Statistics for Data Science Interview Questions and Answers

Sanjay Kumar PhD Following 19 min read · Jan 1, 2025

665 32

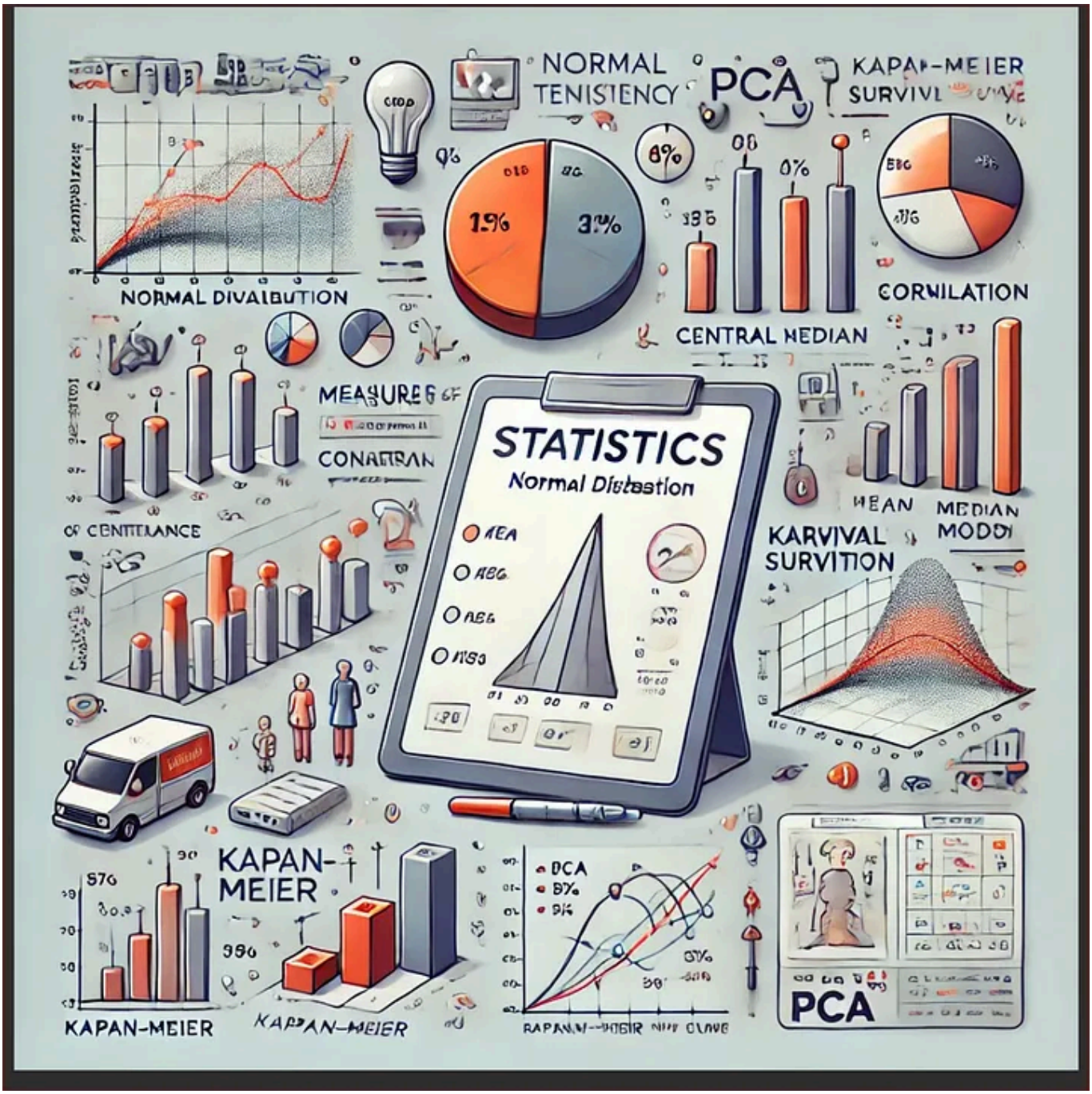


Image generated by author using DALL E

Answer: Some important topics include:

- Measures of central tendency (mean, median, mode)
- Measures of dispersion (variance, standard deviation)

- Covariance and correlation
- Probability distribution functions
- Standardization and normalization
- Central limit theorem
- Population and sample
- Hypothesis testing

2. How would you explain Exploratory Data Analysis (EDA)?

Answer: EDA involves visually and statistically analyzing data to uncover underlying patterns, distributions, and relationships. The goal is to gain insights, identify potential issues, and guide subsequent data processing steps.

3. What is the difference between qualitative and quantitative data?

Answer:

- **Quantitative data:** Numeric, measurable, and analyzed statistically (e.g., age, income). Types include discrete and continuous data.
- **Qualitative data:** Descriptive, non-numeric, and analyzed by grouping (e.g., gender, marital status). Types include nominal and ordinal data.

4. How do you handle datasets with over 30% missing values?

Answer: Choose imputation methods based on the nature of the data:

- **Mean/Median Imputation:** For numerical data with normal distributions.
- **Mode Imputation:** For categorical data.
- **KNN Imputation:** Uses nearest neighbors for estimating missing values.

5. When is the median a better measure than the mean?

Answer: The median is preferred over the mean in the presence of extreme outliers or highly skewed data, as it is less affected by these extreme values.

Example: For incomes {25, 28, 30, 32, 35, 5000}, the mean is heavily skewed by the outlier, while the median provides a more accurate central tendency.

6. What is the Central Limit Theorem?

Answer: The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes larger ($n \geq 30$), regardless of the population's original distribution.

7. How do you detect outliers in a dataset?

Answer: Methods include:

- Box plots: Visualize outliers as points beyond the whiskers.
- Z-scores: Identify values beyond 2 or 3 standard deviations from the mean.
- IQR method: Detect values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$.

8 . What is the Empirical Rule?

Answer: For a normal distribution:

- ~68% of data falls within 1 standard deviation from the mean.
- ~95% falls within 2 standard deviations.
- ~99.7% falls within 3 standard deviations.

9. How do you calculate the coefficient of variation (CV)?

Answer: The CV measures relative variability and is calculated as:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100$$

Where σ is the standard deviation, and μ is the mean.

10. How does one distinguish between descriptive and inferential statistics?

Answer:

- **Descriptive statistics:** Summarize and describe data features (e.g., mean, median).
- **Inferential statistics:** Draw conclusions about a population based on a sample (e.g., hypothesis testing, regression).

11. What is the difference between univariate, bivariate, and multivariate analysis?

Answer:

- **Univariate Analysis:** Examines a single variable. Example: Histogram, mean, median.
- **Bivariate Analysis:** Examines the relationship between two variables. Example: Scatter plots, correlation.
- **Multivariate Analysis:** Examines relationships among multiple variables. Example: PCA, factor analysis.

12. What are the different measures of dispersion in statistics?

Answer: Measures of dispersion include:

- **Range:** Difference between maximum and minimum values.
- **Variance:** Average squared deviation from the mean.
- **Standard Deviation:** Square root of variance, representing spread in original units.
- **Interquartile Range (IQR):** Difference between Q3 and Q1, representing the middle 50% of data.

13. What is Bessel's correction, and why is it used?

Answer: Bessel's correction adjusts the sample variance by dividing by n−1n — 1n−1 instead of nnn to correct bias when estimating population variance from a sample. It ensures an unbiased estimate of the population variance.

14. What are outliers, and how do they impact data analysis?

Answer: Outliers are data points that significantly deviate from others in a dataset.

Impacts:

- **Negative:** Skew central tendency measures, inflate dispersion, and mislead statistical tests.
- **Positive:** Indicate anomalies or rare events, useful in fraud detection or quality control.

15. How do you handle outliers in a dataset?

Answer:

- **Remove:** If outliers are due to errors.
- **Transform:** Use log or square root transformations to reduce impact.
- **Winsorize:** Cap extreme values to a specific percentile.
- **Use Robust Methods:** Replace mean with median or use robust algorithms.

16. What are the key properties of a normal distribution?

Answer:

- Symmetric and bell-shaped.
- Mean = Median = Mode.
- Defined by mean μ\muμ and standard deviation σ\sigmaσ.
- Empirical Rule applies: ~68% within 1σ, ~95% within 2σ, ~99.7% within 3σ.

17. What is the coefficient of variation, and why is it used?

Answer: The coefficient of variation (CV) measures relative variability and is expressed as:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100$$

It's useful for comparing variability between datasets with different units or means.

18. Explain the difference between point estimation and confidence interval estimation.

Answer:

- **Point Estimation:** Provides a single value estimate of a population parameter (e.g., sample mean for population mean).
- **Confidence Interval:** Provides a range of values likely to contain the population parameter, with a specified confidence level (e.g., 95%).

19. What are the conditions required for the Central Limit Theorem to hold?

Answer:

- Random sampling.
- Independence of data points.
- Sufficient sample size ($n \geq 30$).

- Population must have finite variance.

20. What are the main types of probability distributions?

Answer:

1. **Discrete Distributions:** Bernoulli, Binomial, Poisson.
2. **Continuous Distributions:** Normal, Uniform, Exponential.

21. What is the difference between left-skewed and right-skewed distributions?

Answer:

- **Left-skewed (Negative skew):** Longer tail on the left; mean < median < mode.
- **Right-skewed (Positive skew):** Longer tail on the right; mean > median > mode.

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

Answer: A z-score measures how many standard deviations a value is from the mean:

Z=X−μσZ = \frac{X — \mu}{\sigma}Z=σX−μ

Where XXX is the data point, μ\muμ is the mean, and σ\sigmaσ is the standard deviation.

23. What are some common methods to impute missing data?

Answer:

- **Mean/Median Imputation:** Replace with mean or median.
- **Mode Imputation:** For categorical data.
- **KNN Imputation:** Use nearest neighbors for prediction.
- **Regression Imputation:** Predict missing values using regression models.

24. Explain the empirical rule in the context of normal distribution.

Answer: In a normal distribution:

- ~68% of data lies within 1σ of the mean.
- ~95% lies within 2σ.
- ~99.7% lies within 3σ.

25. What is the Pareto Principle, and where is it used?

Answer: The Pareto Principle, or 80/20 Rule, states that 80% of outcomes result from 20% of causes. It's used in economics, business, and quality control to identify key contributing factors.

26. What is the difference between covariance and correlation?

Answer:

- **Covariance:** Measures the extent to which two variables change together. It can take any value and is unit-dependent.
- **Correlation:** Standardized version of covariance, ranging from -1 to 1, indicating the strength and direction of the relationship between two variables. It is unitless.

27. What is skewness, and how is it measured?

Answer:

- **Skewness:** A measure of the asymmetry of a data distribution.
- **Right-skewed (positive):** Long tail on the right.
- **Left-skewed (negative):** Long tail on the left.
- **Symmetric:** Zero skewness.
- **Measures:**
- Pearson's Coefficient of Skewness

- Fisher-Pearson Standardized Moment Coefficient
- Bowley's Coefficient (based on quartiles)

28. What is kurtosis, and what are its types?

Answer:

- **Kurtosis:** Measures the "tailedness" or peakness of a distribution.
- **Types:**
- **Mesokurtic:** Normal distribution (moderate tails).
- **Leptokurtic:** Heavy tails and sharp peak.
- **Platykurtic:** Light tails and flat peak.

29. What are the differences between population and sample?

Answer:

- **Population:** Entire group of interest; contains all possible observations.
- **Sample:** Subset of the population used for analysis; should represent the population.

30. What are the assumptions of a normal distribution?

Answer:

1. Symmetry about the mean.
2. Mean = Median = Mode.
3. Data is continuous.
4. Tails extend to infinity without touching the x-axis.

31. Explain the relationship between confidence level and significance level.

Answer:

- **Confidence Level (1 − α):** Probability that the confidence interval contains the true population parameter.
- **Significance Level (α):** Probability of rejecting a true null hypothesis (Type I error).

Example: At a 95% confidence level, the significance level is 0.05.

32. What is selection bias, and what are its types?

Answer:

- **Selection Bias:** Occurs when the sample is not representative of the population.

- **Types:**
- Sampling bias
- Attrition bias
- Observer selection bias
- Protopathic bias
- Time interval bias

33. What are some methods to deal with missing data?

Answer:

- **Remove missing data** if it is a small proportion.
- **Impute missing data** using:
- Mean, median, or mode.
- Predictive models like regression or KNN.
- **Model with missingness** as a feature.

34. What is the Central Limit Theorem (CLT), and why is it important?

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

Importance:

- Enables inferential statistics.
- Forms the basis of hypothesis testing and confidence intervals.

35. What is a hypothesis test, and what are the steps involved?

Answer:

- **Hypothesis Testing:** A statistical method to test assumptions about a population parameter.
- **Steps:**

1. Formulate null (H0) and alternative (Ha) hypotheses.
2. Choose significance level (alpha α).
3. Select the test (e.g., t-test, chi-square test).
4. Calculate the test statistic and p-value.
5. Compare p-value with α\alphaα and decide to reject or fail to reject H0

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

Answer:

- **Type I Error (alpha α):** Rejecting a true null hypothesis.

- **Type II Error (beta β):** Failing to reject a false null hypothesis.

37. What are the assumptions for linear regression?

Answer:

1. Linearity between independent and dependent variables.
2. Homoscedasticity (constant variance of residuals).
3. Independence of residuals.
4. No multicollinearity.
5. Normally distributed residuals.

38. What is the difference between a t-test and ANOVA?

Answer:

- **T-test:** Compares means of two groups.
- **ANOVA (Analysis of Variance):** Compares means across three or more groups.

39. What are the properties of a binomial distribution?

Answer:

1. Fixed number of trials (nnn).

2. Each trial is independent.
3. Two possible outcomes (success or failure).
4. Constant probability of success (ppp).

40. What is the purpose of Z-scores in standardization?

Answer: Z-scores transform data to have a mean of 0 and a standard deviation of 1. This allows comparison across different datasets or variables on the same scale.

41. What is the Pareto principle (80/20 rule)?

Answer: The Pareto Principle states that roughly 80% of outcomes result from 20% of causes. It's used in business, economics, and quality control to identify key contributors.

Here are additional advanced statistics interview questions and answers based on the document:

26. What is the difference between covariance and correlation?

Answer:

- **Covariance:** Measures the extent to which two variables change together. It can take any value and is unit-dependent.
- **Correlation:** Standardized version of covariance, ranging from -1 to 1, indicating the strength and direction of the relationship between two variables. It is unitless.

27. What is skewness, and how is it measured?

Answer:

- **Skewness:** A measure of the asymmetry of a data distribution.
- **Right-skewed (positive):** Long tail on the right.
- **Left-skewed (negative):** Long tail on the left.
- **Symmetric:** Zero skewness.
- **Measures:**
- Pearson's Coefficient of Skewness
- Fisher-Pearson Standardized Moment Coefficient
- Bowley's Coefficient (based on quartiles)

28. What is kurtosis, and what are its types?

Answer:

- **Kurtosis:** Measures the "tailedness" or peakness of a distribution.

- **Types:**
- **Mesokurtic:** Normal distribution (moderate tails).
- **Leptokurtic:** Heavy tails and sharp peak.
- **Platykurtic:** Light tails and flat peak.

29. What are the differences between population and sample?

Answer:

- **Population:** Entire group of interest; contains all possible observations.
- **Sample:** Subset of the population used for analysis; should represent the population.

30. What are the assumptions of a normal distribution?

Answer:

1. Symmetry about the mean.
2. Mean = Median = Mode.
3. Data is continuous.
4. Tails extend to infinity without touching the x-axis.

31. Explain the relationship between confidence level and significance level.

Answer:

- **Confidence Level (1 − α):** Probability that the confidence interval contains the true population parameter.
- **Significance Level (α):** Probability of rejecting a true null hypothesis (Type I error).

Example: At a 95% confidence level, the significance level is 0.05.

32. What is selection bias, and what are its types?

Answer:

- **Selection Bias:** Occurs when the sample is not representative of the population.
- **Types:**
- Sampling bias
- Attrition bias
- Observer selection bias
- Protopathic bias
- Time interval bias

33. What are some methods to deal with missing data?

Answer:

- **Remove missing data** if it is a small proportion.
- **Impute missing data** using:
- Mean, median, or mode.
- Predictive models like regression or KNN.
- **Model with missingness** as a feature.

34. What is the Central Limit Theorem (CLT), and why is it important?

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

Importance:

- Enables inferential statistics.
- Forms the basis of hypothesis testing and confidence intervals.

35. What is a hypothesis test, and what are the steps involved?

Answer:

- **Hypothesis Testing:** A statistical method to test assumptions about a population parameter.
- **Steps:**

1. Formulate null (H0H_0H0) and alternative (HaH_aHa) hypotheses.
2. Choose significance level (α\alphaα).
3. Select the test (e.g., t-test, chi-square test).
4. Calculate the test statistic and p-value.
5. Compare p-value with α\alphaα and decide to reject or fail to reject H0H_0H0.

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

Answer:

- **Type I Error (α\alphaα):** Rejecting a true null hypothesis.
- **Type II Error (β\betaβ):** Failing to reject a false null hypothesis.

37. What are the assumptions for linear regression?

Answer:

1. Linearity between independent and dependent variables.
2. Homoscedasticity (constant variance of residuals).
3. Independence of residuals.
4. No multicollinearity.
5. Normally distributed residuals.

38. What is the difference between a t-test and ANOVA?

Answer:

- **T-test:** Compares means of two groups.
- **ANOVA (Analysis of Variance):** Compares means across three or more groups.

39. What are the properties of a binomial distribution?

Answer:

1. Fixed number of trials (nnn).
2. Each trial is independent.
3. Two possible outcomes (success or failure).

4. Constant probability of success (ppp).

40. What is the purpose of Z-scores in standardization?

Answer: Z-scores transform data to have a mean of 0 and a standard deviation of 1. This allows comparison across different datasets or variables on the same scale.

41. What is the Pareto principle (80/20 rule)?

Answer: The Pareto Principle states that roughly 80% of outcomes result from 20% of causes. It's used in business, economics, and quality control to identify key contributors.

42. How do you determine the range and interquartile range (IQR)?

Here are additional advanced statistics interview questions and answers based on the document:

26. What is the difference between covariance and correlation?

Answer:

- **Covariance:** Measures the extent to which two variables change together. It can take any value and is unit-dependent.
- **Correlation:** Standardized version of covariance, ranging from -1 to 1, indicating the strength and direction of the relationship between two variables. It is unitless.

27. What is skewness, and how is it measured?

Answer:

- **Skewness:** A measure of the asymmetry of a data distribution.
- **Right-skewed (positive):** Long tail on the right.
- **Left-skewed (negative):** Long tail on the left.
- **Symmetric:** Zero skewness.
- **Measures:**
- Pearson's Coefficient of Skewness
- Fisher-Pearson Standardized Moment Coefficient
- Bowley's Coefficient (based on quartiles)

28. What is kurtosis, and what are its types?

Answer:

- **Kurtosis:** Measures the "tailedness" or peakness of a distribution.

- **Types:**
- **Mesokurtic:** Normal distribution (moderate tails).
- **Leptokurtic:** Heavy tails and sharp peak.
- **Platykurtic:** Light tails and flat peak.

29. What are the differences between population and sample?

Answer:

- **Population:** Entire group of interest; contains all possible observations.
- **Sample:** Subset of the population used for analysis; should represent the population.

30. What are the assumptions of a normal distribution?

Answer:

1. Symmetry about the mean.
2. Mean = Median = Mode.
3. Data is continuous.
4. Tails extend to infinity without touching the x-axis.

31. Explain the relationship between confidence level and significance level.

Answer:

- **Confidence Level (1 − α):** Probability that the confidence interval contains the true population parameter.
- **Significance Level (α):** Probability of rejecting a true null hypothesis (Type I error).

Example: At a 95% confidence level, the significance level is 0.05.

32. What is selection bias, and what are its types?

Answer:

- **Selection Bias:** Occurs when the sample is not representative of the population.
- **Types:**
- Sampling bias
- Attrition bias
- Observer selection bias
- Protopathic bias
- Time interval bias

33. What are some methods to deal with missing data?

Answer:

- **Remove missing data** if it is a small proportion.
- **Impute missing data** using:
- Mean, median, or mode.
- Predictive models like regression or KNN.
- **Model with missingness** as a feature.

34. What is the Central Limit Theorem (CLT), and why is it important?

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

Importance:

- Enables inferential statistics.
- Forms the basis of hypothesis testing and confidence intervals.

35. What is a hypothesis test, and what are the steps involved?

Answer:

- **Hypothesis Testing:** A statistical method to test assumptions about a population parameter.
- **Steps:**

1. Formulate null (H0H_0H0) and alternative (HaH_aHa) hypotheses.
2. Choose significance level (α\alphaα).
3. Select the test (e.g., t-test, chi-square test).
4. Calculate the test statistic and p-value.
5. Compare p-value with α\alphaα and decide to reject or fail to reject H0H_0H0.

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

Answer:

- **Type I Error (α\alphaα):** Rejecting a true null hypothesis.
- **Type II Error (β\betaβ):** Failing to reject a false null hypothesis.

37. What are the assumptions for linear regression?

Answer:

1. Linearity between independent and dependent variables.
2. Homoscedasticity (constant variance of residuals).
3. Independence of residuals.
4. No multicollinearity.
5. Normally distributed residuals.

38. What is the difference between a t-test and ANOVA?

Answer:

- **T-test:** Compares means of two groups.
- **ANOVA (Analysis of Variance):** Compares means across three or more groups.

39. What are the properties of a binomial distribution?

Answer:

1. Fixed number of trials (nnn).
2. Each trial is independent.
3. Two possible outcomes (success or failure).

4. Constant probability of success (ppp).

40. What is the purpose of Z-scores in standardization?

Answer: Z-scores transform data to have a mean of 0 and a standard deviation of 1. This allows comparison across different datasets or variables on the same scale.

41. What is the Pareto principle (80/20 rule)?

Answer: The Pareto Principle states that roughly 80% of outcomes result from 20% of causes. It's used in business, economics, and quality control to identify key contributors.

42. How do you determine the range and interquartile range (IQR)?

Answer:

- **Range**: Difference between maximum and minimum values ($\text{Range} = \text{Max} - \text{Min}$).
- **IQR**: Difference between the 75th percentile ($Q3$) and 25th percentile ($Q1$):

$$IQR = Q3 - Q1$$

43. What are some common probability distributions in statistics?

Answer:

- **Discrete:** Binomial, Poisson.
- **Continuous:** Normal, Uniform, Exponential.

44. What are the different types of sampling methods?

Answer:

1. **Simple Random Sampling**
2. **Stratified Sampling**
3. **Cluster Sampling**
4. **Systematic Sampling**
5. **Convenience Sampling**

45. What is multicollinearity, and how do you detect it?

Answer:

Multicollinearity: Occurs when independent variables in a regression model are highly correlated, making it difficult to isolate their individual effects.

Detection:

- **Variance Inflation Factor (VIF):** VIF > 5 or 10 indicates high multicollinearity.
- **Correlation Matrix:** High pairwise correlations (e.g., > 0.8) among independent variables.
- **Condition Index:** Values > 30 indicate severe multicollinearity.

46. What are the assumptions of ANOVA?

Answer:

1. Independence of observations.
2. Homogeneity of variances (equal variances across groups).
3. Normally distributed residuals.

47. What is the difference between parametric and non-parametric tests?

Answer:

Parametric Tests:

- Assume specific distribution (e.g., normal).
- Examples: t-test, ANOVA.

Non-parametric Tests:

- No distribution assumptions.

- Examples: Mann-Whitney U test, Kruskal-Wallis test.

48. What are the differences between Type II error and statistical power?

Answer:

- **Type II Error (β)** Probability of failing to reject a false null hypothesis.
- **Statistical Power (1−β):** Probability of correctly rejecting a false null hypothesis. Higher power reduces the likelihood of Type II errors.

49. What is the difference between absolute and relative risk?

Answer:

- **Absolute Risk:** The probability of an event occurring in a group (e.g., 5%).
- **Relative Risk:** The ratio of probabilities between two groups (e.g., risk in exposed vs. unexposed).

50. Explain the difference between standard deviation and standard error.

Answer:

- **Standard Deviation (σ):** Measures the spread of data in a sample.
- **Standard Error (SE):** Measures the accuracy of the sample mean as an estimate of the population mean:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Where n is the sample size.

51. What is the difference between R-squared and adjusted R-squared in regression?

Answer:

- **R-squared:** Proportion of variance in the dependent variable explained by the independent variables.
- **Adjusted R-squared:** Adjusts R^2 for the number of predictors, penalizing for overfitting.

52. What is the difference between p-value and confidence interval?

Answer:

- **P-value:** Probability of observing data as extreme as the sample, assuming the null hypothesis is true.
- **Confidence Interval:** Range of values that likely contain the true population parameter, based on a specified confidence level.

53. What is the Kolmogorov-Smirnov test?

Answer:

- A non-parametric test that compares a sample distribution with a reference distribution or two sample distributions to check for differences.
- Used to test for normality or distribution equality.

54. What is heteroscedasticity, and how do you detect it?

Answer:

Heteroscedasticity: Occurs when the variance of residuals is not constant across levels of an independent variable.

Detection:

- Residual plots.
- Breusch-Pagan test.
- White test.

55. What are some common methods to address multicollinearity?

Answer:

- Remove highly correlated predictors.
- Use dimensionality reduction techniques like PCA.
- Apply regularization techniques (e.g., Lasso, Ridge regression).

56. What is the difference between t-tests and z-tests?

Answer:

- **T-tests:** Used when sample size is small (n < 30) and population standard deviation is unknown.
- **Z-tests:** Used when sample size is large and population standard deviation is known.

57. What is the purpose of a chi-square test?

Answer:

- To test the independence of categorical variables (Chi-square test of independence).
- To test goodness-of-fit to a distribution.

58. What is the difference between time series decomposition and ARIMA modeling?

Answer:

- **Time Series Decomposition:** Breaks a time series into trend, seasonal, and residual components.
- **ARIMA:** Combines autoregressive (AR), moving average (MA), and differencing to model time series data.

59. What is the difference between AIC and BIC in model selection?

Answer:

- **AIC (Akaike Information Criterion):** Balances model fit and complexity, penalizing additional parameters.
- **BIC (Bayesian Information Criterion):** Similar to AIC but imposes a heavier penalty for additional parameters.

60. How do you interpret a p-value?

Answer:

- **P-value < α:** Reject the null hypothesis.

- **P-value > α:** Fail to reject the null hypothesis.

Example: At α=0.05\alpha = 0.05α=0.05, a p-value of 0.03 suggests strong evidence against the null hypothesis.

61. What is the F-statistic in ANOVA?

Answer:

- Measures the ratio of variance explained by the model to the variance within groups.
- Higher F-statistic indicates a significant difference among group means.

62. How do you calculate confidence intervals for the mean?

Answer:

$$CI = \bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

Where:

- $\bar{X}$: Sample mean.
- Z: Z-score corresponding to confidence level.
- σ: Standard deviation.
- n: Sample size.

63. What is the difference between a one-sided and two-sided hypothesis test?

Answer:

- **One-sided Test:** Tests if the parameter is greater or less than a specific value.
- **Two-sided Test:** Tests if the parameter is different from a specific value.

64. What is the Durbin-Watson test?

Answer:

- Used in regression to detect autocorrelation in residuals.
- Values close to 2 indicate no autocorrelation; values closer to 0 or 4 indicate positive or negative autocorrelation, respectively.

65. What is the difference between A/B testing and hypothesis testing?

Answer:

- **A/B Testing:** A specific type of hypothesis testing used to compare two versions (A and B) of a product, webpage, or process to determine which performs better.

- **Hypothesis Testing:** A general statistical framework to test assumptions about a population parameter.

66. What are the assumptions of logistic regression?

Answer:

1. Dependent variable is binary.
2. Observations are independent.
3. No multicollinearity among predictors.
4. Linear relationship between independent variables and the log odds of the dependent variable.
5. Large sample size for reliable estimation.

67. What is the difference between overfitting and underfitting?

Answer:

- **Overfitting:** Model performs well on training data but poorly on unseen data due to excessive complexity.
- **Underfitting:** Model performs poorly on both training and test data due to insufficient complexity.

68. What is the Bayesian approach to statistics?

Answer: Bayesian statistics incorporates prior knowledge or beliefs, updates them with evidence (data), and provides a posterior distribution. Bayes' theorem is central to this approach:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

69. How do you interpret an interaction term in regression?

Answer: An interaction term shows how the effect of one independent variable on the dependent variable depends on the level of another independent variable. A significant interaction indicates that the relationship is not additive.

70. What are the differences between a confidence interval and a prediction interval?

Answer:

- **Confidence Interval:** Range of values for a population parameter (e.g., mean).
- **Prediction Interval:** Range of values for an individual prediction.

71. What is bootstrapping, and when is it used?

Answer:

- **Bootstrapping:** A resampling technique to estimate the sampling distribution of a statistic by repeatedly sampling with replacement.
- **Usage:** When the theoretical distribution is unknown or sample size is small.

72. What is the Jackknife resampling method?

Answer: The Jackknife is a resampling method used to estimate the bias and variance of a statistic by systematically leaving out one observation at a time from the dataset and recalculating the statistic.

73. What is the difference between PCA (Principal Component Analysis) and Factor Analysis?

Answer:

- **PCA:** Reduces dimensionality by transforming data into orthogonal principal components.
- **Factor Analysis:** Identifies underlying latent variables (factors) that explain observed correlations.

74. What is the Kullback-Leibler (KL) divergence?

Answer: KL divergence measures the difference between two probability distributions P and Q:

$$D_{KL}(P||Q) = \sum P(x) \log \frac{P(x)}{Q(x)}$$

It is asymmetric and quantifies how much Q diverges from P.

75. What is the Gini coefficient, and how is it used?

Answer: The Gini coefficient measures inequality in a distribution, such as income or model performance (e.g., in classification). Values range from 0 (perfect equality) to 1 (maximum inequality).

76. How do you assess the goodness of fit for a regression model?

Answer:

- **R-squared:** Proportion of variance explained.
- **Adjusted R-squared:** Adjusted for the number of predictors.
- **Residual plots:** Check patterns and variance of residuals.
- **AIC/BIC:** Penalized measures of model complexity.
- **Cross-validation:** Out-of-sample prediction accuracy.

77. What is a survival analysis, and where is it used?

Answer: Survival analysis models time-to-event data. It is used in fields like medicine (time until recovery/death), business (time until churn), and engineering (time until failure).

Key techniques:

- Kaplan-Meier estimator.
- Cox proportional hazards model.

78. What are some common methods to handle imbalanced datasets?

Answer:

1. Resampling: Oversampling the minority class or undersampling the majority class.
2. Synthetic data generation: SMOTE (Synthetic Minority Oversampling Technique).
3. Weight adjustments: Adjust model weights to penalize misclassification of minority classes.
4. Algorithms designed for imbalance: XGBoost, Random Forest with class weights.

79. What is the difference between ROC-AUC and PR-AUC?

Answer:

- **ROC-AUC (Receiver Operating Characteristic — Area Under Curve):** Evaluates the trade-off between true positive rate (sensitivity) and false positive rate.
- **PR-AUC (Precision-Recall Area Under Curve):** Focuses on precision and recall, more informative for imbalanced datasets.

80. What is the Monte Carlo simulation, and where is it used?

Answer: Monte Carlo simulation uses random sampling to estimate probabilistic outcomes. Applications include:

- Risk assessment.
- Pricing financial derivatives.
- Estimating integrals in high dimensions.

81. What are hierarchical models in statistics?

Answer: Hierarchical (or multilevel) models handle data with nested structures (e.g., students within schools). They account for variability at each level using random effects.

82. What is a Markov Chain, and where is it applied?

Answer: A Markov Chain models a system where the next state depends only on the current state (memoryless property). Applications include:

- Weather forecasting.
- PageRank algorithm.
- Stock price modeling.

83. What is a time-varying covariate in survival analysis?

Answer: A time-varying covariate changes over the duration of observation (e.g., blood pressure in a medical study). Models like Cox regression can handle these variables.

84. How do you interpret a QQ-plot?

Answer: A QQ-plot compares the distribution of a dataset with a theoretical distribution (e.g., normal). Points along the diagonal indicate a good fit, while deviations suggest discrepancies.

85. What is Gibbs sampling?

Answer: Gibbs sampling is an MCMC (Markov Chain Monte Carlo) method to sample from a multivariate distribution by iteratively sampling from conditional distributions of each variable.

Data Science | Statistics | Machine Learning | Data Analysis | Interview

Written by Sanjay Kumar PhD

966 followers · 443 following

Following

Data Science |Machine Learning | AI Product | GenAI | RAG | LLM | AI Agents | NLP | Analytics | Data Engineering | Deep Learning | Statistics

Responses (32)

Hlgsagar

What are your thoughts?

Ahmed Abdulwahid him
Jan 24

very helpful. Thanks for sharing

5 Reply

Wise Michael T
Jan 21

Very thorough. Thank you for your efforts putting this together. Pinnned.

2 Reply

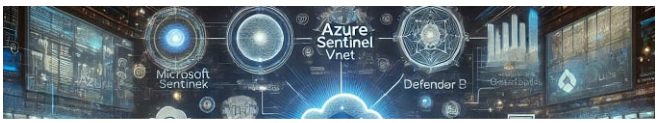
Anjana Mathew
May 10 (edited)

Thankyou sir for this wonderful note...God bless you.

1 Reply

See all responses

More from Sanjay Kumar PhD



Sanjay Kumar PhD

Microsoft Azure Interview Questions and Answers

Core Azure Concepts & Cloud Computing Basics

Feb 27 70

In Artificial Intelligence in Plain... by Sanjay Kumar...

Top 100 AI Agent Interview Questions

What is an AI agent, and how does it work?

Jun 1 67 2

Sanjay Kumar PhD

Data Engineering Interview Questions and Answers

1. What is a Data Warehouse, and how is it different from a Data Lake?

Dec 26, 2024 61 2

Sanjay Kumar PhD

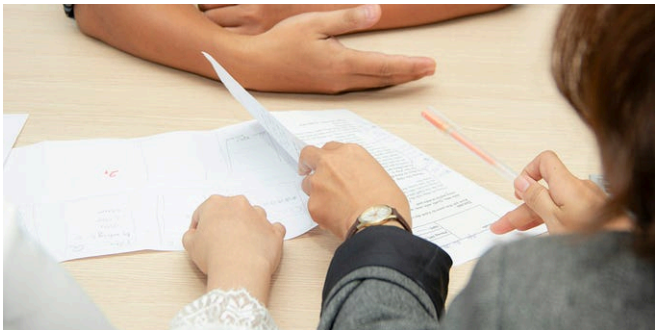
Advanced SQL Interview Questions and Answers

1. What is the difference between RANK(), DENSE_RANK(), and ROW_NUMBER()?

Mar 26 60

See all from Sanjay Kumar PhD

Recommended from Medium



In AI-ML Interview Playbook by Sajid Khan

Top 25 Machine Learning Interview Questions (And How to Answer...

Master these must-know ML questions to ace your next interview—from basics to advanc...

May 4 8 1

In Data Science Collective by Egor Howell

How To Improve Your LinkedIn For Data Scientist

A couple of my posts have recently gone down really well on LinkedIn.

Jun 13 168 1

Vikash Singh

How Much Statistics Do You Really Need for Data Science?

Non-medium members can read here.

Jun 12 1

In Write A Catalyst by Adarsh Gupta

How I Study Consistently While Working a 9–5 Full-Time Job

No, I don't wake up at 5 AM. And yes, I have a life.

Apr 21 6.4K 271

Ruth Yang

Data Clarity #1: Ten Data Strategy Principles for the Age of AI

By Ruth Yang

Damini Vadrevu

Data Analyst/Scientist Interview Questions—Read if you're Scared.

Only the BEST Guide to remove your Interview Anxiety

May 20 2

Mar 8 20 5

See more recommendations