

1. Explain Key Statistical Concepts

1. Mean (Arithmetic Average)

The mean is the sum of all values divided by the total number of values. It represents the central tendency of a dataset.

Formula:

$$\text{Mean} = (x_1 + x_2 + \dots + x_n) / n$$

Example: For the data [10, 20, 30], Mean = $(10 + 20 + 30) / 3 = 20$

Use: Used in calculating average scores, average sales, etc.

2. Median

The median is the middle value in a sorted dataset. It is less affected by extreme values (outliers) than the mean.

Example:

- Odd number of values: [5, 10, 15] -> Median = 10

- Even number of values: [5, 10, 15, 20] -> Median = $(10 + 15) / 2 = 12.5$

Use: Commonly used in income data to avoid skew from very high earners.

3. Mode

The mode is the number that appears most frequently in a dataset.

Example: [2, 3, 4, 3, 3, 5, 6] -> Mode = 3

Use: Useful in categorical data like most common product sold or preferred brand.

4. Variance

Variance shows how spread out the numbers in a dataset are from the mean.

Formula:

$$\text{Variance} = (1/n) * \text{SUM}((x_i - \text{mean})^2)$$

Use: Understanding variability in production, quality checks, etc.

5. Standard Deviation

The standard deviation is the square root of the variance. It tells how much each data point typically deviates from the mean.

Formula:

$$SD = \sqrt{\left(\frac{1}{n}\right) * \text{SUM}((x_i - \text{mean})^2)}$$

Use: Helps in comparing volatility in stock returns, measurement errors.

6. Probability

Probability is the likelihood or chance of an event happening. It ranges from 0 (impossible) to 1 (certain).

Formula:

$$P(E) = (\text{Number of favorable outcomes}) / (\text{Total number of outcomes})$$

Use: Used in predictions, games, risk assessment, and machine learning.

7. Distribution

A distribution describes how data is spread or arranged.

Common Types:

- Normal distribution: Bell-shaped curve
- Binomial distribution: Successes in fixed trials
- Poisson distribution: Count of events in time/space
- Uniform distribution: All outcomes equally likely

8. Hypothesis Testing

A method to make data-driven decisions. Helps in testing assumptions about a population using a sample.

Steps:

1. Null hypothesis (H_0): No effect or difference
2. Alternative hypothesis (H_1): There is an effect
3. Test statistic: z or t value
4. p-value: Probability of such results under H_0

5. Decision: Reject or fail to reject H0

9. Confidence Interval

A range of values that likely contains the population parameter.

Formula:

CI = sample_mean ± z * (s / sqrt(n))

Use: We are 95% confident the true mean lies within the interval.

10. Correlation

Measures the relationship between two variables.

Range: -1 to +1

+1 = perfect positive, -1 = perfect negative, 0 = no correlation

Use: Analyzing trends, e.g., temperature vs electricity usage.

Summary Table:

Concept	Purpose
Mean	Central value of data
Median	Middle value of data
Mode	Most frequent value
Variance	Measure of spread
Standard Deviation	Typical deviation from the mean
Probability	Likelihood of an event
Distribution	Pattern of data spread
Hypothesis Testing	Validate assumptions using samples
Confidence Interval	Estimate population parameters
Correlation	Relationship between variables