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

   a) True

   b) False

   Ans. A) True

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.

10. What do you understand by the term Normal distribution?
    Ans.
    The normal distribution is also called the Gaussian distribution, de Moivre distribution. It is symmetric about its centre as half of data falls to the left of the mean (average) and half falls to the right.
    It represents the symmetric Bell shaped curve. Most of the values are located near the mean. Also the Mean , Median and Mode are equal in Normal distribution.

11. How do you handle missing data? What imputation techniques do you recommend?
    Ans.
    When dealing with missing data, we can use two methods to solve the error: imputation or data removal.
    The imputation method substitutes reasonable guesses for missing data. It's most useful when the missing data is low.
    The other option is to remove data. When dealing with data that is missing at random, the entire data point that is missing information can be deleted to help reduce bias.

    Some imputation techniques are:

- Next or Previous Value : For time-series data or ordered data, The next or previous value inside the time series is typically substituted for the missing value as part of a common method for imputed incomplete data in the time series. This strategy is effective for both nominal and numerical values.

- K Nearest Neighbours: The objective is to find the k nearest examples in the data where the value in the relevant feature is not absent and then substitute the value of the feature that occurs most frequently in the group.

- Maximum or Minimum Value: we can use the minimum or maximum of the range as the replacement for missing values if you are aware that the data must fit within a specific range.

- Missing Value Prediction: Using a machine learning model to determine the final imputation value for characteristic x based on other features is another popular method for single imputation.

- Most Frequent Value: The most frequent value in the column is used to replace the missing values in another popular technique that is effective for both nominal and numerical features.

- Average or Linear Interpolation: The average or linear interpolation, which calculates between the previous and next accessible value and substitutes the missing value, is similar to the previous/next value imputation but only applicable to numerical data.

- (Rounded) Mean or Moving Average or Median Value: In this technique we can replace the null values with mean, rounded mean, or median values determined for that feature across the whole dataset.

- Fixed Value: Fixed value imputation is a universal technique that replaces the null data with a fixed value and is applicable to all data types.

12. What is A/B testing?
    Ans.
    A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element,

etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. Is mean imputation of missing data acceptable practice?
    Ans.
    Mean imputation is not a good solution. Mean imputation does not preserve the relationships among variables. It ignores feature correlation. It 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.

14. What is linear regression in statistics?
    Ans.
    Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line.

15. What are the various branches of statistics?
    Ans.
    The two branches of statistics are descriptive statistics and inferential statistics.
    - **Descriptive Statistics:** Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. It can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations. Descriptive statistics can be categorized into
        - Measures of central tendency
        - Measures of variability

      To easily understand the analysed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.

    - **Inferential Statistics:** Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. It often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyse data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.
      The different types of calculation of inferential statistics include:
        - Regression analysis
        - Analysis of variance
        - Analysis of covariance
        - Statistical significance (t-test)
        - Correlation analysis