Statistics

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) True

b) False

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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans. a) Central Limit Theorem

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

Ans. b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables 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

Ans. d) All of the mentioned

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans c) Poisson

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

a) True

b) False

Ans. b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Ans. b) Hypothesis

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

a) 0

b) 5

c) 1

d) 10

Ans. a) 0

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

An**s. c) Outliers cannot conform to the regression relationship**

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

The normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the 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. The mean, median, and mode of the distribution are all equal. The standard normal distribution has a mean of 0 and a standard deviation of 1.

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

- **Deletion Methods**: Removing rows or columns with missing values.
- **Imputation Methods**: Filling in missing values using:
    - **Mean/Median/Mode Imputation**: Replacing missing values with the mean, median, or mode of the column.
    - **Forward/Backward Fill**: Using preceding or succeeding values to fill in missing data.
    - **Interpolation**: Using linear or polynomial interpolation to estimate missing values.
    - **K-Nearest Neighbors (KNN)**: Using the nearest neighbors to impute missing values.
    - **Multivariate Imputation by Chained Equations (MICE)**: Creating multiple imputations for the missing data.

The choice of technique depends on the nature of the data and the extent of the missing values.

12. What is A/B testing?

A/B testing, also known as split testing, is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants (A and B) are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a simple and commonly used technique, but it has limitations. While it can be useful for maintaining dataset size, it can also distort the data distribution, underestimate variability, and weaken relationships between variables. It is often better to use more sophisticated methods like multiple imputation, KNN imputation, or model-based methods to preserve the statistical properties of the dataset.

14. What is linear regression in statistics?

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. The simplest form, simple linear regression, involves a single independent variable and models the relationship using a straight line. The equation for a simple linear regression line is $y=mx+c$, where y is the dependent variable, x is the independent variable, m is the slope, and c is the y-intercept. Multiple linear regression extends this concept to include multiple independent variables.

15. What are the various branches of statistics?

The various branches of statistics include:

- **Descriptive Statistics**: Summarizing and describing the features of a dataset.
- **Inferential Statistics**: Drawing conclusions from data that are subject to random variation.