Q1. Bernoulli random variables take (only) the values 1 and 0.

- **Correct Answer: a) True**

Q2. Which of the following theorems states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

- **Correct Answer: a) Central Limit Theorem**

Q3. Which of the following is incorrect with respect to the use of Poisson distribution?

- **Correct Answer: b) Modeling bounded count data**

Q4. Point out the correct statement.

- **Correct Answer: d) All of the mentioned**

Q5. _____ random variables are used to model rates.

- **Correct Answer: c) Poisson**

Q6. Usually replacing the standard error by its estimated value does change the CLT.

- **Correct Answer: b) False**

Q7. Which of the following testing is concerned with making decisions using data?

- **Correct Answer: b) Hypothesis**

Q8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

- **Correct Answer: a) 0**

Q9. Which of the following statements is incorrect with respect to outliers?

- **Correct Answer: c) Outliers cannot conform to the regression relationship**

Q10. What do you understand by the term Normal Distribution?

- **Answer:**
    - **Normal Distribution, often referred to as a Gaussian distribution, is a continuous probability distribution characterized by a symmetric bell-shaped curve. The curve is centered around the mean, where most observations cluster, and it has two key parameters: mean (μ\muμ) and standard deviation (σ\sigmaσ).**

- **Properties:**
  - **The mean, median, and mode of a normal distribution are all equal.**
  - **It is defined by its mean and standard deviation.**
  - **About 68% of data falls within one standard deviation of the mean, 95% within two, and 99.7% within three (empirical rule).**
  - **It is commonly used in statistical analyses because many natural phenomena and measurement errors tend to follow this distribution.**
- **The probability density function (PDF) of a normal distribution is given by:**
  f(x)=1σ2πe−12(x−μσ)2f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}f(x)=σ2π1e−21(σx−μ)2
- **Normal distributions are central to many statistical methods and inferences, such as hypothesis testing, confidence intervals, and in the application of the Central Limit Theorem.**

Q11. How do you handle missing data? What imputation techniques do you recommend?

- **Answer:**
  - **Handling missing data is a crucial part of data preprocessing in statistical analysis and machine learning. Here are some common methods for handling missing data:**
    - **Removing Missing Data:**
      - **Listwise Deletion: Remove rows with any missing values. Simple but can lead to loss of valuable data.**
      - **Pairwise Deletion: Use all available data without deleting entire rows.**
    - **Imputation Techniques:**
      - **Mean Imputation: Replace missing values with the mean of the column. Simple but can distort variance.**
      - **Median Imputation: Replace missing values with the median. Useful for skewed data.**
      - **Mode Imputation: Replace missing values with the mode. Suitable for categorical data.**
      - **K-Nearest Neighbors (KNN): Impute based on the nearest data points in the dataset. More sophisticated and takes data distribution into account.**
      - **Multiple Imputation: Use algorithms to generate multiple imputations and average results, preserving uncertainty.**
      - **Regression Imputation: Predict missing values using regression models based on other variables.**
      - **Hot Deck Imputation: Impute missing values using similar records from the same dataset.**
  - **Recommended Approach:**
    - **The choice of imputation technique depends on the data type, missing data pattern, and analysis goals. For most scenarios, starting with mean/median imputation for numerical data and mode for categorical data is practical. For more complex datasets, KNN or Multiple Imputation methods are recommended to better capture data variability.**

Q12. What is A/B testing?

- **Answer:**
  - **A/B Testing is a statistical method used to compare two versions of a variable to determine which one performs better. It involves random experiments with two variants, A and B, which are the control and treatment groups, respectively.**
  - **Process:**
    - **Hypothesis: Formulate a hypothesis to test. For example, "Version B of a webpage will have a higher conversion rate than Version A."**
    - **Experiment Design: Randomly divide the audience into two groups.**
    - **Data Collection: Implement both versions and collect performance data.**
    - **Analysis: Analyze the results using statistical methods to evaluate the significance of the differences observed.**
    - **Conclusion: Determine if the differences are statistically significant to draw conclusions.**
  - **Applications:**
    - **Used extensively in marketing, user experience (UX) design, and website optimization to evaluate changes' impact on metrics such as conversion rates, click-through rates, and engagement levels.**

Q13. Is mean imputation of missing data an acceptable practice?

- **Answer:**
  - **Mean imputation is a basic method where missing values are replaced with the mean of the available data for that variable. While it is easy to implement, it has several limitations and is generally not recommended as the best practice in all situations.**
  - **Advantages:**
    - **Simple to implement and quick for large datasets.**
    - **Maintains the sample size for analysis.**
  - **Disadvantages:**
    - **Reduces variance in the dataset and can lead to biased estimates.**
    - **Does not account for relationships between variables and can distort statistical tests.**
    - **Potentially affects the correlation structure.**
  - **Acceptability:**
    - **Mean imputation is acceptable when the missing data is minimal and randomly distributed. However, more sophisticated methods like Multiple Imputation or K-Nearest Neighbors are preferred to maintain data integrity and variability in more complex datasets.**

Q14. What is linear regression in statistics?

- **Answer:**
  - **Linear regression is a statistical technique used to model and analyze the relationship between two or more variables. The primary aim is to understand the linear relationship between a dependent variable (target) and one or more independent variables (predictors).**

- **Types:**
  - **Simple Linear Regression: Involves one independent variable. The relationship is modeled as a linear equation: $Y=\beta0+\beta1X+\epsilon$ Where YYY is the dependent variable, XXX is the independent variable, β0\beta_0β0 is the intercept, β1\beta_1β1 is the slope, and ϵ\epsilonϵ is the error term.**
  - **Multiple Linear Regression: Involves multiple independent variables. The relationship is modeled as: $Y=\beta0+\beta1X1+\beta2X2+…+\betanXn+\epsilon$**
- **Assumptions:**
  - **Linearity: The relationship between dependent and independent variables is linear.**
  - **Independence: Observations are independent.**
  - **Homoscedasticity: Constant variance of errors.**
  - **Normality: Errors are normally distributed.**
- **Applications:**
  - **Predictive modeling for business, finance, economics, and many scientific fields where relationships between variables need exploration and prediction.**

Q15. What are the various branches of statistics?

- **Answer:**
  - **Statistics can be broadly categorized into two major branches:**
    - **Descriptive Statistics:**
      - **Focuses on summarizing and describing the features of a data set.**
      - **Includes measures such as mean, median, mode, variance, standard deviation, and graphical representations like histograms and box plots.**
      - **Purpose: Provide a quick overview of data characteristics and patterns.**
    - **Inferential Statistics:**
      - **Involves making predictions or inferences about a population based on a sample of data.**
      - **Utilizes probability theory to evaluate hypotheses and derive conclusions.**
      - **Includes methods like hypothesis testing, confidence intervals, regression analysis, and ANOVA.**