation of y on x for given data:

X	Y	XY	X^2	Y^2
65	68	4420	4225	4624
63	66	4158	3969	4356
67	68	4556	4489	4624
64	65	4160	4096	4225
68	69	4692	4624	4761
62	66	4092	3844	4356
70	68	4760	4900	4624
66	65	4290	4356	4225
68	71	4828	4624	5041
67	67	4489	4489	4489
660	673	44445	43616	45325

Equation y on x:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\Rightarrow (y - \frac{\Sigma y}{n}) = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}(x - \frac{\Sigma x}{n})$$

$$\Rightarrow y - \frac{673}{10} = \frac{10\times44445 - 660\times673}{10\times43616 - 660\times660}(x - \frac{660}{10})$$

$$\Rightarrow y - 67.3 = 0.482(x - 66)$$

Equation x on y:

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$\Rightarrow x - \frac{\Sigma x}{n} = \frac{n(\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}\left(y - \frac{\Sigma y}{n}\right)$$

$$\Rightarrow x - \frac{660}{10} = \frac{10\times44445 - 660\times673}{10\times44938 - 673\times673}\left(y - \frac{673}{10}\right)$$

$$\Rightarrow x - 66 = -0.0761(y - 63.3)$$

$$\Rightarrow x - 66 + 0.0761y - 5.12 = 0$$

$$\Rightarrow x + 0.76y - 71.12 = 0 \quad \underline{\text{Ans.}}$$

❖ **Estimate the regression equation of y on x and x on y when regression coefficient is 0.61 for the following dataset**

	Advertising expenditure (Lakh)	Sales (Crore)
Mean	35.00	82.00
Standard deviation	9.03	17.16

Answer:

Given here,

Regression coefficient = 0.61

∴ mean of x, $\bar{x} = 35$

And mean of y, $\bar{y} = 82$

Equation: y on x

$y - \bar{y} = b_{xy}(x - \bar{x})$

=>y – 82=0.61(x–35)

=>y–82–0.61x + 21.35=0

=>0.61x –y +60.65=0 Ans.

And equation x on y:

$x - \bar{x} = b_{xy}(y - \bar{y})$

=>x –35 = 0.61(y – 82)

=>x–35 – 0.617 + 50.02 = 0

=>x–0.61y + 50.02 = 0 Ans.

- **The following data give the test scores and sales made by nine salesmen during the last year of a big departmental store.**

Test scores (y):	**14**	**19**	**24**	**21**	**26**	**22**	**15**	**20**	**19**
Sales in lakh taka (x):	**31**	**36**	**48**	**37**	**50**	**45**	**33**	**41**	**39**

i. **Find the regression model of test scores on sales.**
ii. **Predict the test score when the sale is Tk. 40 lakh.**
iii. **Find the regression line of sales on test scores.**
iv. **Predict the value of sale if the test score is 30.**

Answer:

The given data:

Test scores(y): 14 19 24 21 26 22 15 20 19

Sales in lakh(Tk.): 31 36 48 37 50 35 33 41 33

(i) Find the regression model of test scores on sales:

First compute mean:

$\bar{x} = \frac{31+36+48+37+50+45+33+41+39}{9} = \frac{360}{9} = 40$

$\bar{y} = \frac{14+19+24+21+26+22+15+20+19}{9} = \frac{180}{9} = 20$

Calculate the slope (b):

$\sum_{i=1}^{9}(x_1-\bar{x})\,(y_1-\bar{y}) = 192$

$=>\sum_{i}{}^{9}{}_{=1}(x_1-\bar{x})^2 = 346$

$=>b=\frac{192}{346} \approx 0.5545$

Next find the intercept(a):

$a=\bar{y}-b\times\bar{x}$

$=>a=20-0.5545\times 40\approx 1.78$

The regression equation for test scores on sales is:

$y=-1.78+0.55x$

(ii) Prediction for sales of taka 40 lakhs:

Using the regression equation $y=-1.78+0.55x$

When $x=40$

$\therefore y=-1.78+0.5545\times 40$

$=-1.78+22.18\approx 20.4$

The predicated score when the state is tk. 40 lakhs is approximately 20.4.

(iii) Regression line of scales on that score:

Reverse the variable in the equation $x=\hat{a}+\hat{b}y$

Calculate $\hat{a}$:

$\hat{a}=\bar{x}-b_x\bar{y}$

$\hat{a}=40-o.5545\times 20\approx 28.1$

calculate b'

$b'=1/b$

$b=\frac{1}{0.5545}\approx 1.803$

The regression equation for sales on test score is $x=28.1+1.803y$

(iv) Prediction of scale value if the test score is 30

Using the regression equation

$x=28.1+1.803y$, where $y=30$.

$X=28.1+1.803\times 30$

$=28.1+54.09\approx 82.19$

The predicated value of scale if the score is 30 is approximately tk 82.19 lakh.

-The End-

Department of Computer Science and Engineering
BSc Honors 1st Year 2nd Semester Exam-2023
Exam Session: 2022-23
Admission Session: 2022-23



Lecture Note 04

Md. Ayub Ali
Assistant Professor
Department of Statistics
University of Barishal, Barishal-8254
maali@bu.ac.bu

Set Theory and Key Elements of Probability

Set

❖ **Definition**

Answer: A **set** is a well-defined collection of objects which have some properties in common.
The sets are usually denoted by the capital letters A, B, C or X, Y, Z etc. Elements of a set are denoted by lower case letters such as a, b, x, y etc. The elements of the set are enclosed in second bracket {}. Suppose A is a set with first 10 natural numbers, then A can be written as
$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.
The numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 are the elements of the set Number of elements of a set A is denoted by $N(A)$.
Here $N(A) = 10$.

Example 2.1.1

i. The vowels of English alphabets may be a set A which can be written as
$$A = \{a, e, i, o, u\}.$$
ii. Let B be the set of all natural numbers, then
$$B = \{1,2,3,\ldots\ldots\}.$$
iii. Suppose C is the set of all natural even numbers, then
$$C = \{2,4,6\}.$$
iv. The students Rahim, Karim, Ram, John may constitute a set D a set D as
$$D = \{\text{Rahim, Karim, Ram, John}\}.$$
v. The solution of the equation $x^2 - 6x + 4 = 0$ is a se
$$\text{E} = \{x \mid x^2 - 6x + 8 = 0\}.$$
vi. The capital cities of Asia is a set F can be defined as
$$F = \{x \mid x \text{ is a capital city in Asia}\}.$$
vii. The rivers of Bangladesh is a set (say) G, Then
$$G = \{x \mid x \text{ is river is Bangladesh}\} \text{ etc.}$$

Universal Set

❖ **Definition 2.1.7** A set is called universal set if it contains all possible objects under consideration in a given discussion.

Universal set is generally denoted by Ω, or U or S.

Example 2.1.5 If we throw a die, the universal set is the all possible outcomes which maybe written as,
$$\Omega = \{1,2,3,4,5,6\} = \{x \mid 1 \leq x \leq 6 \text{ and } x \text{ is a natural number}\},$$
When 1, 2, 3, 4, 5 and 6 are the number of points on the faces of the die.

Equality of Sets

❖ **Definition 2.1.9. Equal Set.** Two sets A and B are said equal, if they contain exactly the same elements.

If set A is equal to set B, then every element of set A is also an element of set B and vice-versa. Equality of two sets A and B is denoted by A = B or A ⊂ B and B ⊂ A.

Example 2.1.7.

i. The set A = {1, 4, 7} is equal to set B = {7, 1, 4} since they both have exactly the same elements. It is noted that a set does not change when the order of the elements is rearranged.
ii. A = {1,2,3,4}, B = {1,1,2,2,2,3, 3, 3, 4, 4, 4}
The sets A and B are equal, since they contain the same elements 1, 2, 3 and 4. That is each element of A belongs to B and each element of B belongs to A. Note that a set does not change if its elements are repeated. In practice, set B is written as A. Repetition of elements of a set does not change its operation; even repetition of elements is infinite.
iii. A (a, c, t) = {x | x is alphabet in the word cat}.
However, when the elements of A are not the same as of set B, they are not equal and we write $A \neq B$.
iv. Let $A = \{2,4,6,8\}$, $B = \{6,4,2,8\}$, $C = \{6,4,8,2,7\}$
Then $A = B$ but $A \neq C$ and $B \neq C$.

Subset

❖ **Definition 2.1.10.** A set B is said to be a subset of C if every element of B is also an element of C. Symbolically, we write this $B \subset C$, where $\subset$ means "is a subset of" or "is contained in".

Example 2.1.8.

i. Let A = {1, 2, 3, 4, 5, 6} and B = {1, 3, 5} then the set B is a subset of set A.

ii. Suppose $A = \{x \mid x$ is the height of student in a class$\}$, $B = \{x \mid x$ is the height of girl student in the same class$\}$, then B is the subset of A. According to definition 2.1.10 every set is a subset of itself.

Remark 1. The null set is considered to be a subset of every set.

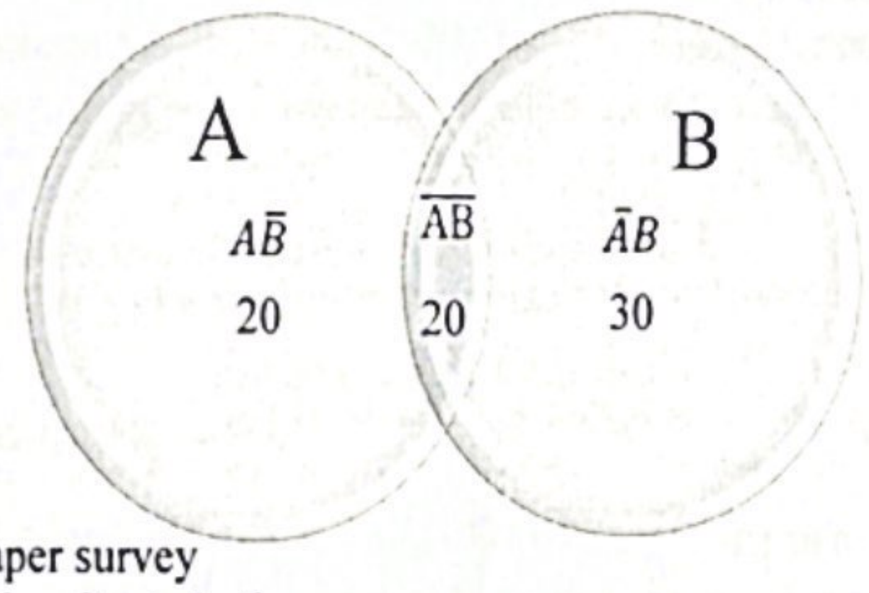
Remark 2. If A is not a subset of B, that is $A = B$ then there is at least one element in A that is not a member of B.

HW: Mutually exclusive sets, Mutually unexclusive sets, Null set, Singleton set, Set differences, Ven-diagram of sets etc.

❖ **Example 2.3.2. Market Research. A city has two daily newspapers, the Azadi and the Purbakon. The following information was obtained from a survey of 100 residents of the city: 50 people subscribe to the Azadi, 60 subscribe to the Purbakon, and 30 subscribe to both newspapers.**

i. **How many people in the survey subscribe to the Azadi but not to the Purbakon?**
ii. **How many subscribe to the Purbakon but not to the Azadi?**
iii. **How many do not subscribe to either paper?**
iv. **How many subscribe at least one paper? (v) Sketch a Venn- diagram and count the sample points.**
v. **Sketch a Venn- diagram and count the sample points.**

Solution. Let Ω be the group of people surveyed. Let A be the set of people subscribe to the newspaper Azadi, and let B be the set of people who subscribe to the newspaper Purbakon. Let $\bar{A}$ and $\bar{B}$ be the complements of A and B. Using the sets, A and B, their complements, and set intersection, we can divide the set Ω into four disjoint subsets defined below and illustrate in the diagram in figure 1.



Venn-diagram for the newspaper survey

$A \cap B$ = Set of people who subscribe to both papers.

$A \cap \bar{B}$ = Set of people who subscribed to the Ajadi but not Purbakon.

$\bar{A} \cap B$ = Set of people who subscribed to the Purbakon but not the Ajadi.

$\bar{A} \cap \bar{B}$ = Set of people who do not subscribe to either paper.

The given survey information can be expressed in terms of set notation as

i. $N(\Omega) = 100, N(A) = 50, N(B) = 60, N(AB) = 30$
ii. $N(A \cap \bar{B}) = N(A) - N(AB) = 50 - 30 = 20.$
iii. $N(\bar{A} \cap B) = N(B) - N(AB) = 60 - 30 = 30.$
iv. $N(\overline{AB}) = N(\bar{A}) - N(\bar{A}B) = 50 - 30 = 20.$

❖ **Suppose A, B and C are three events in Ω. Find the expression for the events noted below, in the context of A, B and C.**

i. **Only A occurs,**
ii. **Both A and B occurs but not C**
iii. **All three events occur**

iv. At least one occurs
v. At least two occurs
vi. One and no more occurs
vii. Two and no more occurs
viii. None occurs and
ix. Exactly A or C occur.

Answer:

i. Only A occurs, means $A\overline{BC} = A \cap \bar{B} \cap \bar{C}$
ii. Both A and B occurs but not C, means $A \cap B \cap \bar{C}$
iii. All three events occur, means $ABC = A \cap B \cap C$
iv. At least one occurs, means $A \cup B \cup C$
v. At least two occurs, means $AB\bar{C} \cup A\bar{B}C \cup \bar{A}BC \cup ABC = A \cap B \cap \bar{C} \cup A \cap \bar{B} \cap C \cup \bar{A} \cap B \cap C \cup A \cap B \cap C$
vi. One and no more occurs, means $A\overline{BC} \cup \bar{A}B\bar{C} \cup \overline{AB}C$
vii. Two and no more occurs, means $AB\bar{C} \cup A\bar{B}C \cup \bar{A}BC$
viii. None occurs, means $\overline{ABC}$
ix. Exactly A or C occur means, $A\Delta C = A\overline{BC} \cup \overline{AB}C$

Probability Theory

❖ **Probability**

Answer:

Probability can be defined as the radio of the number of favorable outcomes to the total number of outcomes of an **event**. For an experiment having n number of outcomes, the number of favorable outcomes can be denoted by x. Hence there are four approaches of defining probability. They are

i. Classical or mathematical probability.
ii. Empirical or statistical or frequency probability.
iii. Subjective probability and
iv. Axiomatic probability.

- **Experiment:** Experiment is an act that can be repeated under given conditions.
- **Deterministic Experiment:** The experiment is called deterministic when the outcome or result is unique or certain.
- **Random Experiment:** An experiment is called random experiment whose outcomes cannot be predicted with certainty.
- **Trial:** Unit experiment is known as trial.
- **Outcomes:** The results of an experiment are known as outcomes.
- **Events:** One or more outcomes of an experiment constitute an event. Events are generally denoted by capital letters A, B, C, etc.
- **Mutually exclusive outcomes:** Outcomes or cases are said to be mutually exclusive if the happening of any one of the precludes the happening of all others.
- **Favorable outcomes:** The outcomes of an experiment are said to be favorable to an event if they entail the happening of the event.
- **Exhaustive outcomes:** Outcomes of an experiment are said to be exhaustive if they include all possible outcomes.
- **Classical or mathematical or a priori probability:** If there are an n mutually exclusive, equally likely and exhaustive outcomes of an experiment and if m of this outcome is favorable to event A, then the probability of the event A which is denoted by $P[A]$ is defined by

$$P[A] = \frac{Favourable\ outcomes\ of\ an\ event\ A}{Total\ number\ outcomes\ of\ the\ experiment} = \frac{m}{n}$$

- **Sure event:** An event is called sure event when it always happens. The probability of a sure event is one.
- **Impossible event:** An event is called impossible event when it never happens.

❖ **Some Examples**

Example 1.2.1. A bag contains 4 white and 6 black balls. If one ball is drawn at random from the bag what is the probability that it is (i) black, (ii) white, (iii) white or black and (iv) red? State the nature of the events. (vi) what are the odds in favor of drawing a white ball? (vii) Also calculate the odds against drawing a white ball.

Solution. Total number of balls are 10. Since one ball is drawn from the bag, there are 10 mutually exclusive, equally likely and exhaustive outcomes of this experiment.

i. Let A be the event that the ball is black, then the number of outcomes favorable to A is 6. Hence,

$$P[A] = \frac{\text{Number of black balls}}{\text{Total number of balls}} = \frac{6}{10}$$

This is an uncertain event.

ii. let B be the event that the ball is white, then the favorable outcomes corresponding to B is 4. Therefore,

$P[B] = \frac{4}{10}$

The event B is also an uncertain event.

iii. Let C be the event that the ball is white or black, then the favorable outcomes corresponding to C is 10. Therefore,

$P[C] = \frac{10}{10} = 1$

This means that the outcome C is a sure outcome.

iv. Let D be the event that the ball is white and black, and then the favorable outcomes to D is zero. Then

$P[D] = \frac{0}{10} = 0$

This means that the outcome D is an impossible event in this since there is no red ball in the bag.

v. Let E be the event that the drawn ball is red. Here the favorable outcomes to E is zero. Hence

$P[E] = \frac{0}{10}$

This means that the outcome E is an impossible event in this experiment, since there is no red ball in the bag.

vi. According to the definition odds in favor of drawing a white ball is,

$$\frac{\text{Number of white balls}}{\text{Number of non white balls}} = \frac{4}{6}$$

vii. According to definition odds against drawing of a white ball is $\frac{6}{4}$.

According to definition probability of a white ball is

$\frac{4}{4+6} = \frac{4}{10}$

Example 1.2.2. A box contains 12 items of which two are defectives. An item is selected at random from this box. Find the probability that the selected item is (i) non-defective, (ii) defective, (iii) defective or non-defective, (iv) defective and non-defective. Comment on the nature of the events.

Solution. Here there are 12 equally likely, mutually exclusive and exhaustive outcomes.

i. Let A be the event that the selected item is non-defective. Then $N(A) = 12 - 2 = 10$. Here $N(A)$ is the number of items in A.

$$P[A] = \frac{\#\text{ of non-defective items}}{\text{Total \# of items}} = \frac{10}{12} = \frac{5}{6}$$

Here the event A is an uncertain event.

ii. Let B be the event that the selected item is defective. Then $N(B) = 2$.

$$P[B] = \frac{\#\ of\ defective\ item}{Total\ \#\ of\ items} = \frac{2}{12} = \frac{1}{6}$$

Here B is also an uncertain event.

iii. Let C be the event that the item is defective or non-defective. Then $N(C) = 12$.

$$P[C] = \frac{12}{12} = 1$$

Here C is a sure event.

iv. Let D be the event that the selected item is defective and non-defective. Then $N(D) = 0$

$$P[D] = \frac{0}{12} = 0$$

Here D is an impossible event.

Example 1.2.3. A card is drawn from a pack of 52 cards. Find the probability that it is (a) a red, (b) a spade, (c) an ace, (d) not a spade and (e) a king or a queen.

Solution. When a card is drawn from a pack of 52 cards, the total number of equally likely, mutually exclusive and exhaustive outcomes is 52. That is, here $n = 52$.

a) Let A be the event of drawing a red card. There are 26 black and 26 red cards in a pack and any one of the red cards can be drawn in 26 ways. That is $m = 26$. Then the probability of a red card is
$P[A] = \frac{26}{52} = \frac{1}{26}$

b) Let B be the event of drawing a spade. There are 13 spades in a pack of 52 cards. That is $m = 13$. Then the probability of a spade is

$$P[B] = \frac{12}{52} = \frac{1}{4}$$

c) Let C be the event of drawing an ace. There are four aces in all, one of each suit. That is, $m = 4$. Then the probability of an ace is

$$P[C] = \frac{4}{52} = \frac{1}{13}$$

d) Let D be the event that the card is not a spade. In 52 cards, only 13 are spades and the remaining 39 are not spades.

$$\therefore m = 39$$

The probability of that the card is not spade is, $P[D] = \frac{39}{52} = \frac{3}{4}$.

Alternatively, $P[D] = 1 - P[B] = 1 - \frac{1}{4} = \frac{3}{4}$

Because B is the complementary of the event D.

e) Let E be the event that the card is a king or a queen. Out of 52 cards there are 4 kings and 4 queens. That is $m = 8$. Hence, $P[E] = \frac{8}{52} = \frac{2}{13}$

Example 1.2.4. There are three events A, B, C one of which must and only one can happen. The odds are 7 to 3 against A and 6 to 4 against B, find the odds against C.

Solution. Here $P[A] = \frac{3}{10}$, $P[B] = \frac{4}{10}$

Since the events A, B and C are mutually exclusive, we have

$$P[A] + P[B] + P[C] = 1$$
$$\text{Or, } \frac{3}{10} + \frac{4}{10} + P[C] = 1$$
$$\therefore P[C] = 1 - \frac{3}{10} - \frac{4}{10} = \frac{3}{10}$$

Hence, the odds are 7 to 3 against C.

Example 1.2.5. Three contractors A, B and C are bidding for the construction of a new cinema hall. Some expert in this industry believes that A has exactly half the chance that B has, B, in turn, is 4/5 as likely as C to win the contract. What is the probability for each to win the contract if the expert's estimates are accurate?

Solution. Let the probability of C's winning the contract be x. Then,

$$P[C] = x,\ P[B] = \frac{4}{5}x,\ P[A] = \frac{4}{5}.\frac{1}{2}x$$

Since total probability is one, $x + \frac{4}{5}x + \frac{2}{5}x = 1 \Rightarrow x = \frac{5}{11}$

Hence, $P[C] = \frac{5}{11}$, $P[B] = \frac{4}{11}$, and $P[A] = \frac{2}{11}$

Example 3.6.5. There are A, B and C are mutually exclusive and their union is the sample space Ω.If $P[A] = \frac{3}{2}P[B]$, $P[B] = 2P[C]$, find the probabilities of. A, B and C

Solution. Since the events are mutually A, B and C exclusive and their union is the sample space Ω, then $A \cup B \cup C = \Omega$ .Thus

$$P[A \cup B \cup C] = P[\Omega] = 1$$

Now, $P[A \cup B \cup C] = P[A] + P[B] + P[C]$, since A, B and C are mutually exclusive. We have,

$$P[A] = \frac{3}{2}P[B] = \frac{3}{2}.2P[C] = 2P[C]$$

Now,

$$P[A] + P[B] + P[C] = 1$$
$$\Rightarrow 3P[C] + 2P[C] + P[C] = 1$$
$$\Rightarrow P[C] = \frac{1}{6}$$

Therefore, $P[B] = 2P[C] = \frac{1}{3}$ and $P[A] = 3P[C] = \frac{1}{2}$

Hence, $P[A] = \frac{1}{2}$, $P[B] = \frac{1}{3}$ and $P[C] = \frac{1}{6}$

Example 3.6.6. A sample survey was undertaken to investigate which paper, A, B and C people read. In a survey of 100 people, the following results were obtained:

60 read A
40 read B
70 read C
30 read A, B and C
32 read A and B
45 read A and C
38 read B and C

Draw a Venn-diagram and count the sample points. If a person is selected at random from this sample, find the probability that (i) he read only Newspaper A, (ii) he read only one Newspaper, (iii) he reads at least two Newspapers, (iv) he read at most one Newspaper, (v) he does not read any Newspaper and (vi). he read at least one Newspaper.

Solution. Let A, B and C be the events that people read Newspaper A, B and C, respectively, then $N(A) = 60, N(B) = 40, N(C) = 70, N(AB) = 32, N(AC) = 45, N(BC) = 38$ and $N(ABC) = 30$.

The events and the sample points are shown in fig 3.6.3 (a) and 3.6.2 (b)

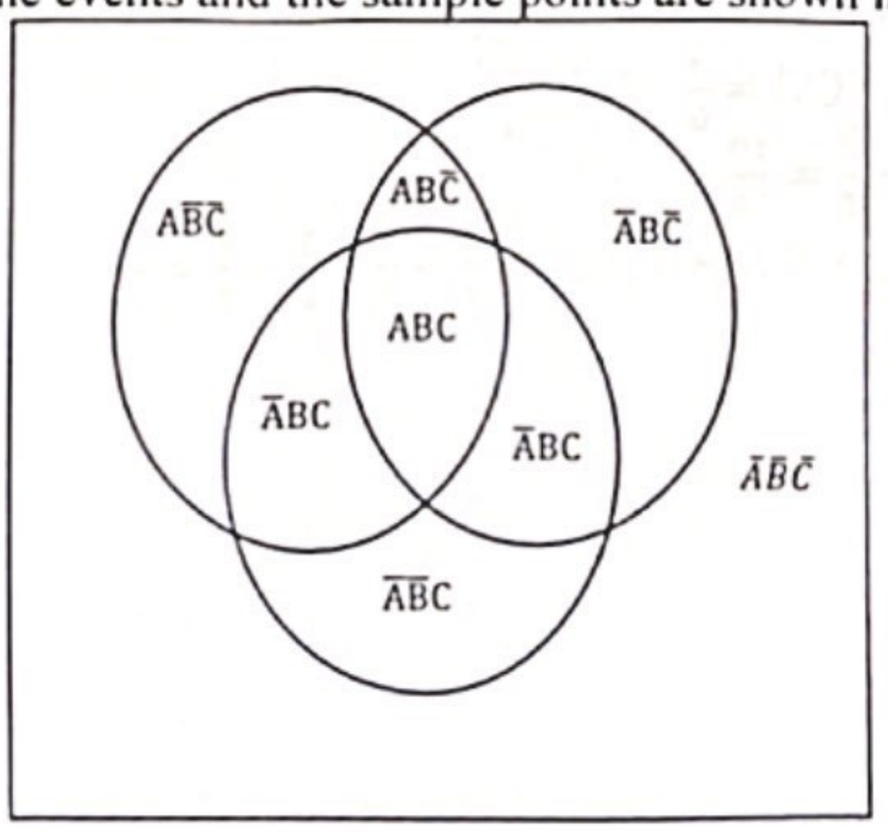


Figure 1: Fig.3.6.2. (a)

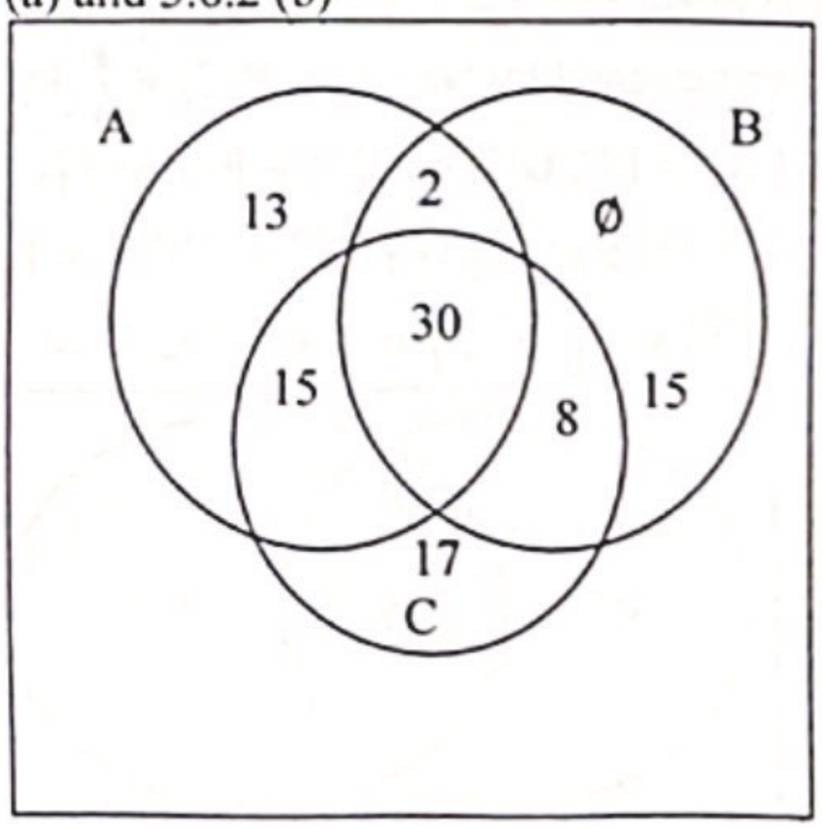


Figure 2: Fig 3.6.2. (b)

The union A ∪ B ∪ C is the sum of the seven mutually exclusive events $A\bar{B}\bar{C}, \bar{A}B\bar{C}, \bar{A}\bar{B}C, AB\bar{C}, \bar{A}BC, A\bar{B}C$ and ABC.
The sample points of these seven events are,

$$N(AB\bar{C}) = N(AB) - N(ABC) = 32 - 30 = 2$$
$$N(\bar{A}BC) = N(BC) - N(ABC) = 38 - 30 = 8$$
$$N(A\bar{B}C) = N(AC) - N(ABC) = 45 - 30 = 15$$
$$N(A\bar{B}\bar{C}) = N(A) - N(AB\bar{C}) - N(A\bar{B}C) - N(ABC)$$
$$= 60 - 2 - 15 - 30 = 13$$
$$N(\bar{A}\bar{B}C) = N(C) - N(AB\bar{C}) - N(\bar{A}BC) - N(ABC)$$
$$= 40 - 2 - 8 - 30 = 0 = \emptyset$$
$$N(\bar{A}B\bar{C}) = N(C) - N(\bar{A}BC) - N(A\bar{B}C) - N(ABC)$$
$$= 70 - 8 - 15 - 30 = 17$$
$$N(\bar{A}\bar{B}\bar{C}) = N(\overline{A \cup B \cup C}) = N(\Omega) - N(A \cup B \cup C)$$
$$= 100 - 85 = 15$$
$$N(A \cup B \cup C) = N(A) + N(B) + N(C) - N(AB) - N(BC) - N(AC) + N(ABC)$$
$$= 85$$

i. $P[A\bar{B}\bar{C}] = P[\text{ He reads only paper A}] = \frac{N(A\bar{B}\bar{C})}{N(\Omega)} = \frac{13}{100} = 0.13$

ii. $P[\text{ He reads only one Newspaper }] = P[A\bar{B}\bar{C} \cup \bar{A}B\bar{C} \cup \bar{A}\bar{B}C]$
$$= P[A\bar{B}\bar{C}] + P[\bar{A}B\bar{C}] + P[\bar{A}\bar{B}C]$$
$$= \frac{13}{100} + \frac{0}{100} + \frac{17}{100} = 0.13$$

iii. $P[\text{He reads at least two Newspapers}] = P[AB\bar{C} \cup \bar{A}BC \cup A\bar{B}C \cup ABC]$
$$= P[AB\bar{C}] + P[\bar{A}BC] + P[A\bar{B}C] + P[ABC]$$
$$= \frac{23}{100} + \frac{8}{100} + \frac{15}{100} + \frac{30}{100} = 0.55$$

iv. $P[\text{He reads at most one Newspaper}] = P[A\bar{B}\bar{C} \cup \bar{A}B\bar{C} \cup \bar{A}\bar{B}C \cup \bar{A}\bar{B}\bar{C}]$
$$= P[A\bar{B}\bar{C}] + P[\bar{A}B\bar{C}] + P[\bar{A}\bar{B}C] + P[\bar{A}\bar{B}\bar{C}]$$
$$= \frac{13}{100} + \frac{0}{100} + \frac{17}{100} + \frac{15}{100} = \frac{45}{100} = 0.45$$

v. $P[\text{He does not read any Newspaper}] = P[\bar{A}\bar{B}\bar{C}] = \frac{15}{100} = 0.15$

vi. $P[\text{He reads at least one Newspaper}] = P[A \cup B \cup C] = \frac{N(A \cup B \cup C)}{N(\Omega)} = \frac{85}{100} = 0.85$

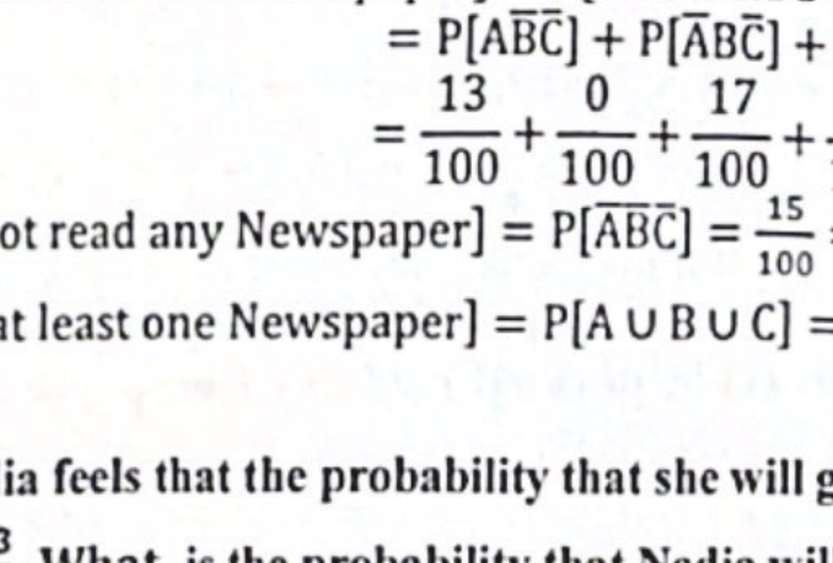
Example 3.6.8. Nadia feels that the probability that she will get A in calculus is $\frac{3}{4}$, A in statistics is $\frac{4}{5}$ and A in both the courses is $\frac{3}{5}$. What is the probability that Nadia will get (i) at least one A , (ii) only one *A* and (iii) no A's?

Solution. Let C be the event that Nadia will get *A* in calculus and *S*be the event that she will get *A* in statistics. The subjective probabilities assigned by Nadia are $P[C] = \frac{3}{4}$, $P[S] = \frac{4}{5}$ and $P[CS] = \frac{3}{5}$

(i) $P[\text{at least one A}] = P[C \cup S] = P[C] + P[S] - P[CS] = \frac{3}{4} + \frac{4}{5} - \frac{3}{5} = \frac{19}{20}$

(ii) $P[\text{Only one A}] = P[C\bar{S} \cup \bar{C}S] = P[C\bar{S}] + P[\bar{C}S] = P[C] - P[CS] + P[S] - P[CS] = \frac{3}{4} - \frac{3}{5} + \frac{3}{5} - \frac{4}{5} = \frac{7}{20}$

(iii) $P[\text{no A's}] = P[\bar{C}\bar{S}] = P[\overline{C \cup S}] = 1 - P[C \cup S] = 1 - \frac{19}{20} = \frac{1}{20}$



➢ **Example 3.6.11. Three horses A, B and C are in a race.A is twice as likely to win as B and B is twice as likely to win as C. What are their respective probabilities of winning, that is P[A], P[B]and P[C]?**

Solution. Let $P[C] = p$, then $P[B] = 2P[C] = 2p$ and $P[A] = 2P[B] = 4P[C] = 4p$

Now, $$P[A] + P[B] + P[C] = 1$$

$$4p + 2p + p = 1 \text{ or } 7p = 1$$

$$\therefore p = \frac{1}{7}. \text{ That is }, P[C] = \frac{1}{7}$$

Hence, $P[B] = 2\,P[C] = {}^{2}/_{7}$ and $P[A] = 4\,P[C] = {}^{4}/_{7}$.

➢ **Example 5.2.6. The personnel department of a company has records which show the following analysis of its 200 engineers.**

Age	Bachelor's degree only	Master's degree	Total
Under 30	90	10	100
30 to 40	20	30	50
Over 40	40	10	50
Total	150	50	200

If one engineer is selected at random from the company find

i. **The probability that he has only a bachelor's degree**
ii. **The probability that he has a master's degree, given that he is over 40**
iii. **The probability that he is under 30, given that he has only a bachelor's degree**

Solution. Let us define the following events:

A: an engineer has a bachelor's degree only
B: an engineer has a master's degree only
C: an engineer is under 30 years of age
D: an engineer is over 40 years of age

We have, $N(\Omega) = 200, N(A) = 150, N(B) = 50, N(C) = 100, N(D) = 50, N(BD) = 10, N(AC) = 90$

i. Now, $P[A] = \frac{N(A)}{N(\Omega)} = \frac{150}{200} = \frac{3}{4}$;

ii. $P[B|D] = \frac{P[BD]}{P[D]} = \frac{{}^{10}/_{200}}{{}^{50}/_{200}} = \frac{1}{5}$;

iii. $P[C|A] = \frac{P[CA]}{P[A]} = \frac{{}^{90}/_{200}}{{}^{150}/_{200}} = \frac{3}{5}$

Independence

Independence of two events

Definition 6.2.1. Let A and B be two events in 'S' of a given probability space (Ω, S, P[.]), then they are said to be mutually independent or simply independent if $P[AB] = P[A].P[B]$. Otherwise, the events A and B are dependent.

➢ **Example 6.2.1. Two fair coins are tossed simultaneously. Let A be the event head on the first coin and B be the event tail on the second coin. Show that A and B are independent.**

Solution. Let H and T denote the head and tail of the coin, respectively.
The sample space of the experiment is $\Omega = \{HH, HT, HT, TT\}$.
The event A will contain the sample points, $A = \{HH, HT\}$.
The sample points in B is $B = \{HT, TT\}$.
The sample point in AB is {HT}. Since the coin is fair, the sample space is simple.
Therefore, $P[A] = 1/2$, $P[B] = 1/2$ and $P[AB] = 1/4$.
Hence, $P[A].P[B] = 1/4 = P[AB]$,
which shows that A and B are independent.

➢ **Example 6.3.1. Two fair coins are tossed simultaneously. Let us define the following events: A_1: head of the first coin, A_2: head on the second coin and A_3: head or tail on both the coins. Prove that pairwise independent events are not always mutually or completely independent.**

Solution. Let H and T denote the head and tail of the coin, respectively. The sample space of the experiment is $\Omega = \{HH, HT, TH, TT\}$.
The sample space of the experiment is simple, since the coins are fair and the probability of each sample point is equal to 1/4.
Now the events $A_1, A_2, A_3, A_1A_2, A_1A_3, A_2A_3$ and $A_1A_2A_3$; contain the following sample points:
$A_1 = \{HH, HT\}, A_2 = \{HH, TH\}, A_3 = \{HH, TT\}$,
$A_1A_2 = \{HH\}, A_1A_3 = \{HH\}, A_2A_3 = \{HH\}$ and
$A_1A_2A_3 = \{HH\}$.
Therefore, $P[A_1] = \frac{1}{2}$, $P[A_2] = \frac{1}{2}$, $P[A_3] = \frac{1}{2}$,
$P[A_1A_2] = \frac{1}{4}$, $P[A_1A_3] = \frac{1}{4}$, $P[A_2A_3] = \frac{1}{4}$ and
$P[A_1A_2A_3] = \frac{1}{4}$.

The events A_1, A_2 and A_3 are pairwise independent, since
$P[A_1A_2] = P[A_1].P[A_2] = \frac{1}{4}$,
$P[A_2A_3] = P[A_2].P[A_3] = \frac{1}{4}$,
$P[A_1A_3] = P[A_1].P[A_3] = \frac{1}{4}$,
But, A_1, A_2, A_3 are not completely independent, since
$P[A_1A_2A_3] = \frac{1}{4} \neq P[A_1].P[A_2].P[A_3] = \frac{1}{8}$
This means that pairwise independent events are not always mutually or completely independent.

➢ **Example 6.3.2. Suppose three fair coins are tossed simultaneously and three events are defined as A_1: head on the first coin, A_2: head on the second coin and A_3: tail on the third coin.**

Test whether the events are completely independent or not.

Solution. Let H and T denote the head and tail of the coin. The sample space of the experiment is Ω = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

The sample space is simple and the probability of each sample point is $\frac{1}{8}$.

The events A_1, A_2, A_3, A_1A_2, A_1A_3, A_2A_3 and $A_1A_2A_3$; contain the sample points as follows:

A_1 = {HHH, HHT, HTH, HTT},
A_2 = {HHH, HHT, THH, THT}
A_3 = {HHT, HTT, THT, TTT},
A_1A_2 = {HHT, HHT},
A_1A_3 = {HHT, THT},
A_2A_3 = {HHT, THT} and
$A_1A_2A_3$ = {HHT}.

The probabilities are

$P[A_1] = P[A_2] = P[A_3] = \frac{1}{2}$,

$P[A_1A_2] = P[A_1A_3] = P[A_2A_3] = \frac{1}{4}$ and

$P[A_1A_2A_3] = \frac{1}{8}$.

Hence, $P[A_1A_2] = P[A_1].P[A_2] = \frac{1}{4}$,

$P[A_2A_3] = P[A_2].P[A_3] = \frac{1}{4}$,

$P[A_1A_3] = P[A_1].P[A_3] = \frac{1}{4}$ and

$P[A_1A_2A_3] = P[A_1].P[A_2].\ P[A_3] = \frac{1}{8}$,

Therefore, the three events A_1, A_2 and A_3 are completely independent.
Now, we show that the fourth condition of definition (Mutual or complete independence) may satisfy but the events are not completely independent.

Definition of Conditional Probability. [HW]

For conditional probability of B given A,

$P[B|A] = \frac{P[AB]}{P[A]}$

Bayes Theorem

❖ **State and prove Bayes theorem.**

Answer:

Statement. Let $B_1, B_2, \ldots, B_n$ be a sequence of n mutually exclusive events in $\mathcal{A}$ of a given probability space (Ω, $\mathcal{A}$, P[.]) such that $\Omega = [\bigcup_{i=1}^{n} B_i]$ and $P[B_i]>0$ for all $i = 1, 2, \ldots, n$ and if A is an event in $\mathcal{A}$ satisfying $P[A] > 0$, then

$$P[B_k|A] = \frac{P[B_K].P[A|B_k]}{\sum_{i=1}^{n} P[B_i].P[A|B_i]}$$

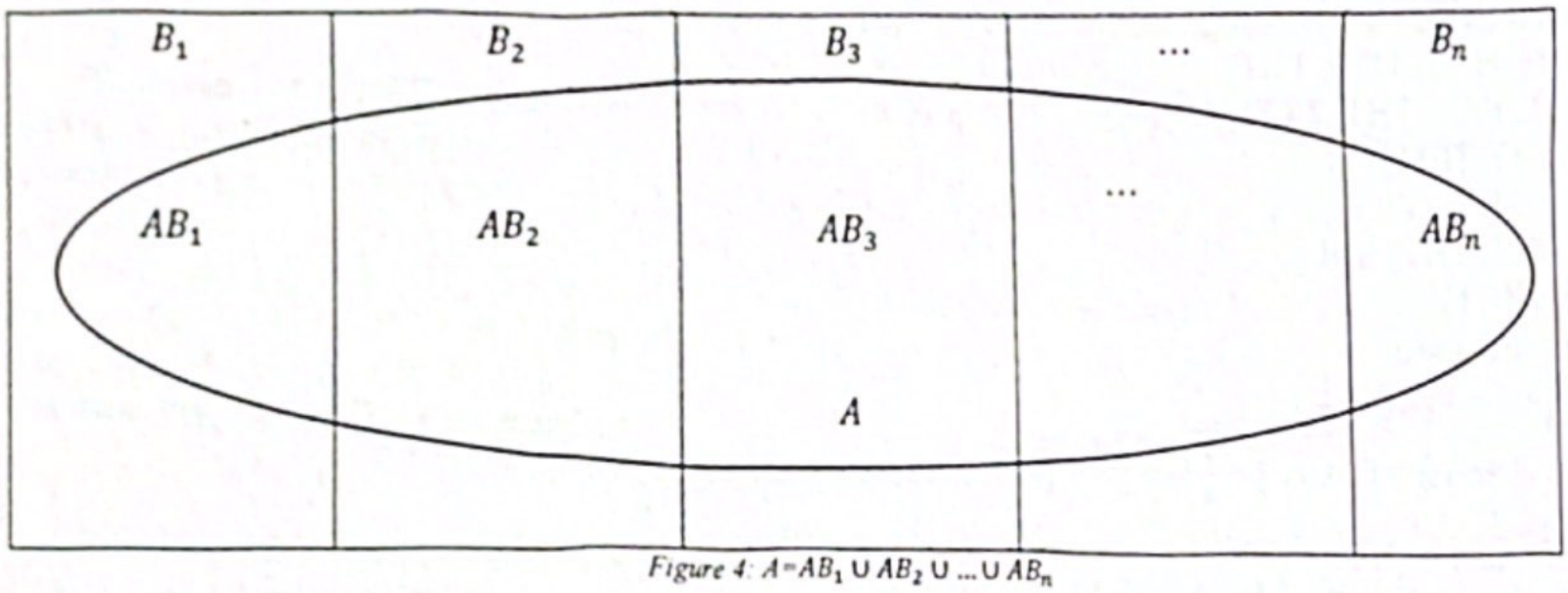


Figure 4: $A = AB_1 \cup AB_2 \cup \ldots \cup AB_n$

Proof:

The Venn-diagram of the theorem is shown in above figure. From the definition of conditional probability,

$$P[B_k|A] = \frac{P[AB_k]}{P[A]} \quad \ldots \quad (1)$$

We know from the theorem of total probabilities,

$$P[A] = \sum_{i=1}^{n} P[B_i].P[A|B_i] \quad \ldots \quad (2)$$

From the theorem of Multiplication, we have

$$P[AB_k] = P[B_k].P[A|B_k] \quad \ldots \quad (3)$$

Using the results (2) and (3) in (1), we have

$$P[B_k|A] = \frac{P[B_k].P[A|B_k]}{\sum_{i=1}^{n} P[B_i].P[A|B_i]}$$

for all k = 1,2,3 ,……., n. This completes the proof of the theorem 1.

The probabilities $P[B_1], P[B_2], \ldots, P[B_n]$ in Bayes theorem are known as prior probabilities, since they are known to the statistician before the experiment on the basis of some prior knowledge or information.

The probabilities $P[B_1|A], P[B_2|A], \ldots, P[B_n|A]$ in Bayes theorem are known as posterior probabilities.

Prior probabilities

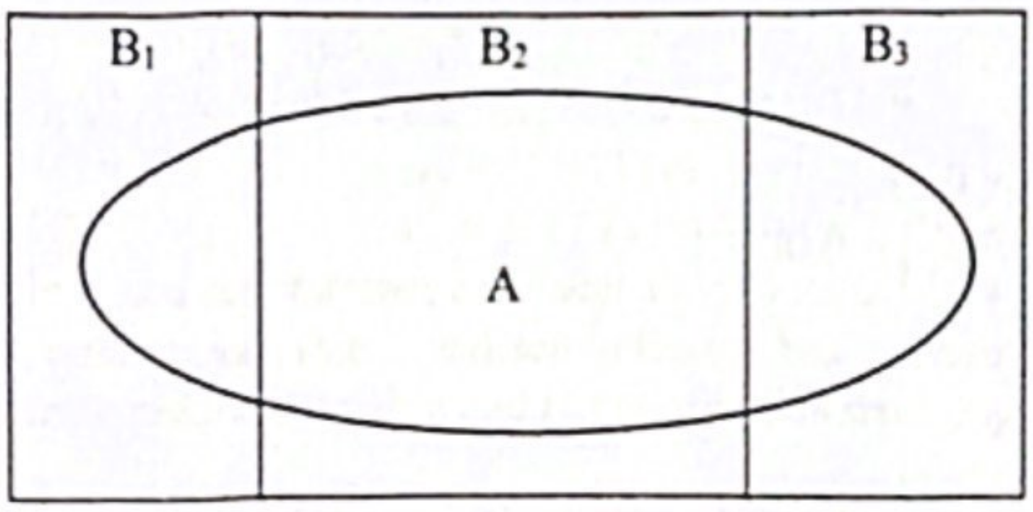
Definition 5.5.1. The probabilities of the events of an experiment are known as prior probabilities if they are known to the statistician before the experiment on the basis of some prior knowledge or information.

Posterior probabilities

Definition 5.5.2. The probabilities of the events of an experiment are known as posterior probabilities if they are obtained after experiment on the basis of prior probabilities.

- **Example (5.5.6) Secret is always obvious. Mr. Karim is the director of Municipality. He has been appointing 30%, 50% and 20% engineers from three countries A, B and C, respectively. He knows from previous experiences that .02%, .01% and .05% engineers from countries A, B and C, respectively are inefficient. Recently, Mr. Karim appointed 500 engineers from these three countries. After one month's working record, it is found that Mr. Zack is an inefficient engineer, what is the probability that he was from country C? Also calculate the probability that he was from A or B?**

Solution.



Let us define the following events:

B_1: an engineer from country A,
B_2: an engineer from country B,
B_3: an engineer from C,
A: an inefficient engineer.

The events are shown in the Venn-diagram of fig. 5.5.4.
According to the information, we have the following probabilities:

$P[B_1] = 0.3$, $P[B_2] = 0.5$, $P[B_3] = 0.2$,
$P[A|B_1] = 0.0002$, $P[A|B_2] = 0.0001$, $P[A|B_3] = 0.0005$
We are asked to find the probabilities

$P[B_3|A]$ and $P[(B_1 \cup B_2) | A]$.

Applying Bayes theorem 5.5.2, we have

$$P[B_3 | A] = \frac{P[B_3].P[A|B_3]}{\sum_{i=1}^{n} P[B_i].P[A|B_i]}$$

Now $P[A] = \sum P[B_i].P[A|B_i]$
$= P[B_1].P[A|B_1] + P[B_2].P[A|B_2] + P[B_3].P[A|B_3]$
$= (0.3)(0.0002) + (0.5)(0.0001) + (0.2)(0.0005)$
$= 0.00021$

Hence, $P[B_3|A] = \frac{P[B_3].P[A|B_3]}{\sum_{i=1}^{n} P[B_i].P[A|B_i]} = \frac{(0.2)(0.0005)}{0.00021} = \frac{10}{21}$

Now, $P[(B_1 \cup B_2)|A] = P[B_1|A] + P[B_2|A]$,
Since B_1 and B_2 mutually exclusive.

Then $P[B_1|A] = \frac{P[B_1].P[A|B_1]}{\sum_{i=1}^{n} P[B_i].P[A|B_i]} = \frac{6}{21}$ And $P[B_2|A] = \frac{P[B_2].P[A|B_2]}{\sum_{i=1}^{n} P[B_i].P[A|B_i]} = \frac{5}{21}$

Therefore, $P[(B_1 \cup B_2)] = \frac{6}{21} + \frac{5}{21} = \frac{11}{21}$

- **Example 5.5.7. Man is not free from error. In a bank 30%, 25% and 45% of the monthly statements are prepared, respectively, by Mr. Ahmed, Mrs. Ali and Miss Karim. These employees are very**

reliable. However, they are in error sometimes. The probabilities that they are in error are 0.01%, 0.05% and 0.001%, respectively. If a monthly statement was found to be in error, what is the probability that it was done by Miss Karim?

Solution:

The events are shown in the Venn-diagram. Let us define the following events:

B_1: statement prepared by Mr. Ahmed,

B_2: statement prepared by Mrs. Ali,

B_3: statement prepared by Miss. Karim,

A: statement with error

We are given

$P[B_1] = 0.30, \quad P[B_2] = 0.25, \quad P[B_3] = 0.45,$

$P[A|B_1] = 0.0001, \quad P[A|B_2] = 0.0005, \quad P[A|B_3] = 0.00001$

We have to find $P[B_3|A]$.

According to Bayes theorem 5.5.2, we have,

$$P[B_3|A] = \frac{P[B_3]P[A|B_3]}{P[B_1]P[A|B_1]+P[B_2]P[A|B_2]+P[B_3]P[A|B_3]}$$

$$= \frac{(0.45)(0.00001)}{(0.3)(0.0001)+(0.25)(0.0005)+(0.45)(0.00001)}$$

$$= \frac{45}{1595} = \frac{9}{319}$$

- **Example 5.5.8. Nature is always against the wills. In winter of country, it rains 45% of the days and shines 55% of the days. Mr. Saad has a barometer which wrongly predicts 5% of the time in rainy days and 10% on shiny days. Mr. Saad does not carry umbrella if the barometer predicts shine. One good shiny morning in winter his barometer predicts "shine" and he does not carry umbrella. What is the probability that Mr. Saad will suffer on that day due to rain?**

Solution:

Let us define the following events:

B_1: a day is rainy,

B_2: a day is shiny,

E: barometer predicts shine, $\bar{E}$: barometer predicts rain.

We are given

$P[B_1] = 0.45, \quad P[B_2] = 0.55, \quad P[E|B_1] = 0.05, \quad P[\bar{E}|B_2] = 0.10$

$P[\bar{E}|B_2] = 1 - 0.10 = 0.90$

We have to find $P[B_1|E]$. We have from Bayes theorem,

$$P[B_1|E] = \frac{P[B_1]P[E|B_1]}{P[B_1]P[E|B_1]+P[B_2]P[E|B_2]}$$

$$= \frac{(0.45)(0.05)}{(0.45)(0.05)+(0.55)(0.90)}$$

$$= \frac{0.0225}{0.5175} = \frac{225}{5175} = \frac{1}{23}$$

Therefore, the probability that Mr. Saad will suffer on that day due to rain is 1/23.

Random Variable

Random Variable: A random variable denoted by $X(.)$ or X for a given probability space $(\Omega, \mathcal{A}, P[.])$ is a function with domain Ω and counter domain the real line.

Example. Let us consider the experiment of tossing two fair coins. The sample space of the experiment is

Ω = (HH, HT, TH, TT) = $\{ \omega_1, \omega_2, \omega_3, \omega_4 \}$

Let the random variable X is defined by the number of heads. Thus

$X\omega_1$ = X (HH) = 2 ω_2 = X (HT) = 1

$X\omega_3$ = X (TH) = 1 ω_4 = X (TT) = 0

So the random variable X is a real number and takes the values 0,1,2, with probabilities 1/4, 1/2 and 1/4, respectively Here, the random variable X is discrete On the basis of sample space, random variable s generally of two types, viz. (i) discrete random variable and (ii) continuous random variable

Discrete random variable: A random variable on Ω is defined to be a discrete random variable if it takes at most a countable number of values. In other words, a real valued function defined on a discrete sample space is called a discrete random variable.

Example. Let us consider the experiment of three-child family in which the probability that a child will be a boy and a girl are equal. Let X be a random variable which is defined by the number of boys. Let B and G denote the boy and girl, respectively. The sample space of the experiment is,

$\Omega = \{ \omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8 \}$ = {BBB, BBG, BGB, GBB, BCG, GBG, GGB, GGG}

Here,

$P[B] = P[G]$=1/2

The sample space is simple and the probability of each outcome is 1/8.

Here, the random variable X is the number of boys and X can take the values 0, 1, 2 and 3 with probabilities 1/8, 3/8, 3/8 and 1/8, respectively which are shown in the table 1.

Table 1

outcomes in Ω:w_i	P [ω_i]	Values of X:x	P[X=x]
BBB	1/8	3	1/8
BBG	1/8	2	
BGB	1/8	2	3/8
GBB	1/8	2	
BGG	1/8		
GBG	1/8	1	3/8
GGB	1/8	1	
GGG	1/8	1	1/8
		0	

Table 2 Probability Distribution

Values of X: x	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

The graph of the probability function is

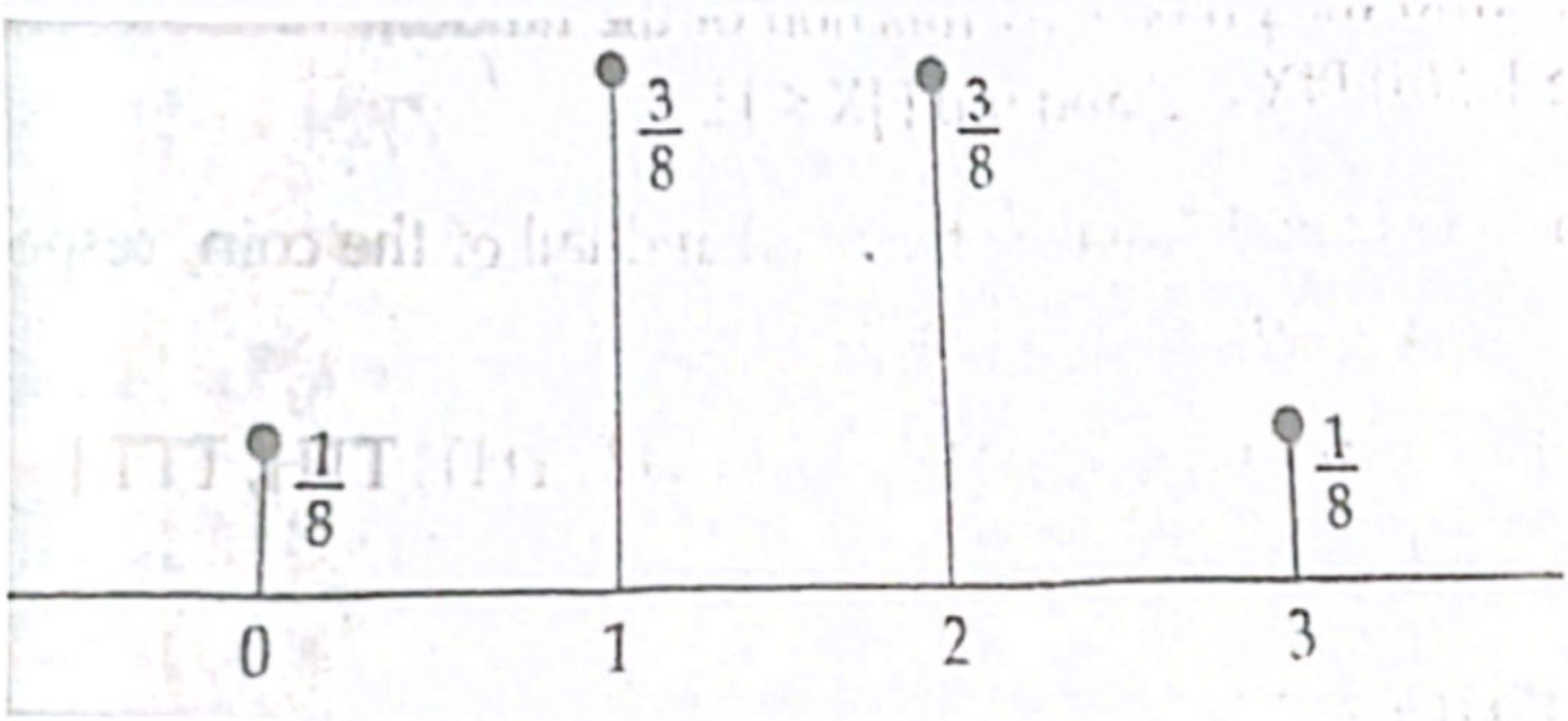


Figure 5: Probability Plot

Graph of probability function for number of boys in three child-family. Mathematically, the probability function can be expressed as

$$P(x) = \binom{3}{x}\left[\frac{1}{2}\right]^3 ; x = 0,1,2,3.$$

Probability mass function or Probability function

If X is a discrete random variable with possible values $x_1, x_2, \ldots, x_n$ and $f(x_1)f(x_2),\ldots f(x_n)$ be their corresponding probabilities, then the set of ordered pair (x, f(x)) is called the probability function or probability mass function of the discrete random variable X. Symbolically,

$$f(x_i) = \begin{cases} [X = x_i], & \text{if } X = x_i, i = 1,2,\ldots\ldots\ldots,n \\ 0, & \text{if } X \neq x_i \end{cases}$$

That is, $f(x)$ is a function with domain the real line and counter domain the interval [0, 1].

Properties of probability function

If f(x) is a probability function of a discrete random variable X, then f(x) satisfies the following two properties:

1) $0 \leq f(x) \leq 1$, for each possible values of X.
2) $\sum f(x_i) = 1$

Summation is taken over all possible values of X. The values of a discrete random variable are often called mass points and $f(x_i)$ denotes the mass associated with the mass point x. That is why f(x) is called probability mass function.

- **Example. A coin is tossed three times in which the probability of head is twice as the probability of tail. If number of heads is a random variable, find the probability function of the random variable. Also find (i) $P[X = 2]$ (ii) $P[X = 2]$ and (iii) $P[X \leq 1]$**

Solution:

Let H and T denote the head and tail of the coin, respectively. The sample space of the experiment is

Ω = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}
= $\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}$

Here P[H] = 2/3 and P[T]=1/3

The probabilities of the different points of the sample space are

P[HHH]=8/27,

P[HHT]= P[HTH] = P[THH]=4/27,
P[HTT] =P[THT] = P[TTH]=2/27,
P[TTT]=1/27

The sample space of the experiment is not simple. Let X be a random variable defined by the number of heads. The different points of the sample space, their probabilities, the different values of the random variable and their respective probabilities are shown in the table 3.

Table 3

Sample Points in Ω: ω_i	Probability of $[\omega_i]$	Number of Heads X: x	Probability F(x)
HHH	8/27	3	8/27
HHT	4/27	2	
HTH	4/27	2	12/27
THH	4/27	2	
HTT	2/27	1	
THT	2/27	1	6/27
TTH	2/27	1	
TTT	1/27	0	1/27

Hence, the probability function of the random variable X is

Table 4

Values of X:x	0	1	2	3
f(x)	1/27	6/27	12/27	8/27

(i) $P[X \geq 1] = 1 - P[X < 1] = 1 - f(0) = 1 - 1/27 = 26/27.$
(ii) $P[X = 2] = f(2) = 12/27.$ and
(iii) $P[X \leq 1] = P[X = 0] + P[X = 1] = f(1) + f(0) = 6/27 + 1/27 = 7/27.$

➢ **Example 7.2.3. Let X be a random variable with probability function defined by f (-2) = 1/10, f (0) = 2/10, f (4) = 4/10 and f (11) = 3/10.**

Find (i) P [-2 ≤ X < 4], (ii) P[X > 0] and (iii) P[X ≤ 4].

Solution:

Here the random variable X takes the values -2, 0, 4 and 11 with respective probabilities 1/10, 2/10, 4/10 and 3/10. That is,

Values of X:x	-2	0	4	11
F(x)	1/10	2/10	4/10	3/10

(i) P [-2 ≤ X < 4] = P [X = - 2] + P [X = 0] = f (-2) + f (0) =1/10 + 2/10 = 3/10.
(ii) P [X > 0] + P [X = 4] + P [X = 11] = f (4) + f (11) = 4/10 + 3/10 = 7/10. and
(iii) P [X ≤ 4] = -1 - P [X = 11] =1-f (11) = 1- 3/10 = 7/10.

➢ **Example 7.2.4. Suppose a discrete random variable X has the following function:**

Values of X:x	-3	0	1	4
F(x)	1/10	3/10	4/10	2/10

Define a random variable for which $f(x)$ is a probability function.

Solution:

Here f(x) satisfies the properties of the probability function since
(1) $f(x) \geq 0$ and (2) $\Sigma f(x) = 1.$

Suppose a box contains ten cards, one has the number -3, three have the number zero, four have the number one and two have the number 4 written on them. If X be the number on the cards selected at random from the box, then the function $f(x)$ given above is clearly the probability function of X.

➢ **Example 7.2.5. A random variable X has the following probability function:**

Values of X:x	0	1	2	3	4	5	6	7	8
F(x)	a	3a	5a	7a	9a	11a	13a	15a	17a

(i) Determine the value of a.

(ii) Find $P[X < 3], P[X \leq 3]$, and $P[0 < X < 5]$.

Solution:

(i) Since $f(x)$ is a probability function, we have $\sum f(x) = 1$. That is,

$f(0) + f(1) + f(2) + f(3) + f(4) + 1(5) + f(6) + f(7) + f(8) = 1$

or, $a + 3a + 5a + 7a + 9a + 11a + 13a + 15a + 17a = 1$

or, $a(1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17) = 1$

$$a = 1/81.$$

(ii) $P[X < 3]\, P[X = 2] + P[X = 1] + P[X = 0] = f(2) + f(1) + f(0) = 1/81 + 3/81 + 5/81 = 1/9.$

$$P[X \geq 3] = 1 - P[X < 3] = 1 - 1/9 = 8/9.$$

$$P[0 < X < 5] = P[X = 1] + P[X = 2] + P[X = 3] + P[X = 4] = f(1) + f(2) + f(3) + f(4) = 8/27.$$

The probability function tells us how likely or probable each of the value of a discrete random variable is. It also enables one to calculate the probability of events described in terms of the discrete random variable X. For example, let x have mass points at $x_1, x_2, \ldots . x_n \ldots \ldots$ with associated mass $f(x_1), f(x_2), \ldots \ldots, f(x_n) \ldots \ldots$, then P[a $< X \leq b] = \sum_{a<x\leq b} f(x_i)$

Bar chart and probability histogram

It is often helpful to look at a probability function in graphic form. One might plot the points $(x;\ f(x))$ of example 7.2.1 to obtain fig. 7.3.1. By joining the points to the x axis, either with a dash or solid line, we obtain what is commonly called a bar chart. Fig. 7.3.1 makes it very easy to see what values of X are most likely to occur and it also indicates a perfectly symmetric situation in this case.

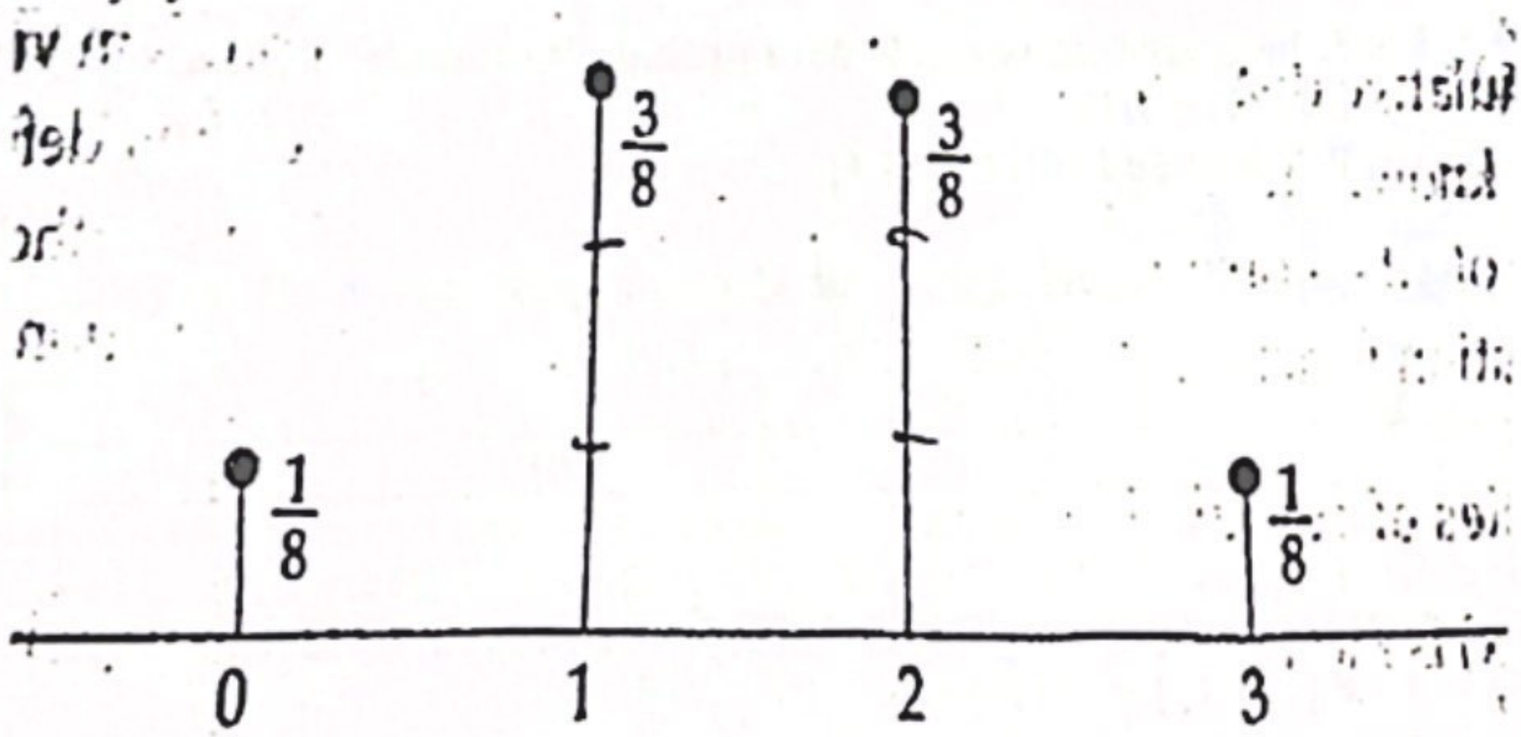


Fig 73.1. Bar chart

Instead of plotting the points $(x, f(x))$ we more frequently construct rectangles as in fig. 73.2. Here the rectangles are constructed so that their bases of equal width are centered at each value x and their heights are equal to the corresponding probabilities given by $f(x)$ of example 7.31. Then bases are constructed so as to leave no space between the rectangles. Fig. 7.3.2 is called probability histogram.

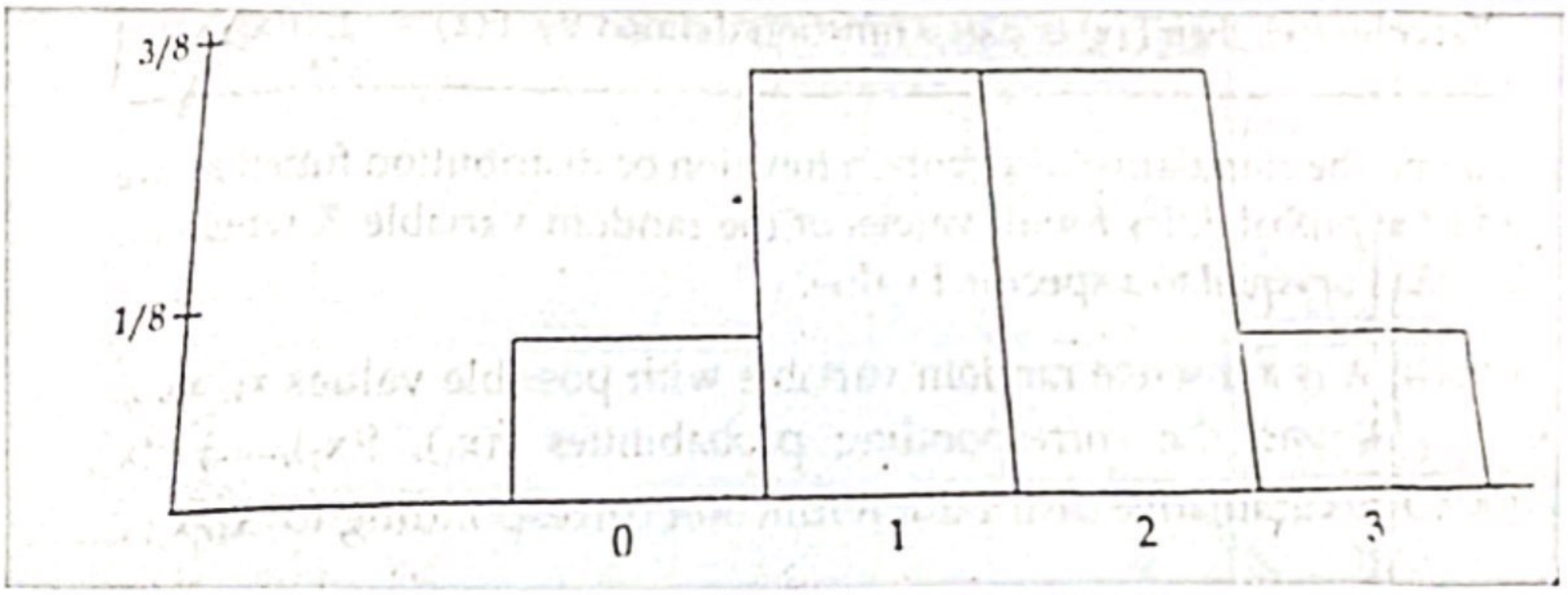


Fig.7.3.2.

7.4. Cumulative distribution function of random variables

When we are interested in finding the probability of a given random variable X which will take on values that are less than or equal to some fixed value x. That is, we wish to find $P[X \leq x]$. This probability considered as a function of X, is called the cumulative distribution function or simple distribution function of the random variable X.

Cumulative distribution function

Definition 7.4.1. The cumulative distribution function of the random variable X, denoted by $F(x)$, is defined to be that function with domain the real line and counter-domain the interval [0, 1] which satisfies

$$F(x) = P[X \leq x]$$

A cumulative distribution is uniquely defined for each random variable. If it is known, it can be used to find probabilities of events defined in terms of its corresponding random variables. Many authors use distribution function in place of cumulative distribution function.

Properties of cumulative distribution function

(i) $F(x)$ is monotonic increasing function; that is, $F(a) \leq F(b)$, when $a \leq b$.

(ii) The limit of $F(x)$ to the left side is 0 and the right side is 1. This is,

(a) $\lim_{x \to -\infty} F(x) = F(-\infty) = 0$ and (b) $\lim_{x \to -\infty} F(x) = F(-\infty) = 1$

(iii) F(x) is continuous from the right, that is, $\lim_{0<h \to \infty} F(x + h) = F(x)$.

Cumulative distribution function of discrete random variable

Definition 7.4.2. If X is a discrete random variable with probability function $f(x)$, then $F(x)$ is a step function defined by $F(x) = \sum_{xi \leq x} f(x_i)$.

To find the cumulative distribution function or distribution function, we add the probabilities for all values of the random variable X which are less than or equal to a specified value. Suppose X is a discrete random variable with possible values $x_1, x_2, \ldots . x_i, \ldots . x_n$ and the corresponding probabilities $f(x_1), \ldots . f(x_i), \ldots f(x_n)$. The cumulative distribution function corresponding to $x_1, x_2, \ldots . x_i, \ldots . x_n$ will be $F(x_1) = f(x_1)$, $F(x_2) = \sum_{i=1}^{2} f(x_j), \ldots . , F(x_i) = \sum_{j=1}^{i} f(x_j), \ldots \ldots F(x_n) = \sum_{j=1}^{n} f(x_j) = 1$

The values of the random variable with their probability and distribution function are shown in the table 7.4.1.

Values of the random variable X:x	Probability Function f(x)	Distribution Function F(x)
x_1	$f(x_1)$	$F(x_1) = f(x_1)$
x_2	$f(x_2)$	$F(x_2) = f(x_1) + f(x_2)$
.	.	.
.	.	.
x_i	$f(x_i)$	$F(x_i) = f(x_1) + \ldots + F(x_i)$
.	.	.
.	.	.
x_n	$f(x_n)$	$F(x_n) = F(x_1) + \cdots + f(x_n) = 1$

Graphical representation of distribution function

The distribution function $F(x)$ is a step function with steps or jumps at $x_1, x_2, \ldots\ldots x_i, \ldots, x_n$. The jump at x_i, is equal to the probability $f(x)$ at that point. The total height of the step at x_i, is $F(x_i) = \sum_{j=1}^{i} f(x_j)$. Between the jumps F the functio n $F(x)$ is constant. The graph of $F(x)$ is shown in fig.

From the graph 7.4.1 it is clear that $F(x)$ is a non-decreasing function. What are its bound? Suppose that x_1, is the smallest value of X and x_n, is the largest value of X. Then if $X < x_1$, the event $A = (X < x_1)$ is the null set. Hence

Continuous random variable

A random variable X is said to be continuous if it can take all possible values within certain limits. In other words, a random variable X is called a continuous random variable if it is defined on a continuous sample space

Example of continuous random variable: X is a continuous random variable if

(i) X is the value of number to be selected at random from the interval of real numbers between a and b.
(ii) X is the distance from the center of a target to the hit position of an arrow that is to be shot at the target.
(iii) X is the arrival time of a student to attend the first lecture.
(iv) X is the arrival time of an airplane from Benghazi to Tripoli.

Probability density function

X is a continuous random variable, then the function $f(x)$, defined on the continuous sample space with the domain the real line and counter domain the interval [0, 1], is called probability density function or simple density function.

The probability density function $f(x)$ satisfies the following properties:

i) $f(x) \geq 0, \forall x \in R$ where R is the collection of all points in the entire range of the variable X.

(ii) $\int_R f(x)\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1$.

(iii) The probability of any event A defined by $X < x$ is

$P[A] = P[X < x] = \int_{-\infty}^{x} f(x)\, dx$.

Moreover, if B is an event defined by $a \leq x \leq b$, then

$P[B] = P[a \leq X \leq b] = \int_a^b f(x)dx$.

Rules for finding probabilities for a continuous random variable

(i) $P[X = x] = 0$ for $-\infty < x < \infty$.
(ii) $P[X < x] = P[X \leq x] = \int_{-\infty}^{x} f(x)dx$. for $-\infty < X < x$.
(iii) $P[a \leq X \leq b] = P[a \leq X < b] = P[a < X < b] = P[a < X < b] = \int_a^b f(x)dx$, for $a < b$

➢ **Example 7.5.1. Let X be a continuous random variable with probability density function**

$f(x) = \begin{cases} \frac{2}{3}x, & if\ 1 < x < 2 \\ 0, & otherwise \end{cases}$

Find $P[X \leq 1.2], P[X > 1.2], P[1.2 < X < 1.6], P[X = 1.5], P[X \leq 0.2]$ and $P[X > 3]$.

Solution:

$P[X \leq 1.2] = \int_1^{1.2} \frac{2}{3}x dx = \frac{1}{3}[1.44 - 1] = 0.1467,$

$P[X > 1.2] = 2, \int_{1.2}^{2} f(x)dx = \frac{2}{3}\int_{1.2}^{2} f(x)dx = \frac{(4-1.44)}{3}$

$P[1.2 \leq x \leq 1.6] = \int_{1.2}^{1.6} \frac{2}{3} f(x)dx = \frac{1}{3}[2.56 - 1.44] = 0.3733,$

$P[X = 1.5] = 0,$

Since, continuous random variable cannot take any isolated value

$P[X \leq 0.2] = \int_{-\infty}^{0.2} f(x)dx = \int_0^{0.2} 0.dx = 0$ and

$P[X > 3] = \int_3^{\infty} f(x)dx = \int_3^{\infty} 0.dx = 0.$

- **Example 7.5.2. The pressure measured in pounds per cm² at a certain Valve is a random variable X whose probability density function is $f(x) = \begin{cases} \frac{2}{9}(3x - x^2), & if\ 0 < x < 3, \\ 0, & otherwise. \end{cases}$**

Find the probability that the pressure at this valve is (i) not more than 2 pounds per cm², (ii) greater than 2 pounds per cm², (iii) between 1.5 and 2.5 pounds per cm² and (iv) less than 1.5 pounds per cm².

Solution.

(i) $P[X \leq 2] = \frac{2}{9}\int_0^2 (3x - x^2)\,dx = \frac{2}{9}\left[\frac{3x^2}{2} - \frac{x^3}{3}\right]_0^2 = \frac{2}{9}\left[6 - \frac{8}{3}\right] = \frac{20}{27} = 0.7407.$

(ii) $P[X > 2] = 1 - P[X \leq 2] = 1 - 0.7407 = 0.2593.$

(iii) $P[1.5 \leq X \leq 2.5] = \int_{1.5}^{2.5} \frac{2}{9}(3x - x^2)\,dx = \frac{2}{9}\left[\frac{3x^2}{2} - \frac{x^3}{3}\right]_{1.5}^{2.5} = \frac{11.50}{27} = 0.4259$

(iv) $P[X \leq 1.5] = = \int_0^{1.5} \frac{2}{9}(3x - x^2)\,dx = = \frac{2}{9}\left[\frac{3x^2}{2} - \frac{x^3}{3}\right]_0^{1.5} = \frac{13.50}{27} = 0.5$

Cumulative distribution or distribution-function of a continuous random variable

Introduction. Like discrete random variable, we are interested in finding the probability of a continuous random variable which is less than or equal to a particular value of the random variable. This probability is called cumulative distribution or distribution function of X.

Definition 7.6.1. If X is a continuous random variable with density function $f(x)$, then cumulative distribution function of X, denoted by $F(x)$, is defined by

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(x)dx \qquad (7.6.1)$$

The cumulative distribution function $F(x)$ of a continuous random variable X is absolutely continuous.

Theorem 7.6.1. If X is a continuous random variable, then $F(x)$ can be obtained from $f(x)$ and vice versa.

Proof. If X is a continuous random variable and $f(x)$ is given, then $F(x)$ is obtained by integrating $f(x)$, that is

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

On the other hand, if F(x) is given, then f(x) can be obtained by dF(x) differentiation, that is, $f(x) = \frac{dF(x)}{dx}$ for those points x for which F(x) is dx differentiable..

Cumulative distribution curve

Graphically, cumulative distribution function F(x) of a continuous random variable X can be shown by a curve which is known as cumulative distribution curve.

➢ **Example 7.6.1. Let X be a continuous variable with distribution function**

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 \leq 1 \\ 1 & 1 < x \end{cases}$$

(i) Sketch the distribution function of X (ii) Find the pdf of X. (iii) Sketch the pdf of X and (iv) Compute P [0.4 ≤ X ≤ 0.6]

Solution:

(i) The graph of the distribution function is

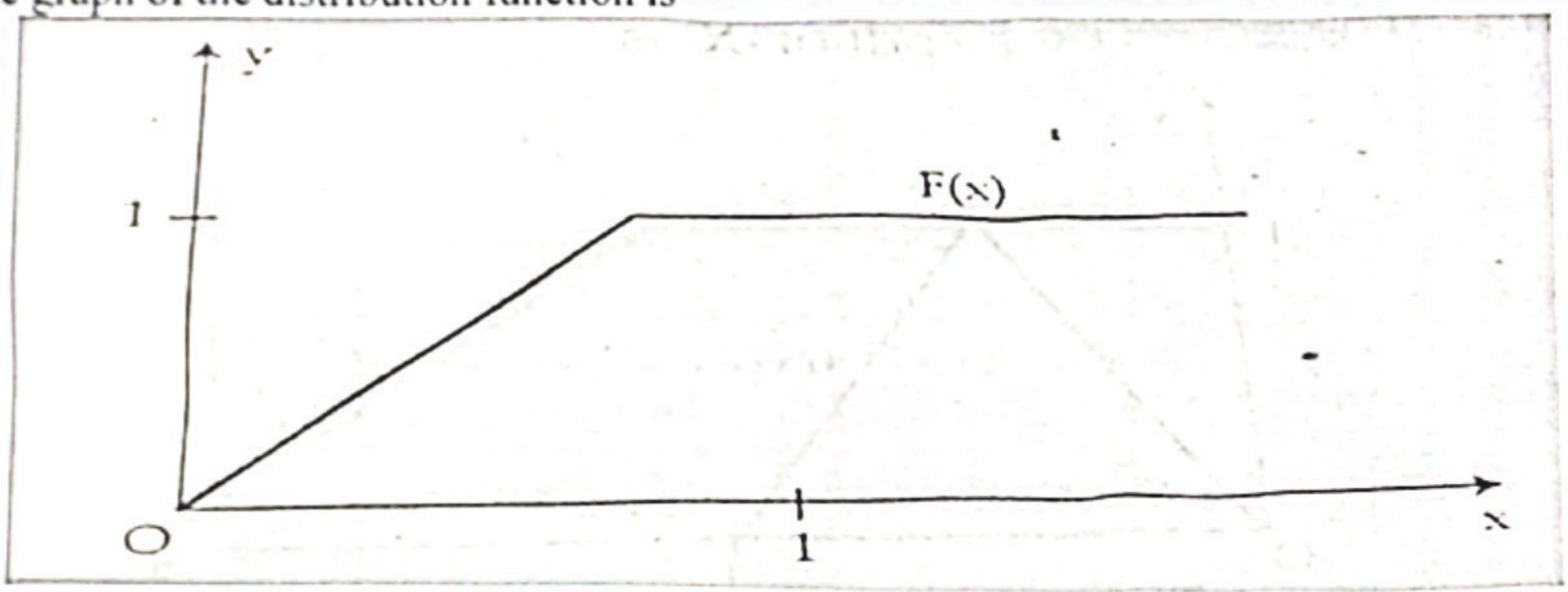


Fig. 7.6.1.

(ii) Differentiating F with respect to x at continuity points of f, we get

$$f(x) = F'(x) = f(x) = \begin{cases} 0, & x < 0 \text{ or } x > 1 \\ 1, & 0 < x < 1 \end{cases}$$

(iii) The function f(x) is not continuous at x = 0 or at x = 1. The graph of the pdf of X is below.

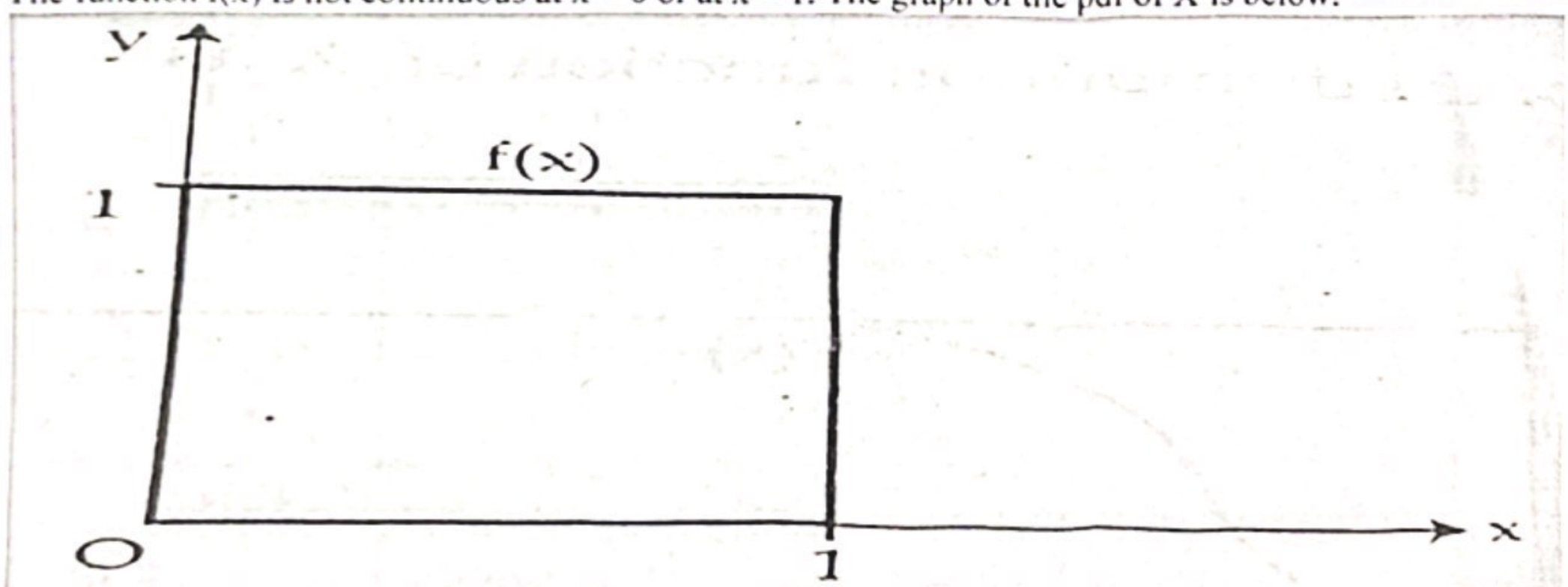


Fig. 7.6.2.

(iv) $P[0.4 < x < 0.6] = \int_{0.4}^{0.6} f(x)dx = [x]_{0.4}^{0.6} = 0.6 - 0.4 = 0.2$

➢ **Example 7.6.2. Let X be a triangular variate with pdf**

$$f(x) = \begin{cases} x, & 0 \le x \le 1 \\ 2 - x, & 1 \le x \le 2 \\ 0 & otherwise \end{cases}$$

(i) Sketch the pdf of X: 6 Find the distribution function of X: (ii) Sketch the distribution function F(x); (iv) Compute $P\left[\frac{1}{2} < X < \frac{3}{2}\right]$ and (v) P[0.3 < X ≤ 1.5]

Solution:

(i) The graph of the pdf of X is

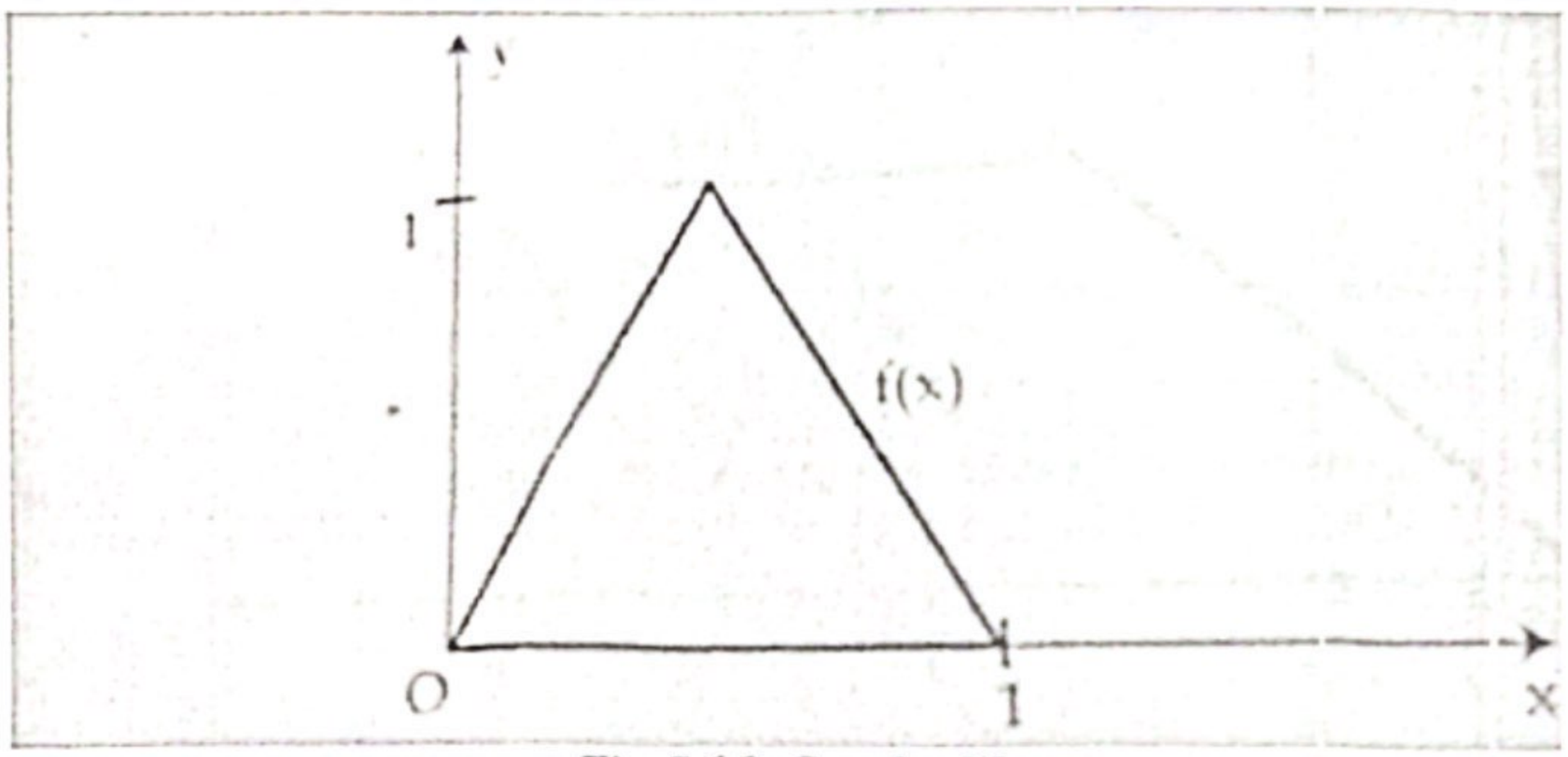


Fig. 7.6.3. Graph of f

(ii) It is easy to check that f(x) is a pdf. The distribution function of X is

$$F(x) = \begin{cases} 0, & if\ x \le 0 \\ \int_0^x x dx = \frac{x^2}{2}, & if\ 0 < x \le 1 \\ \int_0^1 x dx + \int_1^x (2-x) dx = 2x - \frac{x^2}{2} - 1, & if\ 1 < x \le 2 \\ 1,, & if\ x \ge 2 \end{cases}$$

(iii) The graph of the distribution function of X is

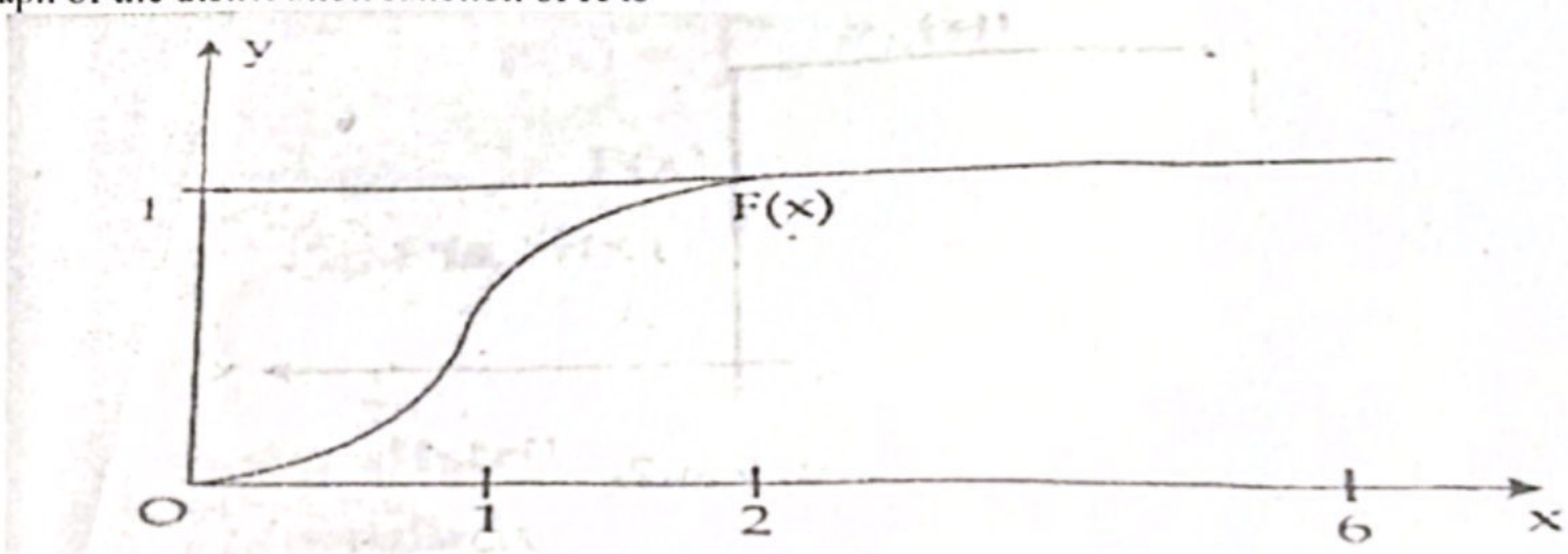


Fig. 7.6.4. Graph of F(x)

(iv) $P\left[\frac{1}{2} < X < \frac{3}{2}\right] = \int_{1/2}^{3/2} f(x)\,dx = \int_{1/2}^{1} f(x)\,dx + \int_1^{3/2} (2 - x)\,dx = \left[\frac{x^2}{2}\right]_{1/2}^{1} + \left[2x - \frac{x^2}{2}\right]_1^{3/2} = 3/4$

(v) $P[0.3 < X \le 1.5] = \int_{0.3}^{1} x\,dx + \int_1^{1.5} (2 - x)\,dx = \left[\frac{x^2}{2}\right]_{0.3}^{1} + \left[2x - \frac{x^2}{2}\right]_1^{1.5} = 0.83$

Mathematical Expectation

Definition 8.2.1 If X is a discrete random variable which can take finite or infinite sequence of different possible values $x_1, x_2, \ldots\ldots x_n, \ldots\ldots$ with corresponding probabilities $f(x_1), f(x_2), \ldots\ldots f(x_n), \ldots\ldots$ then the mathematical expectation of the random variable X, denoted by μ, is defined by $\mu = E[X] = \sum_{i=1} x_i f(x_i)$
Provided that the series (8.2.1) is absolutely convergent.

Example 8.2.2. A fair coin is tossed until a head appears. What is the expected number of tossed required to get the first head?

Solution:

Let X be the number of tosses required to get a head. Let H and T denote the head and tail of the coin, respectively, and P[H] = P[T]= 1/2 since the coin is fair. The sample points of the experiment, number of tosses and their probabilities are given in the table 8.2.1.

Table 8.2.1

Sample points	H	TH	TTH	TTTH	.....	TTT...TH	.....
Number of tosses X:x	1	2	3	4	.....	n	.....
Probability: f(x)	1/2	1/4	1/8	1/16	.....	$\frac{1}{2^n}$	.....

According to definition 8.2.1:

$\mu = E[X] = \sum_{i=1}^{\infty} x_i f(x_i) = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 4 \times \frac{1}{16} + \cdots + n \times \frac{1}{2^n} + \cdots$

This is an arithmetic-geometric series with geometric progression being $\frac{1}{2}$

Let $S = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 4 \times \frac{1}{16} + \cdots$ (8.2.2)

$\frac{1}{2} S = 1 \times \frac{1}{4} + 2 \times \frac{1}{8} + 3 \times \frac{1}{16} + 4 \times \frac{1}{31} + \cdots$ (8.2.3)

By Subtracting (8.2.3) from (8.2.2) we have

$\frac{1}{2} S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = ½ \left[1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots\right] = \frac{1}{2}\left[\frac{1}{1-\frac{1}{2}}\right]$

$\frac{1}{2} S = \frac{1}{2} \times 2 = 1$

Therefore, S = 2 That is $E[X] = \mu = 2$

This means that the coin tossed is fair, the expected number of tosses required to get a head is 2.

- **Example 8.2.3. Suppose that a game is played with a single die assumed fair. In this game a player wins taka 4 if a 2 turns up, taka 8 if a 4 turns up, loses Taka 6 if 6 turns up; while he neither wins nor loses if any other face turns up. Find the expected sum of money won.**

Suppose:

Let X be a random variable giving the amount of money won on any toss. The possible amount won when the die turns up 1, 2, 3, 4, 5, 6 are $x_1, x_2, x_3, x_4, x_5, x_6$ respectively, while the probabilities of these are $f(x_1), f(x_2), \ldots f(x_6)$The probability function for X is displayed in Table 8.2.2.

Table 8.2.2

Value of X: x	0	4	0	8	0	-6
Probability: $f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Thus, the expected value or expectation is

$$E[X] = 0 \times \frac{1}{6} + 4 \times \frac{1}{6} + 0 \times \frac{1}{6} + 8 \times \frac{1}{6} + 0 \times \frac{1}{6} + (-6) \times \frac{1}{6} = \frac{2}{3} + \frac{4}{3} - 1 = 1$$

It follows that the player can expect to win taka one. In a fair game, therefore, he should be expected to pay taka 1 to play the game.

- **Example 8.2.4. A lottery sells 10,000 tickets at Taka 5 per ticket, a prize Haka 5,000 will be given to the winner of the first draw. Suppose. We have bought a ticket; how much should we expect to win?**

Solution:

Here the random variable X has two possible values taka -5 and Taka 4,995. Their respective probabilities are $\frac{9,999}{10,000}$ and $\frac{1}{1000}$

Thus, the expected value of win is

$$E[X] = -5 \times \frac{9,999}{10,000} + 4,995 \times \frac{1}{10,000} = -\frac{149,995}{10,000} + \frac{4,995}{10,000} = -\frac{45,000}{10,000} = -4.5$$

Thus amount, a minus taka 4.5 is the amount we expect to win on the average if we play this game over and over again.

Definition 8.2.2 If X is a continuous random variable with probability density function $f(x)$, then the mathematical expectation of X is defined by

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Provided that the integral in equation (8.2.4) exists.

- **Example 8.2.11. Let X be a discrete random variable with probability function**

Value of X: x	1	4	9
$f(x)$	1	4	5

Find the mean of $nX, X^2, \sqrt{x}$ and $2\sqrt{x}$.

Solution:

By definition, mean of X is $\mu = E[X] = \sum x_i f(x_i)$

$$= 1 \times 0.1 + 4 \times 0.4 + 9 \times 0.5 = 0.1 + 1.6 + 4.5 = 6.2$$

Mean of X^2. Here $g(X) = X^2$

$$E(X^2) = \sum x_i^2 f(x_i) = 1 \times 0.1 + 16 \times 0.4 + 81 \times 0.5 = 47.0$$

Now, $g(X) = \sqrt{X}$

Hence,

$$E[\sqrt{X}] = \sum \sqrt{x_i} f(x_i) = 1 \times 0.1 + 2 \times 0.4 + 3 \times 0.5 = 2.4$$

Now, mean of $2\sqrt{X}$ is

$$E[2\sqrt{x}] = 2[\sum \sqrt{x_i} f(x_i)] = 2 \times 2.4 = 4.8$$

- **Example 8.2.12 Let X be a random variable with density function**

$$f(x) = \begin{cases} \frac{1}{9}(4x - x^2), & for\ 1 < x < 4, \\ 0, & otherwise. \end{cases}$$

Find $E[X], E[X^2]$ and $E[X - 1]$.

Solution:

By definition, mean is

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_1^4 x\frac{1}{9}(4x^2 - x^2)dx = \frac{1}{9}\int_1^4 (4x^2 - x^3)dx$$

$$\frac{1}{9}\left[\frac{4}{3}x^3 - \frac{x^4}{4}\right]_1^4 = \frac{252}{45} = 5.6\ and$$

$$E[X - 1] = E[X] - 1 = 2.25 - 1 = 1.25$$

Properties of expectation.
Suppose that X is a random variable for which expectation E[X] exists. We shall now present a few theorems relating to expectation. All the results that will be presented here are valid for both discrete and continuous random variables. Proofs will be given here for discrete random variable only.

Theorem 8.2.1. If X is a random variable with expectation E[X], then $E[aX + b] = aE[X] + b$ for any numerical constants a and b,

Proof:
Let X be a discrete random variable which can take values $x_1, x_2, \ldots\ldots, x_n$ with probabilities $f(x_1), f(x_2), \ldots\ldots, f(x_n)$. Now, from the definition of expectation, we have

$$\begin{aligned} E[X] &= \sum(ax_i + b)f(x_i) \\ &= (ax_1 + b)f(x_1) + (ax_2 + b)f(x_2) + \cdots + (ax_n + b)f(x_n) \\ &= a\{x_1f(x_1) + x_2f(x_2) + \cdots + x_nf(x_n)\} + b\{f(x_1) + f(x_2) + \cdots + f(x_n)\} \\ &= a\sum x_if(x_i) + b\sum f(x_i) \\ &= aE[X] + b \end{aligned}$$

This completes the proof of the theorem 8.2.1
To prove the above theorem, we assume that X has only finite set of possible values. The result is also true fir a countable infinite sequence of values of the random variable provided $E[X]$ exists.

Corollary 8.2.1 If a = 0, we have from (8.2.5) $E[b] = b$
That is, expectation of a constant quantity is the constant quantity itself.

Corollary 8.2.2 If b = 0, we have from (8.2.5), $E[aX] = aE[X]$.
That is expectation of 'a' times the expectation of $E[X]$.

Theorem 8.2.2. The expected value of the sum or difference of two or more functions of a random variable X is the sum or difference of the expected values of the function, that is

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)]$$

Where both $g(X)$ and $h(X)$ are functions of X.
Proof:
By definition, $E[g(X) + h(X)] = \sum[g(x_i) \pm h(x_i)]f(x_i)$

$$\begin{aligned} &= \sum g(x_i)f(x_i) \pm h(x_i)f(x_i) \\ &= E[g(X)] \pm E(h[X])] \end{aligned}$$

The results can be generalized for more than 2 functions of the random variable X.

- **Example 8.2.14. Let X be a discrete random variable with probability functions:**

Values of X: x	-2	-1	0	1	2
Probability: $f(x)$	0.2	0.1	0.3	0.3	0.1

i. **Compute $E[X]$.**
ii. **Find the probability function of the function $3X - 1$ and then compute $E[3X - 1]$ and compare your answer with $3E(X) - 1$.**
iii. **Find $E[3x + 3]$ and compare your answer with $3E(X) + 3$.**
iv. **Find the probability function for X^2 and compute $E[X^2]$ and $E[X^2 + 1]$.**

Solution:
By the definition of expectation, we have

i. $E[X] = \sum x_if(x_i) = (-2).(.2) + (-1).(.1) + 0.(.3) + 1.(.3) + 2.(.1) = 0$
ii. Let $Y = 3X - 1$
When $x_1 = -2$, $y_1 = -7$ and $p[Y = y_1] = f(y_1) = f(x_1) = .2$,

When $x_2 = -1$, $y_2 = -4$ and $p[Y = y_2] = f(y_2) = f(x_2) = .1$,
When $x_3 = 0$, $y_3 = -1$ and $p[Y = y_3] = f(y_3) = f(x_3) = .3$,
When $x_4 = 1$, $y_4 = 2$ and $p[Y = y_4] = f(y_4) = f(x_4) = .3$,
When $x_5 = 2$, $y_5 = 5$ and $p[Y = y_5] = f(y_5) = f(x_5) = .1$.
Therefore, the probability function of $Y = 3X - 1$ is

Value of Y: y	-7	-4	-1	2	5
Probability: $f(y)$	0.2	0.1	0.3	0.3	0.1

Therefore, $= E[Y] = E[3X - 1]$
$= \sum y_i f(y_i) = (-7).(0.2) + (-4).(0.1) + (-1).(0.3) + 2.(0.3) + 5.(0.1)$
$= -1.4 - 0.4 - 0.3 + 0.6 + 0.5 = -1.$
And $3E[X] - 1 = 3.(0) - 1 = -1$
Therefore, $E[3X - 1] = 3E(X) - 1$; Which follows from the theorem 8.2.1.

iii. Let $Y = 3X + 3$. When X takes the values -2, -1, 0, 1 and 2, Y takes the values -3, 0, 3, 6 and 9 with probabilities 0.2, 0.1, 0.3, 0.3 and 0.1 respectively.

Hence, $E[Y] = E[3X + 3] = (-3).(.2) + 0.(.1) + 3.(.3) + 6.(.3) + 9.(.1)$
$= -.6 + .9 + 1.8 + .9 = 3.0$
And $3E(X) + 3 = 3.(0) + 3 = 3$
Therefore, $E[3X + 3] = 3E(x) + 3$, which follows from the theorem 8.2.1.

iv. The probability function of $Y = X^2$ is

Y	4	1	0	1	4
$f(y)$	0.2	0.1	0.3	0.3	1

$Y = y$	0	1	4
$f(y)$	0.3	0.4	0.3

Therefore, $E[X^2] = E[Y] = 0.(.3) + 1.(.4) + 4.(.3) = 1.6$
$E[X^2 + 1] = E[X^2] + 1 = 1.6 + 1 = 2.6$

➢ **Example 8.2.16. Let X be a continuous random variable with pdf**

$$f(x) = \frac{3}{4}x(2 - x), \qquad 0 \le x \le 2$$

Find mean, median, mode and harmonic mean of X and comment.

Solution:
By definition,

$$\text{Mean} = E[X] = \frac{3}{4}\int_0^2 x.x(2 - x)dx = \frac{3}{4}\int_0^2 (2^2 - x^2)dx = \frac{3}{4}\left[\frac{2x^3}{3} - \frac{x^4}{4}\right]_0^2 =$$

$$\frac{3}{4}\left[\frac{16}{3} - \frac{16}{4}\right] = 4 - 3 = 1$$

If M is the median of X, then

$$\int_0^M f(x)dx = \frac{1}{2}$$
$$\Rightarrow \frac{3}{4}\int_0^M x(2 - x)dx = \frac{1}{2}$$
$$\Rightarrow \left[\frac{2x^2}{2} - \frac{x^3}{3}\right]_0^M = \frac{2}{3}$$
$$\Rightarrow M^2 - \frac{M^3}{3} = \frac{2}{3}$$
$$\Rightarrow 3M^2 - M^2 = 2$$
$$\Rightarrow M^3 - 3M^2 + 2 = 0$$

$$\Rightarrow (M-1)(M^2-2M^2-2)=0$$

The only values of M lying in [0,2] is M = 1

Hence median is 1.

Mode of the distribution

Now, $f'(x)=\frac{3}{4}(2-2x)=0 \Rightarrow x=1$

And $f''(x)=\frac{3}{2}<0.$

Hence the mode of the distribution is 1.

Harmonic mean H is given by

$$\frac{1}{H}=\int_0^2 \frac{1}{x}f(x)dx=\frac{3}{4}\int_0^2 \frac{x(2-x)}{x}dx=\frac{3}{4}\int_0^2 (2-x)dx=\frac{3}{4}\left[2x-\frac{x^2}{2}\right]_0^2$$

$$=\frac{3}{4}[4-2]=\frac{3}{2}\Rightarrow H=\frac{3}{2}$$

Comments: For this distribution mean = median = mode = 1.

Hence the distribution is symmetrical.

Definition 8.3.1. If X is a random variable with mean , then the variance of X, is denoted by var $[X]$, is defined by

$$\text{Var}[X]=E[X-E(X)]^2=E[X-\mu]^2$$

$$=\begin{cases}\sum(x_i-\mu)^2 f(x_i), & \text{if } X \text{ is a discrete random variable}\\ \int_{-\infty}^{\infty}(x_i-\mu)^2 f(x)dx, & \text{if } X \text{ is a continuous random variable}\end{cases}$$

Provided that the expectation $E[X-\mu]^2$ exists.

- **Example 8.3.2. Let X be a continuous random variable with density function**

$$f(x)=\begin{cases}2x & \text{if } 0<x^2<1\\ 0 & \text{otherwise}\end{cases}$$

Find the variance and standard deviation of X.

Solution:

By definition of variance, we have

$$\sigma^2=V[X]=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx=\int_0^1(x-\mu)^2 2xdx$$

Here, $\mu=E[X]=\int_0^1 x2xdx=\frac{2}{3}$

Hence, $\sigma^2=\int_0^1\left[x-\frac{2}{3}\right]^2 2xdx=\int_0^1\left[x^2-\frac{4}{3}x-\frac{4}{9}\right]dx$

$$=\int_0^1 2x^3dx-\frac{8}{3}\int_0^1 x^2dx+\frac{8}{9}\int_0^1 xdx=\frac{1}{2}-\frac{8}{9}+\frac{4}{9}=\frac{1}{18}$$

Therefore, standard deviation $=\sigma=\sqrt{\frac{1}{18}}=0.23$

- **Example 8.3.4. Let X be a continuous random variable with density function**

$$f(x)=\begin{cases}\frac{1}{8}, & \text{if } 0\le x\le 8\\ 0, & \text{otherwise}\end{cases}$$

(a) Find the variance of X, (b) compute var $[3x+2]$ and compare with value of 9Var$[X]$

Solution:

By definition, variance of X is

$$Var[X]=\sigma^2=E[X-E(X)]^2=E[X^2]-(E[X])^2$$

Which follows the theorem 8.3.3.

Now, $E[X^2]=\int_0^8 x^2 f(x)dx=\frac{1}{8}\int_0^8 x^2dx=\frac{1}{8}\left[\frac{x^3}{3}\right]_0^8=\frac{64}{3}$

And, $E[X] = \int_0^8 xf(x)dx = \frac{1}{8}\left[\frac{x^2}{3}\right]_0^8 = \frac{64}{16} = 4$

Therefore, $var[X] = \frac{64}{3} - 4^2 = \frac{64}{3} - 16 = \frac{16}{3}$

According to theorem 8.3.2. $var[3X + 2] = 9var[X] = 9 \times \frac{16}{3} = 48$

Now, $Var[3X + 2] = E[3X + 2 - E(3x + 2)] = 9Var[X] = 9 \times \frac{16}{3} = 48$

➢ **Example 8.3.5. Let X be a random variable with $E[X] = 1$ and $E[X(X-1)] = 4$. Find $Var[X]$ and $Var[2 - 3X]$.**

Solution:

According to corollary 8.3.1,

$Var[X] = E[X(X-1)] + \mu - \mu^2 = 4 + 1 - 1 = 4$

$Var[2 - 3X] = var[-3X] = 9var[-x] = 9var[X] = 9 \times 4 = 36$

Mean deviation

Like variance, mean deviation is also an important measure of dispersion.

Definition 8.3.3. Let X be a random variable with probability function $f(x)$ or pdf $f(x)$ if X is a discrete or continuous, then mean deviation (M.D.) about mean (μ) is defined as

$$\text{M.D. [mean]} = \text{M.D. } (\mu) = E[|X - \mu|] = \begin{cases} \sum|x_i - \mu|f(x_i) \\ \int_{-\infty}^{\infty}|x - \mu|f(x)dx \end{cases}$$

Provided the expectation $E[|X - \mu|]$ exists.

➢ **Example 8.3.6. If X is a continuous random variable with pdf**

$$f(x) = \frac{3}{4} \times (2 - x) \qquad 0 \le x \le 2$$

Find the mean deviation about mean.

Solution:

Mean $= E[X] = \int_0^2 xf(x)dx = \frac{3}{4}\int_0^2 x(2-x)dx = 1$

By definition, mean deviation about mean

$$= \int_0^2 |x-1|f(x)dx$$
$$= \int_0^1 |x-1|f(x)dx + \int_1^2 |x-1|f(x)dx$$
$$= \frac{3}{4}\left[\int_0^1 (1-x) \times (2-x)\,dx + \int_1^2 (1-x) \times (2-x)\,dx\right]$$
$$= \frac{3}{4}\left[\int_0^1 (2x - 3x^2 + x^3)dx + \int_1^2 (3x^2 - x^3 - 2x)dx\right]$$
$$= \frac{3}{4}\left[\frac{2x^2}{2} - \frac{3x^3}{3} + \frac{x^4}{4}\right]_0^1 + \frac{3}{4}\left[\frac{3x^3}{3} - \frac{x^4}{4} - \frac{2x^2}{2}\right]_1^2$$
$$= \frac{3}{16} + \frac{3}{16} = \frac{6}{16} = \frac{3}{8}$$

➢ **Example 8.3.7. If X is a continuous random variable with pdf**

$$f(x) = \frac{1}{2}e^{-|x|}; \qquad -\infty < x < \infty$$

Show that the mean deviation about mean is 1.

Solution:

By definition,

$$\text{Mean} = \mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} xe^{-|x|}\,dx$$
$$= \frac{1}{2}\int_{-\infty}^{0} xe^{x}\,dx + \frac{1}{2}\int_0^{\infty} xe^{-x}dx$$
$$= \frac{1}{2}\left[-\int_0^{\infty} xe^{-x}\,dx + \int_0^{\infty} xe^{-x}\,dx\right]$$

(By putting $x = -x$ in the 1st integral)

$= 0$

M.D. about mean $= M.D.[\mu] = \int_{-\infty}^{\infty} |x - mean| f(x) dx$

$= \frac{1}{2}\int_{-\infty}^{\infty} |x| e^{-|x|}\, dx$ (Mean $= \mu = 0$)

$= \frac{1}{2}\left[\int_{-\infty}^{0} -xe^{x}\, dx + \int_{0}^{\infty} xe^{-x}\, dx\right]$

$= \frac{1}{2}\left[\int_{-\infty}^{0} xe^{-x} dx + \int_{0}^{\infty} xe^{-x}\right]$ (By putting $x = -x$)

$= \int_{0}^{\infty} xe^{-x} dx = 1$

Joint Distribution

Introduction

If x and y are discrete random variable the function given by f(x, y) = P(x=x, y=y) for each pair of values (x, y) within the range of x is called the joint probability distribution of x and y.

Joint discrete probability function.

Let x and y be two discrete random variables of a given probability space (A, Ω[.]) then the joint probability function of (x, y), denoted by f(x, y) is defined by $F(x, y) = P(X = x, Y = y)$,
Where (x, y) is a value of (x, y)

It is to be noted that, like probability function of a single random variable, the joint probability function of (x, y) satisfies the properties:
(1) F(x, y) ≥0 for all i and j, and (2) $\sum_i \sum_j f(x, y) = 1$

Marginal and conditional probability function

Marginal probability function

Let x and y be two discrete random variables with joint probability function $f(x, y)$ then the marginal probability function of x, denoted by $g(x)$ is defined by
$G(x) = P[X = x] = \sum_{j=1}^{n} f(x, y)$, for all x.

values of X:x / values Y:y	0	1	2	3	4
0	0.23	0.23	0.11	0.04	0.01
1	0.11	0.11	0.05	0.02	0.01
2	0.03	0.03	0.01	0.01	00

Example 9.3.1. Suppose two discrete random variables x and y have the following joint probability function (1) Find the marginal probability function of x and y. (2) find the probability that $x = 4$ and $y = 2$ (3) Find $P[x = 3], P[y = 1]$, (4) find $P[x \geq 2, y = 2], P[x = 2, y = 2], P[x > 2, y < 1]$, (5) find $f(3, 2)$ and $f(2, 1)$.

Solution:

(1) The marginal probability function of x and y is

Value of X:x / Value of Y:y	0	1	2	3	4	Probability : h(y)
0	0.23	0.23	0.11	0.04	0.01	0.62
1	0.11	0.11	0.05	0.02	0.01	0.30

2	0.03	0.03	0.01	0.01	00	0.08
Probability: g(x)	0.37	0.37	0.17	0.07	0.02	1

(2)

$$P[x = 4, y = 2] = f(4,2) = .00 = 0.$$

(3)

$$P[x = 3] = g[x = 3] = g(3) = \sum y\ f(3,y) = f(3,0) + f(3,1) + f(3,2) = 0.04 + 0.02 + 0.01 = 0.07$$

$$P[y = 1] = h(y = 1) = h(1) = \sum f(x,1) = f(0,1) + f(1,1) + f(2,1) + f(3,1) + f(4,1)$$
$$= 0.11 + 0.11 + 0.05 + 0.02 + 0.01 = 0.3$$

The probabilities P[x=3] and P[y=1] can also be found from the marginal probability functions of X and Y.

(4)

$$P[x \geq 2, y = 2] = P[x = 2, y = 2] + P[x = 3, y = 2] + P[x = 4, y = 2] = f(2,2) + f(3,2) + f(4,2)$$
$$= 0.01 + 0.01 + 00 = 0.02$$
$$P[x = 2, y = 2] = f(2,2) = 0.01$$
$$P[x > 2, y < 1] = P[x = 3, y = 0] + P[x = 4, y = 0] = f(3,0) + f(4,0) = 0.04 + 0.01 = 0.05$$

(5)

$f(3,2) = \sum_x^3 \sum_y^2 f(x,y) = f(0,0) + f(0,1) + f(0,2) + f(1,0) + f(1,1) + f(1,2) + f(2,0) + f(2,1) +$
$f(2,2) + f(3,0) + f(3,1) + f(3,2) = 0.23 + 0.11 + 0.03 + 0.23 + 0.11 + 0.03 + 0.11 + 0.05 + 0.01 + 0.04 +$
$0.02 + 0.01 = 0.98$ and
$F(2,1) = \sum_x^2 \sum_y^1 f(x,y) = f(0,0) + f(0,1) + f(1,0) + f(1,1) + f(2,0) + f(2,1) = 0.23 + 0.11 + 0.23 +$
$0.11 + 0.11 + 0.05 = 0.84$

Conditional Probability Function.

Definition

Let X and Y be two discrete random variables with marginal probability functions $g(x)$ and $h(y)$ respectively and let $f(x,y)$ by the joint probability function of X and Y. Then conditional probability function of X for given $Y = y$, denoted by $f(x|y)$, is defined by, $f(x/y) = \frac{f(x,y)}{h(y)} = \frac{p[X=x, Y=y]}{p[Y=y]}$

- **Example 9.3.2. Let x and y be two discrete random variables with joint probability function as follow:**

value of: Y:y / Value of: X:x	0	1	2
-1	0.2	0.1	0.05
0	0.05	0.2	0.3
3	0.02	0.04	0.04

(1) Find the marginal probability function of X and Y
(2) Find the conditional probability function of x for given $y = 2$
(3) Find the conditional probability function of y for given $x = 0$,
(4) Find f(x=0/y=1)

Solution:

(1) The marginal probability function of x and y is

Value of Y:y / Value of X:x	0	1	2	Marginal probability: g(x)
-1	0.2	0.1	0.05	0.35
0	0.05	0.2	0.3	0.55

3	0.02	0.04	0.04	0.1
Probability : h(y)	0.27	0.34	0.39	1

(2) According to definition,

$F(x/y=2) = \frac{f(x,2)}{h(y=2)}$, for x = -1,0,3

When x=-1, $f(x=-1/y=2) = \frac{f(-1,2)}{h(y=2)} = \frac{.05}{.39} = \frac{5}{39}$

When x=0, $f(x=0/y=2) = \frac{f(0,2)}{h(y=2)} = \frac{.3}{.39} = \frac{30}{39}$

When x=3, $f(x=3/y=2) = \frac{f(3,2)}{h(y=2)} = \frac{.04}{.39} = \frac{4}{39}$

Therefore, the conditional probabilities function of x for given y=2 is

Value of X:x	-1	0	3
F(x/y=2)	5/39	30/39	4/39

(3) According to the definition of conditional probability function of y given x=0 is,

$F(y/x=0) = \frac{f(y,0)}{g(x=0)}, for\ all\ y = 0,1,2.$

Now, when y=0, $f(y=0/x=0) = \frac{f(0,0)}{g(x=2)} = \frac{.05}{.55} = \frac{1}{11}$

When y=1, $f(y=1/x=0) = \frac{f(1,0)}{g(x=0)} = \frac{.2}{.55} = \frac{4}{11}$

When y=2, $f(y=2/x=0) = \frac{f(2,0)}{g(x=0)} = \frac{.3}{.55} = \frac{6}{11}$

Therefore, the conditional probability function of y for given x=0 is

Values of Y:y	0	1	3
h(y/x=0)	1/11	4/11	6/11

(4) According to definition.

$F(x=0/y=1) = \frac{f(0,1)}{h(y=1)} = \frac{.2}{.34} = \frac{10}{17}$

Independence of two discrete random variable

Definition 9.4.1. Let X and Y be two discrete random variable with marginal probability functions g(x) and h(y), respectively. Let f(x, y) be the joint probability function of X and Y, then X and Y are said to be independent if and only if F(x, y) =g(x).h(y). For all values of X and Y.

- **Example 9.4.3. Consider the following joint probability function of x and y.**

Values of Y:y / Values of X:x	-4	2	7	g(x)
1	1/8	1/4	1/8	1/2
5	1/4	1/8	1/8	1/2
h(y)	3/8	3/8	1/4	1

Are x and y independent?

Solution:

The marginal probability function of x is

$g(x) = \sum_y f(x,y); for\ x = 1,5.$

When x=1, $g(x=1) = g(1) = \sum_{-4}^{7} f(1,y) = f(1,-4) + f(1,2) + f(1,7) = \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}$

When x=5, g(x=5) = g(5) = $\sum_{-4}^{7} f(5,y) = f(5,-4) + f(5,2) + f(5,7) = \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$

Similarly, the marginal probability function of y is

When y= -4, h(y= -4) = $\sum_x f(x,-4)$ = f(1,-4)+f(5,-4)= $\frac{1}{8} + \frac{1}{4} = \frac{3}{8}$

When y=2, h(y= 2) = $\sum_x f(x,2) = f(1,2) + f(5,2) = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$

When y=7, h(y=7) = $\sum_x f(x,7) = f(1,7) + f(5,7) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$

The random variables x and y will be independent if

F(1,-4) = g(1).h(4)

Here, f(1,-4) = $\frac{1}{8}$ *and* $g(1).h(-4) = \frac{1}{2}\frac{3}{8} = \frac{3}{16}$

Thus, f(1,-4) ≠ g(1).h(-4)

Therefore, x and y are not independent.

- **Example 9.4.4.**

Let x and y be two random variables with the joint probability function.

values of X:x \ values of Y:y	-2	5	8
1	.21	.35	.14
2	.09	.15	.06

Are x and y independent?

Solution:

The marginal probability functions of x and y will be as follows:

Values of X:x \ Values of Y:y	-2	5	8	Probability: g(x)
1	0.21	0.35	0.14	0.7
2	0.09	0.15	0.06	0.3
Probability: h(y)	0.30	0.50	0.20	1

The random variables x and y will be independent if (say)

F(1,5) =g(1).h(5)

Here, f(1,5)=.35 and g(x=1)=g(1)=.7 ,h(y=5)=h(5)=.5

Which show that

F(1,5)=g(1).h(5)=.35.

Therefore, x and y are independent.

Joint continuous random variable

Definition: Let X and Y be two random variables defined on a continuous sample space Ω, then (X,Y) is said to be a joint or two dimensional continuous random variable if and only if there exists a function f(x,y)≥0 such that,

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy.$$

- **Example 9.5.1. Suppose (X,Y) is a joint continuous random variable with joint density function**

$$f(x,y) = \begin{cases} cxy, & if\ 1<x<2,\ 1<y<2 \\ 0, & otherwise \end{cases}$$

(a) Find the value of c for which f(x,y) is a joint probability density function.

(b) Find the joint cumulative distribution function of X and Y.
(c) Find P[X≤1,Y≤5], P[X≥1.2,Y≥1.4], P[1.5≤X≤1.75, 1.5≤Y≤1.75].
(d) Find the probability that the value of X will be greater than the value of Y.
Solution:
(a) If f(x,y) is a joint density function, then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

Thus, $1=\int_1^2\int_1^2 cxy\,dx\,dy = c\int_1^2[\frac{x^2}{2}]_1^2\, y\,dy = \frac{3}{2}c\,[\frac{y^2}{2}]_1^2 = \frac{9}{4}c$

Therefore, $c=\frac{4}{9}$.

Hence if $c=\frac{4}{9}$, then $f(x,y)\geq 0$ for all values of X and Y in the xy-plane as

$1<X<2$ and $1<Y<2$.

(b) The joint cumulative distribution function of X and Y, according to definition, is

$$F(x,y) = \frac{4}{9}\int_1^x\int_1^y xy\,dx\,dy = \frac{4}{9}\int_1^y y\,dy\,[\frac{x^2}{2}]_1^x\,dx$$
$$= \frac{4}{9}\frac{(x^2-1)}{2}[\frac{y^2-1}{2}] = \frac{1}{9}(x^2-1)(y^2-1), \quad \text{if } 1<x<2, 1<y<2.$$

(c) $P[X\leq 1, Y\leq 5]\ \int_1^1\int_1^5 \frac{4}{9}xy\,dx\,dy = 0$, since $\int_1^1 \frac{4}{9}xydx = 0$

$$P[Y\geq 1.2, Y=1.4] = \int_{1.2}^2\int_{1.4}^2 \frac{4}{9}xy\,dx\,dy = \frac{4}{9}\,[\frac{x^2}{2}]_{1.2}^2\,[\frac{y^2}{2}]_{1.4}^2$$
$$= \frac{4}{9}\,(2-0.72)(2-0.98) = \frac{4}{9}\times 1.28\times 1.02 = 0.5803.$$

$$P[1.5\leq X\leq 1.75,\ 1.5\leq Y\leq 1.75] = \frac{4}{9}\int_{1.5}^{1.75}\int_{1.5}^{1.75} xy\,dx\,dy = \frac{4}{9}[\frac{x^2}{2}]_{1.5}^{1.75}\,[\frac{y^2}{2}]_{1.5}^{1.75}$$
$$= \frac{4}{9}\left[\frac{49}{32}-\frac{9}{8}\right]\left[\frac{49}{32}-\frac{9}{8}\right] = \frac{4}{9}\cdot\frac{13}{32}\cdot\frac{13}{32} = \frac{169}{3204}.$$

(d) $P[Y\leq X\leq 2,\ 1\leq Y\leq 2] = \frac{4}{9}\int_1^2\int_y^2 xy\,dx\,dy = \int_1^2\frac{4}{9}[\frac{x^2}{2}]_y^2\,y\,dy$

$$= \frac{4}{9}\int_1^2[2-\frac{y^2}{2}]y\,dy = \frac{4}{9}[y^2-\frac{y^4}{8}]_1^2$$
$$= \frac{4}{9}\left[4-2-1+\frac{1}{8}\right] = \frac{4}{9}\cdot\frac{9}{8} = \frac{1}{2} = 0.5$$

Marginal probability density functions
Definition: If X and Y are two continuous random variable with joint density function f(x,y), then the marginal density function of X and Y, denoted as g(x) and h(y), are defined by

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad for -\infty < x < \infty \text{ and}$$
$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx, \quad for -\infty < y < \infty.$$

Conditional probability density functions
Definition: Let X and Y be two continuous random variables with joint probability density function f(x,y) and if g(x) and h(y) be their marginal probability density functions, then the conditional probability density function of X given Y=y, denoted by f(x|y), is defined by

$$f(x|y) = \frac{f(x,y)}{h(x)}, \qquad -\infty < x < \infty,$$

provided $h(y) > 0$ and undefined at points where $h(y) = 0$.

- **Example 9.5.3. Suppose X and Y have a continuous joint probability density function as follows**

$$f(x,y) = c(x+y^2); \qquad 0\leq x\leq 1 \text{ and } 0\leq y\leq 1$$

(1) Find the value of c.
(2) Determine the conditional probability density function of X for any given of Y and find

$$P[X < \tfrac{1}{2} \mid Y = \tfrac{1}{2}].$$

Solution:

(1) The value of c can be found from the relation $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$, since f(x,y) is a joint probability density function.

Therefore, $\int_0^1\int_0^1 c(x+y^2)\,dx\,dy = 1,$

or $c[\frac{x^2}{2}]_0^1 + c[\frac{y^3}{3}]_0^1 = 1$

or $c.\frac{5}{6} = 1.$

Therefore, $c = \frac{6}{5}$.

(2) Now, by definition, the conditional probability density function of X for given Y=y is

$$f(x|y) = \frac{f(x,y)}{h(y)}, \qquad 0 < x \le 0.$$

Here, $h(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{6}{5}\int_0^1[x+y^2]dx = \frac{6}{5}[\frac{x^2}{2} + y^2x]_0^1 = \frac{6}{5}[\frac{1}{2} + y^2]$

Therefore, $f(x|y) = \frac{\frac{6}{5}[x+y^2]}{\frac{6}{5}[\frac{1}{2}+y^2]} = \frac{2(x+y^2)}{1+2y^2}; \qquad 0 \le x \le 1$

Now, $P[X<\tfrac{1}{2} \mid Y = \tfrac{1}{2}] = \int_1^{1/2} f[x|y=\tfrac{1}{2}] = \int_0^{1/2} \frac{2[x+\frac{1}{4}]}{[1+2.\frac{1}{4}}dx$

$$= \frac{2}{3}\int_0^{1/2}\left[2x+\frac{1}{2}\right]dx = \frac{2}{3}[x^2 + \frac{1}{2}x]_0^{1/2} = \frac{2}{3}[\frac{1}{4}+\frac{1}{4}] = \frac{1}{3}.$$

➢ **Example 9.5.4. Let X and Y be continuous random variable with joint probability density function**

$$f(x,y) = \begin{cases} 2xy+\frac{3y^2}{2}; & 0<x<1,\ 0<y<1, \\ 0, & otherwise. \end{cases}$$

(i) Find the marginal probability density function of X and Y.

(ii) Find the conditional probability density function of X for given value of Y.

(iii)Find the conditional probability density function of Y for given value of X.

(iv)If Y=0.3, what is the probability that X will take on a value that is between 0 and 0.5?

(v) If X=0.2, what is the probability that Y will take on a value that is between 0 and 0.4?

Solution:

(i) The marginal probability density function of X is

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1\left[2xy + \frac{3}{2}y^2\right]dy = [2\frac{y^2}{2}.x + \frac{3}{2}.\frac{y^3}{3}]_0^1 = x + \frac{1}{2}, \quad \text{for } 0 < x < 1$$

Similarly, the marginal probability density function of Y is

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dy; \qquad 0 < y < 1.$$

$$h(y) = \int_0^1\left[2xy + \frac{3}{2}y^2\right]dy = [2y.\frac{x^2}{2} + \frac{3}{2}y^2x]_0^1 = y + \frac{3}{2}y^2, \qquad 0 \le y \le 1.$$

(ii) the conditional density of X for given Y=y is,

$$f(x|y) = \frac{f(x,y)}{h(y)} = \frac{2xy+\frac{3}{2}y^2}{y+\frac{3}{2}y^2} = \frac{2x+\frac{3}{2}y}{1+\frac{3}{2}y} = \frac{4x+3y}{2+3y}, \qquad \text{for } 0 < x < 1.$$

(iii) The conditional density function of Y for given value of X=x is,

$$f(y|x) = \frac{f(x,y)}{g(x)} = \frac{2xy+\frac{3}{2}y^2}{x+\frac{1}{2}} = \frac{4xy+3y^2}{2x+1}, \qquad \text{for } 0 < x < 1.$$

(iv) $P[0 \le X \le 0.5 \mid Y = 0.3]$

$$= \int_0^{0.5} f(x|y=0.3)dx = \int_0^{0.5} \frac{4x+0.9}{2.9}dx = [\frac{4x^2+0.9x}{2.9}]_0^{0.5} = \frac{19}{58}.$$

Similarly, $P[0 \le Y \le 0.4|X = 0.2] = \int_0^{0.4} \frac{0.8y+3y^2}{0.4+1}dy = [\frac{0.4y^2+y^3}{1.4}]_0^{0.4} = \frac{46}{175}.$

➢ **Example 9.5.8. Suppose X and Y are continuous random variables with joint probability density function,**

$$f(x,y) = \begin{cases} 2xy+\frac{3y^2}{2}, & 0<x<1,\ 0<y<1, \\ 0, & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Solution:

By definition, the marginal probability density function of X is,

$$g(x) = \int_0^1 \left[2xy + \frac{3}{2}y^2\right] dy = x+\frac{1}{2}; \qquad 0<x<1.$$

And $$h(y) = \int_0^1 \left[2xy + \frac{3}{2}y^2\right] dx = y+\frac{3}{2}y^2; \qquad 0<x<1.$$

$$f(x,y) = (2xy+\frac{3y^2}{2}); \qquad 0<x<1,\ 0<y<1.$$

Therefore,

$$f(x,y) \neq g(x).h(y).$$

Thus X and Y are not independent.

➢ **Example 9.5.10. The joint density function of two continuous random variables X and Y is,**

$$f(x,y) = \begin{cases} cxy, & 0<x<4,\ 1<y<5, \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the value of the constant c.

(ii) Find P[1<X<2, 2<Y<3],

(iii) Find P[X≥ 3, Y ≤ 2]

(iv) Find the marginal density functions of X and Y and comment on the nature of the random variables.

Solution:

(i) We must have the total probability equal 1,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dx\,dy = 1$$

We have,

$$\int_{x=0}^{4}\int_{y=1}^{5} c\,xy\,dx\,dy = c\int_{x=0}^{4}[\frac{xy^2}{2}]_1^5\,dx = c\int_{x=0}^{4}\frac{25x-x}{2}dx$$

$$= 12c\int_{x=0}^{4} xdx = 12c[\frac{x^2}{2}]_0^4 = \frac{12c.16}{2} = 96c.$$

Then, $96c = 1$ and $c = \frac{1}{96}$.

(ii) Using the value of c found in (i), we have,

$$P[1<X<2, 2<Y<3] = \frac{1}{96}\int_{x=1}^{1}\int_{y=2}^{3} xy\,dx\,dy = \frac{1}{96}\int_{x=1}^{2}[\frac{xy^2}{2}]_2^3\,dx$$

$$= \frac{1}{96}\int_{x=1}^{2}\frac{5x}{2}dx = \frac{5}{192}[\frac{x^2}{2}]_1^2 = \frac{5}{128}.$$

(iii) $P[X \geq 3, Y \leq 2] = \int_{x=3}^{4}\int_{y=1}^{2}\frac{xy}{96}\,dx\,dy = \frac{1}{96}\int_{x=3}^{4}[\frac{xy^2}{2}]_1^2 dx$

$$= \frac{1}{96}\int_{x=3}^{4}\frac{3x}{2}dx = \frac{7}{128}.$$

(iv) By definition, the marginal density function of X is,

$$g(x) = \int_{y=1}^{5}\frac{xy}{96}\,dy = \frac{1}{96}[\frac{xy^2}{2}]_1^2 = \frac{1}{96}[\frac{25x}{2} - \frac{x}{2}] = \frac{1}{8}x, \qquad 0<x<4.$$

By definition, the marginal density of Y is,

$$h(y) = \int_{x=0}^{4}\frac{xy}{96}dx = \frac{1}{96}[\frac{x^2}{2}y]_0^4 = \frac{1}{96}(8y) = \frac{y}{12}, \qquad 1<y<5.$$

It is obvious that $g(x).h(y) = \frac{1}{96}dx = f(x,y)$.

Which shows that the random variable X and Y are independent.

Moment Generating Function and Probability Generating Function

Moment generating function: The moment generating function (MGF) of a random variable X is defined as:

$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$ (for continuous variables)

$M_X(t) = \sum_x e^{tx} p(x)$ (for discrete variables)

Probability Generating Function (PGF)

The probability generating function (PGF) of a discrete random variable X taking non-negative integer values is defined as: $P_X(s) = E[s^X] = \sum_{k=0}^{\infty} s^k P(x)$

The PGF can be used to find the probabilities of the distribution and can also be used to find moments.

Binomial Distribution

For a binomial random variable $X \sim Binomial(n, p)$:

Moment Generating Function:

$$M_X(t) = (q + pe^t)^n$$

Probability Generating Function:

$$P_X(s) = (q + ps)^n$$

Poisson Distribution

For a Poisson random variable $X \sim Poisson(\lambda)$:

Moment Generating Function:

$$M_X(t) = e^{\lambda(e^t - 1)}$$

Probability Generating Function:

$$P_X(s) = e^{\lambda(s-1)}$$

Normal Distribution

For a normal random variable $X \sim N(\mu, \sigma^2)$:

Moment Generating Function:

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Probability Generating Function: The PGF is not typically defined for continuous distributions like the normal distribution.

Probability Distributions

Bernoulli Distribution and its formulation

❖ **Bernoulli distribution with example.**

Answer:

Bernoulli distribution: A discrete random variable X is said to have a Bernoulli distribution if its probability function is

$$P(x;1,p) = \begin{cases} p^x q^{1-x}; x = 0,1 \\ 0; \ Otherwise \end{cases}$$

Where p is the parameter of the distribution satisfying $0 \leq p \leq 1$ and $p + q = 1$

Example. A coin is tossed in which the outcome head is a success and the probability of head is p. Then $q = 1 - p$ is the probability of failure or tail. If the number of heads or successes is a random variable X, then X can take values 0 or 1 according to the outcome is a tail or head.

❖ **Moment generating function of Bernoulli distribution and find skewness and kurtosis. Also, comment on the shape of distribution.**

Answer:

By definition,

$$M_x(t) = E[e^{tx}] = \sum_{x=0}^{1} e^{tx} p^x q^{1-x} = \sum_{x=0}^{1} (pe^t)^x q^{1-x} = (q + pe^t)$$

By definition, the cumulant generating function of x is

$K_x(t) = \log M_x(t)$

$= log(q + pe^t) = log[pe^t + q] = log\left[p\left\{1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots\right\} + q\right] = log\left[1 + p\left\{t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots\right\}\right]$

$= \left[p\left\{t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots\right\} - \frac{p^2}{2}\left\{t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots\right\}^2 + \frac{p^3}{3}\left\{t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots\right\}^3 - \frac{p^4}{4}\left\{t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots\right\}^4\right]$

Now,

K_r *is the coefficient of* $\frac{t^r}{r!}$ *in* $K_x(t)$

$Mean = k_1 = Coefficient\ of\ t\ in\ K_x(t) = p$

$\mu_2 = k_2 = Coefficient\ of \frac{t^2}{2!}\ in\ K_x(t) = (p - p^2) = pq$

Now,

$$\mu_3 = k_3 = Coefficient\ of \frac{t^3}{3!}\ in\ K_x(t)$$

$$= (p - 3p^2 + 2p^3) = (1 - 3p + 2p^2) = p(1-p)(1-2p) = pq(q-p)$$

$k_4 = Coefficient\ of \frac{t^4}{4!}\ in\ K_x(t)$

$= [p - 7p^2 + 12p^3 - 6p^4] = p(1-p)(1-6p+6p^2) = pq[1-6p(1-p)] = pq(1-6pq)$

Hence,

$\mu_4 = k_4 + 3{k_2}^2 = pq(1-3pq)$. [on simplification]

##Comment

Binomial distribution

❖ **Binomial distribution with example.**

Answer:

A discrete random variable X is said to have a Binomial distribution if its probability function is defined by

$$P(X;n,p) = \begin{cases} \binom{n}{x} p^x q^{n-x} & ;x = 0,1,2,\ldots.,n \\ 0 & ;\ Otherwise \end{cases}$$

where the two parameters n and p satisfy $0 \leq p \leq 1$ and n is a positive integer and $q = 1 - p$. A distribution defined by the probability function of X given in the above expression is called binomial distribution.

- Mean and variance of binomial distribution, np and npq respectively, and $np > npq$ that is mean > variance

Example. A coin is tossed n times in which the outcome head is a success and the probability of head is p. Then $q = 1 - p$ is the probability of failure or tail. If the number of heads or successes is a random variable X, then X can take values 0, 1, 2, … n according to the outcome is a tail or head.

Again for example, if we toss two coins and we are interested in the probability of getting exactly one head, we can model this situation using the binomial distribution with $n = 2$ (the number of trials) and $p = \frac{1}{2}$ (the probability of success, i.e., getting a head on each trial).

Using the binomial distribution formula, we can calculate the probability of getting exactly one head in two tosses:

$$P\left(1;2,\tfrac{1}{2}\right) = \binom{2}{1}\left(\tfrac{1}{2}\right)^1\left(1-\tfrac{1}{2}\right)^{2-1} = \tfrac{1}{2}$$

So, the probability of getting exactly one head in two-coin tosses is $\frac{1}{2}$, or 50%. We can also calculate the probabilities of getting zero or two heads using the same formula:

$$P\left(0;2,\tfrac{1}{2}\right) = \binom{2}{1}\left(\tfrac{1}{2}\right)^0\left(1-\tfrac{1}{2}\right)^{2-0} = \tfrac{1}{4}$$

$$P\left(1;2,\tfrac{1}{2}\right) = \binom{2}{1}\left(\tfrac{1}{2}\right)^1\left(1-\tfrac{1}{2}\right)^{2-1} = \tfrac{1}{2}$$

So, the probabilities of getting zero or two heads in two-coin tosses are also $\frac{1}{4}$ or 25% each.

- Relation of binomial distribution with other distributions:
  - Transformation of binomial to Poisson:
    - p, the probability of success in a Bernoulli trial is very small, that is $p \to 0$
    - n, the number of trials is very large, that is $n \to \infty$
    - $np = \lambda$ is finite constant, that is average number of successes is finite.
  - When $n = 1$, binomial distribution reduces to Bernoulli distribution with parameter p
  - When p and q are not so small and n tends to infinity, then binomial distribution tends to normal distribution.

❖ **Mean and variance of Binomial distribution and mean is greater than variance for binomial distribution.**

$$mean = E(x) = \sum x.p(x) = \sum_{x=0}^{n} x\binom{n}{x}p^x q^{n-x} \qquad [p^x = p^{x-1}.p]$$

$$= \sum_{x=0}^{n} x.\frac{n!}{x!(n-x)!}\,p^x q^{n-x} \qquad \left[\sum_{x=0}^{n}\frac{n!}{x!(n-x)!}\,p^x q^{n-x} = 1\right]$$

$$= 0.\frac{n!}{0!(n-0)!}p^0 q^{n-0} + \sum_{x=1}^{n} x.\frac{n(n-1)!}{x(x-1)!(n-x)!}p^x q^{n-x}$$

$$= np\sum_{x=1}^{n}\frac{(n-1)!}{(x-1)}p^{x-1}q^{n-x}$$

$$= np(p+q)^{n-1} = np.1 = np$$

This is the mean of binomial distribution.

$$Variance, V(x) = E(x^2) - \{E[(x)]\}^2$$

$$E(x^2) = \sum_{x=0}^{n} x^2 P(x) \qquad [x^2 = x^2 - x + x]$$

$$= \sum_{x=0}^{n}\{x(x-1)+x\}P(x)$$
$$= \sum_{x=0}^{n}\{x(x-1)P(x)\} + \sum_{x=0}^{n} x\,.P(x)$$

Now,

$$\sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x q^{n-x} = \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x q^{n-x} = \sum_{x=2}^{n}\frac{x(x-1)(n(n-1)(n-2)!)}{x(x-1)(x-2)!\,(n-x)\,!}p^{x-2}p^2 q^{n-x}$$
$$= p^2 n(n-1)\sum_{x=2}^{n}\frac{(n-2)!}{(x-2)\,!\,(n-x)\,!}p^{x-2}q^{n-x} = p^2 n(n-1)(p+q)^{n-2}$$
$$= n(n-1)\,p^2$$
$$E[x^2] = n(n-1)p^2 + np$$

We have,

$$V(x) = n(n-1)p^2 + np - n^2p^2$$
$$= n^2p^2 - np^2 + np - n^2p^2$$
$$= np - np^2 = np(1-p) = npq$$

Which is the variance of binomial distribution.

$$E(x) = np;\ V(x) = npq$$

Since p and q are functions npq will always be less than np where n is fixed for both cases.

- **Moment generating function of binomial distribution and find skewness and kurtosis. Also, comment on the shape of distribution.**

Proof:

By definition,

$$M_x(t) = E[e^{tx}] = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x q^{n-x} = \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x q^{n-x} = (q+pe^t)^n$$

By definition, the cumulant generating function of x is

$$K_x(t) = \log M_x(t) = \log(pe^t+q)^n = nlog[pe^t+q] = nlog\left[p\left\{1+t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots\right\}+q\right]$$
$$= nlog\left[1+p\left\{t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots\right\}\right]$$
$$= n\left[p\left\{t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots\right\}-\frac{p^2}{2}\left\{t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots\right\}^2+\frac{p^3}{3}\left\{t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots\right\}^3-\frac{p^4}{4}\left\{t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots\right\}^4\right]$$

Now,

K_r *is the coefficient of* $\frac{t^r}{r!}$ *in* $K_x(t)$

$Mean = k_1 = Coefficient\ of\ t\ in\ K_x(t) = np$

$\mu_2 = k_2 = Coefficient\ of \frac{t^2}{2!}\ in\ K_x(t) = n(p-p^2) = npq$

Now,

$$\mu_3 = k_3 = Coefficient\ of \frac{t^3}{3!}\ in\ K_x(t)$$
$$= n(p-3p^2+2p^3) = n(1-3p+2p^2) = np(1-p)(1-2p) = npq(q-p)$$
$$k_4 = Coefficient\ of \frac{t^4}{4!}\ in\ K_x(t)$$

$= n[p-7p^2+12p^3-6p^4] = np(1-p)(1-6p+6p^2) = npq[1-6p(1-p)] = npq(1-6pq)$

Hence,

$\mu_4 = k_4 + 3{k_2}^2 = npq(1-6pq) + 3n^2p^2q^2 = npq[1+3(n-2)pq]$. [on simplification]

Shape characteristics of the distribution

β_1 and β_2 measure the skewness and kurtosis of a distribution,

$$Hence, \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(npq(q-p))^2}{(npq)^3} = \frac{(1-2p)^2}{npq}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq(1-6pq)+3n^2p^2q^2}{(npq)^2} = \frac{1-6pq}{npq}$$

Now, $\gamma_1 = \sqrt{\beta_1} = \frac{1-2p}{\sqrt{npq}}$

Remarks

❖ **Theorem 01**

If X_1 and X_2 are two independent binomial variates with parameters (n_1, p) and (n_2, p) then $X = X_1 + X_2$ is a binomial variate with parameter $(n_1 + n_2, p)$

Answer:

We know that the $m.g.f.$ of x_1 (binomial variate) is $(q + pe^t)^{n_1}$

And again,

the $m.g.f.$ of x_2(binomial variate) is $(q + pe^t)^{n_2}$

$$M_{x_1+x_2}(t) = E\left(e^{(x_1+x_2)t}\right) = E(e^{tx_1}).E(e^{tx_2}) = (q + pe^t)^{n_1}.(q + pe^t)^{n_2} = (q + pe^t)^{n_1+n_2}$$

$$= (q + pe^t)^n$$

Which is the $m.g.f.$ of binomial variate with $n_1 + n_2$ and p. [proved]

❖ **Theorem 02**

If X_1 and X_2 are two independent binomial variates with parameters (n_1, p) and (n_2, p) then the conditiona distribution of X_1 for given $X = X_1 + X_2$ is heper-geometric distribution.

Answer:

HW

Some Related Examples

Mathematical Problems

❖ **A coin is tossed 5 times. Find the probability of**

a) **exactly two heads**
b) **at least three heads**
c) **no heads**
d) **at most two heads**

Solution:

Given that,

$$n = 5; \qquad x = 0,1,2 \ldots\ldots 5$$

$$\left(x; 5, \frac{1}{2}\right) = \binom{5}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x}$$

(a) $P\ [exactly\ two\ heads] = p(x = 2)$

$$= \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = 10.\frac{1}{32} = \frac{5}{16}$$

(b) $P\ [at\ least\ 3\ heads] = p[x \geq 3]$

$$= p[x = 3] + p[x = 4] + p[x = 5]$$

$$= \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^5 = \frac{10}{32} + \frac{5}{32} + \frac{1}{32} = \frac{16}{32} = \frac{1}{2}$$

(c) $P\ [No\ heads] = p[x = 0]$

$$= \binom{5}{0}\left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

(d) $P[At\ most\ two\ heads] = p[x \leq 2]$

$$= p[x = 2] + p[x = 1] + p[x = 0]$$

$$= \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{3}\right)^3 + \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{3}\right)^4 + \left(\frac{1}{2}\right)^5 = \frac{10}{32} + \frac{5}{32} + \frac{1}{32} = \frac{16}{32} = \frac{1}{2}$$

❖ **Binomial distribution whose mean and standard deviation are 4 and $\sqrt{3}$ respectively.**

Solution: Given,

$$mean, np = 4$$

$$\text{And } standard\ deviation\ \sqrt{npq} = \sqrt{3}$$

$$\Rightarrow npq = 3$$

$$\Rightarrow q = \frac{3}{4}$$

$$p + q = 1$$

$$p + \frac{3}{4} = 1$$

$$p = 1 - \frac{3}{4}$$

$$\therefore p = \frac{1}{4}$$

Again,

$$np = 4$$

$$\therefore n = 4 * \frac{4}{1} = 16$$

Hence the required binomial distribution is,

$$P\left(X; 16, \tfrac{1}{4}\right) = \begin{cases} \binom{16}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{16-x} & ; x = 0,1,2,\ldots,n \\ 0 & ; Otherwise \end{cases}$$

❖ **Example 03**

Over a long period of time it has been observed that a given rifleman can hit a target on a single trial with probability equal to 0.8. Suppose that he fires four shots at the target. (i) what is the probability he will hit the target exactly two times? (ii) what is the probability that he will hit the target at least once?
Answer: HW

❖ **Example 04**

The probability that a MBA graduate from a public university gets an executive job is 0.6. Four MBA graduates applied for executive jobs. What is the probability that (i) exactly 3 will get the job? (ii) at least two will get the job? (iii) all will get the job? (iv) none will get the job?
Answer: HW

❖ **Example 05**

The probability that the patient recovers from a rare blood disease is 0.4. If 5 people are known to have this disease, what is the probability that (i) exactly 3 survive, (ii) at least two survive, (iii) at most two survive and (iv) none survive?
Answer: HW

❖ **Example 06**

In a locality 30% people are iterate. 10 persons are selected at random from that area. What is the probability that (i) at most 3 are literate, (ii) at least 3 are literate, (iii) exactly 3 are literate?
Answer: HW

Poisson Distribution

❖ **Poisson distribution. The conditions by which binomial converges to Poisson distribution.**

Answer:

A discrete random variable X is said to have a Poisson distribution if its probability function is,

$$P(x;\ \lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & ; x = 0,1,2 \ldots . \infty \\ 0 & ;\quad Otherwise \end{cases}$$

Where $e = 2.718$ and λ is the only parameter of the distribution which is the mean number of success.

Poisson distribution is the limiting case of the binomial distribution under the following conditions:

a. The probability of success or failure in Bernoulli trial is very small. That is $p \to 0$ or $q \to 0$
b. n, the number of trials is very large.
c. $np = \lambda$ is a finite constant.

❖ **Derivation of Poisson distribution from Binomial distribution.**

Answer:

By the definition we have

$np = \lambda$ so, $p = \lambda/n$ and $q = 1 - p = 1 - \lambda/n$

The probability function of Binomial variate X with parameters n and p is,

$$P(X; n, p) = \frac{n!}{x!(n-x)!}\, p^x q^{n-x} = \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^n \frac{n!}{n^x(n-x)!}$$

Now for fixed x; $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1$

and

$$\lim_{n\to\infty} \frac{n!}{n^x(n-x)!} = \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots[n-(x-1)](n-x)!}{n^x(n-x)!} = \lim_{n\to\infty} \frac{n^x\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)}{n^x} = 1$$

While, $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\frac{\lambda}{n}.n} = e^{-\lambda}$

Hence, $\lim_{n\to\infty} P(X; n, p) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$; $x = 0,1,2 \ldots . \infty$

Which is the probability function of a Poisson distribution with parameter λ.

❖ **Theorem: Let x be a Poisson variate with pdf, $f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$; $x = 0, 1, 2 \ldots . \infty$**

Then, $M_x(t) = e^{\lambda}(e^t - 1)$ and $P(s) = e^{-\lambda(1-s)}$. Find the skewness and kurtosis by using $M_x(t)$.

Proof:

By the definition of moment generating function,

$$M_x(t) = E[e^{tx}] = \sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} e^{-\lambda} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

And, the probability generating function of x is,

$$P[S] = E[S^x] = \sum_{x=0}^{\infty} S^x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty} \frac{(S\lambda)^x e^{-\lambda}}{x!} = e^{-\lambda} e^{S\lambda} = e^{-\lambda(1-S)} \quad \text{(proved)}$$

By definition, r^{th} raw moment of the distribution is,

$$\mu_r' = \left[\frac{d^r M_x(t)}{dt^r}\right]_{t=0}$$

Thus,

$$\mu_1' = \left[\frac{dM_x(t)}{dt}\right]_{t=0} = \left[\lambda e^t e^{\lambda(e^t-1)}\right] = [\lambda e^t M_x(t)]_{t=0} = \lambda, \quad since [M_x(t)]_{t=0} = 1$$

$$\mu_2' = \left[\frac{d^2 M_x(t)}{dt^2}\right]_{t=0} = [\lambda e^t M_x(t) + \lambda e^t \frac{dM_x(t)}{dt}]_{t=0} = \lambda^2 + \lambda$$

$\mu_3' = [\frac{d^3M_x(t)}{dt^3}]_{t=0} = [\lambda e^t M_x(t) + \lambda e^t \frac{dM_x(t)}{dt} + \lambda e^t \frac{dM_x(t)}{dt} + \lambda e^t \frac{d^2M_x(t)}{dt^2}]_{t=0} = \lambda + \lambda^2 + \lambda^2 + \lambda(\lambda + \lambda^2)$

$= \lambda + 3\lambda^2 + \lambda^3$

$\mu_4' = [\frac{d^4M_x(t)}{dt^4}]_{t=0}$

$= [\lambda e^t M_x(t) + \lambda e^t \frac{dM_x(t)}{dt} + 2\lambda e^t \frac{dM_x(t)}{dt} + 2\lambda e^t \frac{d^2M_x(t)}{dt^2} + \lambda e^t \frac{d^2M_x(t)}{dt^2} + \lambda e^t \frac{d^3M_x(t)}{dt^3}]_{t=0}$

$= \lambda + \lambda^2 + 2\lambda^2 + 2\lambda(\lambda + \lambda^2) + \lambda(\lambda + \lambda^2) + \lambda(\lambda + 3\lambda^2 + \lambda^3) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$

Now, $\mu_1 = \lambda$

$\mu_2 = \lambda$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = \lambda + 3\lambda^2 + \lambda^3 - 3\lambda(\lambda^2 + \lambda) + 2\lambda^3 = \lambda$

And, $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$= \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda - 4\lambda(\lambda + 3\lambda^2 + \lambda^3) + 6\lambda^2(\lambda^2 + \lambda) - 3\lambda^4$

$= 3\lambda^2 + \lambda$

Therefore,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\lambda^2 + \lambda}{\lambda^2} = 3 + \frac{1}{\lambda}$$

Theorem: Sum of k independent Poisson variates is also a Poisson variate.

Proof:

Let $x_1, x_2 \ldots\ldots x_k$ be k independent Poisson variates with parameters $\lambda_1. \lambda_2 \ldots\ldots \lambda_k$ that is

$$f(x_j, \lambda_j) = \frac{e^{-\lambda j}(\lambda_j)^{x_j}}{x_j!} \qquad x_j = 0,1,2 \ldots\ldots \infty$$

Then, we have to prove that $\sum x_j = x$ is also a Poisson variate with parameter $\sum \lambda_j = \lambda$

The mgf of x_j is

$M_{x_j}(t) = E[e^{tx_j}] = e^{\lambda_j(e^t-1)}$ for all j = 1,2……,k

The mgf of $\sum x_j = x$ is

$M_x(t) = E[e^{t(x_1+x_2\ldots\ldots+x_k}] = E[e^{tx_1}]E[e^{tx_2}] \ldots\ldots\ldots E[e^{tx_k}] = e^{\lambda_1(e^t-1)}. e^{\lambda_2(e^t-1)} \ldots\ldots\ldots e^{\lambda_k(e^t-1)}$

$= e^{(\lambda_1\lambda_2\ldots\ldots\lambda_k)(e^t-1)} = e^{\lambda(e^t-1)}$

Which is the mgf of a Poisson distribution with parameter $\lambda = \sum \lambda_j$ then pdf of x is

$f(x;\lambda) = \frac{e^{-\lambda}(\lambda)^x}{x!}$ Where $\lambda = \sum_{j=1}^k \lambda_j$ *and* $x = \sum_{j=1}^k x_j$

Theorem: If x_1 and x_2 are two independent poisson variates with parameters t_1 and t_2. Then $x=x_1 - x_2$ is not a Poisson variate

Proof:

By definition, the mgf $x = x_1 - x_2$

$$M_x(t) = E[e^{t(x_1-x_2)}] = E[e^{tx_1}].E[e^{-tx_2}] = M_{x_1}(t)M_{x_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^{-t}-1)}$$

Which cannot be put in the form $e^{\lambda(e^t-1)}$ therefore $x = x_1 - x_2$ is not a Poisson variate .Although $x = x_1 + x_2$ is a Poisson variate. It is to be noted that $x = x_1 - x_2$ cannot be a Poisson variate because x can take both positive and negative values but Poisson variate can take only positive value.

Some Related Examples

- **Example-11.4.4: The average number of traffic accidents on a certain section of highway is two per week. Assume that the number of accidents follows a Poisson distribution with λ =2.**

(i) Find the probability of no accident on this section of highway during a 1-week period,
(ii) Find the probability of at most three accidents on this section of highway during a 2-week period.

Solution:

i) The average number of accident per week is λ =2.Therefore, the probability of no accidents on this section of highway during a given week is

$$P[X=0]=f(0;2)=\frac{2^0e^{-2}}{0!}=e^{-2}=0.135335.$$

ii) During a 2-week period, the average number of accidents on this section of highway is 2*2=4.The probability of at most three accidents during a 2-week period is

$$P[X\le 3]=f(0;4)+f(1;4)+f(2;4)+f(3;4)$$

$$=\frac{e^{-4}4^0}{0!}+\frac{e^{-4}4^1}{1!}+\frac{e^{-4}4^2}{2!}+\frac{e^{-4}4^3}{3!}=0.018316+0.073263+0.146525+0.195367=0.433471.$$

- **Example-11.4.6: A car hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as Poisson variate with mean 1.5. Calculate the probability of days on which**

i) neither car is used and
ii) some demand is refused.

Solution:

Let X be the number of cars demand per day. It is a Poisson variate with parameter λ=1.5. Then the probability function of X is

$$f(x;\ \lambda)=f(x;1.5)=\frac{e^{-1.5}(1.5)^x}{x!}$$

i) $P[X=0]=f(0,1.5)=e^{-1.5}=0.2231.$

ii) $P[X>2]=1-P[X\le 2]$

$=1-[P(X=0)+P(X=1)+P(X=2)]$

$=1\text{-}e^{-1.5}\left[1+1.5+\frac{(1.5)^2}{2!}\right]$

=1-(0.2231*3.625)

=0.19126

- **Example 11.4.12: if X is a Poisson variate such that $P[X=2]=9\,P[X=4]+90\,P[X=6]$ Find λ, the mean of X, β_1, the measure of the skewness.**

Solution:

If X is a Poisson variate with parameter λ, then

$$P[X=x]=\frac{e^{-\lambda}\lambda^x}{x!};\ x=0,1,2,\ldots..;\ \lambda>0.$$

We are given,

$$P[X=2]=9\,P[X=4]+90\,P[X=6]$$

$$\frac{e^{-\lambda}\lambda^2}{2!}=9\,\frac{e^{-\lambda}\lambda^4}{4!}+90\frac{e^{-\lambda}\lambda^6}{6!}=e^{-\lambda}\left[\frac{9\lambda^4}{4!}+\frac{90\lambda^6}{6!}\right]=e^{-\lambda}\left[3\frac{\lambda^4}{8}+\frac{\lambda^6}{8}\right]=\frac{e^{-\lambda}\lambda^2}{8}[3\lambda^2+\lambda^4]$$

$$\Rightarrow 3\lambda^2+\lambda^4=4$$

$$\text{Or, } \lambda^4+3\lambda^2-4=0.$$

Solving as a quadratic in λ^2,we get

$$\lambda^2=\frac{-3\pm\sqrt{9+16}}{2}=\frac{-3\pm 5}{2}$$

Since $\lambda>0$,we get $\lambda^2=1$

$\therefore \lambda=1.$

Hence Mean= $\lambda=1$,variance= $\lambda=1$ and $\mu_3=1$

$$\therefore\ \beta_1=\frac{1}{\lambda}=1.$$

Normal Distribution

Definition: A continuous random variable X is said to have a normal distribution if its density function is given by

$$f(X;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad ;-\infty<x<\infty$$

Where the parameters μ and σ^2 satisfy $-\infty<x<\infty$ and $\sigma^2>0$ .It is denoted by $X\sim N(\mu,\sigma^2)$ and its graphical representation will be

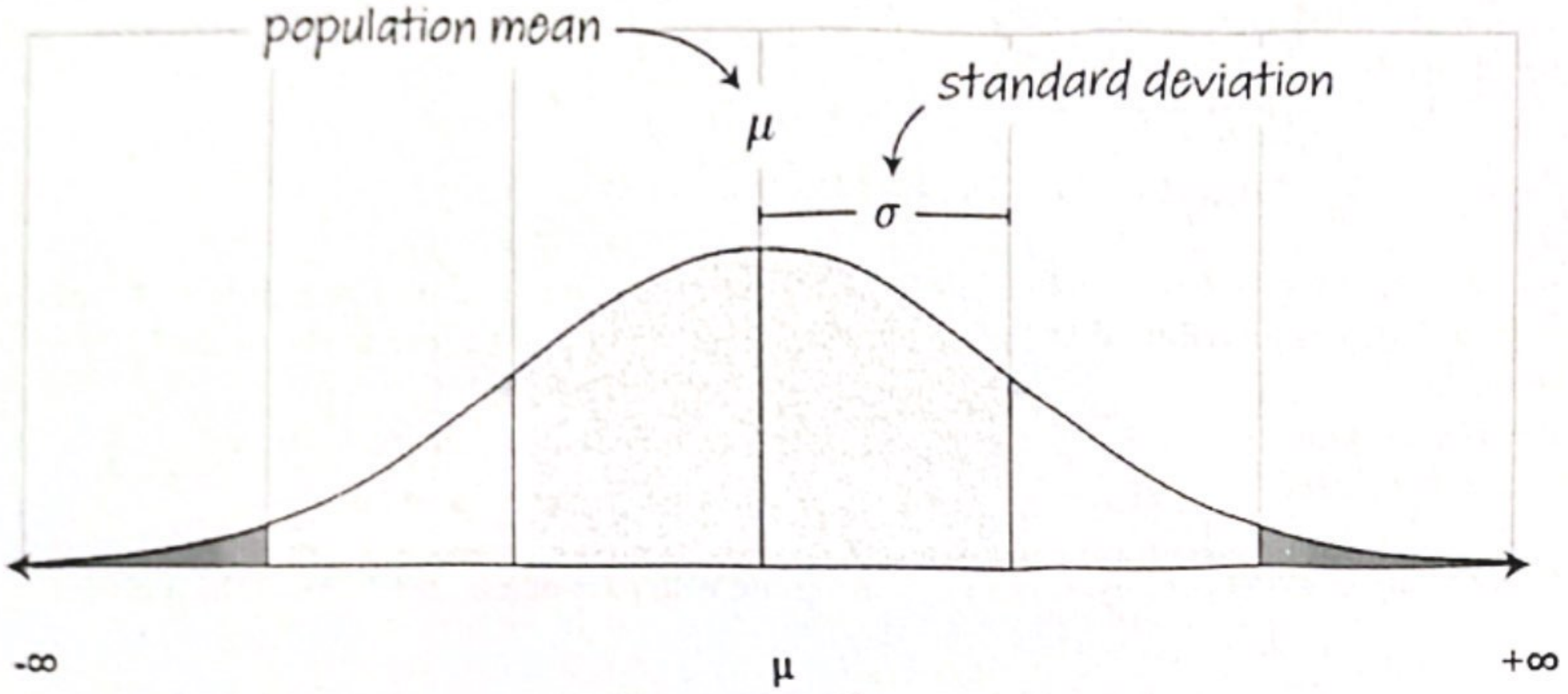
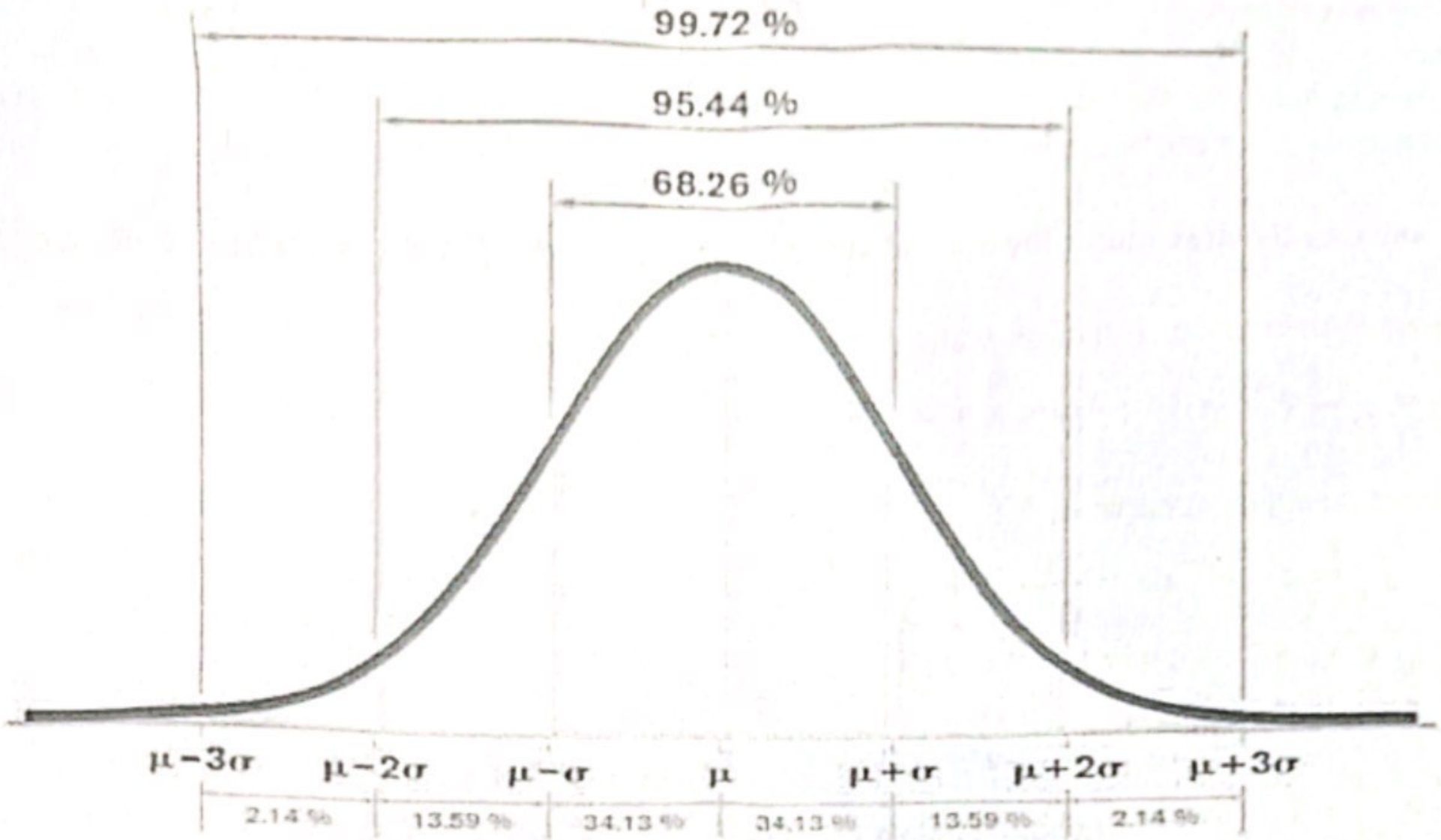


Figure: Normal Distribution Curve

Some characteristics of normal distribution

i. The curve of the distribution is symmetrical about point $x=\mu$ and it is bell shaped.
ii. As the distribution the distribution is symmetrical about the point $x=\mu$,it devides the whole distribution into two equal parts, which indicates the median of the distribution.
iii. In this distribution mean = median = mode for all values of x =1,2,3, ----------, n.
iv. There are two point of inflexion which are at $x=\mu\pm\sigma$.
v. All odds moment about mean is zero because the curve is symmetric.
vi. The mean deviation about mean is approximately equal to $\frac{4}{5}\sigma$.
vii. The quartile deviation of the distribution is $\frac{2}{3}\sigma$.
viii. Linear combination of independent normal variates is also a normal variate.
ix. The skewness and kurtosis are $\beta_1=0$ and are $\beta_2=3$ respectfully for this distribution.
x. If Z is a standard normal variate, then the pdf of Z is $f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \quad ;-\infty<x<\infty$

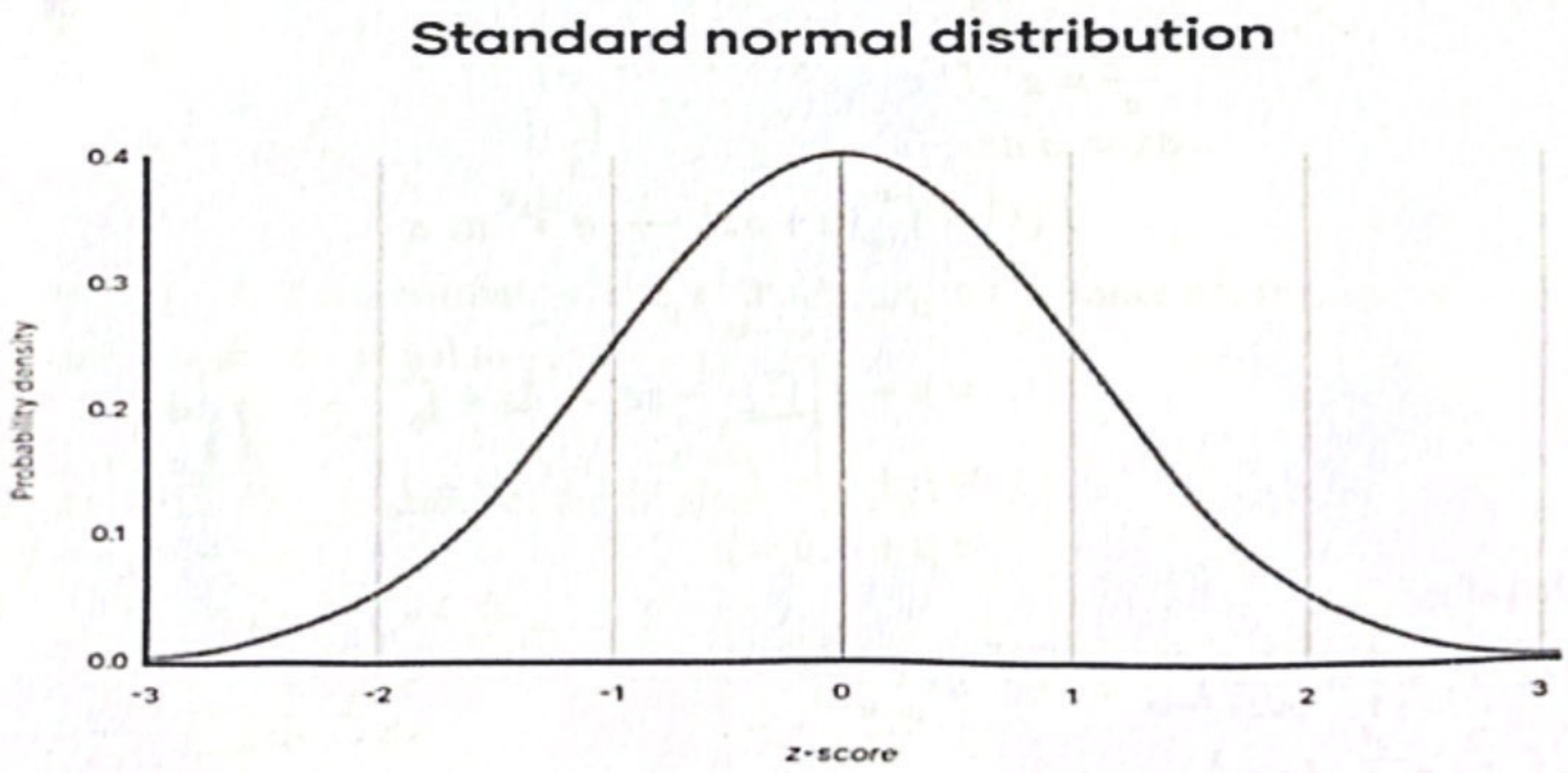


Standard Normal Distribution

Definition: A continuous random variable Z is said to have a normal variate if its density function is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad ; -\infty < x < \infty$$

And its graphical representation will be



Application of normal distribution:

i. In practice, under certain conditions most of the probability and sampling distributions can be approximated by normal distribution.

ii. In the theory of probability it plays an important role. The central limit theorem proves that the distribution of most of the standardized variates turns into normal distribution as the sample size tends to infinity.

iii. Normal distribution is the basis of all the sampling distributions. Without the assumption of normality, sampling distribution has no existence.
iv. Assumption of normality is the basis of all the tests of significance in parametric list.
v. Its application in industrial statistics such as quality control.
vi. Theory of normal curve can be applied to the graduation of the curves which are not normal.

Theorem-Show that the area under the normal curve is unity/one/1 or $\int f(x)dx = 1$ *when* $X \sim N(\mu, \sigma^2)$.
Solution:
The pdf of normal distribution with mean μ and σ^2 variance is

$$f(X;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; -\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0$$

Total area under the moment curve is,

$$\int \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{2}{\sigma\sqrt{2\pi}} \int_0^\infty e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{1}{2}z^2} dz \quad ; \text{when } \frac{x-\mu}{\sigma} = z \text{ then } dx = \sigma dz$$

$$= \frac{1}{\sqrt{\pi}} \int_0^\infty e^{-t} t^{\frac{1}{2}-1} dt \quad ; \text{when } t = \frac{z^2}{2}$$

$$= \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}} = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1 \quad \text{[proved]}$$

<u>Theorem. 12.6.1.(b).</u> If X is a normal variate with pdf

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} ; \qquad -\infty < x < \infty; -\infty < \mu < \infty \text{ and } \sigma^2 > 0$$

Then prove the mean=median=mode= μ and variance σ^2 .
Proof:
By definition, Mean = $E[X] = \int_{-\infty}^{\infty} x f(x) dx$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let,

$$\frac{x-\mu}{\sigma} = z$$

$$\therefore dx = \sigma\, dz$$

$$E[X] = \int_{-\infty}^{\infty} (\mu + \sigma z) . \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz\, \sigma$$

$$= \mu + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, e^{-\frac{1}{2}z^2} dz$$

$$= \mu + \sigma \left[\int_{-\infty}^{o} \frac{1}{\sqrt{2\pi}} z e^{-\frac{1}{2}z^2} dz + \int_0^\infty \frac{1}{\sqrt{2\pi}} z e^{-\frac{1}{2}z^2} dz \right]$$

$$= \mu + \sigma \left[-\int_0^\infty \frac{1}{\sqrt{2\pi}} z e^{-\frac{1}{2}z^2} dz + \int_0^\infty \frac{1}{\sqrt{2\pi}} z e^{-\frac{1}{2}z^2} dz \right]$$

$$= \mu + \sigma . 0 = \mu$$

Now, by definition of median,

$$\text{Median} = \int_{-\infty}^{M} f(x) dx = \frac{1}{2} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{M-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz = \frac{1}{2}$$

We know,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{z^2}{2}} dz = \frac{1}{2}$$

So, we can write,

$$\frac{x-\mu}{\sigma} = 0$$

And $M = \mu$ when $(\sigma > 0, \sigma \neq 0)$
Again, by definition of mode, we can write,

$$f'(x) = 0 \text{ and } f''(x) < 0$$

We have,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Then,

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} [\frac{(x-\mu)}{\sigma^2}][-\frac{2}{2}] = 0$$

$$\Rightarrow -f(x)\frac{(x-\mu)}{\sigma^2} = 0$$

$$\Rightarrow \frac{(x-\mu)}{\sigma^2} = 0 \qquad \text{since, } f(x) \neq 0$$

$$\Rightarrow x - \mu = 0 \quad \therefore x = \mu$$

Now, differentiate $f'(x)$ again with respect to, we have

$$f''(x) = -f'(x)\frac{(x-\mu)}{\sigma^2} - \frac{(x-\mu)}{\sigma^2}$$

$$= +f(x)\frac{(x-\mu)^2}{\sigma^4} - \frac{f(x)}{\sigma^2} = -\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$$

Here $f''(x)$ at $x = \mu$ is $[f''(x)]_{x=\mu} = -\frac{1}{\sigma^3\sqrt{2\pi}} < 0$
Therefore $x = \mu$ is the mode of the distribution. Hence mean = median = mode.

By definition,

$$V[X] = E[X - E(X)]^2 = E[X - \mu]^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty}(x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let,

$$\frac{x-\mu}{\sigma} = z, \text{ then } dx = \sigma\, dz$$

$$= \sigma^2 . 2\int_0^{\infty} z^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.\sigma$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} e^{-u} u^{\frac{3}{2}-1} du \qquad \left[\frac{z^2}{2} = u\right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \Gamma(\frac{3}{2}) \qquad \left[\int_0^{\infty} e^{-ax} x^{n-1} dx = \frac{\Gamma n}{a}\right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} . \frac{1}{2}\sqrt{\pi} \qquad \left[\Gamma\frac{1}{2}\right]$$

$$= \sigma^2$$

❖ **<u>Theorem 12.6.4.</u> If X is a normal variate with mean μ and variance σ^2, then mean deviation about mean is approximately equal to $\frac{4}{5}\sigma$.**

Proof:
Let η be the mean deviation about mean of the distribution, then

$$\eta = E[|X - \mu|]$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x-\mu| e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-\frac{z^2}{2}} dz \qquad \text{if } z = \frac{x-\mu}{\sigma}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \left[-\int_{-\infty}^{0} z e^{-\frac{z^2}{2}} dz + \int_0^{\infty} z e^{-\frac{z^2}{2}} dz\right]$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_0^{\infty} z e^{-\frac{z^2}{2}} dz$$

$$= \sigma\sqrt{\frac{2}{\pi}} \int_0^{\infty} t^{1-1} e^{-1} dt, \qquad \text{when } t = \frac{1}{2}z^2$$

$= \sigma\sqrt{\frac{2}{\pi}}\int_0^{\infty}\Gamma 1 = \frac{4}{5}\sigma$ approximately.

Mathematical Problems

❖ **Example 12.6.1. An unbiased die is thrown 6000 times. Find the probability that the three points faces will appear 980 to 1030 times.**

Solution.
For an unbiased die,

$$p = \frac{1}{6} \text{ and } q = \frac{5}{6}$$

$$P(980 \le x \le 1030) = P\left(\frac{980-\mu}{\sigma} \le \frac{x-\mu}{\sigma} \le \frac{1030-\mu}{\sigma}\right)$$

$$Mean = np = 6000 \times \frac{1}{6} = 1000 \text{ and}$$

$$variance = npq = 6000 \times \frac{1}{6} \times \frac{5}{6} = 833.33$$

The standard deviation is

$$\sigma = \sqrt{npq} = \sqrt{833.33} = 28.87$$
$$= P\left(\frac{980-1000}{28.87} \le z \le \frac{1030-1000}{28.87}\right)$$
$$= P(-0.693 \le z \le 1.039)$$
$$= \Phi(1.04) - \Phi(0.69)$$
$$= 0.8508 - 0.7549$$
$$= 0.0959$$

❖ **Example 12.6.3. The reliability of an electrical fuse is the probability that a fuse, chosen at random from production, will function under the conditions for which it has been designed. A random sample of 1000 fuses were tested and $x = 27$ defectives were observed. Calculate the approximate probability of observing 27 or more defectives, assuming that the fuse reliability is 98.**

Solution.
The probability of observing a defective when a simple fuse is tested is $p = 0.02$ since the fuse reliability is 0.98.
Then

$$\mu = np = 1000(0.02) = 20$$
$$\sigma = \sqrt{npq} = \sqrt{1000(0.02)(0.98)} = 4.43$$

It is appropriate to use the normal approximation to the binomial probability because

$$np = 20 \text{ and } nq = 1000(0.98) = 980$$

Both np and nq are greater than 5.

$$P[X \ge 27] = P\left[\frac{X-\mu}{\sigma} \ge \frac{27-\mu-\frac{1}{2}}{\sigma}\right]$$
$$= P\left[Z \ge \frac{27-20-\frac{1}{2}}{4.43}\right] = P[Z \ge 1.47] = 1 - P[Z \le 1.47]$$
$$= 1 - 0.9297 = 0.0708$$

❖ **Example 12.6.5. Students are divided into three categories according to the marks obtained by a public examination. Students are classified into A category if he gets mark more than 80%, B category if he gets less than 80 but more than 60 and C category if he gets less than 60%. Percentages of students in A, B and C categories are 5%, 40% and 55% respectively. Suppose the marks obtained by the students follow normal distribution, find the mean and standard deviation of the marks.**

Solution.
Let the marks of the student X follows normal distribution with mean μ and variance. That is

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \qquad -\infty < x < \infty$$

According to the problem,

$$P[X < 60] = 0.55 \quad (1)$$
$$P[X > 80] = 0.05 \quad (2)$$
$$P[60 \leq X \leq 80] = 0.40 \quad (3)$$

From (1), we have,

$$P\left[\frac{X-\mu}{\sigma} < \frac{60-\mu}{\sigma}\right] = 0.55$$
$$P\left[Z < \frac{60-\mu}{\sigma}\right] = 0.55$$
$$\Phi\left[\frac{60-\mu}{\sigma}\right] = 0.55$$

From the area of the standard normal distribution

$$\Phi(0.13) = 0.55$$
$$\therefore \frac{60-\mu}{\sigma} = 0.13 \quad \Rightarrow \mu + 0.13\sigma = 60 \quad (4)$$

From (2), we get $P\left[\frac{X-\mu}{\sigma} > \frac{80-\mu}{\sigma}\right] = 0.05$

$$P\left[Z > \frac{80-\mu}{\sigma}\right] = 0.05$$
$$P\left[Z < \frac{80-\mu}{\sigma}\right] = 0.95$$
$$\Phi\left[\frac{80-\mu}{\sigma}\right] = 0.95 \quad (5)$$

From the area of the standard normal integral table, we have

$$\Phi(1.645) = 0.95 \quad (6)$$

Comparing (5) and (6), we get

$$\frac{80-\mu}{\sigma} = 1.645 \quad \Rightarrow \mu + 1.645\sigma = 80 \quad (7)$$

From (4) and (7), we get

$$\mu = 58.29 \quad \text{and} \quad \sigma = 13.20$$

The variance of the distribution is

$$\sigma^2 = 174.24$$

❖ **<u>Example 12.6.10.</u> Suppose the yield of wheat per one-acre plot in north Bengal follows normal distribution with mean yield 662 kilos and standard deviation 32 kilos. How many one-acre plot in a batch of 1000 plots would you expect to have yield (i) over 700 kilos; (ii) below 650 kilos and (iii) what is the lowest yield of the best 100 plots?**

Solution.
Let X be the yield in kilos per one-acre plot. Hence

$$X \sim N(662, 32)$$

(i) The probability that a plot has a yield over 700 kilos is

$$P[X > 700] = P\left[Z > \frac{700-662}{32}\right]$$
$$= P[Z > 1.19]$$
$$= 1 - P[Z \leq 1.19]$$
$$= 1 - 0.8830$$
$$= 0.1170$$

Hence in a batch of 1000 plots, the expected number of plots with yield over 700 kilos are

$$1000 \times 0.117 = 117$$

(*ii*) The probability that a plot has a yield below 650 is

$$P[X < 650] = P\left[Z < \frac{650-662}{32}\right]$$
$$= P[Z < -0.38]$$
$$= 0.352$$

The required number of plots will yield below 650 kilos are

$$1000 \times 0.3520 = 352$$

(*iii*) Let x' be the lowest yield of the best 100 plots.

Then $$P[X > x'] = \frac{100}{1000} \times 0.1$$

$$P\left[\frac{X-\mu}{\sigma} > \frac{x'-\mu}{\sigma}\right] = 0.01$$

$$P\left[Z > \frac{x'-662}{32}\right] = 0.01$$

$$P\left[Z < \frac{x'-662}{32}\right] = 0.99$$

$$\Phi\left[\frac{x'-662}{32}\right] = 0.99 \qquad (1)$$

From the normal integral table, we have

$$\Phi(1.28) = 0.9 \qquad (2)$$

Comparing (1) and (2), we have

$$\frac{x'-662}{32} = 1.28$$

$$\therefore \quad x' = 662 + 1.28 \times 32$$
$$= 702.96$$

Hence the lowest yield of the best 100 plots is 702.96 kilos.

❖ Moment and cumulant generating function of Normal distribution

Moment generating function of normal distribution about origin is

$$M_x(t) = E[e^{tx}] = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$M_x(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{t(\mu+\sigma z)} e^{-\frac{1}{2}z^2} dz \qquad \text{when } z = \frac{x-\mu}{\sigma}$$

$$= e^{t\mu}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz\sigma-\frac{1}{2}z^2} dz$$

$$= e^{t\mu+\frac{1}{2}t^2\sigma^2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t\sigma)^2} dz$$

$$= e^{t\mu+\frac{1}{2}t^2\sigma^2}; \qquad \text{Since } \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t\sigma)^2} dz = 1$$

Cumulant generating function of the normal variate X is

$$K_x(t) = \log M_x(t) = \mu t + \frac{1}{2}t^2\sigma^2$$

Thus $k_1 = \mu, k_2 = \sigma^2, k_3 = k_4 = \cdots \ldots\ldots\ldots\ldots = k_r = \cdots \ldots\ldots\ldots\ldots\ldots = 0$

Hence except the first two cumulants all the cumulants of the normal distribution are zero.

Moment generating function about mean is defined as

$$M_x(t) = E\left[e^{t(X-\mu)}\right] = e^{-t\mu}E[e^{tx}] = e^{-t\mu}e^{+t\mu+\frac{1}{2}t^2\sigma^2} = e^{\frac{1}{2}t^2\sigma^2}$$

$$= 1 + \frac{t^2\sigma^2}{2} + \frac{(t^2\sigma^2/2)^2}{2!} + \frac{(t^2\sigma^2/2)^3}{3!} + \cdots \ldots\ldots + \frac{\left(\frac{t^2\sigma^2}{2}\right)^r}{r!} \ldots\ldots\ldots$$

Hence $$\mu_{2r} = \frac{\sigma^{2r}(2r)!}{2^r.r!}$$

-The End-