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

Ans. a) True

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

Ans. a) Central Limit Theorem

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

Ans. b) Modeling bounded count data

4. Point out the correct statement

Ans. c) The square of a standard normal random variable follows what is called chi-squared distribution

5. _____ random variables are used to model rates.

Ans. c) Poisson

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

Ans. b) False

7. Which of the following testing is concerned with making decisions using data? a) Probability

Ans. b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans. a) 0

9. Which of the following statement is incorrect with respect to outliers?

Ans. c) Outliers cannot conform to the regression relationshi

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

Ans. The normal distribution, often referred to as the Gaussian distribution, is a continuous probability distribution that is symmetrical around its mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, the normal distribution will appear as a bell curve.

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

Ans. Handling missing data is crucial in data analysis to avoid bias and loss of information. Here are some common techniques for handling missing data:

1. **Deletion**: Remove observations with missing data. This can be done if missing data is small and random, but it reduces sample size.
2. **Mean/Median/Mode Imputation**: Replace missing values with the mean, median, or mode of the observed data for that variable. This is simple but may distort the statistical properties of the data.
3. **Hot Deck Imputation**: Replace missing values with values from similar observations based on some distance metric (e.g., Euclidean distance).
4. **Multiple Imputation**: Generate multiple plausible values for each missing value, considering the uncertainty around the imputations. This involves creating several datasets with different imputed values and averaging results.
5. **Regression Imputation**: Predict missing values using a regression model based on other variables that are not missing.
6. **K-nearest neighbors (KNN) Imputation**: Replace missing values with the average of nearest neighbors' values.

The choice of imputation technique depends on the nature of the data and the extent of missingness. Multiple imputation is often preferred as it captures the uncertainty associated with missing data, but it requires more computational effort.

12. What is A/B testing?

Ans. A/B testing, also known as split testing, is a method used in marketing, product development, and user experience design to compare two versions (A and B) of a webpage, app, email, or other digital asset to determine which one performs better. It involves randomly assigning users to either version A or B and measuring key metrics such as conversion rates, click-through rates, or user engagement. Statistical analysis is then used to determine if any observed differences in performance are statistically significant, helping businesses make data-driven decisions to optimize their offerings and achieve better outcomes.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation, where missing values are replaced with the mean of the observed data, is a commonly used method for handling missing data due to its simplicity. However, its acceptability depends on the context and the nature of the data being analyzed. Here are some considerations:

7. **Impact on Variability**: Mean imputation can artificially reduce the variance of the dataset because it assigns the same value (the mean) to all missing values. This can lead to underestimation of standard errors and confidence intervals.
8. **Distortion of Relationships**: Mean imputation can distort relationships between variables, especially if missingness is not completely random. It assumes that the missing data have the same mean as the observed data, which may not always be the case.
9. **Applicability**: Mean imputation is more suitable for variables where missing data are few and missing completely at random (MCAR) or missing at random (MAR). If data are missing not at random (MNAR), where the missingness is related to the unobserved value itself, mean imputation may bias results.
10. **Alternatives**: Depending on the data and context, other imputation methods such as multiple imputation, regression imputation, or nearest neighbor imputation may be more appropriate as they can preserve the distributional characteristics of the data and account for uncertainty.

In conclusion, while mean imputation is widely used due to its simplicity, it should be applied cautiously and with an understanding of its implications on data analysis and interpretation. It may be acceptable in certain contexts with appropriate justification, but

researchers and analysts should consider alternative methods depending on the specific characteristics of their data and the goals of their analysis.

14. What is linear regression in statistics?
Ans. Linear regression is a statistical method used to study the relationship between two continuous variables. It models the relationship between a dependent variable (often denoted as YYY) and one or more independent variables (often denoted as XXX) by fitting a linear equation to observed data. The goal of linear regression is to find the best-fitting straight line through the data points that minimizes the sum of squared residuals (the vertical distances between the observed data points and the fitted line).

15. What are the various branches of statistics?

Ans. Statistics is a broad field with several branches, each focusing on different aspects of data collection, analysis, interpretation, and presentation. The main branches of statistics include:

11. **Descriptive Statistics**:
    - **Definition**: This branch deals with summarizing and describing the features of a dataset.
    - **Key Concepts**: Measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and graphical representations (histograms, bar charts, pie charts).
12. **Inferential Statistics**:
    - **Definition**: This branch focuses on making inferences and predictions about a population based on a sample of data.
    - **Key Concepts**: Hypothesis testing, confidence intervals, p-values, significance levels, and various statistical tests (t-tests, chi-square tests, ANOVA).
13. **Probability Theory**:
    - **Definition**: This branch provides the mathematical foundation for statistics, dealing with the likelihood of different outcomes.
    - **Key Concepts**: Probability distributions (normal, binomial, Poisson), random variables, expected value, variance, and the law of large numbers.
14. **Biostatistics**:
    - **Definition**: This branch applies statistical methods to biological and health sciences.

- **Key Concepts**: Clinical trials, epidemiology, survival analysis, and bioinformatics.

15. **Econometrics**:
    - **Definition**: This branch combines economic theory with statistical methods to analyze economic data.
    - **Key Concepts**: Regression analysis, time series analysis, panel data, and forecasting.
16. **Experimental Design**:
    - **Definition**: This branch focuses on designing experiments to ensure that the data collected can provide valid and objective conclusions.
    - **Key Concepts**: Randomization, control groups, factorial designs, and blocking.
17. **Time Series Analysis**:
    - **Definition**: This branch deals with analyzing data collected over time to identify trends, seasonal patterns, and cyclical behavior.
    - **Key Concepts**: Autocorrelation, moving averages, ARIMA models, and exponential smoothing.
18. **Multivariate Statistics**:
    - **Definition**: This branch involves analyzing data that includes more than one variable to understand relationships and patterns.
    - **Key Concepts**: Principal component analysis (PCA), factor analysis, cluster analysis, and discriminant analysis.
19. **Bayesian Statistics**:
    - **Definition**: This branch incorporates prior knowledge or beliefs into the statistical analysis using Bayes' theorem.
    - **Key Concepts**: Prior and posterior distributions, Bayesian inference, Markov Chain Monte Carlo (MCMC) methods.
20. **Nonparametric Statistics**:
    - **Definition**: This branch deals with methods that do not assume a specific distribution for the data.
    - **Key Concepts**: Rank tests, sign tests, and nonparametric regression.

These branches of statistics provide a comprehensive toolkit for analyzing data in various fields, helping to extract meaningful insights and make informed decisions.