PROBABILITY AND STATISTICS

LECTURE NO. 2

TOPICS: MEAN, MEDIAN, MODE, VARIANCE, STANDARD DEVIATION

DR. SAAD MALIK
dr.saad@szabist-isb.edu.pk

MEAN (DEFINITION & FORMULA)

- **Definition:** The arithmetic average of a dataset.
- **Formula:**

$$\text{Mean}(\mu) = \frac{\sum X_i}{N}$$

where X_i are data points and N is the number of observations.

MEAN (EXAMPLE & WHEN TO USE)

- **Example:**
  - Test scores: $80, 85, 90, 95, 100$
  - Mean = $(80 + 85 + 90 + 95 + 100)/5 = 90$
- **When to Use?**
  - Best for normally distributed data without outliers.

MEDIAN (DEFINITION & FORMULA)

- **Definition:** The middle value of a sorted dataset.
- **Formula:**
  - If N is odd:

$$\text{Median} = X_{\frac{N+1}{2}}$$

  - If N is even:

$$\text{Median} = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2}$$

MEDIAN (EXAMPLE & WHEN TO USE)

- Example:
  - Data: $10, 20, 30, 40, 50 \rightarrow$ Median = 30
  - Data: $10, 20, 30, 40 \rightarrow$ Median = (20+30)/2 = 25
- When to Use?
  - Best for skewed data or datasets with outliers.

MODE (DEFINITION & FORMULA)

- **Definition:** The most frequently occurring value in a dataset.
- **Formula:**
  - No standard formula, just identify the most frequent number.

MODE (EXAMPLE & WHEN TO USE)

- Example:
  - Data: $2, 3, 3, 5, 7, 8, 8, 8 \rightarrow$ Mode = 8
- When to Use?
  - Best for categorical data (e.g., survey responses).

COMPARISON OF MEAN, MEDIAN & MODE

- **Mean:** Affected by outliers, best for normal data.
- **Median:** Not affected by outliers, best for skewed data.
- **Mode:** Used for categorical data.

SO WHAT IS NORMAL OR SKEWED DATA?

❑ **Normal Data (Symmetrical Distribution)**

Definition: When data is evenly distributed around the mean, forming a bell-shaped curve (Gaussian distribution).

Characteristics: Mean ≈ Median ≈ Mode

Symmetrical around the center

Example: Heights of people, IQ scores

Graph Shape: Looks like a smooth, symmetrical bell curve.

❑ **Skewed Data (Asymmetrical Distribution)**

Definition: When data is not evenly distributed, meaning the tail on one side is longer than the other.

❑ **Types of Skewness:**

Right-Skewed (Positive Skew):Long tail on the right (higher values).Mean > Median > Mode

Example: Income distribution (most people earn lower, few earn very high).

Left-Skewed (Negative Skew):Long tail on the left (lower values).Mean < Median < Mode

Example: Test scores (most students score high, few fail).

VARIANCE (DEFINITION & FORMULA)

- **Definition:** Measures how data points deviate from the mean.
- **Formula (Population Variance):**

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

- **Formula (Sample Variance):**

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{N - 1}$$

VARIANCE (WHY DIVIDE BY N-1 IN SAMPLE??)

- **Why Divide by N?**
  - Since we know the true mean (μ), we get an exact measure of variance.
  - No need for correction.
- **Why Divide by $N-1$ Instead of N?**
  - **Sample mean ($\bar{X}$) is just an estimate** of the true mean (μ), so it tends to underestimate variability.
  - **Dividing by $N-1$ (Bessel's correction) makes the variance an unbiased estimate** of the population variance.

VARIANCE (EXAMPLE & WHEN TO USE)

- **Example:**
  - Data: $5, 7, 9$, Mean = 7
  - Variance = $[(5-7)^2 + (7-7)^2 + (9-7)^2]/3 = 2.67$
- **When to Use?**
  - Measures overall spread, important for risk analysis.

STANDARD DEVIATION (DEFINITION & FORMULA, EXAMPLE & USE)

- **Definition:** The square root of variance, showing data dispersion.
- **Formula:**

$$\sigma = \sqrt{\sigma^2}$$

- **Example:**
  - If variance = 2.67, then Standard Deviation = $\sqrt{2.67} \approx 1.63$
- **When to Use?**
  - Useful for understanding consistency in data.