**Practice Questions**

1. Define the following measures of central tendency: mean, median, and mode. Provide an example dataset and compute each measure.

Ans:

Mean (Average): The sum of all values divided by the number of values.

Median: The middle value when data is arranged in order. If even count, average of two middle values.

Mode: The value that appears most frequently.

Example dataset: [2, 4, 4, 6, 8]

Mean = (2+4+4+6+8)/5 = 24/5 = 4.8

Median = middle value = 4

Mode = most frequent = 4

1. Explain the concepts of variance and standard deviation. How do they differ, and what do they tell us about a dataset?

Ans:

Variance: The average of squared differences from the mean.

Standard deviation (SD): The square root of variance.

Difference: Variance is in squared units; SD is in original units.

Interpretation: They show how spread out or dispersed the data is.

Small SD → data is close to the mean.

Large SD → data is more spread out.

1. What is skewness? Describe positive and negative skewness with appropriate examples.

Ans:

Skewness: A measure of asymmetry in a distribution.

Positive skew (Right skew): Tail on the right is longer; mean > median.

Example: Income distribution.

Negative skew (Left skew): Tail on the left is longer; mean < median.

Example: Age at retirement.

1. Discuss the concept of kurtosis. What does it indicate about a data distribution?

Ans:

Kurtosis: Describes the "tailedness" or peak of a distribution.

Types:

Leptokurtic: High peak, heavy tails (outliers likely).

Mesokurtic: Normal distribution (benchmark).

Platykurtic: Flat peak, light tails (fewer outliers).

Use: Helps understand whether data has extreme values (outliers).

1. Differentiate between discrete and continuous data. Provide two examples of each.

Ans:

Discrete data: Countable, whole numbers.

Example: Number of students in a class, number of cars.

Continuous data: Measurable, can take any value within a range.

Example: Height of students, temperature.

1. Describe the interquartile range (IQR).

Ans:

IQR: The difference between the 75th percentile (Q3) and 25th percentile (Q1).

Formula: IQR = Q3 – Q1

Use: Measures the spread of the middle 50% of data, reduces effect of outliers.

1. Differentiate between descriptive and inferential statistics. Provide scenarios where each would be appropriately applied.

Ans:

Descriptive statistics: Summarizes data (mean, median, charts).

Example: Reporting average test scores in a class.

Inferential statistics: Makes predictions or generalizations from a sample to a population.

Example: Predicting election results from survey samples.

1. Define the following terms:

Ans:

Population : The entire group under study.

Sample: A subset of the population.

Parameter : A numerical summary that describes the population (e.g., population mean μ).

1. What is the Central Limit Theorem (CLT)? Why is it fundamental in inferential statistics?

Ans:

CLT: States that the sampling distribution of the sample mean approaches a normal distribution as sample size increases, regardless of population distribution.

Importance:

Enables use of normal distribution for hypothesis testing.

Justifies confidence intervals and p-values.

1. Explain the concept of a confidence interval.

Ans:

Confidence interval (CI): A range of values, derived from sample data, that likely contains the population parameter.

Example: “We are 95% confident that the true mean lies between 50 and 60.”

Use: Gives both estimate and uncertainty.

1. What is a p-value? How is it used in the context of hypothesis testing?

Ans:

p-value: Probability of observing the test results, or more extreme, assuming the null hypothesis is true.

Usage:

If p < significance level (α, e.g., 0.05) → reject null hypothesis.

If p ≥ α → fail to reject null.

1. Explain the difference between simple random sampling and stratified sampling. Provide examples of when each would be used.

Ans:

Simple random sampling: Every individual has equal chance.

Example: Randomly picking 100 students from a university.

Stratified sampling: Population divided into groups (strata), and samples taken proportionally.

Example: Selecting students proportionally from each department.

1. Differentiate between discrete and continuous probability distributions. Provide two examples of each.

Ans:

Discrete probability distributions: Deals with countable outcomes.

Examples: Binomial, Poisson.

Continuous probability distributions: Deals with measurable outcomes.

Examples: Normal distribution, Exponential distribution.

1. What is a probability mass function (PMF) and how does it differ from a probability density function (PDF)?

Ans:

PMF: Gives probability for discrete random variables.

PDF: Gives probability density for continuous random variables (not exact probability at a point, but area under curve).

Difference: PMF → probabilities of exact values; PDF → probabilities over intervals.

1. Describe the 68-95-99.7 rule in the context of the normal distribution.

Ans:

In a normal distribution:

68% of data lies within 1 SD of mean.

95% within 2 SDs.

99.7% within 3 SDs.

Use: Quick way to understand spread and probabilities in normally

1. Discuss the importance of understanding probability distributions in the field of data science and analytics.

Ans:

Helps model real-world uncertainty.

Forms basis for hypothesis testing, predictive models, and risk analysis.

Choosing the right distribution improves accuracy of models (e.g., normal for heights, Poisson for event counts).