STATISTICS

Q1. What is Statistics? Explain its main types.

Statistics is the study of collecting, organizing, analyzing, and interpreting data.Descriptive Statistics: Summarizes data (mean, median, charts, tables).Iferential Statistics: Makes predictions/conclusions about a population using a sample (hypothesis testing, regression).

Q2. Define population and sample with examples.P

opulation: The complete set of individuals/items under study. Example: All students in a college.Sample: A subset taken from the population for analysis. Example: 100 randomly selected students from the college.

Q3. What is the difference between descriptive and inferential statistics?

Descriptive Statistics: Describes and summarizes data.

Inferential Statistics: Draws conclusions or predictions from data.

Descriptive = what happened, Inferential = what might happen.

Q4. Explain data types (qualitative vs quantitative, discrete vs continuous).

Qualitative (Categorical): Non-numerical, like gender or colors.

Quantitative (Numerical): Numbers, like marks or height.

Discrete: Countable (number of students).

Continuous: Measurable (weight, time).

Q5. What is a variable in statistics? Give examples.

A variable is a characteristic that can take different values.

Example: Age, income, height, grades.

Can be independent (cause) or dependent (effect) in analysis.

Q6. Define mean, median, and mode. How are they different?

Mean: Average value = (sum of values ÷ number of values).

Median: Middle value when data is arranged in order.

Mode: Value that occurs most frequently.

They differ in how they represent the “central tendency.”

Q7. How do you calculate the range of a dataset?

Range = Maximum value – Minimum value.

Example: For data [5, 8, 12, 20], range = 20 – 5 = 15.

It shows the spread of the data

Q8. What is the standard deviation, and why is it important?

Standard deviation measures how much values deviate from the mean.

A low SD means data is close to mean.

A high SD means data is spread out.

It helps understand data consistency and variability.

Q9. Explain variance and how it relates to standard deviation.

Variance measures the average squared deviation from the mean.

Formula: $\Sigma(x - \text{mean})^2 / n$.

Standard deviation is the square root of variance.

So, SD is easier to interpret since it is in original units

Q10. What is a frequency distribution? Give an example.

A frequency distribution shows how often each value occurs in a dataset.

Example: Test scores of 10 students:

5 students scored 40–50,

3 students scored 50–60,

2 students scored 60–70.

Q11. Explain the concept of normal distribution and its characteristics.

A normal distribution is a bell-shaped, symmetric curve.

Characteristics:

Mean = Median = Mode.

68% of data lies within 1 SD, 95% within 2 SD, 99.7% within 3 SD.

Used in probability and hypothesis testing

Q12. What is skewness, and how does it affect data interpretation?

Skewness measures the asymmetry of data distribution.

Positive skew: Tail on right, mean > median.

Negative skew: Tail on left, mean < median.

It affects how averages represent the data.

Q13. What is kurtosis, and what does it tell us about a dataset?

Krtosis measures the “peakedness” of data distribution.

High kurtosis: More outliers, sharper peak.

Low kurtosis: Flatter distribution.

It shows how much data is concentrated in tails.

Q14. Differentiate between probability and statistics.

Probability: Predicts likelihood of events (future-oriented).

Statistics: Analyzes collected data (past-oriented).

Probability starts with a model → predicts data.

Statistics starts with data → builds conclusions.

Q15. What is a z-score, and how is it calculated?

A z-score tells how many standard deviations a value is from the mean.

Formula: z = (x – mean) / SD.

Example: If mean = 50, SD = 10, x = 70 → z = 2.

It standardizes data for comparison.

Q16. Explain the difference between population standard deviation and sample standard deviation.

Population SD: Uses all data (divided by N).

Sample SD: Uses sample (divided by n–1, called Bessel’s correction).

Sample SD is used to estimate population SD more accurately.

Q17. What is the Central Limit Theorem, and why is it important?

The CLT states: When sample size is large, the sampling distribution of the sample mean becomes approximately normal, regardless of population distribution.

It’s important because it allows use of normal distribution in inferential statistics.

Q18. What is correlation? Differentiate between positive and negative correlation.

Correlation measures the strength and direction of a relationship between two variables.

Positive correlation: As X ↑, Y ↑ (e.g., height & weight).

Negative correlation: As X ↑, Y ↓ (e.g., speed & travel time).

Q19. Explain the difference between correlation and causation.

Correlation: Two variables move together, but one may not cause the other.

Causation: One variable directly influences the other.

Example: Ice cream sales and drowning deaths are correlated (summer) but not causal.

Q20. What is regression analysis, and when is it used?

Regression shows the relationship between dependent and independent variables.

Simple regression: One independent variable.

Multiple regression: Many independent variables.

It’s used for prediction and forecasting.

Q21. Explain hypothesis testing and its steps.

Hypothesis testing is a procedure to test assumptions about a population.

Steps:

1. State null (H0) and alternative (H1).

2. Choose significance level (α).

3. Collect data & calculate test statistic.

4. Compare with critical value or p-value.

5. Accept/reject H0.

Q22. What is a null hypothesis and an alternative hypothesis?

Null hypothesis (H0): No effect or no difference. Example: “Average height = 170 cm.”

Alternative hypothesis (H1): There is an effect or difference. Example: “Average height ≠ 170 cm.”

Q23. Explain p-value in hypothesis testing.

The p-value shows the probability of obtaining test results if H0 is true.

Small p-value (< α): Reject H0, strong evidence against it.

Large p-value (> α): Fail to reject H0.

Q24. What is the difference between Type I and Type II errors?

Type I error (α): Rejecting H0 when it is true (false positive).

Type II error (β): Failing to reject H0 when it is false (false negative).

Q25. What is a confidence interval, and how is it interpreted?

A confidence interval gives a range of values where the true population parameter lies with a certain confidence level.

Example: 95% CI = [45, 55] → We are 95% confident true mean lies between 45 and 55.

Q26. Explain t-test and when to use it.

A t-test compares means of two groups to see if they are significantly different.

One-sample t-test: Compare sample mean with population mean.

Two-sample t-test: Compare two sample means.

Used when sample size is small (n < 30).

Q27. Explain chi-square test and its applications.

Chi-square test checks the relationship between categorical variables.

Goodness of fit test: Checks if data fits expected distribution.

Independence test: Checks if two variables are related.

Example: Checking if gender and voting preference are related.

Q28. What is ANOVA, and when is it used?

ANOVA (Analysis of Variance) compares means of 3 or more groups.

It tests if at least one group mean is different.

Example: Comparing average marks of students from 3 different schools

Q29. How do you handle missing data in statistics?

Ways to handle missing data:

Remove rows with missing values.

Replace with mean/median/mode.

Use forward/backward fill.

Apply advanced methods like regression or imputation.

Q30. What is sampling bias, and how can it be reduced?

Sampling bias happens when the sample does not represent the population.

It leads to inaccurate results.

Reduce it by:

Random sampling.

Large sample size.

Avoiding selective or convenient samples.