Statistics Interview Questions for Data Analysts

1. What is a population in statistics?

- A population is the entire group we want to study or collect data from.

2. What is a sample?

- A sample is a smaller group selected from the population.

3. What is the mean?

- The mean is the average of a set of numbers, calculated by adding them up and dividing by the count.

4. What is the median?

- The median is the middle value in a list of numbers sorted in ascending or descending order.

5. What is the mode?

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

6. What is range?

- The range is the difference between the highest and lowest values in a dataset.

7. What is standard deviation?

- Standard deviation measures how spread out the numbers are in a dataset.

8. What is variance?

- Variance is the average of the squared differences from the mean.

9. What is a histogram?

- A histogram is a bar graph that represents the frequency distribution of numerical data.

10. What is a scatter plot?

- A scatter plot is a graph used to study the relationship between two variables.

INTERMEDIATE QUESTIONS

11. What is correlation?

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

12. What is the difference between positive and negative correlation?

- Positive correlation means that as one variable increases, the other also increases. Negative correlation means that as one variable increases, the other decreases.

13. What is regression analysis?

- Regression analysis is used to predict the value of a dependent variable based on the value of one or more independent variables.

14. What is the difference between correlation and causation?

- Correlation is a relationship between two variables, while causation means that one variable directly affects the other.

15. What is a null hypothesis?

- The null hypothesis is a statement that there is no effect or no difference.

16. What is an alternative hypothesis?

- The alternative hypothesis is a statement that there is an effect or a difference.

17. What is a p-value?

- A p-value indicates the probability of obtaining the observed results assuming the null hypothesis is true.

18. What is a confidence interval?

- A confidence interval is a range of values that is likely to contain the population parameter.

19. What is a Type I error?

- A Type I error occurs when we reject the null hypothesis when it is actually true.

20. What is a Type II error?

- A Type II error occurs when we fail to reject the null hypothesis when it is actually false.

ADVANCED QUESTIONS

21. What is ANOVA?

- ANOVA (Analysis of Variance) is used to compare the means of three or more groups.

22. What is a t-test?

- A t-test is used to compare the means of two groups.

23. What is the Central Limit Theorem?

- The Central Limit Theorem states that the distribution of the sample mean approaches a normal distribution as the sample size increases.

24. What is heteroscedasticity?

- Heteroscedasticity occurs when the variance of errors is not constant across observations.

25. What is multicollinearity?

- Multicollinearity occurs when independent variables in a regression model are highly correlated.

26. What is logistic regression?

- Logistic regression is used for predicting binary outcomes.

27. What is a chi-square test?

- A chi-square test is used to determine if there is a significant association between two categorical variables.

28. What is the difference between a parametric and non-parametric test?

- Parametric tests assume underlying statistical distributions; non-parametric tests do not.

29. What is R-squared?

- R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

30. What is a residual in regression analysis?

- A residual is the difference between the observed value and the predicted value of the dependent variable.