Q1-Q9: Single Correct Answer

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

- a) True

- b) False

Answer: a) True

A Bernoulli random variable is a discrete random variable that takes only two possible values, typically 1 (for success) and 0 (for failure).

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?**

- a) Central Limit Theorem

- b) Central Mean Theorem

- c) Centroid Limit Theorem

- d) All of the mentioned

Answer: a) Central Limit Theorem

The Central Limit Theorem (CLT) states that the distribution of the sample mean of independent and identically distributed (iid) random variables approaches a normal distribution as the sample size becomes large.

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

- a) Modeling event/time data

- b) Modeling bounded count data

- c) Modeling contingency tables

- d) All of the mentioned

Answer: b) Modeling bounded count data

The Poisson distribution is used to model count data that is not bounded, such as the number of events occurring in a fixed interval of time or space. It is not typically used for bounded count data.

4. **Point out the correct statement.**

- a) The exponent of a normally distributed random variable follows what is called the log-normal distribution

- b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

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

- d) All of the mentioned

Answer: d) All of the mentioned

All statements are correct:

- The exponent of a normally distributed random variable follows a log-normal distribution.

- Sums of normally distributed random variables are normally distributed regardless of dependence.

- The square of a standard normal random variable follows a chi-squared distribution.

5. **_____ random variables are used to model rates.**

- a) Empirical

- b) Binomial

- c) Poisson

- d) All of the mentioned

Answer: c) Poisson

The Poisson distribution is used to model the rate at which events occur in a fixed interval of time or space.

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

- a) True
- b) False

Answer: b) False

The Central Limit Theorem (CLT) is based on the idea that the distribution of sample means approaches a normal distribution regardless of the standard error estimate used.

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

- a) Probability
- b) Hypothesis
- c) Causal
- d) None of the mentioned

Answer: b) Hypothesis

Hypothesis testing is a statistical method used to make decisions about data by testing a null hypothesis against an alternative hypothesis.

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

- a) 0
- b) 5

- c) 1

- d) 10

Answer: a) 0

Normalized data, often resulting from standardization, are centered at 0 with a standard deviation of 1.

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

- a) Outliers can have varying degrees of influence

- b) Outliers can be the result of spurious or real processes

- c) Outliers cannot conform to the regression relationship

- d) None of the mentioned

Answer: c) Outliers cannot conform to the regression relationship

Outliers can conform to the regression relationship but may significantly influence the results or indicate deviations from the norm.

Q10-Q15: Subjective Answer Type Questions

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

Answer: The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution characterized by its symmetric bell-shaped curve. It is defined by its mean (μ) and standard deviation (σ). About 68% of the data falls within one standard deviation of the mean, 95% within two, and 99.7% within three standard deviations. It is widely used in statistics due to its properties and the Central Limit Theorem.

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

Answer: Handling missing data can be approached in several ways:

- **Deletion:** Removing records with missing values.

- **Mean/Median/Mode Imputation:** Replacing missing values with the mean, median, or mode of the observed values.

- **Interpolation:** Estimating missing values based on the trends in the data.

- **Predictive Modeling:** Using algorithms to predict missing values based on other available data.

- **Multiple Imputation:** Creating multiple datasets with different imputations and combining the results.

Recommendations depend on the nature and extent of the missing data, but mean or median imputation is common for simple cases, while more complex methods are used for intricate data patterns.

12. **What is A/B testing?**

Answer: A/B testing, also known as split testing, is a method used to compare two versions of a variable to determine which performs better. In a typical A/B test, two variants (A and B) are randomly presented to users, and their performance is measured using predefined metrics. The goal is to determine which variant leads to better outcomes, such as higher conversion rates or user engagement.

13. **Is mean imputation of missing data acceptable practice?**

Answer: Mean imputation is a common practice for handling missing data, especially when the missing data is missing at random. It is simple and maintains the sample size. However, it can introduce bias, reduce variance, and may not be suitable if

the data is not missing at random or if the proportion of missing data is high. For more robust solutions, other imputation techniques or models might be preferred.

14. **What is linear regression in statistics?**

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables. The goal is to find the best-fitting line (or hyperplane) that minimizes the sum of squared differences between observed values and predicted values. The linear regression model is represented by the equation $Y = \beta_0 + \beta_1X + \epsilon$, where β_0 is the intercept, β_1 is the slope, and ϵ represents the error term.

15. **What are the various branches of statistics?**

Answer: The main branches of statistics include:

- **Descriptive Statistics:** Involves summarizing and organizing data to describe its main features, including measures like mean, median, mode, and standard deviation.

- **Inferential Statistics:** Involves making predictions or inferences about a population based on a sample of data, using techniques such as hypothesis testing and confidence intervals.

- **Probability Theory:** Studies the likelihood of different outcomes and the mathematical foundation for statistical inference.

- **Regression Analysis:** Focuses on modeling and analyzing relationships between variables.

- **Bayesian Statistics:** Uses Bayes' theorem to update the probability of a hypothesis as more evidence or information becomes available.

- **Multivariate Statistics:** Deals with the analysis of data involving multiple variables to understand relationships and patterns.