MAA OMWATI DEGREE COLLEGE, HASSANPUR

EXAM NOTES

CLASS-B.C.A

SEM- 3RD (NEP)

SUBJECT- PROBABILTY AND STATISTICS

UNIT 1

Statistics is the branch of mathematics focused on collecting, analyzing, interpreting, presenting, and organizing data. It helps us make sense of raw information, uncover patterns, and make informed decisions under uncertainty.

1. Key Terminology

- **Population:** The entire group of individuals, items, or events you want to study (e.g., all registered voters in a country).
- **Sample:** A smaller, manageable subset of the population chosen for measurement (e.g., 1,000 registered voters surveyed for an election poll).
- **Data:** The actual facts, measurements, or observations collected (can be qualitative/categorical like eye color, or quantitative/numerical like height or income).
- **Variable:** Any characteristic, number, or quantity that can be measured or counted.

2. Main Branches of Statistics

Statistics is broadly divided into two main categories:

Branch	Description	Example
Descriptive Statistics	Summarizes, organizes, and describes the main features of a dataset using tables, graphs, and summary metrics.	Calculating the average test score of a single classroom.
Inferential Statistics	Uses a random sample of data taken from a population to make generalizations,	Using the average test score of 50 students to estimate the average

Branch	Description	Example
	predictions, or inferences about the wider population.	score of all students in the state.

3. Core Measures of Data

To understand datasets, statisticians rely on key summary values:

A. Measures of Central Tendency (Where is the middle?)

- **Mean:** The arithmetic average of all data points.
- **Median:** The exact middle value when data is ordered from least to greatest.
- **Mode:** The most frequently occurring value in the dataset.

B. Measures of Variability / Dispersion (How spread out is the data?)

- **Range:** The difference between the highest and lowest values.
- **Variance:** Measures how far each number in the set is from the mean.
- **Standard Deviation:** The square root of the variance; shows how much variation exists from the average. A low standard deviation means data points are close to the mean, while a high standard deviation indicates data is spread out.

4. Why Statistics Matters

- **Decision Making:** Helps businesses forecast sales, manage risks, and target marketing.
- **Scientific Research:** Used in medicine and science to test hypotheses and determine if treatment results are statistically significant or due to chance.
- **Policy Making:** Governments use statistical data on unemployment, inflation, and demographics to shape public policy.

1. Qualitative Data (Categorical)

Qualitative data describes qualities, characteristics, or categories. It answers questions like **"what kind?"**, **"how?"**, or **"which category?"**. You cannot perform meaningful mathematical operations on qualitative data (e.g., you cannot calculate the "average" eye color).

- **Characteristics:** Non-numerical, descriptive, and divided into labels or groups.
- **Common Examples:**
  - **Colors:** Red, blue, green.
  - **Names or Places:** Paris, Tokyo, Canada.
  - **Feedback/Opinions:** "Satisfied," "Neutral," "Dissatisfied."
  - **Marital Status:** Single, married, divorced.

Subtypes of Qualitative Data:

- **Nominal:** Categories with **no inherent order or ranking** (e.g., blood types: A, B, AB, O; car brands: Toyota, Ford, Honda).
- **Ordinal:** Categories that have a **specific logical order or ranking**, but the exact distance between the ranks cannot be measured (e.g., survey ratings like "Poor, Fair, Good, Excellent" or academic grades like A, B, C).

2. Quantitative Data (Numerical)

Quantitative data represents numerical values, quantities, or amounts. It answers questions like **"how much?"** or **"how many?"**. Mathematical operations (addition, subtraction, averages) can be performed on this data.

- **Characteristics:** Numerical, measurable, and objective.
- **Common Examples:**
  - **Height or Weight:** 175 cm, 68 kg.
  - **Temperature:** 22°C.
  - **Countable items:** Number of cars in a parking lot, number of website visitors.

Subtypes of Quantitative Data:

- **Discrete Data:** Consists of **countable, distinct whole numbers**. You cannot have fractions of these values (e.g., number of children in a family, number of goals scored in a match).
- **Continuous Data:** Consists of **measurable values that can be broken down into infinite decimals or fractions** along a continuum (e.g., weight, time, height, exact speed).

3. Comparison at a Glance

Feature	Qualitative Data	Quantitative Data
Focus	Descriptions, attributes, and categories	Numbers, quantities, and measurements
Question Answered	Which category? What kind?	How much? How many?
Analysis Method	Counting frequencies, proportions, text analysis	Statistical formulas (mean, median, standard deviation, regression)
Common Visualizations	Bar charts, pie charts, word clouds	Histograms, scatter plots, line graphs

CLASSIFICATION OF DATA

Data classification is the systematic arrangement or grouping of data into distinct categories based on shared characteristics, sources, structure, or time frames. Properly classifying data helps researchers and organizations manage, store, and analyze information more efficiently.

1. Classification Based on Source (Origin)

This classification looks at **how and where** the data was originally gathered.

- **Primary Data:** First-hand data collected directly by the researcher for a specific study or purpose.
    - *Examples:* Surveys, interviews, focus groups, or clinical trial results.
- **Secondary Data:** Data that has already been collected, processed, and published by someone else for another purpose.
    - *Examples:* Government census reports, books, historical statistics, or financial records from a company.

2. Classification Based on Time Dimension

This category focuses on **when** the data is collected relative to time.

- **Cross-Sectional Data:** Data collected from multiple subjects, entities, or individuals at a **single, specific point in time**.
    - *Example:* A survey measuring the employment status of 5,000 people in January 2026.
- **Time-Series Data:** Data collected from a single subject or entity **over regular, sequential intervals of time**.
    - *Example:* Tracking the daily closing price of a stock every day for an entire year.
- **Pooled / Panel Data:** A combination of both—tracking multiple subjects over multiple time periods.

3. Classification Based on Structure

In modern data science and computing, data is often classified by how organized or "formatted" it is.

- **Structured Data:** Highly organized data that fits neatly into traditional relational databases, rows, and columns. It is easily searchable by machines.
    - *Examples:* Spreadsheets, SQL databases, customer tables.
- **Semi-Structured Data:** Data that does not reside in a relational database, but contains organizational markers, tags, or hierarchies that make it easier to parse.
    - *Examples:* JSON files, XML files, HTML code, emails.

- **Unstructured Data:** Raw, unorganized data that lacks a predefined data model or structure, making it harder to analyze traditionally.
    - *Examples:* Audio recordings, video files, PDFs, social media posts, satellite imagery.

4. Classification Based on Measurement Scale (Levels of Measurement)

Statisticians also classify data based on the mathematical nature of the scale used to measure it (often referred to as **Stevens' scales of measurement**):

Scale Type	Description	Mathematical Properties	Example
Nominal	Categories with no order or ranking.	Can only count frequencies/modes.	Eye color, gender, zip codes.
Ordinal	Categories with a logical order, but unknown intervals.	Can determine greater/lesser, but no precise math.	Likert scales ("Satisfied", "Neutral"), rankings (1st, 2nd, 3rd).
Interval	Ordered numerical values where the difference between values is meaningful, but there is **no true zero** (zero is arbitrary).	Addition and subtraction are valid.	Temperature in Celsius or Fahrenheit (0°C does not mean "no temperature").
Ordered / Ratio	Numerical values with an order, equal intervals, and an **absolute, non-arbitrary zero** (zero means complete absence of the variable).	All mathematical operations (addition, subtraction, multiplication, division) are valid.	Weight, height, income, age, time duration.

A **frequency distribution** is a table that organizes raw data into classes (or categories) along with the corresponding number of observations (frequencies) that fall into each category. It transforms a confusing mass of numbers into a structured, readable summary.

Depending on the nature of the data, frequency distributions can be **ungrouped (discrete)** or **grouped (continuous)**.

1. Key Terms to Know

- **Class Intervals (Classes):** The sub-ranges into which data is grouped (e.g., 10-14, 15-19).
- **Class Limits:** The lowest and highest values that can fit into a class.
    - *Lower Class Limit:* The smallest value in a class.
    - *Upper Class Limit:* The largest value in a class.
- **Class Boundaries:** The true dividing limits used to separate classes without gaps (calculated by subtracting 0.5 from lower limits and adding 0.5 to upper limits for whole numbers).
- **Class Size (Width):** The difference between the upper and lower class boundaries (or limits).
- **Class Mark (Midpoint):** The center of a class interval, calculated as:
- text{Class Mark} = {Lower Limit} + {Upper Limit}}/{2}
- **Frequency (f):** The number of data values residing within a specific class.

2. Steps to Construct a Grouped Frequency Distribution

When dealing with a large continuous dataset, follow these step-by-step procedures to build a frequency distribution table:

Step 1: Find the Range (R)

Calculate the difference between the highest value (HS) and the lowest value (LS) in the dataset:

R = HS - LS

Step 2: Determine the Number of Classes (k)

Decide how many groups/classes you want. Ideally, a table should have between **5 to 15 classes**. You can also use **Sturges' Formula** to estimate the ideal number of classes:

k = 1 + 3.322(\log_{10} N)

(where N is the total number of observations). Always round k to the nearest whole number.

Step 3: Calculate the Class Width / Size (c)

Divide the Range by the number of classes:

c = {R}/{k}

- **Rule of Thumb:** Always **round up** the class width to the next convenient whole number to ensure all data points fit comfortably, regardless of decimals.

Step 4: Set Up the Class Limits

- Choose the **lowest value** (or a slightly lower, convenient number) as the lower class limit of the first class.
- Add the class width (c) successively to find the lower limits of the subsequent classes.
- Establish corresponding upper limits so that classes are **mutually exclusive** (meaning a data point can only belong to one class).

Step 5: Tally and Count Frequencies

- Go through your raw data value by value, assigning tally marks (text{IIII}) to their respective classes.
- Bundle tallies in groups of five ({HI}) for easy counting.
- Convert the tallies into numerical values to form your final **Frequency (f)** column.

3. Example of a Frequency Distribution Table

Imagine we recorded the scores of 30 students on a quiz and organized them using classes of width 6:

Class Intervals	Tally	Frequency (f)	Class Mark (Midpoint)
14 - 19	{IIII}	4	16.5
20 - 25	{HII}	5	22.5
26 - 31	{III}	7	28.5
32 - 37	{HII I}	6	34.5

Class Intervals	Tally	Frequency (f)	Class Mark (Midpoint)
38 - 43	{IIII}	4	40.5
44 - 49	{III}	3	46.5
50 - 55	{I}	1	52.5
Total		**sum f = 30**	

4. Types of Derived Frequencies

Beyond basic counts, tables often include extended analytical columns:

- **Cumulative Frequency (cf):** The running total of frequencies as you move down the table. Found by successively adding each class frequency to the sum of its predecessors.
- **Relative Frequency:** The proportion of the total data falling into a class, calculated as {f}{N}.
- **Percentage Frequency:** The relative frequency expressed as a percentage

Diagrammatic representation is the technique of visually expressing statistical data using geometric figures, pictures, lines, and charts. It transforms complex, raw numbers from tables into appealing, easy-to-understand formats that facilitate quick comparisons and reveal underlying trends at a glance.

Slideshare+ 1

1. Main Types of Diagrams

Statistics utilizes various categories of diagrams depending on the nature of the data and the objective of the analysis:

A. One-Dimensional Diagrams (Bar Diagrams)

In these diagrams, **only the length (height)** of the bar is considered, while the width remains constant. They are primarily used to compare discrete categories.

- **Simple Bar Diagram:** Consists of single bars representing a single variable across different categories (e.g., total sales over five years).

- **Multiple Bar Diagram:** Features two or more adjacent bars grouped together to compare multiple related variables side-by-side (e.g., comparing imports and exports from 2022 to 2025).
- **Sub-Divided (Component) Bar Diagram:** Bars are split into segments stacked on top of each other to show how a total value breaks down into sub-components.

- **Percentage Bar Diagram:** A sub-divided bar diagram where each bar is scaled to a total of 100%, making proportional comparisons across different total sizes easy.

- **Deviation Bar Diagram:** Used to represent net positive or negative changes, such as net profits, net losses, or trade deficits.

B. Two-Dimensional Diagrams (Area Diagrams)

When data varies significantly or involves multiple components of size, **both length and width** are utilized to form geometric shapes where the **total area** represents the magnitude of the data.

- **Rectangles:** Areas are proportional to the values they represent.
- **Pie Chart (Circular Diagram):** A circle divided into sectors, where the area of each sector (and its angle) is proportional to the percentage share of a component relative to the whole (360∘).

C. Three-Dimensional Diagrams

Used for very large datasets where **length, width, and height** are all factored in, creating volume-based shapes like cubes, cylinders, or spheres.

D. Pictograms and Cartograms

- **Pictograms:** Uses small pictures or symbols to represent quantities (e.g., drawing a picture of a car to represent 10,000 vehicles produced).

- **Cartograms (Map Diagrams):** Uses geographical maps to display statistical data or distributions across regions (e.g., population density or election results by state).

2. Advantages of Diagrammatic Representation

- **Simplifies Complex Data:** Large blocks of confusing numbers become instantly readable.
- **Quick Comparisons:** Makes it effortless to compare different categories, periods, or groups at a glance.
- **Eye-Catching and Engaging:** Attracts attention faster than text or tables and stays memorable longer.
- **Universal Appeal:** Easily understood by non-experts, professionals, and decision-makers alike.

3. General Rules for Constructing Good Diagrams

1. **Appropriate Title:** Every diagram must have a clear, concise title explaining what data is displayed.

2. **Proper Scale:** Choose a neat, uniform scale (e.g., multiples of 5, 10, or 100) and avoid irregular or odd intervals.
3. **Index / Key:** If using multiple colors, shades, or patterns (such as in a multiple or sub-divided bar chart), always include a legend explaining them.
4. **Neatness and Proportion:** Keep a balanced proportion between the height and width of the chart so it looks visually clean and undistorted.

5. **Source Note:** Mention the source of the data at the bottom of the diagram for credibility and verification.

Sample Dataset: Monthly Sales of a Store

Imagine a small electronics shop recorded its total sales (in thousands of dollars) over a span of five months:

Month	Sales (in thousands)

Month	Sales (in thousands)
January	15
February	22
March	18
April	30
May	25

Example 1: Simple Bar Chart

A **simple bar chart** uses vertical or horizontal bars of equal width, where the height of each bar corresponds directly to the data value.

Visual Representation (Text-Based Bar Chart)

Plaintext

```
Sales ($ in thousands)
 35 |
 30 |                      [  ]
 25 |                      [  ]               [  ]
 20 |          [  ]        [  ]               [  ]
 15 |  [  ]    [  ]        [  ]        [  ]   [  ]
 10 |  [  ]    [  ]        [  ]        [  ]   [  ]
  5 |  [  ]    [  ]        [  ]        [  ]   [  ]
  0 +-----------------------------------------------
     Jan     Feb         Mar         Apr    May
```

- **Key Takeaway:** At a single glance, you can immediately tell that **April** had the highest sales (30,000) while **January** had the lowest (15,000).

A **measure of central tendency** is a single value that attempts to describe a set of data by identifying the central position within that data set. They are sometimes called summary statistics. The three most common measures are the mean, median, and mode.

MEAN

The **arithmetic mean** (most commonly called the **mean** or **average**) is the most popular and widely used measure of central tendency. It is calculated by adding up all the individual values in a data set and then dividing that sum by the total number of values.

Formulas

Depending on whether you are working with a sample or an entire population, the notation differs slightly, though the calculation method is the same.

1. For Raw (Ungrouped) Data

- **Sample Mean:** $\bar{x}=n\sum x$
- **Population Mean:** $\mu=N\sum x$

Where:

- $\sum x$ = Sum of all data values
- n = Number of values in the sample
- N = Total number of values in the population

2. For Grouped Data (Frequency Distribution)

When data is organized into a frequency table, multiply each value (or midpoint) by its frequency, sum them up, and divide by the total frequency.

- **Formula:** $\bar{x}=\sum f\sum(f\cdot x)$

Where:

- f = Frequency of each class or value
- x = Value (or midpoint of a class interval)

Step-by-Example (Raw Data)

Data Set: 4,8,6,5,7

1. **Count the values (n):** There are 5 numbers in this data set.
2. **Sum the values ($\sum x$):** $4+8+6+5+7=30$
3. **Divide the sum by n:**

$\bar{x} = \frac{30}{5} = 6$

Result: The arithmetic mean of this data set is **6**.

Properties and Characteristics

- **Mathematical Rigor:** Every single value in the data set is used in the calculation.
- **Sensitivity to Outliers:** Because every value counts, introducing an extremely high or low value (an outlier) will pull the mean toward it, potentially skewing how "central" it feels.
- **Sum of Deviations:** The sum of the differences (deviations) between each data value and the mean is always equal to zero: $\sum(x - \bar{x}) = 0$.

Summary of Advantages and Limitations

Advantages	Limitations
* Easy to calculate and understand intuitively. • Uses every data point, making it mathematically stable. • Essential for further statistical calculations (like variance and standard deviation).	* Highly sensitive to extreme values (outliers or skewed distributions). • Can result in a value that doesn't actually exist within the original data set (e.g., an average of 2.4 children).

MEDIAN

In statistics, the **median** is a measure of central tendency that represents the middle value of a dataset when it is ordered from least to greatest (or greatest to least). It divides the data into two equal halves, meaning 50% of the observations lie above the median and 50% lie below it.

Testbook+ 1

How to Calculate the Median

To find the median, follow these steps:

1. **Sort the data** in ascending or descending order.
2. **Identify the middle position** depending on whether the total number of data points (n) is odd or even:

- **If n is Odd:** The median is the exact middle number.

- *Formula:* Value at the (2n+1)th position.

  NCBI - NIH

- *Example:* For the set {3,1,7,9,5} → Sorted: {1,3,5,7,9} → **Median = 5** (the 3rd term).

- **If n is Even:** There is no single middle number, so the median is the average (mean) of the two central numbers.
  - *Formula:* Average of the (2n)th and (2n+1)th positions.

    NCBI - NIH

  - *Example:* For the set {4,1,8,6} → Sorted: {1,4,6,8} → Middle numbers are 4 and 6 → **Median = 24+6=5**.

Key Characteristics & Advantages

- **Resistance to Outliers:** Unlike the mean (average), the median is not heavily skewed by extremely high or low values (outliers). For example, in the dataset {2,3,4,5,100}, the median is **4**, which much better reflects the "typical" value than the mean of **22.8**.
-
- **Use in Skewed Data:** It is widely preferred for skewed distributions, such as household incomes or housing prices, where a small number of extremely wealthy individuals can artificially inflate the average (mean).

Formulas to Calculate the Median

The formula to find the median depends on whether you are working with raw (ungrouped) data or grouped frequency distribution data, and whether the number of observations is odd or even.

1. Ungrouped Data (Raw Data)

For a dataset containing n observations arranged in ascending or descending order:

- **When n is Odd:**

  Median=(2n+1)th term

- **When n is Even:**

  Median=2(2n)th term+(2n+1)th term

2. Grouped Data (Frequency Distribution)

For continuous data organized into intervals/classes, the median is calculated using the following formula:

$$\text{Median} = L + \left(\frac{\frac{n}{2} - CF}{f}\right) \times h$$

Where:

- L = Lower boundary of the median class (the class interval where the cumulative frequency (2n) falls)
- n = Total number of observations (sum of all frequencies)
- CF = Cumulative frequency of the class *preceding* the median class
- f = Frequency of the median class
- h = Class width (size of the class interval)

MODE

In statistics, the **mode** is a measure of central tendency that represents the value or values that appear most frequently in a dataset. Unlike the mean and median, a dataset can have one mode, multiple modes, or no mode at all.

Types of Modes

- **Unimodal:** A dataset with only **one** mode (e.g., $\{1,2,2,3,4\} \rightarrow \text{Mode}=2$).
- **Bimodal:** A dataset with **two** modes, meaning two different values tie for the highest frequency (e.g., $\{1,2,2,3,3,4\} \rightarrow \text{Modes}=2 \text{ and } 3$).
- **Multimodal:** A dataset with **three or more** modes.
- **No Mode:** A dataset where every value appears the exact same number of times (e.g., $\{1,2,3,4\} \rightarrow$ No mode).

How to Calculate the Mode

1. Ungrouped Data (Raw Data)

For a raw dataset, the mode is found by simply counting the frequency of each value and identifying the one(s) with the highest count.

- *Example:* For the set $\{4,1,4,2,4,3,5\}$, the number 4 appears three times, which is more than any other number.
- **Mode = 4**

2. Grouped Data (Frequency Distribution)

For continuous data organized into class intervals, the mode is calculated using the following formula:

Mode=L+(2f1−f0−f2f1−f0)×h

Where:

- L = Lower boundary of the modal class (the class interval with the highest frequency)
- f1 = Frequency of the modal class
- f0 = Frequency of the class *preceding* the modal class
- f2 = Frequency of the class *succeeding* the modal class
- h = Class width (size of the class interval)

Key Characteristics & Advantages

- **Easy to Identify:** For raw data, it can often be found by quick inspection without performing calculations.
- **Applicable to Categorical Data:** Unlike the mean and median, the mode can be used with non-numeric (categorical) data (e.g., finding the most popular shoe color or brand).
- **Unaffected by Outliers:** Extreme high or low values do not distort the mode.
- **Limitation:** It is not based on all values in the dataset and can be unstable if the sample size is small.

MEASURES OF DISPERSION

In statistics, the **range** is the simplest and most basic measure of dispersion. It measures the total spread of a dataset by calculating the difference between the highest and lowest values.

Formulas FOR the Range

1. Ungrouped Data (Raw Data)

For a raw dataset, the range is the difference between the maximum value and the minimum value.

Range=Xmax−Xmin

- *Example:* For the dataset {4,8,12,15,21}
    - Maximum value = 21
    - Minimum value = 4
    - Range=21−4=17

2. Grouped Data (Frequency Distribution)

For continuous data organized into class intervals, the range is defined as the difference between the **upper-class boundary of the highest interval** and the **lower-class boundary of the lowest interval**.

Range=Umax−Lmin

- **Where:**
  - Umax = Upper-class boundary of the highest class interval
  - Lmin = Lower-class boundary of the lowest class interval

Key Characteristics & Limitations

- **Simplicity:** It is extremely easy to calculate and quick to understand at a glance.

- **Extreme Sensitivity to Outliers:** Because the range depends *only* on the two extreme values (the highest and lowest numbers), a single outlier can drastically distort the result, making it an unreliable measure of overall spread on its own.
- **Ignores Distribution:** It provides no insight into how the data points are distributed between the maximum and minimum values (e.g., two datasets can have the exact same range but completely different internal variations).

In statistics, the **mean deviation** (also known as the **mean absolute deviation**) is a measure of dispersion that tells you the average distance between each data point and a central value (such as the mean or median).

Unlike variance or standard deviation, which square the differences, mean deviation uses the **absolute values** of the differences. This ensures that positive and negative deviations do not cancel each other out.

Formulas for Mean Deviation

The formula depends on whether you are measuring the deviation from the **mean** or the **median**, and whether the data is ungrouped (raw) or grouped.

1. Ungrouped Data (Raw Data)

For a dataset of n observations (x_1, x_2, , x_n):

- **About the Mean** {Mean Deviation} = {sum (x_i - bar{x})/{n}
- **About the Median (M):**

{Mean Deviation} = {sum {x_i - M}/{n}

2. Grouped Data (Frequency Distribution)

For data organized into frequencies (f_i) or class intervals with midpoints (x_i):

- **About the Mean (bar{x}):**

  {Mean Deviation} = {sum f_i .x_i - bar{x}/{sum f_i}

- **About the Median**
- {Mean Deviation} = {sum f_i .x_i – M/{sum f_i}

Key Characteristics & Advantages

- **Easy to Interpret:** Because it represents a straightforward average distance from the center, it is often more intuitive to understand than standard deviation.
- **Less Affected by Extreme Outliers:** Squaring deviations (as done in variance and standard deviation) heavily penalizes large outliers. Mean deviation uses absolute values, making it slightly more robust against extreme values.
- **Choice of Center:** While it is most commonly calculated around the arithmetic mean, calculating it around the **median** actually yields the minimum possible mean deviation mathematically.

In statistics, the **quartile deviation** (also known as the **semi-interquartile range**) is a measure of dispersion that represents half of the interquartile range. It measures the spread of the middle 50% of a dataset.

Because it focuses strictly on the middle portion of the data, any extreme values (outliers) at the very top or bottom of the dataset are completely ignored.

Formula to Calculate Quartile Deviation

For any dataset, the quartile deviation is calculated using the first quartile (Q_1) and the third quartile (Q_3):

{Quartile Deviation (QD)} = {Q_3 - Q_1}/{2}

Where:

- Q_1= First quartile (the 25th percentile, or the median of the lower half of the data)
- Q_3 = Third quartile (the 75th percentile, or the median of the upper half of the data)
- (Q_3 - Q_1) = Interquartile Range (IQR)

Coefficient of Quartile Deviation

To compare the dispersion of two or more datasets that have different units or drastically different scales, statisticians use the **relative measure** known as the coefficient of quartile deviation:

{Coefficient of QD} = {Q_3 - Q_1}/{Q_3 + Q_1}

Key Characteristics & Advantages

- **Immune to Outliers:** Because it completely disregards the top 25% and bottom 25% of the data, extreme outliers will not distort the calculation.
- **Ideal for Skewed Data:** It is the preferred measure of dispersion when dealing with heavily skewed distributions or open-ended frequency tables (where the exact lowest or highest values are unknown).
- **Limitation:** Like the range and median, it ignores 50% of the dataset entirely (the outer tails), meaning it does not reflect the behavior of every individual data point.

VARIANCE AND STANDARD DEVIATION

1. Definitions

- **Variance (σ^2 or s^2):** Variance is a measure of dispersion that calculates how far each value in a data set is spread out from its mean (average) value. It is computed by taking the average of the squared differences from the mean.

- **Standard Deviation (σ or s):** Standard deviation is the positive square root of the variance. It measures the absolute variability or spread of a distribution in the **same units** as the original data set, making it easier to interpret than variance.

  Testbook

2. Formulas

A. For Ungrouped Data (Raw Data)

- **Population Variance (σ^2):**

  σ^2=∑i=1 TO N,(Xi−μ)2/N

(Where Xi = individual data points, μ = population mean, N = total number of observations)

- **Sample Variance (s^2):**

  S^2=∑i=1 TO n ,(xi−x¯)2/ (N -1)

  (Where xi = sample data points, x¯ = sample mean, n = sample size)

- **Standard Deviation (σ or s):**

  Standard Deviation=Variance

  σ=∑(Xi−μ)2/N OR

  s=∑(xi−x¯)2/(N-1)

B. For Grouped Data (Frequency Distribution)

When data is given with frequencies (fi), the formulas change to:

- **Variance (σ2):**

  σ2=∑fi(xi−x¯)2/∑fi

- **Standard Deviation (σ):**

  σ= √(∑fi(xi−x¯)2 /∑fi) or use the shortcut formula:σ= √(∑fixi^2/∑fi −∑fixi)^2/(∑fi))

3. Properties of Variance and Standard Deviation

1. **Non-Negative:** Both variance and standard deviation can never be negative (≥0).

2. **Unit Dependency:** Variance is expressed in **squared units** of the original data (e.g., cm2), whereas standard deviation is expressed in the **same units** (e.g., cm).

3. **Effect of Change of Origin:** Standard deviation is independent of change of origin (adding or subtracting a constant from data points does not change the standard deviation).
4. **Effect of Change of Scale:** If each observation is multiplied by a constant c, the new standard deviation becomes |c| times the original standard deviation.

4. Solved Example (Exam Style)

Problem: Find the variance and standard deviation for the following discrete data set:

2,4,5,6,8

Step 1: Find the Mean (x^-)

$x^-=n\sum xi=52+4+5+6+8=525=5$

Step 2: Create a table for deviations

xi	**(xi−x⁻)**	**(xi−x⁻)2**
2	2−5=−3	(−3)2=9
4	4−5=−1	(−1)2=1
5	5−5=0	02=0
6	6−5=1	12=1
8	8−5=3	32=9
Total	∑(xi−x⁻)=0	∑(xi−x⁻)2=20

Step 3: Calculate Variance (s2)

S^2=n−1∑(xi−x⁻)2(Using sample variance formula)
S^2=5−120=420=5

(Note: If treating it as a population variance, divide by N instead of N−1: σ2=520=4)

Step 4: Calculate Standard Deviation (s)

s=Variance=5≈2.236

UNIT 2

1. Definition of Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It tells us whether a change in one variable is accompanied by a change in another variable, indicating the direction and strength of the relationship.

1. Definition of Scatter Diagram

A **Scatter Diagram** is a two-dimensional graphical representation of bivariate data (data involving two variables).

- The independent variable (cause) is plotted on the **X-axis**.
- The dependent variable (effect) is plotted on the **Y-axis**.
- Each pair of observations (X, Y) is marked as a dot on the coordinate plane. The resulting cluster of dots visually reveals whether a relationship exists between the two variables.

2. Interpretation of Patterns in Scatter Diagrams

By observing how the dots are distributed on the graph, we can infer the type and strength of correlation:

Pattern / Observation	**Type of Correlation**	**Numerical Equivalent (r)**
Points slope closely upwards from bottom-left to top-right	High Positive Correlation	r is close to +1
Points slope upwards loosely from bottom-left to top-right	Low Positive Correlation	r is positive, closer to 0
Points slope closely downwards from top-left to bottom-right	High Negative Correlation	r is close to -1

Pattern / Observation	Type of Correlation	Numerical Equivalent (r)
Points slope downwards loosely from top-left to bottom-right	Low Negative Correlation	r is negative, closer to 0
Points are scattered randomly with no specific direction	Zero Correlation (No Correlation)	r = 0
Points cluster along a curve rather than a straight line	Non-Linear (Curvilinear) Correlation	Cannot be measured by standard linear r

3. Merits (Advantages) of Scatter Diagrams

1. **Simplicity:** It is the easiest and most user-friendly method to understand data correlation without complex mathematical calculations.
2. **Visual Inspection:** Provides an immediate, intuitive view of the relationship at a glance.
3. **Outlier Detection:** Abnormal values (outliers) stick out immediately on the graph, helping researchers spot data entry errors.
4. **Non-Linearity:** It can easily reveal non-linear patterns that standard mathematical formulas might miss.

4. Demerits (Disadvantages) of Scatter Diagrams

1. **Non-Mathematical:** It is a qualitative or visual tool; it **does not give an exact numerical value** or coefficient representing the strength of correlation (unlike Karl Pearson's method).
2. **Subjectivity:** Interpretation can sometimes be subjective. Two different people might view the same scattered plot and interpret the degree of correlation slightly differently.

2. Types of Correlation

- **Based on Direction:**
  - **Positive (Direct) Correlation:** When both variables move in the *same* direction. (e.g., Higher study hours lead to higher grades).
  - **Negative (Inverse) Correlation:** When variables move in *opposite* directions. (e.g., Higher speed of a vehicle results in less travel time).
  - **Zero Correlation:** No relationship exists between the variables. (e.g., Shoe size and intelligence level).
- **Based on Proportion/Linearity:**
  - **Linear Correlation:** When the amount of change in one variable tends to bear a constant ratio to the amount of change in the other (forms a straight line on a graph).
  - **Non-Linear (Curvilinear) Correlation:** When the rate of change is not constant.

3. Degree of Correlation (Karl Pearson's Coefficient, r)

The coefficient of correlation is denoted by **r** and always lies between **−1 and +1** (

$-1 \leq r \leq +1$

).

Value of r	Degree of Relationship	Direction
+1	Perfect Positive	Complete direct relationship
+0.75 to +0.99	High Degree Positive	Strong positive association
+0.25 to +0.74	Moderate Degree Positive	Moderate association
0 to +0.24	Low / Absent	Weak or no association
0	Zero Correlation	No linear relationship
−0.25 to −0.74	Moderate Degree Negative	Moderate inverse association
−0.75 to −0.99	High Degree Negative	Strong inverse association
−1	Perfect Negative	Complete inverse relationship

4. Methods of Measuring Correlation

A. Karl Pearson's Coefficient of Correlation (r)

This is the most widely used mathematical method, also known as the product-moment correlation coefficient.

- **Formula:**

  (Where N = number of pairs of observations, X = values of first variable, Y = values of second variable)

- **Alternative Form (using actual means $\bar{X}$ and $\bar{Y}$):**

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

B. Spearman's Rank Correlation Coefficient (R or ρ)

Used when variables cannot be measured quantitatively (like beauty, honesty, ranking in a competition), but can be arranged in order/ranks.

- **Formula (when ranks are not tied):**

  $R=1-n(n2-1)6\sum d2$

  (Where $d=R1-R2$ is the difference between the ranks of two variables, and n is the number of observations).

- **Formula (when ranks are tied/repeated):** When the same rank is given to two or more items, a correction factor is added to $\sum d2$:

  $R=1-n(n2-1)6[\sum d2+\sum 12m3-m]$

  (Where m is the number of times an item is repeated).

5. Properties of Correlation Coefficient (r)

1. **Unit-free:** r is independent of units of measurement. If X is in kilograms and Y is in meters, r has no unit.
2. **Symmetric:** The correlation between X and Y (rxy) is identical to the correlation between Y and X (ryx).
3. **Origin and Scale Independent:** r is unaffected by the change of origin (shifting/subtracting a constant) and change of scale (multiplying by a constant).
4. **Bounds:** The value of r always falls strictly within the range of −1 and +1.

1. Definition of Regression

The term "regression" literally means **"stepping back towards the average"** (first coined by Sir Francis Galton).

- In statistics, **Regression Analysis** is a mathematical tool used to estimate or predict the value of a dependent variable (target/effect) based on the known value of an independent variable (predictor/cause).

2. Types of Regression Analysis

1. **Simple Linear Regression:** Involves only two variables—one independent (X) and one dependent (Y)—and the relationship can be represented by a straight line.
2. **Multiple Regression:** Involves one dependent variable and *two or more* independent variables (e.g., predicting a student's marks based on study hours, attendance, and previous grades).
3. **Non-Linear (Curvilinear) Regression:** When the relationship between variables does not form a straight line.

3. Regression Lines

Because regression is used for prediction, we have two lines of regression depending on what we want to predict:

1. **Line of Y on X:** Used to estimate the value of dependent variable Y for a given value of X.
2. **Line of X on Y:** Used to estimate the value of independent variable X for a given value of Y.

Note: The two regression lines do not coincide (they intersect at the mean values $\bar{X}, \bar{Y}$) unless the correlation coefficient $r = \pm 1$.

4. Regression Equations and Coefficients

A. Regression Equation of Y on X

$$Y - \bar{Y} = byx(X - \bar{X})$$

- Where byx is the **regression coefficient of Y on X**.
- Formula for byx:

  $$byx = r\sigma x\sigma y = N\sum X2 - (\sum X)2N\sum XY - (\sum X)(\sum Y)$$

B. Regression Equation of X on Y

$$X - \bar{X} = bxy(Y - \bar{Y})$$

- Where bxy is the **regression coefficient of X on Y**.

- Formula for bxy:

  $$b_{xy} = r\frac{\sigma_y}{\sigma_x} = \frac{N\sum Y^2 - (\sum Y)^2}{N\sum XY - (\sum X)(\sum Y)}$$

5. Important Properties of Regression Coefficients

1. **Relationship with Correlation (r):** The correlation coefficient is the geometric mean of the two regression coefficients:

   $$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

2. **Sign Agreement:** Both regression coefficients (b_{yx} and b_{xy}) will always have the same algebraic sign. They are either both positive or both negative (matching the sign of r).
3. **Value Limits:** If one regression coefficient is greater than unity (>1), the other must be less than unity (<1). Their product cannot exceed 1 if $r \leq 1$.
4. **Origin Independence:** Regression coefficients are independent of the change of origin (shifting data by a constant doesn't affect them), but they **depend** on the change of scale.

6. Quick Comparison: Correlation vs. Regression (Frequent Exam Question)

Feature	Correlation	Regression
Meaning	Measures the degree of co-variation between variables.	Estimates the values of a variable based on another.

UNIT-3

1. Experiment

An **experiment** is an action or process that produces one of several possible outcomes, where the exact outcome is uncertain.

Examples:

- Tossing a coin
- Rolling a die
- Drawing a card from a deck
- Picking a ball from a bag

2. Sample Space

The **sample space** is the set of all possible outcomes of an experiment. It is usually denoted by **S**.

Examples:

- **Experiment:** Tossing a coin
  Sample Space: S={H,T}
- **Experiment:** Rolling a six-sided die
  Sample Space: S={1,2,3,4,5,6}
- **Experiment:** Tossing two coins
  Sample Space: S={HH,HT,TH,TT}

In simple words:

- **Experiment** = The action you perform.
- **Sample Space** = All the possible results of that action.
- **Event and Operations with Events**
- In probability, an **event** is a collection of one or more outcomes from the sample space.
- **1. Event**
- An **event** is a subset of the sample space. It is usually denoted by **A, B, C**, etc.
- **Example:**
  When a die is rolled,
- S={1,2,3,4,5,6}
- Let event A be getting an even number:
- A={2,4,6}
- So, A⊆S.
- **2. Operations with Events**
- The main operations are:

- **a) Union (A∪B)**
  Means **A or B or both** occur.
- Example:
- A={1,2,3},B={3,4,5} A∪B={1,2,3,4,5}
- **b) Intersection (A∩B)**
  Means **both A and B** occur.
- A∩B={3}
- **c) Complement (A′ or Ac)**
  Means **A does not occur**.
- If
- S={1,2,3,4,5,6},A={2,4,6}
- then
- A′={1,3,5}
- **d) Difference (A−B)**
  Means outcomes that are in **A but not in B**.
- A−B={1,2}
- **Quick Summary**

Operation	Symbol	Meaning
Union	A∪B	A or B
Intersection	A∩B	A and B
Complement	A′	Not A
Difference	A−B	A but not B

- **Remember:**
  Sample space → all possible outcomes
  Event → selected outcomes from the sample space

The **probability of an event** is the measure of how likely the event is to occur.

It is defined as:

Probability of an event (E)=Total number of equally likely outcomes
/Number of favorable outcomes

where:

- $0 \leq P(E) \leq 1$
- P(E)=0: the event is impossible.
- P(E)=1: the event is certain.

Example

If a fair six-sided die is rolled, the probability of getting a **4** is:

P(getting 4)=61

Since the probability formula is the central concept here, here's an interactive visualization:

Rolls for each result
0
1
0
2
0
3
0
4
0
5
0
6
No rolls recorded yet.

Another example:
If a bag contains 3 red balls and 2 blue balls, the probability of drawing a red ball is:

P(Red)=53

So, in general,

P(E)=n(S)n(E)

where:

- n(E) = number of favorable outcomes
- n(S) = total number of outcomes in the sample space.

Basic Probability Rules

Here are the most important rules you should know:

1. Probability is between 0 and 1

0≤P(A)≤1

- P(A)=0 → impossible event
- P(A)=1 → certain event
- 0<P(A)<1 → event may or may not occur

---

2. Probability of an event

P(A∩B)

P(A∩B)=0.60·0.35=0.21P(A\cap B)=\text{0.60}\cdot\text{0.35}=\text{0.21}P(A∩B)=0.60·0.35=0.21

P(A)P(A)P(A)

0.

P(A)P(A)P(A)

P(B|A)P(B\mid A)P(B|A)

0.

P(B|A)P(B\mid A)P(B|A)

P(A) = 0.60P(B|A) = 0.350.21

Give feedback

For equally likely outcomes:

P(A)=Total number of outcomesNumber of favorable outcomes

Example: Probability of getting a 6 on a fair die:

P(6)=61

---

3. Complement Rule

The probability that an event **does not happen** is:

P(A')=1−P(A)

Example: If the probability of rain is 0.3:

P(no rain)=1−0.3=0.7

---

4. Addition Rule — "A OR B"

P(A∪B)=P(A)+P(B)−P(A∩B)

P(A∪B)=0.55+0.45−0.20=0.80P(A\cup B)=\text{0.55}+\text{0.45}-\text{0.20}=\text{0.80}P(A∪B)=0.55+0.45−0.20=0.80

P(A)P(A)P(A)

0.

P(A)P(A)P(A)

P(B)P(B)P(B)

0.

P(B)P(B)P(B)

P(A∩B)P(A\cap B)P(A∩B)

0.

P(A∩B)P(A\cap B)P(A∩B)

AB0.20

Give feedback

For two events:

P(A∪B)=P(A)+P(B)−P(A∩B)

Use this when the question contains **OR**.

---

5. Multiplication Rule — "A AND B"

For **independent events**:

P(A∩B)=P(A)P(B)

P(A∩B)=P(A)P(B)=(0.60)(0.45)=0.27

P (A) = 0.60 P(B) = 0.45 P(A ∩ B) =0.27

P(A∩B)=P(A)×P(B)

Use this when events are independent and the question contains **AND**.

Example: Tossing a coin twice:

$P(\text{Head AND Head}) = \frac{2}{1} \times \frac{2}{1} = \frac{4}{1}$

---

6. Conditional Probability

Probability of A happening **given that B has already happened**:

$P(A|B) = P(B)P(A \cap B)$

The symbol A|B means **"A given B."**

Quick Memory Table

Rule	Formula	Key word
Basic probability	$P(A) = \text{totalfavorable}$	—
Complement	$P(A') = 1 - P(A)$	NOT
Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	OR
Multiplication	$P(A \cap B) = P(A)P(B)$	AND
Conditional	$(P(A/B) =$	$\{P(A \cap B)\}/\{P(B)\})$

Applications of Probability Rules

Probability rules are used to **measure uncertainty and make decisions** in many real-life situations.

1. Business and Finance

Businesses use probability to estimate:

- Future sales
- Investment returns

- Market risks
- Loan default probabilities

Example:
If there is a 20% probability that an investment will lose money, the investor can consider this risk before investing.

---

2. Insurance

Insurance companies use probability to estimate the likelihood of:

- Accidents
- Illness
- Property damage
- Death

They use these probabilities to calculate **insurance premiums**.

Example:
If a particular type of accident is more likely, the insurance premium may be higher.

---

3. Medicine

Probability helps doctors and researchers:

- Evaluate medical tests
- Estimate disease risk
- Compare treatments
- Analyze clinical trials

For example, conditional probability can answer:

What is the probability that a person has a disease **given that the test is positive**?

$P(A|B)=P(B)P(A\cap B)$

$P(A|B)=P(A\cap B)P(B)=0.300.65=0.46$

$P(B)=0.65P(A\cap B)=0.30P(A|B) \approx 0.46A\cap B$ is the part of B where A also happens

4. Weather Forecasting

Meteorologists use probability to predict weather.

Example:

P(rain)=0.70

means there is a **70% probability of rain** under the forecasting model.

---

5. Quality Control

Factories use probability to determine the likelihood of defective products.

Example:
If the probability of a defective item is 0.02, then approximately **2 out of every 100 items** may be defective over a large production run.

---

6. Games and Gambling

Probability is used to calculate the chances of:

- Getting a particular card
- Rolling a particular number
- Winning a game
- Drawing a particular lottery number

Example:
For a fair die:

P(getting 6)=61

---

7. Sports

Probability can help estimate:

- A team's chance of winning

- A player's chance of scoring
- The probability of a particular result

For example, if Team A has a 70% estimated chance of winning, then:

P(Team A wins)=0.70

---

8. Science and Research

Researchers use probability to:

- Analyze experimental results
- Test hypotheses
- Estimate uncertainty
- Draw conclusions from samples

---

9. Engineering and Reliability

Engineers use probability to determine whether machines and systems will work successfully.

Example:
If a component has a 0.98 probability of working correctly, its reliability is **98%**.

---

10. Everyday Decision-Making

We use probability without always realizing it:

- Should I carry an umbrella?
- Which route is less likely to have traffic?
- Should I buy an extended warranty?
- What is the chance my flight will be delayed?

Quick Summary

Area	Application of Probability
Business	Sales and market risk
Insurance	Accident and loss risk

Area	Application of Probability
Medicine	Disease and test probabilities
Weather	Rain and temperature forecasts
Manufacturing	Defective products
Games	Winning chances
Sports	Winning/scoring probabilities
Science	Experimental uncertainty
Engineering	System reliability
Daily life	Decision-making under uncertainty

onditional Probability

Conditional probability is the probability of an event happening **when we already know that another event has happened**.

It is written as:

$P(A|B)=P(A\cap B)\ P(B)$

$P(A|B)=P(A\cap B)P(B)=0.300.65=0.46$

$P(B)=0.65$ $P(A\cap B)=0.30$ $P(A|B) \approx 0.46A\cap B$ is the part of B where A also happens

$P(A|B)=P(B)P(A\cap B)$

Here:

- $P(A|B)$ = probability of **A given B**
- $P(A\cap B)$ = probability that **A and B both occur**
- $P(B)$ = probability that **B occurs**
- $P(B)>0$

Simple Example

Suppose a class has **10 students**:

- 6 are boys
- 4 are girls
- 3 of the boys wear glasses
- 1 of the girls wears glasses

If we know that a randomly selected student is a **boy**, what is the probability that the student wears glasses?

We are looking for:

$$P(\text{Glasses}|\text{Boy})$$

There are **6 boys**, and **3 of them wear glasses**.

Therefore:

$$P(\text{Glasses}|\text{Boy})=63=21=0.5$$

So the answer is:

50%

Important Idea

Conditional probability **changes the sample space**.

Instead of considering all 10 students, we consider **only the 6 boys**, because we already know the student is a boy.

Another Useful Formula

The multiplication rule can be obtained from conditional probability:

$$P(A\cap B)=P(B)\times P(A|B)$$

Key Words to Look For

In questions, phrases such as:

- **"given that"**
- **"if we know that"**
- **"assuming that"**
- **"among those who..."**

usually indicate **conditional probability**.

Discrete and Continuous Random Variables

A **random variable** is a variable whose value is determined by the outcome of a random experiment.

There are two main types:

1. Discrete Random Variable

A **discrete random variable** can take **countable, separate values**.

Think: **you count it**.

Examples

- Number of heads when tossing 3 coins: 0,1,2,3
- Number of students in a classroom
- Number of defective products
- Number of goals scored in a match
- Number shown when rolling a die: 1,2,3,4,5,6

For a discrete random variable X, we can assign a probability to each individual value:

$P(X=x)$

The probabilities must satisfy:

$\sum P(X=x)=1$

Example

Suppose X is the number obtained when a fair die is rolled.

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

---

2. Continuous Random Variable

A **continuous random variable** can take **any value within an interval**.

Think: **you measure it**.

Examples

- Height of a person
- Weight of a person
- Temperature
- Time taken to complete a task
- Distance traveled
- Blood pressure

For example, a person's height could be:

170.1,170.15,170.153,170.1537,...

There are infinitely many possible values.

For a continuous random variable:

$P(X=x)=0$

for any exact single value. We instead calculate probabilities over **intervals**, such as:

$P(170<X<180)$

Key Difference

Feature	Discrete	Continuous
Meaning	Countable values	Measurable values
Values	Separate/countable	Any value in an interval
Example	Number of children	Height
Individual probability	Can be >0	P(X=x)=0
Probability	Sum of probabilities	Area under a density curve
Typical data	0, 1, 2, 3...	2.5, 2.51, 2.513...

Easy Trick to Remember

DISCRETE = COUNT
CONTINUOUS = MEASURE

Example:

- Number of cars in a parking lot → **Discrete**
- Weight of a car → **Continuous**
- Number of calls received → **Discrete**
- Duration of a phone call → **Continuous**

Probability Distribution of a Random Variable

A **probability distribution** tells us **all the possible values of a random variable and the probability associated with each value**.

In simple words:

Probability distribution = possible values + their probabilities

---

1. Discrete Probability Distribution

For a **discrete random variable**, we use a **probability mass function (PMF)**:

$P(X=x)$

The following conditions must be satisfied:

$P(X=x)\geq 0$

and

$\sum P(X=x)=1$

Example: Tossing Two Coins

Let X = number of heads obtained.

Possible values are:

$X=0,1,2$

?

Bars: cumulative observed results

0%

50%

Theoretical 50%

100%

Heads

Tails

Flip to compare observed results with the fair-coin benchmark

Number of Heads X	**Probability P(X)**
0	1/4
1	2/4
2	1/4

Check:

41+42+41=1

Therefore, this is a valid probability distribution.

---

2. Continuous Probability Distribution

For a **continuous random variable**, probabilities are described using a **probability density function (PDF)**.

Examples of continuous random variables:

- Height
- Weight
- Temperature
- Time
- Distance

For a continuous variable, the probability of an exact value is:

P(X=x)=0

Instead, we find probabilities over an interval:

$P(a<X<b)$

The probability corresponds to the **area under the probability density curve** between a and b.

---

3. Main Properties

A probability distribution must satisfy:

Property 1: Probability cannot be negative

$P(X=x) \geq 0$

Property 2: Total probability equals 1

For a discrete variable:

$\sum P(X=x)=1$

For a continuous variable:

$\int_{-\infty}^{\infty} f(x)dx=1$

---

4. Example of a Probability Distribution

Suppose X represents the number of customers arriving at a shop in one hour.

X	P(X=x)
0	0.10
1	0.20
2	0.30
3	0.25
4	0.15
Total	**1.00**

Because:

$0.10+0.20+0.30+0.25+0.15=1$

it is a valid probability distribution.

---

5. Expected Value from a Probability Distribution

A probability distribution can also be used to calculate the **expected value (mean)**:

$E(X)=\sum xP(X=x)$

For the example above:

$E(X)=(0)(0.10)+(1)(0.20)+(2)(0.30)+(3)(0.25)+(4)(0.15) =0+0.20+0.60+0.75+0.60$ $E(X)=2.15$

So, the **expected number of customers per hour is 2.15**.

Quick Summary

Type	Distribution	Example
Discrete	PMF $P(X=x)$	Number of customers
Continuous	PDF $f(x)$	Customer waiting time
Total probability	Must equal 1	$\sum P(X=x)=1$
Mean	Expected value	$E(X)=\sum xP(X=x)$

Probability Mass Function (PMF)

A **Probability Mass Function (PMF)** gives the probability that a **discrete random variable** takes a particular value.

It is written as:

P(X=x)

where:

- X = discrete random variable
- x = a particular value of X
- P(X=x) = probability that X equals x

Main Conditions of a PMF

E[X]=∑xx P(X=x)

E[X]=2(0.25)+8(0.75)=6.5

Chance of outcome 8, P(X=8)

28P=0.25P=0.75E[X]

A valid PMF must satisfy:

P(X=x)≥0

for every possible value, and

x∑P(X=x)=1

That means all probabilities must be **non-negative** and their total must be **1**.

Example

Suppose X is the number of heads when **two coins** are tossed.

The possible values are:

X=0,1,2

The PMF is:

x P(X=x)

0 1/4

x P(X=x)

1 2/4

2 1/4

Check:

41+42+41=1

Therefore, this is a valid PMF.

Finding a Probability

Suppose we want the probability of getting **at least one head**:

P(X≥1)

We add the probabilities for X=1 and X=2:

P(X≥1)=42+41

P(X≥1)=43

PMF vs PDF

PMF	PDF
Used for **discrete** variables	Used for **continuous** variables
Gives P(X=x)	Gives probability density
Individual values can have positive probability	P(X=x)=0
Example: number of customers	Example: customer waiting time

Definition

PMF is a function that assigns a probability to each possible value of a discrete random variable.

Remember:
Discrete → PMF → P(X=x)

Probability Density Function (PDF)

A **Probability Density Function (PDF)** is a function used to describe the probability distribution of a **continuous random variable**.

It is usually written as:

$f(x)$

Definition

For a continuous random variable X, the probability that X lies between a and b is the **area under the PDF curve** between those values.

$E[X]=\sum xxP(X=x)$

$E[X]=2(0.25)+8(0.75)=6.5$

Chance of outcome 8, $P(X=8)= 0$

$28P=0.25 \quad P=0.75E[X]$

More generally,

$P(a\leq X\leq b)=\int abf(x)dx$

Conditions of a PDF

A valid probability density function must satisfy:

1. It cannot be negative:

$f(x)\geq 0$

2. Total area under the curve must equal 1:

$\int -\infty\infty f(x)dx=1$

Important Point

For a continuous random variable:

$P(X=x)=0$

This does **not** mean the value x is impossible. It means that probability is assigned to **intervals**, not individual points.

For example:

$P(10<X<20)$

can be greater than zero, while

$P(X=15)=0$.

Example

Suppose the PDF is:

$f(x)=101, 0\le x\le 10$

and $f(x)=0$ otherwise.

The probability that X is between 2 and 5 is:

$P(2<X<5)= 2 \text{ TO } 5 \int 101 dx = 105-2=0.3$

So there is a **30% probability** that X lies between 2 and 5.

PMF vs PDF

PMF	PDF
Discrete random variable	**Continuous** random variable
P(X=x) can be positive	P(X=x)=0
Uses individual probabilities	Uses area under a curve
Probabilities are summed	Probabilities are integrated
Example: number of calls	Example: waiting time

Easy Memory Trick

PMF → Discrete → Count

PDF → Continuous → Measure

Expectation and Variance of a Random Variable

Two important characteristics of a random variable are **expectation** and **variance**.

1. Expectation

The **expectation** or **expected value** of a random variable is its **long-run average value**.

It is denoted by:

$E(X)$ or μ

For a **discrete random variable**:

$E(X)=x\sum xP(X=x)$

In simple words:

Multiply each possible value by its probability, then add them all.

Example

Suppose:

X P(X=x)

1 0.2

2 0.5

3 0.3

Then:

$E(X)=(1)(0.2)+(2)(0.5)+(3)(0.3) =0.2+1.0+0.9$ $E(X)=2.1$

So the expected value is **2.1**.

Original amounts

Equal shares

2

4

7

x¯=2+4+73=133=4 13\bar{x} = {2 + 4 + 7}/{3} ={13}/{3} = 4 x̄ is the arithmetic mean (simple average)

x1x_1x1

2

x1x_1x1

x2x_2x2

4

x2x_2x2

x3x_3x3

7

x3x_3x3

Give feedback

---

2. Variance

Variance measures how much the values of a random variable **spread out around the expected value**.

It is denoted by:

Var(X)

For a discrete random variable:

Var(X)=E[(X−μ)2]

where:

μ=E(X)

An easier formula is:

Var(X)=E(X2)−[E(X)]^2

Var(X)=E(X2)−[E(X)]^2

Var(X)=(1.4)2=1.96

σ

μ-σ+σVar(X) ≈ 1.96

Give feedback

Example of Variance

Using the previous distribution:

X P(X=x)

1 0.2

2 0.5

3 0.3

We already found:

E(X)=2.1

First calculate E(X2):

E(X2)=12(0.2)+22(0.5)+32(0.3) =0.2+2.0+2.7=4.9

Therefore:

Var(X)=E(X2)−[E(X)^s]2 =4.9−(2.1)2 =4.9−4.41 Var(X)=0.49

---

3. Standard Deviation

The **standard deviation** is the square root of variance:

σ=Var(X)

For our example:

σ=0.49 σ=0.7

σ=N1∑i=1N(xi−μ)2

The points have mean μ = 5 and σ = 1.5.

Difference Between Expectation and Variance

Concept	Meaning	Formula
Expectation	Average/central value	E(X)=∑xP(X=x)
Variance	Spread around the mean	Var(X)=E(X2)−[E(X)]^2
Standard deviation	Spread in original units	σ=Var(X)

Easy Memory Trick

Expectation → Where is the center?

Variance → How spread out are the values?

Standard deviation → Square root of the spread.

A **binomial probability distribution** is a discrete probability distribution that gives the probability of obtaining exactly **x successes** in **n independent trials**, where each trial has only **two possible outcomes** (success or failure) and the probability of success remains constant.

6 independent yes/no trialsppp = 50%
12345620406080Success count kProbability (%)3 successes, 31.3%
nnn
6.
nnn
ppp
5(
%
ppp
Give feedback

Conditions for a Binomial Distribution

A random experiment follows a binomial distribution if:

1. There is a fixed number of trials (n).
2. Each trial has only two outcomes: success or failure.
3. The probability of success (p) is the same for every trial.
4. The trials are independent of each other.

Binomial Probability Formula

The probability of getting exactly x successes is:

$P(X=x) = {}^{n}c_{x}\ \ p\text{^}x\ .(1-p)\text{^}n-x$

- n = total number of trials
- x = number of successes
- p = probability of success
- $1-p=q$ = probability of failure
- ${}^{n}c_{x}$ = n**!** / x!(n−x)!

Mean and Variance

- **Mean (Expected Value):**

  $\mu=np$

- **Variance:**

  $\sigma 2=npq=np(1-p)$

- **Standard Deviation:**

  $\sigma=\sqrt{npq}$

Example

A fair coin is tossed **5 times**. Find the probability of getting exactly **3 heads**.

Here,

- n=5
- x=3
- p=21

P(X=3)=(35)(21)3(21)2 =3!2!5!×321=10×321=3210=165=0.3125

Answer: The probability of getting exactly **3 heads** is **0.3125** (or **31.25%**).

Applications

Binomial distribution is commonly used in:

- Quality control (defective/non-defective items)
- Medical studies (success/failure of treatments)
- Survey analysis (yes/no responses)
- Coin tossing and other probability experiments with two outcomes.

Poisson Probability Distribution

The **Poisson probability distribution** is a discrete probability distribution used to find the probability of a certain number of events occurring in a **fixed interval of time or space**, when these events occur independently at a known average rate.

Formula

E[X]=∑xxP(X=x)

E[X]=2(0.25)+8(0.75)=6.5\mathbb{E}[X]=2(\text{0.25})+8(\text{0.75})=\text{6.5}E[X]=2(0.25)+8(0.75)=6.5

Chance of outcome 8, P(X=8)

0.

Chance of outcome 8, P(X=8)

28P=0.25P=0.75E[X]

Give feedback

The Poisson probability formula is:

P(X=x)=x!e−λλx

where:

- X = number of events
- x = specific number of events
- λ = average number of events in the given interval
- e = approximately 2.71828
- x! = factorial of x

Conditions for Poisson Distribution

A situation generally follows a Poisson distribution when:

1. Events occur independently.
2. Events occur at a constant average rate.
3. Two events cannot occur at exactly the same instant.
4. We count events over a fixed interval of **time, distance, area, volume, etc.**

Mean and Variance

For a Poisson distribution:

Mean=λ Variance=λ Standard Deviation= √λ

Example

Suppose a call center receives an average of **4 calls per hour**. What is the probability of receiving exactly **2 calls in one hour**?

Here:

λ=4,x=2 P(X=2)=2!e−442 =20.018315×16 P(X=2)≈0.1465

So, the probability of receiving exactly **2 calls is approximately 14.65%**.

Binomial vs Poisson

Binomial	Poisson
Fixed number of trials	Fixed time/space interval
Two outcomes per trial	Number of events is counted
Parameters: n,p	Parameter: λ
Mean: np	Mean: λ
Variance: np(1−p)	Variance: λ

In simple terms:

- **Binomial:** "Out of 10 attempts, how many successes?"
- **Poisson:** "In 1 hour, how many events will occur?"

Normal Probability Distribution

The **normal probability distribution** is a continuous probability distribution that is **symmetric and bell-shaped**. It is one of the most important distributions in statistics and is widely used to model measurements such as height, weight, test scores, and measurement errors.

Main Features

- It is **bell-shaped** and symmetric about the mean.
- **Mean = Median = Mode**.
- The total area under the curve is **1** (or 100%).
- The curve extends indefinitely in both directions.
- It is determined by two parameters:
    - μ = mean
    - σ = standard deviation

Probability Density Function

$$f(x)=\sigma 2\pi 1e-21(\sigma x-\mu)2$$

where:

- x = observed value
- μ = population mean
- σ = population standard deviation

Standard Normal Distribution

When:

$$\mu=0,\sigma=1$$

the distribution is called the **standard normal distribution**.

The standardized value is called the **Z-score**:

$$Z=\sigma X-\mu$$

The Z-score tells us how many standard deviations an observation is above or below the mean.

Empirical Rule: 68–95–99.7 Rule

For a normal distribution:

Range Approximate probability

μ±1σ **68%**

μ±2σ **95%**

μ±3σ **99.7%**

E[X]=∑xxP(X=x)

E[X]=2(0.25)+8(0.75)=6.5\mathbb{E}[X]=2(\text{0.25})+8(\text{0.75})=\text{6.5}E[X]=2(0.25)+8(0.75)=6.5

Chance of outcome 8, P(X=8)

0.

Chance of outcome 8, P(X=8)

28P=0.25P=0.75E[X]

Give feedback

Example

Suppose students' test scores are normally distributed with:

μ=70,σ=10

Find the Z-score for a student who scored 85.

Z=1085−70 Z=1.5

So, a score of **85 is 1.5 standard deviations above the mean**.

Normal vs. Binomial vs. Poisson

Distribution	Type	Main use	Parameters
Binomial	Discrete	Number of successes in fixed trials	n,p
Poisson	Discrete	Number of events in an interval	λ
Normal	Continuous	Measurements/values around a mean	μ,σ

In simple terms: A normal distribution describes data that tend to **cluster around an average value**, with fewer observations occurring as we move farther away from the mean.

Unit 4

Population:
A **population** is the entire group of individuals, objects, or items that a researcher wants to study.

Example: If a researcher wants to study the heights of all students in a school, then **all the students in the school** are the population.

Sample:
A **sample** is a smaller group selected from the population to represent it in the study.

Example: If the researcher measures the heights of **100 students** chosen from the school, those 100 students are the sample.

In simple terms:

- **Population = the whole group**
- **Sample = a part of the whole group**
- **Parameter and Statistic**
- **Parameter:**
  A **parameter** is a numerical value that describes a characteristic of the **entire population**.
- **Example:** The average height of **all students in a school** is a population parameter.
- **Statistic:**
  A **statistic** is a numerical value that describes a characteristic of a **sample** taken from the population.
- **Example:** The average height of **100 selected students** is a sample statistic.
- **Simple difference**

Parameter	Statistic
Describes the **population**	Describes the **sample**

Parameter	Statistic
Usually fixed but often unknown	Calculated from sample data
Example: Population mean (μ)	Example: Sample mean ($\bar{x}$)

- **Easy way to remember:**
  Population → Parameter
  Sample → Statistic

Sampling Distribution of the Sample Mean and Sample Proportion

A **sampling distribution** is the probability distribution of a statistic obtained from **all possible samples of the same size** from a population.

1. Sampling Distribution of the Sample Mean

Suppose a population has:

- Population mean = **μ**
- Population standard deviation = **σ**
- Sample size = **n**
- Sample mean = $\mathbf{X^-}$

Then the sampling distribution of X^- has:

Mean:

$\mu_{X^-}=\mu$

Standard deviation (standard error):

$\sigma_{X^-}=n\sigma$

So, as the sample size increases, the sample mean becomes **less variable**.

If the population is normally distributed, X^- is normally distributed. Even if the population is not normal, the **Central Limit Theorem** says that X^- becomes approximately normal when n is sufficiently large.

Example:
If μ=50, σ=10, n=100,

$\mu_{X^-}=50$ $\sigma_{X^-}=10010=1$

Therefore,

$\bar{X} \sim N(50,1)$

approximately.

---

2. Sampling Distribution of the Sample Proportion

Suppose a population has:

- Population proportion = **p**
- Sample size = **n**
- Sample proportion = **p^**

The sampling distribution of p^ has:

Mean:

$\mu_{\hat{p}} = p$

Standard deviation (standard error):

$\sigma_{\hat{p}} = \sqrt{p(1-p)}$

For sufficiently large samples, p^ is approximately normally distributed when:

$np \geq 10 \text{ and } n(1-p) \geq 10$

Example:
If p=0.60 and n=100,

$\mu_{\hat{p}} = 0.60 \quad \sigma_{\hat{p}} = \sqrt{\frac{0.60(0.40)}{100}} \approx 0.049$

So the sampling distribution is approximately:

$\hat{p} \sim N(0.60, 0.049)$

Quick Comparison

Feature	Sample Mean $\bar{X}$	Sample Proportion p^
Population parameter	μ	p
Mean	μ	p

Feature	Sample Mean $\bar{X}$	Sample Proportion $\hat{p}$
Standard error	$n\sigma$	$np(1-p)$
Approx. normality	CLT / normal population	$np \geq 10$, $n(1-p) \geq 10$
Statistic	Numerical average	Fraction/percentage in a category

In one sentence: The sampling distribution tells us **how a sample statistic varies from sample to sample**, and its standard error tells us **how much it typically varies around the true population parameter**.

Estimation and Hypothesis Testing

Both **estimation** and **hypothesis testing** are major parts of **inferential statistics**. They use sample data to make conclusions about a population.

1. Estimation

Estimation means using sample data to estimate an unknown **population parameter**.

There are two main types:

A. Point Estimation

A **point estimate** gives a single value as an estimate of a population parameter.

Population parameter	Point estimate
Population mean μ	Sample mean $\bar{X}$
Population proportion p	Sample proportion $\hat{p}$
Population variance σ^2	Sample variance s^2

Example:
A sample of 100 students has an average height of 165 cm.

Then:

$\bar{X} = 165$

So **165 cm is the point estimate of the population mean μ**.

B. Interval Estimation

Instead of giving one value, we give a **range of plausible values** for the population parameter.

This is called a **confidence interval**.

For example:

$\bar{X} \pm z\alpha/2n\sigma$

A 95% confidence interval might be:

$162 \text{ cm} < \mu < 168 \text{ cm}$

This means we use the sample to construct an interval that is intended to capture the true population mean.

---

2. Hypothesis Testing

Hypothesis testing is a statistical procedure used to determine whether there is enough evidence in a sample to support a claim about a population.

There are two hypotheses:

Null Hypothesis H0

The **null hypothesis** represents the existing assumption or a statement of no effect/no difference.

Example:

$H0: \mu = 50$

Alternative Hypothesis H1 or Ha

The **alternative hypothesis** represents what we want to find evidence for.

For example:

$Ha: \mu \cdot = 50$

Steps in Hypothesis Testing

Step 1: State the hypotheses

$H_0: \mu = 50$ $H_a: \mu \cdot = 50$

Step 2: Choose the significance level

Common choices are:

$\alpha = 0.05$

or

$\alpha = 0.01$

Step 3: Calculate the test statistic

For a population mean when σ is known:

$z = \sigma / n\bar{X} - \mu_0$

Step 4: Find the p-value or critical value.

Step 5: Make a decision

- If **p-value < α → Reject H0**
- If **p-value ≥ α → Fail to reject H0**

Step 6: State the conclusion in words.

Estimation vs Hypothesis Testing

Estimation	Hypothesis Testing
Estimates a population parameter	Tests a claim about a population parameter
Gives a point or interval	Gives a decision/conclusion
Uses point estimates and confidence intervals	Uses H0, Ha, test statistic, p-value
Example: Estimate average income	Example: Test whether average income is ₹30,000
Main question: **"What is the parameter?"**	Main question: **"Is there sufficient evidence for this**

Estimation **Hypothesis Testing**

claim?"

Easy way to remember

Estimation:

"What is the population value likely to be?"

Hypothesis testing:

"Is there enough evidence to support or reject a particular claim?"

Connection: Both use **sample statistics and sampling distributions** to make conclusions about the population.

Z-test

A **Z-test** is a statistical hypothesis test used to determine whether there is a significant difference between sample data and a population parameter when the sampling distribution is approximately normal.

It is commonly used when:

- The population standard deviation (σ) is known, or
- The sample size is large (typically $n \geq 30$).

The test statistic is:

$Z = \sigma / n\bar{x} - \mu$

where:

- $\bar{x}$ = sample mean
- μ = population mean (hypothesized)
- σ = population standard deviation
- n = sample size

Steps in a Z-test

1. State the null hypothesis (H0) and alternative hypothesis (H1).
2. Choose a significance level (α), such as 0.05.
3. Calculate the Z-statistic.
4. Compare the calculated Z-value with the critical Z-value (or compute the p-value).
5. Reject H0 if the Z-value falls in the rejection region or if the p-value is less than α.

Example

Suppose:

- Population mean = 50
- Population standard deviation = 10
- Sample size = 100
- Sample mean = 52

Then:

$Z=10/10052-50=12=2$

At the 5% significance level for a two-tailed test, the critical values are ±1.96. Since 2>1.96, we reject the null hypothesis and conclude that the sample mean is significantly different from the population mean.

Z-tests can be used for:

- Testing a single population mean.
- Comparing two population means (under appropriate conditions).
- Testing population proportions.

T-test

A **t-test** is a statistical test used to determine whether there is a significant difference between means. It is especially useful when the **population standard deviation is unknown** and/or the sample size is relatively small.

Formula for one-sample t-test

$t= \bar{x}-\mu/ s/n$

Where:

- $\bar{x}$ = sample mean
- μ = hypothesized population mean
- s = sample standard deviation
- n = sample size

Types of t-test

1. **One-sample t-test** — compares a sample mean with a known/hypothesized population mean.
2. **Independent two-sample t-test** — compares the means of two independent groups.
3. **Paired t-test** — compares two measurements from the same subjects, such as before and after treatment.

Example

Suppose:

- Sample mean = 52
- Population mean = 50
- Sample SD = 8
- Sample size = 16

$t = 8/16 52 - 50 = 22 = 1$

The calculated t-value is **1**. We then compare it with the appropriate **critical t-value** using the degrees of freedom:

$df = n - 1 = 15$

Z-test vs T-test

Z-test	T-test
Population SD (σ) is known	Population SD is usually unknown
Uses Z-distribution	Uses t-distribution
Generally used for large samples	Particularly useful for small samples
Critical values are fixed for a given α	Critical values depend on degrees of freedom

In short: Use a **t-test** when you are testing means and the population standard deviation is unknown.