CHAPTER 2 – STATISTICAL ANALYSIS

2.1 Descriptive Statistics

Descriptive Statistics is the branch of statistics that deals with collecting, summarizing, presenting, and describing data in a meaningful form.
It helps to convert raw data into simple summaries, such as tables, charts, and numerical measures like mean, median, and standard deviation.

Definition

Descriptive statistics refers to the methods used to organize and simplify data so that patterns can be identified easily.
It provides tools to summarize large datasets using numerical values and graphical methods.

Purpose of Descriptive Statistics

The main purposes are:

1. To summarize large datasets in a meaningful way.
2. To simplify complex data into understandable formats.
3. To provide a clear picture of data distribution and variation.
4. To serve as a basis for inferential statistics, which involves conclusions and predictions.

Functions of Descriptive Statistics

Function	Description
Data Collection	Gathering information from surveys, experiments, or observations.
Data Organization	Arranging data systematically (tables, frequency distributions).
Data Summarization	Representing data through measures (mean, median, mode, etc.).
Data Presentation	Displaying data visually (graphs, charts, diagrams).
Data Interpretation	Drawing basic insights and identifying trends.

2.1.1 Types of Descriptive Statistics

Descriptive statistics are broadly classified into three categories**:**

1. **Measures of Central Tendency**
   - Represent the center or average value of a dataset.
   - Examples: Mean, Median, Mode
2. **Measures of Dispersion (Variability)**
   - Indicate how much the data values differ from each other.
   - Examples: Range, Variance, Standard Deviation
3. **Measures of Shape (Skewness & Kurtosis)**
   - Describe the distribution pattern of the data.
   - Skewness**:** Measures symmetry or asymmetry.

- **Kurtosis:** Measures peakedness or flatness.

2.1.2 Data Representation in Descriptive Statistics

Data can be represented in two forms:

Form	Explanation	Examples
Tabular Form	Data arranged in rows and columns for clarity.	Frequency distribution tables
Graphical Form	Data displayed visually for easy comparison.	Bar charts, histograms, pie charts

2.1.3 Steps in Descriptive Statistical Analysis

1. **Collect Data:**
   - Gather relevant data through surveys, experiments, or reports.
2. **Organize Data:**
   - Sort and classify the collected data.
3. **Summarize Data:**
   - Compute descriptive measures like mean, median, range, etc.
4. **Present Data:**
   - Use tables, charts, or diagrams for visualization.
5. **Interpret Results:**
   - Identify key insights, patterns, and relationships.

2.1.4 Importance of Descriptive Statistics

- Helps in understanding the basic characteristics of data.
- Enables comparison between datasets.
- Provides foundation for further statistical analysis.
- Useful in research, business, and data analytics.
- Simplifies complex raw data into concise and readable formats.

2.1.5 Tools Used in Descriptive Statistics

Tool	Purpose
Mean / Median / Mode	Measure of central tendency (average).
Range / Variance / SD	Measure of variability or spread.
Tables & Charts	Visual presentation of data.
Box Plot	Displays data distribution, median, and outliers.
Skewness / Kurtosis	Describes data shape and pattern.

2.1.6 Applications of Descriptive Statistics

- **Business:** Analyze sales performance and profit trends.
- **Education:** Evaluate student marks or attendance data.
- **Healthcare:** Study patient recovery statistics.

- **Government:** Represent population, census, or crime data.
- **Research:** Summarize experimental or survey results.

2.1.7 Advantages

- Simplifies large datasets into understandable summaries.
- Helps visualize trends through tables and graphs.
- Provides quick numerical indicators for comparison.
- Reduces data complexity for decision-making.
- Forms the groundwork for advanced statistical analysis.

2.1.8 Limitations

Limitation	Explanation
No Prediction	It only describes data; does not predict outcomes.
Lack of Detail	Summaries may hide individual variations.
Possible Misinterpretation	Poor representation may lead to wrong conclusions.
Limited Scope	Cannot determine cause-and-effect relationships.

2.2 Definition of Basic Terms

Before performing statistical analysis, it is essential to understand a few basic terms commonly used in statistics.
These terms — such as data, population, sample, variable, class interval, and frequency — form the foundation of descriptive and inferential statistical methods.

Each concept helps us understand how data is collected, organized, and analyzed for meaningful conclusions.

2.2.1 Data

Definition:
Data refers to raw facts, figures, or observations collected for analysis.
They can be quantitative (numerical) or qualitative (categorical) in nature.

Example:
Marks obtained by students: 65, 78, 80, 92, 88.

Types of Data:

Type	Description	Example
Primary Data	Collected first-hand by the researcher.	Data from a survey or experiment.
Secondary Data	Already collected and published by someone else.	Census data, reports, journals.

2.2.2 Population

Definition:
A population is the complete set of all items or individuals relevant to a particular study or experiment.
It includes every element that shares a common characteristic being studied.

Examples:

- All students in a college.
- All voters in a state.
- All manufactured items in a factory batch.

Types of Population:

Type	Explanation
Finite Population	A population with a limited number of elements. (e.g., 100 students)
Infinite Population	A population with uncountable or limitless elements. (e.g., stars in the sky)

2.2.3 Sample

Definition:
A sample is a subset of the population selected for analysis.
Instead of studying the entire population, we study a representative part (sample) to make inferences about the whole.

Example:
If a college has 2000 students and we study 200 of them, those 200 represent a sample of the population.

Importance of Sampling:

- Saves time and cost.
- Easier to handle and analyze.
- Provides quick results with acceptable accuracy.

Relation between Population and Sample:

Aspect	Population	Sample
Definition	Complete set of all items.	Subset chosen for study.
Size	Large or infinite.	Smaller portion.
Purpose	Provides full information.	Used for estimation or inference.

2.2.4 Variable

Definition:
A variable is a characteristic or attribute that can take different values for different individuals or items in a dataset.

Example:

- Age, height, and weight of students.
- Income of families.
- Marks in an exam.

Types of Variables:

Type	Description	Example
Quantitative Variable	Represents measurable quantities.	Height, weight, income.
Qualitative Variable	Represents non-numeric categories.	Gender, color, brand.
Continuous Variable	Can take any value within a range.	Height = 165.4 cm, 170.2 cm.
Discrete Variable	Takes only whole number values.	Number of students = 50, 51, 52.

2.2.5 Class Interval

Definition:
A class interval represents a range of values into which data is divided when creating a frequency distribution table.

Example:
If marks are grouped as 0–10, 10–20, 20–30, etc., each range (e.g., 0–10) is a class interval.

Terms related to Class Interval:

- **Lower Class Limit (LCL):** The smallest value in the interval.
- **Upper Class Limit (UCL):** The largest value in the interval.
- **Class Width (or Size):** Difference between upper and lower limits.

Example:
For the class 20–30:

- LCL = 20, UCL = 30
- Class Width = 10

2.2.6 Frequency

Definition:
Frequency refers to the number of times a particular value or class occurs in a dataset.

Example:
If 5 students scored between 60–70 marks, then the frequency of the class interval 60–70 is 5.

Types of Frequency:

Type	Description
Absolute Frequency	Actual number of observations in a class.
Relative Frequency	Proportion or percentage of observations.
Cumulative Frequency	Running total of frequencies up to a class.

2.2.7 Parameter and Statistic

Term	Definition	Example
Parameter	A numerical value describing a characteristic of a population.	Population mean (μ), population standard deviation (σ).
Statistic	A numerical value describing a characteristic of a sample.	Sample mean ($\bar{x}$), sample standard deviation (s).

Note:
A statistic is used to estimate a parameter.

2.2.8 Attributes and Characteristics

- Attributes are qualitative features of data such as color, gender, or category.
- Characteristics are quantitative or measurable features like height, weight, or age.

2.2.9 Differences Between Qualitative and Quantitative Data

Aspect	Qualitative Data	Quantitative Data
Nature	Descriptive or categorical.	Numerical or measurable.
Examples	Gender, color, type of car.	Height, income, marks.
Statistical Use	Represented through counts and percentages.	Used for arithmetic analysis and computation.

2.2.10 Importance of Understanding Basic Terms

- Helps in proper data collection and classification.
- Essential for designing statistical experiments.
- Aids in constructing accurate frequency distributions.
- Simplifies understanding of statistical measures.
- Builds foundation for advanced topics like hypothesis testing or regression.

2.3 Classification of Data and Frequency Distribution

Introduction

When data is collected, it is often raw and unorganized, making it difficult to analyze or interpret.
To make it useful, data is first classified into meaningful groups and then arranged into a frequency distribution.

- Classification means grouping data into categories based on similarities.
- Frequency Distribution shows how often each value or range of values occurs.

Both processes simplify large datasets, highlight patterns, and form the foundation of all descriptive analysis.

2.3.1 Classification of Data

Classification is the process of arranging raw data into groups or classes according to shared characteristics.
This helps to simplify, compare, and analyze the data effectively.

Example:
If students' marks are: 25, 45, 55, 62, 78, 81, 92
They can be classified as:

- 0–20, 20–40, 40–60, 60–80, 80–100.
  Each range is a class interval.

Objectives of Classification

1. To simplify large and complex data.
2. To make comparison and analysis easier.
3. To help identify patterns and trends.
4. To provide a basis for tabulation and graphical presentation.
5. To summarize information for statistical computation.

Rules for Classification

1. **Exhaustiveness:** Every observation must belong to some class.
2. **Mutual Exclusiveness:** Each item must belong to only one class.
3. **Uniformity:** The same classification principle must be followed throughout.
4. **Clarity:** Class limits must be well-defined and non-overlapping.
5. **Suitability:** The classification should match the nature and purpose of the data.
6. **Flexibility:** The method should allow for future data expansion.

Types of Classification

Type	Basis	Description / Example
1. Qualitative Classification	Attributes	Based on characteristics like gender, color, or religion. Example: Classifying employees as male/female.
2. Quantitative Classification	Numerical values	Based on measurable variables. Example: Classifying students by marks or employees by income.
3. Temporal Classification	Time	Based on time periods. Example: Yearly rainfall, monthly sales.
4. Spatial Classification	Place	Based on location. Example: Population by city, district, or state.

2.3.2 Frequency Distribution

A Frequency Distribution is a statistical table showing how many observations fall into each class or category.
It organizes data to show the distribution and frequency of values.

Example:
If 10 students scored between 40–50 marks, then the frequency of the class 40–50 is 10.

Objectives of Frequency Distribution

1. To condense and organize raw data.
2. To make patterns and variations visible.
3. To provide a basis for computing averages and measures of dispersion.
4. To simplify graphical representation.
5. To compare datasets effectively.

2.3.3 Components of a Frequency Table

Component	Description
Class Interval	The range of values in which data is grouped (e.g., 0–10, 10–20).
Class Limits	The smallest and largest values defining a class.
Class Width (Size)	The difference between upper and lower limits of a class. **Formula:** Class Width = UCL − LCL
Class Boundary	The midpoint between the upper limit of one class and the lower limit of the next. Example: Between 10–20 and 20–30 → boundary = 19.5.
Midpoint (Class Mark)	The average of the lower and upper class limits. **Formula:** Class Mark = (Lower Limit + Upper Limit) / 2
Tally Marks	Used for counting observations quickly.
Frequency	Number of observations in a class interval.
Cumulative Frequency	Running total of frequencies up to a certain class.
Relative Frequency	Proportion or percentage of data in each class. **Formula:** Relative Frequency = (Class Frequency / Total Frequency) × 100

2.3.4 Types of Frequency Distributions

1. Discrete Frequency Distribution

Used for data with distinct, separate values (usually integers).

Example:

No. of Books	**1**	**2**	**3**	**4**	**5**
Frequency	2	4	5	3	6

2. Continuous Frequency Distribution

Used when data covers a continuous range and is grouped into intervals.

Example:

Marks Range	**0–10**	**10–20**	**20–30**	**30–40**
Frequency	3	7	10	5

2.3.5 Terms Related to Frequency Distribution

Term	Meaning
Class Limit	The smallest and largest values defining a class (e.g., 10–20).
Class Boundary	The midpoint between the upper limit of one class and the lower limit of the next (e.g., 9.5, 19.5).
Class Width (Size)	The difference between class limits. **Formula:** Class Width = UCL − LCL
Midpoint (Class Mark)	The average of the lower and upper class limits. **Formula:** Class Mark = (Lower Limit + Upper Limit) / 2

2.3.6 Cumulative Frequency Distribution

This distribution shows the running total of frequencies up to a certain class.
It helps in finding medians, quartiles, and percentiles.

There are two types:

1. **Less Than Cumulative Frequency** – cumulative total moving upward from the lowest class.
2. **More Than Cumulative Frequency** – cumulative total moving downward from the highest class.

Example:

Marks	Frequency	Less Than CF	More Than CF
0–10	3	3	27
10–20	5	8	24
20–30	9	17	19
30–40	10	27	10

2.3.7 Relative Frequency Distribution

Shows the percentage or proportion of total observations in each class.

Formula:

$$\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Total Frequency}}$$

Example:

$$\text{Relative Frequency} = \frac{10}{50} = 0.2 \text{ or } 20\%$$

2.3.8 Advantages of Frequency Distribution

- Converts large data into an easy-to-read form.
- Shows patterns and variations clearly.
- Useful for computing averages and dispersion.
- Facilitates graphical representation (histograms, polygons, ogives).
- Makes comparison between datasets simple.

2.3.9 Limitations

Limitation	Explanation
Loss of Detail	Grouping hides individual data points.
Choice of Class Interval	Improper intervals can misrepresent data.
Subjectivity	Requires judgment while selecting range or number of classes.

2.3.10 Applications

- Used in research, business, and education for summarizing data.
- Essential in graphical analysis (bar charts, histograms, ogives).
- Foundation for computing mean, median, mode, and standard deviation.
- Helps visualize distribution trends and comparisons.

2.4 Measures of Central Tendency

In any dataset, it's useful to know a single representative value that describes the overall characteristics of the data.
This value is called a Measure of Central Tendency.

It represents the center or average of the dataset and indicates where most of the observations tend to cluster.

The three main measures of central tendency are:

- Mean
- Median
- Mode

Each provides a different way of describing the "average" or central point of a dataset.

A Measure of Central Tendency is a single value that represents the entire distribution and gives a quick summary of the data.
It locates the center around which data values are distributed.

Objectives

1. To summarize large amounts of data with a single value.
2. To compare two or more datasets easily.
3. To provide a base for further statistical analysis (e.g., variance, correlation).
4. To describe the typical or most representative value of a dataset.

Characteristics of a Good Average

A good measure of central tendency should:

1. Be rigidly defined (no ambiguity).
2. Be based on all observations.
3. Be simple to calculate and understand.
4. Be not affected much by extreme values.
5. Be capable of further algebraic treatment.
6. Be stable and representative of the dataset.

2.4.1 Mean

The Mean (also called Arithmetic Mean) is the sum of all observations divided by the total number of observations.
It is the most commonly used measure of central tendency.

Formula (Individual Series):

$$\bar{X} = \frac{\sum X}{N}$$

Where,
X = Mean
∑X = Sum of all observations
N = Number of observations

For Discrete Series:

$$\bar{X} = \frac{\sum fX}{\sum f}$$

Where,
f = Frequency of each observation
X = Observation value

For Continuous Series:

$$\bar{X} = \frac{\sum fm}{\sum f}$$

Where,
m = Midpoint (Class Mark) = (Lower Limit + Upper Limit) / 2
f = Frequency of each class

Shortcut Methods

1. Assumed Mean Method

$$\bar{X} = A + \frac{\sum fd}{\sum f}$$

Where,
A = Assumed mean
d = X - A

2. Step-Deviation Method

$$\bar{X} = A + \left(\frac{\sum fd'}{\sum f}\right) \times c$$

Where,
d' = (X-A)/c
c = Class width (common difference)

Merits of Mean

Easy to calculate and understand.

Based on all values in the dataset.

Suitable for further mathematical treatment.

Provides a stable and precise average.

Demerits of Mean

Limitation	Explanation
Affected by Extreme Values	Very high or low values can distort the mean.
Not Suitable for Qualitative Data	Can only be used for numerical data.
Misleading in Skewed Data	May not represent the actual central tendency.

2.4.2 Median

The Median is the middle value of an ordered dataset.
It divides the data into two equal halves, where 50% of the observations are below and 50% are above the median.

Formula (Individual Series):

Arrange data in ascending order.

If (N) = number of observations:

$$\text{Median} = \begin{cases} \text{Value of } \left(\frac{N+1}{2}\right)^{th} \text{ item, if N is odd} \\ \text{Average of } \left(\frac{N}{2}\right)^{th} \text{ and } \left(\frac{N}{2}+1\right)^{th} \text{ items, if N is even} \end{cases}$$

For Discrete Series:

1. Find cumulative frequencies.
2. Locate the class where N/2 lies.

Formula:

$$\text{Median} = \text{Value of the } \frac{N+1}{2}^{th} \text{ observation}$$

For Continuous Series:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - F}{f} \right) \times C$$

Where,
(L) = Lower boundary of median class
(F) = Cumulative frequency before median class
(f) = Frequency of median class
(C) = Class width

Merits of Median

Not affected by extreme values.

Can be used for open-end classes.

Easy to calculate for uneven or skewed data.

Represents the middle position of the data clearly.

Demerits of Median

Limitation	Explanation
Ignores Extreme Values	Does not use all data values.
Not Suitable for Further Calculation	Cannot be used in algebraic operations.
Approximate Result	Depends on data grouping and class intervals.

2.4.3 Mode

The Mode is the value that occurs most frequently in a dataset.
It represents the most typical observation or the value with the highest frequency.

Formula (For Continuous Series):

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times C$$

Where,
(L) = Lower boundary of the modal class
(f1) = Frequency of the modal class
(f0) = Frequency of the class before the modal class

(f2) = Frequency of the class after the modal class
(C) = Class width

Merits of Mode

Simple and easy to identify.

Can be used for qualitative data (e.g., most common color or category).

Not affected by extreme values.

Represents the most typical value in a dataset.

Demerits of Mode

Limitation	Explanation
May Not Be Unique	Data may have more than one mode (bimodal/multimodal).
Less Accurate	Not based on all data points.
Difficult to Calculate in Continuous Data	Requires precise identification of modal class.

2.4.4 Relationship Between Mean, Median, and Mode

For a moderately skewed distribution, the following relationship exists:

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

or

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

This relationship helps estimate one measure when the others are known.

2.4.5 Comparison Between Mean, Median, and Mode

Basis	Mean	Median	Mode
Definition	Arithmetic average of all observations.	Middle value of the dataset.	Most frequent value.
Type of Data	Quantitative only.	Quantitative and ordered data.	Quantitative or qualitative.
Affected by Extreme Values	Yes	No	No
Mathematical Use	Suitable for further algebraic work.	Not suitable.	Not suitable.
Ease of Calculation	Simple and direct.	Moderate.	Easy (for discrete data).

2.4.6 Advantages of Measures of Central Tendency

Represent large data with a single value.

Help in comparison between datasets.

Assist in decision-making and forecasting.

Provide foundations for further analysis like variance and correlation.

Useful in business, economics, and research.

2.4.7 Limitations

Limitation	Explanation
Affected by Skewness	Central values may not represent data accurately if distribution is skewed.
Different Methods → Different Results	Each measure gives a different idea of "average."
Cannot Describe Full Dataset	Single value may hide variation among data points.

2.5 Measures of Dispersion

While measures of central tendency (like Mean, Median, Mode) tell us where the center of data lies, they don't tell us how spread out the data is.
To understand the variability or consistency in data, we use Measures of Dispersion.

Dispersion means the extent to which data values differ from the central value.
It indicates how much variation exists in a dataset.

A Measure of Dispersion is a statistical tool used to quantify the spread, scatter, or variability of a set of data values.
It tells us how far the values deviate from the mean or other central measures.

Objectives

1. To measure the degree of variation in data.
2. To compare the consistency of two or more datasets.
3. To support further statistical analysis (like correlation or regression).
4. To assess reliability of an average or mean.

Types of Measures of Dispersion

Type	Measure	Description
Absolute Measures	Range, Quartile Deviation, Mean Deviation, Standard Deviation	Expressed in the same units as the data.
Relative Measures	Coefficient of Range, Coefficient of Variation, etc.	Expressed as ratios or percentages — used for comparing datasets.

2.5.1 Range

The Range is the simplest measure of dispersion.
It is the difference between the highest and lowest values in a dataset.

Formula:

$$R = X_{max} - X_{min}$$

Where,
(R) = Range
(Xmax) = Maximum value
(Xmin) = Minimum value

Coefficient of Range:

$$\text{Coefficient of Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

Merits of Range

- Easy to understand and calculate.
- Quick measure for small datasets.
- Useful for preliminary comparison.

Demerits of Range

Limitation	Explanation
Based on Two Values Only	Ignores all intermediate data points.
Affected by Extreme Values	Not reliable if dataset has outliers.
Not Suitable for Large Datasets	Provides limited information.

2.5.2 Quartile Deviation (Q.D.)

The Quartile Deviation (Semi-Interquartile Range) measures the spread of the middle 50% of the data.
It is half the difference between the third and first quartile.

Formula:

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Where,
(Q3) = Third Quartile (75th percentile)
(Q1) = First Quartile (25th percentile)

Coefficient of Quartile Deviation:

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Merits of Quartile Deviation

- Not affected by extreme values.
- Simple to calculate.
- Suitable for skewed distributions.

Demerits of Quartile Deviation

Limitation	Explanation
Ignores 50% of Data	Only considers middle half of data.
Not Algebraically Usable	Can't be used in advanced statistical formulas.
Less Accurate	Not ideal for highly precise analysis.

2.5.3 Mean Deviation (M.D.)

The Mean Deviation is the average of absolute deviations of each observation from a central value (Mean, Median, or Mode).

Formula:

$$M.D. = \frac{\sum |X - A|}{N}$$

Where,
(A) = Mean, Median, or Mode
(|X - A|) = Absolute deviation of each observation from the average

For Discrete Series:

$$M.D. = \frac{\sum f|X - A|}{\sum f}$$

For Continuous Series:

$$M.D. = \frac{\sum f|m - A|}{\sum f}$$

Where,
(m) = Midpoint (Class Mark)

Coefficient of Mean Deviation:

$$\text{Coefficient of M.D.} = \frac{M.D.}{A}$$

Merits of Mean Deviation

- Uses all data values.
- Simple and easy to understand.
- Less affected by extreme values than range.

Demerits of Mean Deviation

Limitation	Explanation
Algebraically Inconvenient	Uses absolute values, making calculations complex.
Limited Application	Rarely used in advanced statistics.

2.5.4 Standard Deviation (S.D.)

The Standard Deviation (σ) is the most widely used and accurate measure of dispersion. It shows the average amount by which data values deviate from the mean.

It indicates consistency — smaller SD means data points are closer to the mean; larger SD means more variability.

Formula (Individual Series):

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

For Discrete Series:

$$\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{\sum f}}$$

For Continuous Series:

$$\sigma = \sqrt{\frac{\sum f(m - \bar{X})^2}{\sum f}}$$

Where,
(m) = Midpoint of class interval
(X') = Mean
(f) = Frequency

Shortcut Methods

1. Assumed Mean Method

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

Where,
(d = X - A) or (m - A)

2. Step-Deviation Method

$$\sigma = \sqrt{\left(\frac{\sum fd'^2}{\sum f}\right) - \left(\frac{\sum fd'}{\sum f}\right)^2} \times c$$

Where,
(d' = X-A/c)
(c) = Class width

Coefficient of Standard Deviation:

$$\text{Coefficient of S.D.} = \frac{\sigma}{\bar{X}}$$

2.5.5 Variance

The Variance measures the average of squared deviations from the mean.
It is the square of standard deviation.

$$\text{Variance} = \sigma^2 = \frac{\sum(X - \bar{X})^2}{N}$$

It provides a measure of total variability but in squared units.

Merits of Standard Deviation and Variance

- Based on all observations.
- Not affected by algebraic sign of deviations.
- Suitable for further analysis (e.g., correlation, regression).
- Most reliable measure of dispersion.

Demerits

Limitation	Explanation
Complex Calculation	Involves squaring and square roots.
Affected by Extreme Values	Sensitive to outliers.
Interpretation Difficulty	Variance expressed in squared units.

2.5.6 Coefficient of Variation (C.V.)

The Coefficient of Variation (C.V.) is a relative measure of dispersion expressed as a percentage of the mean.

It helps compare consistency between two or more datasets, even if their means differ.

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

Interpretation

- Lower C.V. → More consistent and stable data.
- Higher C.V. → More variable and less consistent data.

Example:

Dataset	Mean	S.D.	C.V.
A	50	5	10%
B	50	8	16%

→ Dataset A is more consistent because its C.V. is smaller.

2.5.7 Relative Measures of Dispersion

Relative measures express dispersion as a ratio or percentage, allowing comparison between datasets.

Measure	Formula
Coefficient of Range	$\frac{X_{max}-X_{min}}{X_{max}+X_{min}}$
Coefficient of Q.D.	$\frac{Q_3-Q_1}{Q_3+Q_1}$
Coefficient of M.D.	$\frac{M.D.}{A}$
Coefficient of S.D.	$\frac{\sigma}{X}$
Coefficient of Variation (C.V.)	$\frac{\sigma}{X} \times 100$

2.5.8 Comparison Between Measures of Dispersion

Measure	Based On	Best For	Remarks
Range	Extreme values	Quick comparison	Too simple, affected by outliers
Q.D.	Middle 50% of data	Skewed data	Ignores extremes
M.D.	All deviations	Moderate accuracy	Algebraically less convenient
S.D.	All deviations squared	Precise analysis	Most reliable
C.V.	S.D. and Mean	Comparing datasets	Expressed in %

2.5.9 Applications of Dispersion

- Used in quality control and risk assessment.
- Measures consistency and reliability of performance.
- Helps in economic forecasting and business planning.
- Basis for correlation, regression, and hypothesis testing.

2.6 Correlation

In real-world data, two or more variables often change together, for example, as income increases, expenditure also increases.
This relationship between variables is studied using Correlation Analysis.

Correlation is a statistical technique that measures the degree and direction of relationship between two or more variables.

It tells us:

- Whether variables move together (positive correlation),
- Move in opposite directions (negative correlation), or
- Have no connection (zero correlation).

Example:

Variable X (Temperature)	20	25	30	35	40
Variable Y (Cold Drink Sales)	100	200	300	400	500

→ As temperature increases, cold drink sales also increase — showing a positive correlation.

Objectives of Correlation Analysis

1. To measure the strength and direction of the relationship between variables.
2. To predict the value of one variable based on another.
3. To support decision-making and forecasting.
4. To provide a basis for regression analysis.

Types of Correlation

Type	Description	Example
1. Positive Correlation	When one variable increases, the other also increases.	Income ↑ → Expenditure ↑
2. Negative Correlation	When one variable increases, the other decreases.	Price ↑ → Demand ↓
3. Zero Correlation	When change in one variable does not affect the other.	Height and intelligence
4. Linear Correlation	Constant rate of change between two variables.	Straight-line relationship
5. Non-Linear (Curvilinear) Correlation	Rate of change is not constant.	Learning curve, growth rate

Methods of Studying Correlation

There are two major categories:

Category	Methods
1. Graphic / Diagrammatic Methods	- Scatter Diagram - Karl Pearson's Graphical Method
2. Mathematical / Statistical Methods	- Karl Pearson's Coefficient of Correlation - Spearman's Rank Correlation Coefficient

2.6.1 Scatter Diagram (Graphical Method)

A Scatter Diagram is a graphical representation of the relationship between two variables. Each pair of values (X, Y) is plotted as a point on a graph.

Interpretation

Type of Correlation	Pattern on Graph
Positive Correlation	Points form an upward-sloping line.
Negative Correlation	Points form a downward-sloping line.
No Correlation	Points are scattered randomly.

Advantages:

- Simple and visual method.
- Indicates direction and strength of correlation.
- Useful for initial analysis before numerical calculation.

Limitations:

- Does not give exact numerical value.
- Difficult to interpret when data points overlap.

2.6.2 Karl Pearson's Coefficient of Correlation (r)

Developed by Karl Pearson, this is the most common quantitative measure of correlation. It shows both direction and degree of correlation between two variables.

Symbol: (r)
Range: $-1 \leq r \leq +1$

- r = +1: Perfect Positive Correlation
- r = -1: Perfect Negative Correlation
- r = 0: No Correlation

Formula (Actual Mean Method):

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Where,
(X, Y) = Variables
(X', Y') = Their respective means

Shortcut Formula (for simpler calculations):

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

Interpretation of 'r'

Value of r	Type of Correlation
+1	Perfect Positive Correlation
0.7 to 0.9	High Positive Correlation
0.3 to 0.6	Moderate Positive Correlation
0.0 to 0.2	Low / Negligible Correlation
0	No Correlation
-0.3 to -0.6	Moderate Negative Correlation
-0.7 to -0.9	High Negative Correlation
-1	Perfect Negative Correlation

Merits of Karl Pearson's r

Gives both magnitude and direction of correlation.

Based on all observations.

Useful for quantitative prediction and further analysis.

Mathematically precise and reliable.

Demerits

Limitation	Explanation
Affected by Extreme Values	Outliers can distort correlation.
Assumes Linear Relationship	Not suitable for non-linear data.
Complex Calculation	Requires computation of deviations and products.

2.6.3 Spearman's Rank Correlation Coefficient (ρ)

When the data are in the form of ranks (not actual values), we use Spearman's Rank Correlation Coefficient.
It measures the degree of association between two ranked variables.

Symbol: ρ (rho)
Range: $-1 \leq \rho \leq +1$

Formula:

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

Where,
(D) = Difference between ranks of each pair
(N) = Number of pairs

Steps to Calculate:

1. Assign ranks to both variables.
2. Calculate the difference (D = R1 - R2).
3. Square each difference → (D^2).
4. Substitute values in the formula.

Interpretation

Value of ρ	Meaning
+1	Perfect Positive Rank Correlation
0	No Correlation
-1	Perfect Negative Rank Correlation

Merits

- Simple and less time-consuming.
- Can be used for qualitative data (e.g., rankings).
- Not affected by extreme values.
- Suitable for small samples.

Demerits

Limitation	Explanation
Less Accurate	Not suitable for large numerical datasets.
Cannot Handle Tied Ranks Easily	Needs correction formula.

Not Algebraically Usable	Can't be used for further mathematical operations.

2.6.4 Difference Between Karl Pearson's and Spearman's Correlation

Basis	Karl Pearson's r	Spearman's ρ
Nature of Data	Quantitative (actual values)	Qualitative / Ranked data
Type of Relationship	Linear	Monotonic
Formula Complexity	More complex	Easier
Use of Ranks	Not used	Used
Accuracy	High	Moderate
Sensitivity to Outliers	High	Low

2.6.5 Uses and Applications of Correlation

- In business, to study relationships between price and demand, income and expenditure, etc.
- In economics, to measure dependence of variables like investment and growth.
- In research, for identifying cause-and-effect tendencies.
- In finance, to study risk and return correlation.
- In marketing, to evaluate sales trends vs. advertising efforts.

2.6.6 Limitations

Limitation	Explanation
Correlation ≠ Causation	Correlation only shows relationship, not cause-effect.
Sensitive to Extreme Values	Outliers can mislead results.
Applicable Only to Measurable Variables	Qualitative factors (e.g., emotions) cannot be correlated.

2.7 Regression Analysis

While correlation only tells us that two variables are related, regression analysis goes a step further — it tells us how much one variable changes when another variable changes.

In simple words, correlation measures relationship, but regression predicts.

Regression Analysis is a statistical method used to estimate the relationship between two or more variables, where one is dependent (whose value we predict) and the others are independent (which influence it).

It helps to:

- Predict unknown values.
- Understand how variables depend on each other.

Example

Hours Studied (X)	2	4	6	8
Marks Scored (Y)	40	50	60	70

→ As study hours increase, marks also increase.
Regression helps us predict marks for any given number of study hours.

Objectives of Regression Analysis

1. To predict the value of one variable based on another.
2. To measure the functional relationship between variables.
3. To analyze cause-and-effect relationships.
4. To provide a basis for forecasting future trends.
5. To enable statistical estimation and modeling.

2.7.1 Types of Regression

Type	Description
1. Simple Regression	Involves two variables — one dependent (Y) and one independent (X).
2. Multiple Regression	Involves more than one independent variable predicting a dependent variable.
3. Linear Regression	Relationship between variables is a straight line.
4. Non-Linear Regression	Relationship follows a curve (not a straight line).

Simple Linear Regression

When there are two variables (X and Y), their relationship can be represented by a straight-line equation:

$$Y = a + bX$$

Where:

- (Y) = Dependent variable
- (X) = Independent variable
- (a) = Intercept (value of Y when X = 0)
- (b) = Regression coefficient (rate of change in Y for one unit change in X)

Interpretation

- (b) represents the slope — i.e., how much Y changes when X increases by 1 unit.

- If (b) is positive, Y increases as X increases.
- If (b) is negative, Y decreases as X increases.

2.7.2 Regression Lines

There are two regression lines in a bivariate analysis:

1. **Regression Line of Y on X**
   - Used to predict Y when X is known.
   - Equation: $Y-Y'=bYX(X-X')$
2. **Regression Line of X on Y**
   - Used to predict X when Y is known.
   - Equation: $X-X'=bXY(Y-Y')$

Regression Coefficients (b_{xy} and by_x)

These coefficients indicate the rate of change in one variable with respect to another.

Formulas:

1. **Regression Coefficient of Y on X:**

$$b_{YX} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

2. **Regression Coefficient of X on Y:**

$$b_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}$$

Relationship Between r and Regression Coefficients

$$r = \sqrt{b_{XY} \times b_{YX}}$$

And

$$b_{YX} = r \times \frac{\sigma_Y}{\sigma_X} \quad ; \quad b_{XY} = r \times \frac{\sigma_X}{\sigma_Y}$$

Where:

- r = Correlation coefficient

- σX,σY = Standard deviations of X and Y respectively

Properties of Regression Coefficients

1. Both byx and bxy have the same sign as (r).
2. Regression coefficients are independent of the change of origin, but not of scale.
3. If one regression coefficient is greater than 1, the other must be less than 1.
4. The geometric mean of the two regression coefficients equals the correlation coefficient.
5. Both regression lines intersect at the point (X', Y')

2.7.3 Methods of Calculating Regression

1. Arithmetic Mean Method

Steps:

1. Find the means of X and Y (X', Y').
2. Calculate deviations (X – X') and (Y – Y')
3. Use the formulas for byx and bxy.
4. Substitute in regression line equations.

2. Assumed Mean Method

If actual mean is not convenient, take an assumed mean (A) near the middle of data and find deviations.

Formulas remain the same, but deviations are calculated from A instead of (\bar{X}).

3. Direct Method (for small data)

Simple substitution of values into:

Y = a + bX

by solving the normal equations:

$$\sum Y = Na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Finding a and b

Solve the above equations simultaneously to find:

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum Y - b\sum X}{N}$$

Then substitute (a) and (b) in (Y = a + bX).

2.7.4 Difference Between Correlation and Regression

Basis	Correlation	Regression
Meaning	Measures the degree and direction of relationship between variables.	Studies cause-and-effect relationship between variables.
Purpose	To find association.	To predict one variable based on another.
Nature	Symmetrical.	Asymmetrical (depends on dependent variable).
Output	Correlation coefficient (r).	Regression equation (Y = a + bX).
Interpretation	Shows how strongly variables move together.	Shows how one variable changes with another.

2.7.5 Uses of Regression Analysis

- **Prediction:** Estimate future values (e.g., sales forecast).
- **Business Planning:** Helps in budgeting, resource allocation.
- **Economics:** Analyze relationship between income and expenditure.
- **Finance:** Study risk-return relationships.
- **Research:** Determine dependency among factors.

2.7.6 Limitations

Limitation	Explanation
Assumes Linear Relationship	Not accurate for non-linear data.
Affected by Extreme Values	Outliers can distort results.
Does Not Prove Causation	Only shows association, not direct cause-effect.
Complex for Large Data	Requires computational effort for multiple variables.

2.8 Time Series Analysis

In the real world, data is often collected over time — like monthly sales, yearly rainfall, or daily temperature.
Such data, when arranged in chronological (time) order, is called a Time Series.

Time Series Analysis helps us understand how data changes over time, identify patterns and trends, and make future forecasts.

A Time Series is a set of observations recorded at regular time intervals (daily, monthly, yearly, etc.) for a particular variable.

Example:
Yearly production of a company, monthly sales revenue, or daily stock prices.

Example Table:

Year	Sales (₹ in Lakhs)
2018	50
2019	55
2020	60
2021	70
2022	80

→ This is a Time Series because it shows data in a sequence of years.

Objectives of Time Series Analysis

1. To study patterns and trends in past data.
2. To forecast future values based on past behavior.
3. To identify seasonal variations or periodic fluctuations.
4. To measure the effect of external factors (e.g., inflation, policy changes).
5. To help in planning and decision-making.

2.8.1 Components of a Time Series

A time series is usually influenced by four main components:

Component	Meaning	Example
1. Secular Trend (T)	Long-term upward or downward movement of data.	Increase in population, technological growth.
2. Seasonal Variation (S)	Regular and repeating patterns within a year.	Ice cream sales rise in summer.
3. Cyclical Variation (C)	Long-term fluctuations due to business cycles.	Recession and boom phases.
4. Irregular Variation (I)	Random or unpredictable changes.	Natural disasters, strikes, pandemics.

Graphical Representation

If plotted, a time series typically looks like a wave pattern — showing trend, seasonal ups and downs, and irregular movements.

Formula for Time Series

$Y = T + S + C + I$

or

$Y = T \times S \times C \times I$

Where,
(Y) = Observed value at time 't'
(T) = Trend component
(S) = Seasonal component
(C) = Cyclical component
(I) = Irregular component

2.8.2 Types of Variations in Time Series

Variation Type	Description
Secular (Trend)	Long-term steady increase or decrease in data over many years.
Seasonal	Regular periodic changes within a fixed time period (e.g., months, quarters).
Cyclical	Repeating up-and-down movements occurring over long intervals (3–10 years).
Irregular	Sudden or short-term fluctuations caused by unexpected events.

2.8.3 Methods of Measuring Trend

The Trend (T) represents the general movement of data over a long time.
There are several methods to estimate this trend.

1. Free-Hand or Graphical Method

- Data values are plotted on a graph with time on the X-axis and the variable on the Y-axis.
- A smooth curve or line of best fit is drawn to represent the general trend.

Merits:
Simple and visual.

Good for small data.

Demerits:
Subjective — depends on personal judgment.

Not suitable for precise forecasting.

2. Semi-Average Method

- Divide the data into two equal parts.
- Find the average of each part.

- Draw a line joining the two points representing these averages — this line represents the trend.

Steps:

1. Divide data into two halves.
2. Compute average of each half.
3. Plot averages and draw a straight line joining them.

Merits:
Simple to use.
Gives approximate trend.

Demerits:
Not very accurate for large or irregular datasets.

3. Moving Average Method

- A moving average smooths out short-term fluctuations to show the long-term trend clearly.

Formula:

$$\text{n-year Moving Average} = \frac{\text{Sum of n consecutive values}}{n}$$

Example:
To find a 3-year moving average for sales data, add 3 consecutive yearly sales and divide by 3, then move ahead one year and repeat.

Merits:
Eliminates random variations.
Suitable for large datasets.

Demerits:
Loses some original data points.
Not suitable when trend changes frequently.

4. Least Squares Method

The most accurate method for finding a mathematical trend line.

The trend line (straight line) is represented as:

$Y = a + bX$

Where:
(Y) = Dependent variable (e.g., sales)
(X) = Time (coded as -2, -1, 0, +1, +2, etc.)
(a) = Intercept (mean value of Y)
(b) = Slope (rate of change per unit of time)

Formulas:

$\sum Y = Na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

After solving, substitute (a) and (b) in (Y = a + bX) to get the trend equation.

Merits:

Most accurate and objective.
Gives a precise equation for forecasting.

Demerits:
Requires mathematical calculations.
Assumes a linear trend only.

2.8.4 Measurement of Seasonal Variation

Seasonal variations can be measured using several methods:

Method	Description
1. Ratio-to-Trend Method	Compare actual values to estimated trend values.
2. Ratio-to-Moving Average Method	Compare observed values to moving averages.
3. Percentage Method	Express each season's value as a percentage of the yearly average.

Steps for Ratio-to-Moving Average Method:

1. Compute moving averages (trend values).
2. Divide actual values by trend values × 100 → gives seasonal indices.
3. Average the indices for each season.
4. Adjust to make their average = 100.

Interpretation:

- If seasonal index = 120, values are 20% above average.
- If seasonal index = 80, values are 20% below average.

2.8.5 Utility and Importance of Time Series

Helps in forecasting and future planning.
Useful in policy-making and decision-making.
Assists in identifying cyclical and seasonal patterns.
Helps businesses plan production and inventory.
Important in economics, meteorology, finance, and research.

2.8.6 Limitations

Limitation	Explanation
Assumes Past Patterns Continue	May fail during sudden economic or natural changes.
Affected by Irregular Variations	Unexpected events distort trends.
Difficult for Small Datasets	Needs sufficient past observations.
Complex Calculations	Some methods (like least squares) require computation.

2.9 Index Numbers

In economics and business, it's important to measure changes over time — like prices, output, wages, or cost of living.
However, since data involves many items and time periods, direct comparison becomes difficult.

To simplify this, we use Index Numbers, which show how much a variable has changed in comparison to a base period.

Definition

An Index Number is a statistical measure that shows the relative change in the value or quantity of a variable over time.
It expresses data as a percentage of a base value.

Example:
If the price of sugar increases from ₹20 to ₹30 per kg,
then the price index = 30/20 X 100 = 150
→ Prices increased by 50%.

Formula:

$$\text{Index Number} = \frac{\text{Value in Current Year}}{\text{Value in Base Year}} \times 100$$

Features of Index Numbers

1. **Relative Measure:** Expresses data as a percentage relative to a base period.
2. **Simplifies Comparison:** Shows changes in price, quantity, or cost easily.
3. **Statistical Tool:** Used widely in economics and business analysis.

4. **Dynamic Indicator:** Reflects economic trends and inflation/deflation levels.
5. **Composite Indicator:** Can combine multiple variables into one measure.

Objectives of Index Numbers

1. To measure changes in variables over time (e.g., price, production, wages).
2. To study economic trends like inflation or growth.
3. To compare data between two periods or regions.
4. To serve as a barometer of economic conditions.
5. To help in policy formulation and forecasting.

2.9.1 Types of Index Numbers

Type	Purpose / Measures	Example
1. Price Index	Measures changes in prices of goods and services.	Consumer Price Index (CPI)
2. Quantity Index	Measures changes in physical quantities.	Industrial Production Index
3. Value Index	Measures changes in total monetary value.	Exports or Imports Value Index
4. Cost of Living Index	Shows changes in cost of maintaining standard of living.	Wage and Salary Adjustments
5. Wholesale Price Index (WPI)	Measures average changes in wholesale prices.	Inflation Measurement

2.9.2 Steps in the Construction of Index Numbers

1. **Define Purpose and Scope**
   → Decide what to measure (price, output, cost of living, etc.).
2. **Selection of Base Year**
   → A normal year with stable economic conditions is chosen.
   → The base year index is always taken as 100.
3. **Selection of Items**
   → Choose representative commodities or data items relevant to the study.
4. **Collection of Data**
   → Gather accurate and comparable data for base and current years.
5. **Selection of Method**
   → Choose appropriate formula (simple or weighted).
6. **Calculation and Interpretation**
   → Compute index number and analyze changes.

Base Year Types

Type	Meaning
Fixed Base	Compare all years with one fixed base year.
Chain Base	Compare each year with the immediately preceding year.

2.9.3 Methods of Constructing Index Numbers

A. Simple Index Numbers

These are calculated without considering weights.

(i) Simple Aggregative Method

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Where,
(P0) = Price in base year
(P1) = Price in current year

Merits:
Simple and easy to compute.
Useful for rough comparisons.

Demerits:
Does not consider importance (weights) of items.
Distorted if units or quantities differ.

(ii) Simple Average of Price Relatives Method

$$P_{01} = \frac{\sum \left(\frac{P_1}{P_0} \times 100 \right)}{N}$$

Where,
(N) = Number of items.

Merits:
Avoids unit differences.
Simple and quick to calculate.

Demerits:
Each item is given equal importance regardless of value or use.

B. Weighted Index Numbers

Weights reflect the relative importance of each item.

(i) Weighted Aggregative Method

$$P_{01} = \frac{\sum P_1 W}{\sum P_0 W} \times 100$$

Where,
(W) = Weight (quantity, expenditure, or importance)

Types of Weighted Index Numbers

Method	Formula	Remarks
Laspeyres Method	$P_{01} = \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100$	Uses **base year quantities** as weights.
Paasche's Method	$P_{01} = \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100$	Uses **current year quantities** as weights.
Fisher's Ideal Method	$P_{01} = \sqrt{\left(\frac{\sum P_1 Q_0}{\sum P_0 Q_0}\right) \times \left(\frac{\sum P_1 Q_1}{\sum P_0 Q_1}\right)} \times 100$	**Geometric mean** of Laspeyres and Paasche — most accurate.

Interpretation:

- If index number > 100 → Increase over base year.
- If index number < 100 → Decrease over base year.

2.9.4 Tests of Consistency of Index Numbers

To ensure reliability, an index should satisfy certain tests:

Test	Condition / Formula	Satisfied By
1. Time Reversal Test	$P_{01} \times P_{10} = 1$	Fisher's Ideal Index
2. Factor Reversal Test	$P_{01} \times Q_{01} = \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$	Fisher's Ideal Index
3. Circular Test	$P_{01} \times P_{12} = P_{02}$	Not usually satisfied by all methods

2.9.5 Cost of Living Index (Consumer Price Index - CPI)

Definition

The Cost of Living Index measures changes in the cost of purchasing goods and services by consumers over time.

It reflects how much more or less expensive life becomes compared to the base year.

Formula (Weighted Average Method):

$$CPI = \frac{\sum P_1 W}{\sum P_0 W} \times 100$$

Where,
(W) = Weights based on expenditure pattern of consumers.

Uses:

- Measure inflation rate.
- Decide wage and salary adjustments.
- Analyze standard of living.
- Guide government policies on prices and income.

2.9.6 Uses of Index Numbers

Measure changes in prices, output, or cost of living.
Useful for economic planning and policy decisions.
Helps business forecasting (sales, demand, production).
Indicates inflation and deflation trends.
Enables comparison between different time periods or places.

2.9.7 Limitations

Limitation	Explanation
Selection Bias	Wrong choice of base year or items affects accuracy.
Data Errors	Inaccurate or incomplete data leads to wrong results.
Changing Quality / Consumption	Items and their importance change over time.
Regional Differences	Same index may not apply equally everywhere.
Complex Calculation	Weighted indices require detailed data.

2.10 Probability

In daily life and data analysis, many events occur by chance — like getting heads in a coin toss, drawing a red card, or a machine failing.
We can't predict outcomes with certainty, but we can measure the likelihood of these events using Probability.

Probability is a numerical measure of the likelihood that a particular event will occur.
It always lies between 0 and 1.

- 0 means the event is impossible.
- 1 means the event is certain.
- Any value between 0 and 1 shows degree of uncertainty.

Example

- Tossing a coin → Probability of getting heads = ½
- Throwing a die → Probability of getting a 4 = 1/6

Formula:

$$P(E) = \frac{\text{Number of favourable outcomes}}{\text{Total number of outcomes}}$$

Where,
(P(E)) = Probability of an event (E).

Features of Probability

1. Probability is a numerical measure (ranges between 0 and 1).
2. The sum of probabilities of all possible outcomes = 1.
3. Helps in quantifying uncertainty.
4. Based on logical reasoning and experimentation.
5. Used for prediction and decision-making.

Terms Used in Probability

Term	Meaning
Experiment	Any activity that can be repeated under identical conditions and gives a set of outcomes.
Trial	Each repetition of an experiment.
Outcome	The result of a single trial (e.g., head or tail).
Sample Space (S)	The set of all possible outcomes.
Event (E)	A subset of the sample space (one or more outcomes).
Favourable Outcomes	Outcomes that represent the desired event.

Example:

Tossing a coin once:

- Sample space, (S = {H, T})
- Probability of head, (P(H) = 1/2)
- Probability of tail, (P(T) = 1/2)

2.10.1 Types of Events

Type of Event	Description / Example
1. Simple Event	An event with only one outcome. *Example:* Getting a 3 on a die.

2. Compound Event	Contains two or more simple events. *Example:* Getting an even number (2, 4, 6).
3. Certain Event	Always occurs. *Example:* Getting a number between 1–6 on a die.
4. Impossible Event	Cannot occur. *Example:* Getting a 7 on a die.
5. Mutually Exclusive Events	Events that cannot occur together. *Example:* Getting head and tail in one toss.
6. Exhaustive Events	All possible outcomes of an experiment.
7. Independent Events	Occurrence of one does not affect the other. *Example:* Two separate coin tosses.
8. Dependent Events	Occurrence of one affects the other. *Example:* Drawing cards without replacement.
9. Complementary Events	Two events where one happens if the other does not. *Example:* Rain and No Rain.

2.10.2 Rules of Probability

Rule 1 – Range of Probability

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

Probability can never be negative or more than 1.

Rule 2 – Sum of All Probabilities

$P(S)=1$

The sum of probabilities of all possible outcomes is always 1.

Rule 3 – Complementary Events

$P(\text{not } E)=1-P(E)$

If probability of rain is 0.3, then
probability of no rain = (1 - 0.3 = 0.7)

Rule 4 – Addition Rule (for Mutually Exclusive Events)

If events A and B cannot occur together, then:

$P(A \text{ or } B)=P(A)+P(B)$

Example:
Probability of getting a 2 or 5 when a die is thrown =
$P(2) + P(5) = 1/6 + 1/6 = 1/3$

Rule 5 – Addition Rule (for Non-Mutually Exclusive Events)

If events can occur together:

P(A or B)=P(A)+P(B)−P(A and B)

Example:
Probability of drawing a red card or a king from a deck.

Rule 6 – Multiplication Rule (for Independent Events)

P(A and B)=P(A)×P(B)

Example:
Probability of getting heads in two consecutive tosses
= (½ × ½ = ¼)

Rule 7 – Multiplication Rule (for Dependent Events)

P(A and B)=P(A)×P(B|A)

Where (P(B|A)) = Probability of B given that A has already occurred.

2.10.3 Conditional Probability

The probability that event B occurs given that event A has already occurred is called conditional probability.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Example:
Probability that a card is king given that it is a face card.

Properties

1. $P(A|A) = 1$
2. $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$
3. $P(A \text{ and } B) = P(A) \times P(B|A)$

2.10.4 Theorems on Probability

1. Addition Theorem (General Form)

For any two events A and B:
P(A or B)=P(A)+P(B)−P(A and B)

2. Multiplication Theorem

For independent events:
$P(A \text{ and } B) = P(A) \times P(B)$

For dependent events:
$P(A \text{ and } B) = P(A) \times P(B|A)$

3. Bayes' Theorem

Bayes' Theorem is used to revise probabilities when new information becomes available.

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum P(A_j)P(B|A_j)}$$

Application:
Used in machine learning, medical testing, and risk prediction.

2.10.5 Probability Distribution

A probability distribution shows how probabilities are distributed among all possible outcomes.

Example:

Tossing a coin twice:

X (No. of Heads)	0	1	2
P(X)	¼	½	¼

$\sum P(X) = 1$

2.10.6 Importance of Probability

Basis of all statistical inference.
Helps in decision-making under uncertainty.
Used in risk management and forecasting.
Fundamental in machine learning, AI, and data science.
Useful in insurance, finance, and quality control.

2.10.7 Limitations

Limitation	Explanation
Depends on Assumptions	Results depend on fairness and uniform probability.
Not Always Objective	May differ from real-life outcomes.
Difficult for Complex Events	Large sample spaces increase complexity.