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

   Answer is A

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 is A

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 is B

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

   Answer is D

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

   Answer is C

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

   Answer is B

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 is B

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 is A

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 is D

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

Q10 What do you understand by the term Normal Distribution?

Answer: It is a set of continuous variable spread across a normal curve or in the shape of a bell curve. It can be considered as a continuous probability distribution and is useful in statistics. It is the most common distribution curve and it becomes very useful to analyze the variables and their relationships when we have the normal distribution curve.

The normal distribution curve is symmetrical. The non-normal distribution approaches the normal distribution as the size of the samples increases. It is also very easy to deploy the Central Limit Theorem. This method helps to make sense of data that is random by creating an order and interpreting the results using a bell-shaped graph.

Q11 How do you handle missing data? What imputation techniques do you recommend?

Answer: Handling missing data is an important step in data preprocessing, as many machine learning algorithms cannot handle missing values. There are several common techniques for handling missing data, including:

1. Deleting rows or columns: If the amount of missing data is small, it may be possible to simply delete the rows or columns with missing data. However, this approach can lead to a loss of information and may not be appropriate if a large amount of data is missing.

2. Mean or median imputation: In this approach, missing values are replaced with the mean or median value of the non-missing values in the same column. This method is simple and easy to implement, but it can lead to biased results if the missing data is not missing at random.

3. Mode imputation: This method replaces missing values with the most common value in the column. It is commonly used for categorical variables.

4. Regression imputation: In this approach, a regression model is used to predict the missing values based on the other variables in the dataset. This method can be more accurate than mean or median imputation, but it requires a significant amount of computational resources.

5. Multiple imputation: This method involves creating multiple imputed datasets based on the observed data, and then analysing each imputed dataset separately. This approach can be more accurate than the other methods, but it is also computationally intensive.

It is important to choose the appropriate imputation method based on the nature of the data and the problem at hand. It is also important to carefully evaluate the imputed data to ensure that the imputation method does not introduce bias or inaccuracies in the analysis. Additionally, it is recommended to report the proportion of missing data, the imputation method used, and any assumptions made in the analysis to ensure transparency in the results.

Q12 What is A/B testing?

Answer: A/B testing—also called split testing or bucket testing—compares the performance of two versions of content to see which one appeals more to visitors/viewers. It tests a control (A) version against a variant (B) version to measure which one is most successful based on your key metrics.

Q13 Is mean imputation of missing data acceptable practice?

Answer: The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly

greater fitness level than he actually does.
Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Q14 What is linear regression in statistics?

Answer: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Q15 What are the various branches of statistics?

Answer: Statistics: Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics

Inferential Statistic.

Descriptive Statistic.

Inferential Statistics: Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population.

Descriptive Statistics: Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.