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: 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) Modelling 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. 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. 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

Ans: 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?

Ans: Normal distribution refers to the probability distribution where the values of a random variable are distributed symmetrically, that means the data points occur within a small range of values with fewer outliers on the high and low ends of the data range thus forming a bell-shaped curve.

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?

Ans: There are several ways to deal with missing data. Based on data we can decide.

If the portion of missing data is low, we can use the fillna() method or by imputation techniques we can predict the missing values and fill them.

If the portion of missing data is too high, the results lack natural variation, so its better to drop the entire column. when we have a huge dataset deleting 1-2 column will not make much difference to the output.

Based on the dataset I can recommend the imputation techniques. It can be Mean, median or mode imputation, knn imputer or iterative imputer can be used.

12. What is A/B testing?

Ans: A/B testing is a statistical method to determine which set of variants performs better based on a given metric. It is also called a controlled experiment or a randomized control trial.

13. Is mean imputation of missing data acceptable practice?

Ans: It is not a good solution, as it ignores the distribution and correlation of the data, hence potentially create unrealistic values.

14. What is linear regression in statistics?

Ans: Linear regression is a type of statistical analysis which is used to predict the value of a variable based on the value of other variables.

The variable we want to predict is called label(dependent variable) and the variables we are using to predict it are called features(Independent variables).

15. What are the various branches of statistics?

Ans: The two main branches of statistics are

Descriptive statistics:

It is the branch of statistics that involves the organization, summarization and display of data.

Inferential statistics:

It is the branch of statistics that involves drawing inferences/conclusion using sample data. It performs estimations and hypothesis tests to determine relationships among variables and make predictions.