Data Science - Statistics Quiz (Medium)

Q1. The Central Limit Theorem (CLT) is fundamental to inferential statistics. Which statement best describes it?

A) For any population, the sampling distribution of the sample mean is always exactly normal.
B) For a large enough sample size, the sampling distribution of the sample mean will be approximately normal, regardless of the population's distribution.
C) The mean of the sample is always equal to the mean of the population.
D) As the sample size increases, the variance of the population decreases.

Q2. In hypothesis testing, what is the correct interpretation of a p-value of 0.03?

A) There is a 3% probability that the null hypothesis is true.
B) There is a 3% probability of observing the data, or something more extreme, if the null hypothesis is true.
C) There is a 3% probability that the alternative hypothesis is false.
D) The probability of making a Type I error is 3%.

Q3. A researcher concludes that a new drug is effective, but in reality, it has no effect. The null hypothesis, which stated the drug has no effect, was rejected. What type of error has been made?

A) Type I Error
B) Type II Error
C) Standard Error
D) Sampling Error

Q4. A 95% confidence interval for the average height of a population is calculated to be [165 cm, 175 cm]. What is the correct interpretation of this interval?

A) There is a 95% probability that the true population mean height is between 165 cm and 175 cm.
B) 95% of the individuals in the population have a height between 165 cm and 175 cm.
C) If we were to repeat the sampling process many times, 95% of the calculated confidence intervals would contain the true population mean.
D) The sample mean has a 95% chance of being the true population mean.

Q5. In a simple linear regression model, what does the R-squared value represent?

A) The correlation between the predicted values and the residuals.
B) The proportion of the variance in the dependent variable that is predictable from the independent variable.
C) The probability that the linear relationship is due to random chance.
D) The average squared distance between the observed values and the regression line.

Q6. What is the primary problem caused by high multicollinearity in a multiple linear regression model?

A) It introduces significant bias into the coefficient estimates, making them systematically wrong.
B) It decreases the R-squared value, making the model appear less effective than it is.
C) It inflates the variance of the coefficient estimates, making them unstable and hard to interpret.
D) It violates the assumption of normality of residuals, invalidating p-values.

Q7. A machine learning model performs exceptionally well on the training data but poorly on unseen test data. This phenomenon is known as:

A) Underfitting, characterized by high bias.
B) Overfitting, characterized by high variance.
C) The bias-variance trade-off.
D) A well-generalized model.

Q8. A data scientist wants to test if there is a statistically significant association between a person's favorite color (a categorical variable) and their profession (a categorical variable). Which statistical test is most appropriate?

A) Two-sample t-test
B) ANOVA
C) Pearson Correlation
D) Chi-squared test for independence

Q9. When a dataset contains extreme outliers, which measure of central tendency is generally the most robust and reliable?

A) Mean
B) Median
C) Mode
D) Geometric Mean

Q10. If the standard deviation of a dataset is 15, and every value in the dataset is multiplied by 2, what will be the new standard deviation?

A) 15
B) 7.5
C) 30
D) 225

Q11. The Poisson distribution is most suitable for modeling which of the following scenarios?

A) The number of defective items in a batch of 100.
B) The waiting time until the next customer arrives.
C) The number of emails received by an office in one hour.
D) The distribution of student scores on an exam.

Q12. In the context of Bayes' Theorem, P(A) is referred to as the:

A) Posterior probability

B) Likelihood
C) Marginal probability
D) Prior probability

Q13. A researcher wants to compare the average test scores of students from four different schools to see if there is a significant difference. What is the most appropriate statistical test?

A) Multiple two-sample t-tests
B) Analysis of Variance (ANOVA)
C) Chi-squared test
D) Simple linear regression

Q14. If the correlation coefficient between two variables, X and Y, is -0.95, what can be concluded?

A) X causes Y to decrease.
B) There is a strong, negative linear relationship between X and Y.
C) There is a weak, negative linear relationship between X and Y.
D) Y causes X to decrease.

Q15. The assumption of homoscedasticity in linear regression implies that:

A) The residuals have a constant variance.
B) The residuals are normally distributed.
C) The independent variables are not correlated.
D) The relationship between X and Y is linear.

Q16. In multiple regression, why is adjusted R-squared often preferred over R-squared?

A) It is always a larger value, indicating a better model fit.
B) It penalizes the addition of irrelevant predictor variables to the model.
C) It is simpler to calculate and interpret.
D) It can be used for non-linear models, unlike R-squared.

Q17. A polling company divides the population into several geographic areas. They randomly select a few areas and then interview every household within those selected areas. This is an example of:

A) Simple Random Sampling
B) Stratified Sampling
C) Cluster Sampling
D) Systematic Sampling

Q18. Which of the following is NOT a necessary condition for a Binomial distribution?

A) The number of trials is fixed.
B) Each trial must be independent.
C) The number of trials must be large (e.g., > 30).
D) Each trial has only two possible outcomes.

Q19. If the covariance between two variables is zero, what can be definitively concluded?

A) The two variables are independent.
B) There is no linear relationship between the two variables.
C) There is no relationship of any kind between the two variables.
D) Both variables have a variance of zero.

Q20. What does the standard error of the mean (SEM) quantify?

A) The standard deviation of the data points in the sample.
B) The systematic error or bias in the measurement process.
C) The precision of the sample mean as an estimate of the population mean.
D) The range of the middle 50% of the data.

Q21. A data point has a z-score of -1.5. What does this signify?

A) The data point is 1.5 units smaller than the mean.
B) The data point is 1.5 standard deviations below the mean.
C) The data point is 1.5 times the mean.
D) The data point is in the 15th percentile.

Q22. In an A/B test comparing a new website layout (B) to an old one (A), the null hypothesis is that the conversion rates are equal. The test results in a p-value of 0.45. What is the appropriate conclusion?

A) The new layout B is significantly better than A.
B) The old layout A is significantly better than B.
C) We fail to reject the null hypothesis; there is no statistically significant difference in conversion rates.
D) The test is invalid and must be repeated with a larger sample size.

Q23. In a multiple regression model, Y = β■ + β■X■ + β■X■, how is the coefficient β■ interpreted?

A) The predicted change in Y for a one-unit increase in X■, ignoring X■.
B) The predicted change in Y for a one-unit increase in X■, holding X■ constant.
C) The correlation between Y and X■.
D) The total change in Y caused by both X■ and X■.

Q24. A probability distribution that has 'fatter tails' than a normal distribution is described as:

A) Mesokurtic
B) Platykurtic
C) Leptokurtic
D) Negatively skewed

Q25. The Interquartile Range (IQR) is often used to identify outliers because it is:

A) Based on the two most extreme values in the dataset.

B) Easy to calculate by hand for any sample size.
C) Directly related to the mean of the distribution.
D) Resistant to the influence of extreme values.

Q26. The Law of Large Numbers states that as the sample size of a random variable increases, the:

A) Sample mean gets closer to the population mean.
B) Sample distribution becomes approximately normal.
C) Sample variance approaches the population mean.
D) Number of outliers decreases to zero.

Q27. What is the statistical power of a hypothesis test?

A) The probability of correctly rejecting a false null hypothesis.
B) The probability of making a Type I error (α).
C) The probability of making a Type II error (β).
D) The confidence level of the test ($1 - \alpha$).

Q28. A scientist wants to test if a new fertilizer *increases* plant growth compared to the current average height of 20 cm. What is the correct alternative hypothesis (H■)?

A) H■: $\mu = 20$
B) H■: $\mu \neq 20$
C) H■: $\mu > 20$
D) H■: $\mu < 20$

Q29. For a standard independent two-sample t-test, a key assumption is that the two populations being compared have equal:

A) Means
B) Sample sizes
C) Medians
D) Variances

Q30. In Principal Component Analysis (PCA), what does the first principal component represent?

A) The feature with the highest variance.
B) The direction in the data that captures the most variance.
C) The average value of all data points.
D) The component that is least correlated with the target variable.

Q31. A simple linear model is used to describe a complex, non-linear relationship in the data. This model will likely suffer from:

A) High variance and low bias.
B) Low variance and low bias.
C) Low variance and high bias.
D) High variance and high bias.

Q32. A log transformation is often applied to a variable in a regression model when:

A) The variable is categorical.
B) The variable's distribution is strongly skewed.
C) There are missing values in the variable.
D) The variable has a negative correlation with the target.

Q33. For a single roll of a fair six-sided die, what is the expected value?

A) 3
B) 3.5
C) 4
D) It cannot be calculated.

Q34. To include a categorical variable with 5 distinct categories in a linear regression model, how many dummy variables should be created?

A) 1
B) 5
C) 4
D) 2

Q35. In a residual plot from a linear regression, a 'fan' or 'cone' shape, where the spread of residuals increases as the predicted value increases, is a sign of:

A) Multicollinearity
B) Autocorrelation
C) Heteroscedasticity
D) Non-normality of residuals

Q36. Given P(A) = 0.6, P(B) = 0.5, and P(A and B) = 0.3. What is the conditional probability P(A|B)?

A) 0.5
B) 0.6
C) 0.9
D) 0.3

Q37. Bootstrapping is a resampling technique primarily used to:

A) Increase the size of a small dataset to improve model accuracy.
B) Remove outliers from a dataset.
C) Estimate the sampling distribution of a statistic, such as the mean or a coefficient.
D) Test for the normality of a dataset.

Q38. In a two-sample t-test comparing two independent groups of 20 and 25 individuals respectively, what are the degrees of freedom?

A) 45
B) 44
C) 43

D) 19

Q39. In a positively skewed (right-skewed) distribution, what is the typical relationship between the mean and the median?

A) Mean < Median
B) Mean > Median
C) Mean = Median
D) The relationship is unpredictable.

Q40. What does the 'box' in a standard box-and-whisker plot represent?

A) The range of the entire dataset.
B) The range containing the middle 50% of the data (the Interquartile Range).
C) The values that fall within one standard deviation of the mean.
D) The 95% confidence interval for the median.

Q41. In hypothesis testing, what does the significance level, alpha (α), represent?

A) The probability of a Type II error.
B) The power of the test.
C) The pre-determined threshold for the probability of a Type I error.
D) The effect size of the result.

Q42. Why is feature scaling (e.g., standardization) crucial for algorithms like K-Means clustering?

A) It ensures all feature values are positive.
B) It converts categorical features into a numerical format.
C) It prevents features with larger scales from dominating the distance calculations.
D) It is only required if the data is not normally distributed.

Q43. What is the primary statistical concern when performing many pairwise t-tests to compare the means of several groups, instead of using ANOVA?

A) It is less statistically powerful than ANOVA.
B) It is computationally much slower.
C) It inflates the family-wise error rate (the probability of making at least one Type I error).
D) It requires that all groups have the same number of observations.

Q44. A customer's star rating for a product (1, 2, 3, 4, or 5 stars) is an example of what type of data?

A) Nominal
B) Ordinal
C) Interval
D) Ratio

Q45. Chebyshev's Inequality is useful because it applies to:

A) Only normal distributions.

B) Only symmetric distributions.
C) Any probability distribution, regardless of its shape.
D) Only distributions with a known mean and median.

Q46. In a time-series analysis, if the residuals of a regression model show a pattern where consecutive residuals are correlated, which assumption is violated?

A) Linearity
B) Normality of residuals
C) Independence of errors (autocorrelation)
D) Homoscedasticity

Q47. A permutation test is a non-parametric method often used to:

A) Estimate confidence intervals by resampling with replacement.
B) Test a hypothesis by calculating a p-value based on random relabeling of data.
C) Assume the data follows a specific theoretical distribution like the t-distribution.
D) Correct for measurement errors in the dataset.

Q48. While a p-value indicates statistical significance, an 'effect size' measure (like Cohen's d) is important because it quantifies the:

A) Probability that the null hypothesis is false.
B) Number of participants needed for adequate statistical power.
C) Magnitude or practical importance of the observed effect.
D) Standard deviation of the sampling distribution.

Q49. In logistic regression, the logit function transforms a probability (p) into log-odds. What is the primary purpose of this transformation?

A) To ensure the output is always positive.
B) To map the [0, 1] probability range to the entire real number line [-∞, +∞], which can then be modeled linearly.
C) To calculate the R-squared value for the model.
D) To normalize the input features before training the model.

Q50. The coefficient of a predictor in a multiple linear regression model is found to be statistically significant. This means:

A) The predictor has a causal relationship with the response variable.
B) The predictor is a practically important and influential variable.
C) The predictor has a statistically significant linear relationship with the response variable, after accounting for other predictors.
D) The predictor is not correlated with any other predictors in the model.

Answer Key:

Q1: B
Q2: B
Q3: A
Q4: C
Q5: B
Q6: C
Q7: B
Q8: D
Q9: B
Q10: C
Q11: C
Q12: D
Q13: B
Q14: B
Q15: A
Q16: B
Q17: C
Q18: C
Q19: B
Q20: C
Q21: B
Q22: C
Q23: B
Q24: C
Q25: D
Q26: A
Q27: A
Q28: C
Q29: D
Q30: B
Q31: C
Q32: B
Q33: B
Q34: C
Q35: C
Q36: B
Q37: C
Q38: C
Q39: B
Q40: B
Q41: C
Q42: C
Q43: C
Q44: B
Q45: C
Q46: C
Q47: B
Q48: C
Q49: B
Q50: C