Descriptive Statistics : Descriptive statistics is a branch of statistics that **summarizes** and **organizes** data to provide meaningful insights. It helps in understanding **patterns, distributions, and trends** in data without making predictions.

1. **Mean (μ or x̄):** The **mean** is the **average** of all numbers in a dataset. It represents the central tendency of the data.
   - **Example:** If exam scores are {50, 60, 70, 80, 90}, the mean is: $\frac{50+60+70+80+90}{5} = 70$

2. **Median:** The **median** is the middle value of a dataset when arranged in ascending order.
   - Example: In {10, 20, 30, 40, 50}, the median is **30**.

3. **Mode:** The **mode** is the value that appears **most frequently** in a dataset. Unlike the mean and median, the mode is useful for categorical and discrete data
   - Example: In {2, 3, 3, 4, 5}, the mode is **3**.

4. **Variance : Variance** measures how far data points are from the **mean**. It shows the **spread** or **dispersion** of the dataset.

Example:

Dataset: [2, 4, 6, 8, 10]

1. Find the mean:

$$\bar{X} = \frac{2+4+6+8+10}{5} = 6$$

2. Find the squared differences:

$$(2-6)^2 = 16, \quad (4-6)^2 = 4, \quad (6-6)^2 = 0, \quad (8-6)^2 = 4, \quad (10-6)^2 = 16$$

3. Compute the variance:

$$\sigma^2 = \frac{16+4+0+4+16}{5} = \frac{40}{5} = 8$$

5. **Standard Deviation: Standard deviation (SD)** measures how **spread out** the data is relative to the **mean.** It is the **square root** of variance and gives a more interpretable measure of dispersion.

Example:

Dataset: [2, 4, 6, 8, 10]

1. Find the mean:

$$\bar{X} = \frac{2+4+6+8+10}{5} = 6$$

2. Find squared differences:

$$(2-6)^2 = 16, \quad (4-6)^2 = 4, \quad (6-6)^2 = 0, \quad (8-6)^2 = 4, \quad (10-6)^2 = 16$$

3. Compute variance:

$$\sigma^2 = \frac{16+4+0+4+16}{5} = 8$$

4. Find standard deviation:

$$\sigma = \sqrt{8} \approx 2.83$$

6. **Skewness: Skewness** measures the **asymmetry** of a dataset's distribution. It tells us whether data is **symmetrically distributed** or **leaning** to one side.

Types of Skewness:

1. **Positive Skew (Right-Skewed, Skewness > 0)**
   a. Tail on the **right** (higher values are more spread out).
   b. Mean > Median > Mode
   c. Example: **Income distribution (few very high salaries)**.
2. **Negative Skew (Left-Skewed, Skewness < 0)**
   a. Tail on the **left** (lower values are more spread out).
   b. Mode > Median > Mean
   c. Example: **Exam scores (most students score high, few fail)**.
3. **Zero Skew (Symmetrical, Skewness = 0)**
   a. Perfectly **balanced** distribution.
   b. Mean = Median = Mode
   c. Example: **Normally distributed height data**.

7. **Kurtosis: Kurtosis** measures the **tailedness** of a probability distribution. It tells us how **extreme values (outliers)** affect the shape of the distribution.

Formula for Kurtosis:

$$K = \frac{\sum (X - \bar{X})^4}{N \cdot \sigma^4}$$

8. **Percentiles & Quartiles:** Percentiles and quartiles are used to understand the **distribution** of data by dividing it into parts.

A **percentile** is the value below which a given percentage of data points fall.

Example:

Dataset: **[10, 20, 30, 40, 50, 60, 70, 80, 90, 100]**

- **50th percentile (median)** = 50
- **25th percentile (Q1)** = 30
- **90th percentile** = 90

◇ **Used in**: Exam scores, salaries, height measurements.

Quartiles divide data into 4 equal parts:

1. Q1 (25th percentile) → Lower quartile (25% of data is below).
2. Q2 (50th percentile) → Median (50% of data is below).
3. Q3 (75th percentile) → Upper quartile (75% of data is below).
4. IQR (Interquartile Range) → $IQR = Q3 - Q1 IQR = Q3 - Q1$ (Measures spread).

Example:

Dataset: **[5, 10, 15, 20, 25, 30, 35, 40, 45, 50]**

- **Q1 (25th percentile)** = 15
- **Q2 (50th percentile, median)** = 25
- **Q3 (75th percentile)** = 35
- **IQR = Q3 - Q1 = 35 - 15 = 20**

◇ **Used in**: Outlier detection, box plots, statistical summaries.

9. **Inferential Statistics :** Inferential statistics allows us to make predictions or generalizations about a population based on a sample of data. It is used when collecting data from an entire population is impractical or impossible

- **Hypothesis Testing:** Used to make statistical decisions.
    - Example: Checking if a new drug improves recovery rate (using T-test).

- **Confidence Intervals:** Range within which the true parameter likely lies.
    - Example: Estimating average salary of employees in a company.
- **p-values & Significance Levels:** Measures probability of results under null hypothesis.
    - Example: If p-value < 0.05, we reject null hypothesis.

10. A **probability distribution** is a mathematical function that describes the likelihood of different outcomes in a random experiment. It provides the probabilities of all possible values of a random variable.

Probability distributions are classified into two types:

1. **Discrete Probability Distribution** – For discrete random variables (countable outcomes).
2. **Continuous Probability Distribution** – For continuous random variables (infinite possible values within a range).

Normal Distribution (Gaussian Distribution):

- A continuous probability distribution with a symmetric, bell-shaped curve.
- Defined by **mean (μ)** and **standard deviation (σ)**.
- Most values cluster around the mean, with probabilities decreasing as values move further away.

Example: Heights of people, IQ scores, and measurement errors follow a normal distribution.

Binomial Distribution

- A **discrete** probability distribution representing the number of successes in **n** independent trials.
- Each trial has two possible outcomes: **success (p)** or **failure (1 - p)**.

Example:

Tossing a coin **n = 3** times, probability of getting **exactly 2 heads**

Poisson Distribution

- A **discrete** distribution that models the number of events occurring in a fixed interval (time or space).
- Events occur **randomly** and **independently**

Example:

- The number of customer arrivals at a bank per hour.
- The number of calls received in a call center per minute.

Exponential Distribution

- A **continuous** distribution that models the time between successive events in a Poisson process.
- Used for **waiting times** and **lifetimes of products**.

Example:

- Time between bus arrivals.
- Time before a machine fails.

Uniform Distribution

- A **continuous** or **discrete** distribution where all outcomes are equally likely.

Example:

- Rolling a fair die (discrete case).
- Generating random numbers in a given range (continuous case).

11. **Bayes' Theorem** describes how to update the probability of a hypothesis based on new evidence. It is fundamental in probability theory and statistics, especially in fields like machine learning, medical diagnosis, and spam filtering.

Formula:

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

where:

- *P(A/B)*$P(A|B)$ = Probability of event **A** given that **B** has occurred (posterior probability).
- *P(B/A)*$P(B|A)$ = Probability of event **B** given that **A** has occurred (likelihood).
- *P(A)*$P(A)$ = Prior probability of **A.**
- *P(B)*$P(B)$ = Total probability of **B** (marginal probability).

12. Conditional Probability

Conditional probability measures the probability of an event occurring given that another event has already occurred. It is written as **P(A | B)**, meaning "the probability of event A given that event B has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

13. Random Variables

A **random variable** is a function that assigns numerical values to the outcomes of a random experiment. It can be **discrete** or **continuous**.

Types of Random Variables

1. **Discrete Random Variable**
    a. Takes **countable** values (e.g., number of heads in a coin toss).
    b. Example:
        i. X = {0, 1, 2} for the number of heads in 2 coin flips.

2. **Continuous Random Variable**

- Takes **infinite** values within a range (e.g., height of a person).

- Example:
    - X = Time taken to complete a task (e.g., 2.3 sec, 2.31 sec).