STATISTICS WORKSHEET-1

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. 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. 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. 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. All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans. poisson

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

a) True

b) False

Ans. 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. 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. 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

Ans. 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?

Ans. Normal distribution, also known as the Gaussian distribution, is a symmetric, bell-shaped distribution where most of the observations cluster around the central peak (mean). The further from the mean, the less likely observations become. It is fully described by its mean and standard deviation.

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

Ans. There are some techniques which are mostly used to handle missing data from a dataset which are following:

1. **Mean/Median Imputation: Replacing missing values with the mean or median of the column.**
2. **Mode Imputation: For categorical data, replacing missing values with the mode (most frequent value).**
3. **Forward or Backward Fill: In time-series data, using the previous or next known value to fill missing data.**
4. **K-Nearest Neighbors (KNN) Imputation: Using neighboring data points to estimate missing values.**
5. **Multiple Imputation: Generating multiple possible values for missing data and combining the results.**

12. What is A/B testing?

Ans. A/B testing, also known as split testing, is a method of comparing two versions of a variable (A and B) to determine which one performs better. It is widely used in marketing, web design, and other fields to test changes to a webpage, product, or strategy against a control to see if the change improves outcomes.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is acceptable in certain situations. It is generally used for numerical data when the proportion of missing values is small. However, it may not be appropriate if the data are not missing completely at random or if the mean is not representative of the population.

14. What is linear regression in statistics?

Ans. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the linear equation that best predicts the dependent variable.

15. What are the various branches of statistics?

Ans. The major branches are:

1. **Descriptive Statistics: Summarizing and organizing data using measures like mean, median, mode, variance, etc.**
2. **Inferential Statistics: Drawing conclusions from data samples and making predictions about populations, using methods such as hypothesis testing, confidence intervals, and regression analysis.**
3. **Probability Theory: The mathematical framework for quantifying uncertainty and making predictions about random events.**
4. **Bayesian Statistics: A branch of statistics where probability expresses a degree of belief in an event, updated as more evidence or information becomes available.**