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

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

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

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]

5. _______ random variables are used to model rates.

a) Empirical

b) Binomial

[✓ c) Poisson]

d) All of the mentioned

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

a) True

[✓ 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

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

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

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

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

Answer: In a normal distribution, the mean, median and mode are the exact same. It is a proper term of probability bell curve; half the values fall above the mean and the remaining half falls below the mean i.e. normal distribution is symmetric around the mean. In a normal distribution, the mean is zero and the standard deviation is 1 with zero skew.

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

Answer: When working with huge dataset, dealing with the missing data is a common issue during data collection. One of the easiest ways is to remove the entire row of the missing data, but this can lead to losing of valuable information another way is the imputation method i.e. replacing the missing values by the mean, median, mode and KNN of the relevant feature, but the estimated values might not be accurate. Every imputation technique has its own advantages as well as disadvantages so is the multiple imputation technique is one of the techniques that can be better than the single imputation. Multiple imputation is not an easy technique for the missing data.

12. What is A/B testing?

Answer: A/B testing is a way to evaluate the performance of two variant of a feature to find out which variant will perform better in a controlled environment. In this testing, we compare the output before and after the modifications are done so that adjustment done can have a positive impact on the result.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is a technique in which the null values are replaced with the data's mean value in a dataset, this reduces the variance of the dataset making it less accurate of the actual population and will also underestimate all the standard errors as well as it destroys the relation between the variables. Let's consider an example of the height and weight dataset of people, some times the people intentionally will not mention the weight, after the mean imputation of the height and weight, their positive correlation will be weaker. The chances of producing the bias result is much higher when we predict the mean value by using data values.

14. What is linear regression in statistics?

Answer: Linear regression in statistics is an analysis to find the out the value of a feature based on the value of another feature i.e. it is game between the dependent continuous variable and independent continuous variable where we can use linear regression to develop a more formal understanding relationship between the variables.

15. What are the various branches of statistics?

Answer: DESCRIPTIVE STATISTICS: This is usually the first part of an analysis in statistics. This deals with the collection of the data and how to present the data.

INFERENTIAL STATISTICS: this is the next part, drawing inferences based on the data collected, summarised in the descriptive statistics .This involves using a sample to draw conclusions about a population