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 : c) Modeling contingency tables

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: c) The square of a standard normal random variable follows what is called chi squared distribution

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: b) 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: 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: The normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric and bell-shaped. It is widely used in statistics and data analysis to model a wide range of natural phenomena and measurements. In a normal distribution, the data is centered around the mean, with the majority of values falling close to the mean and fewer values further away. The distribution is characterized by its mean and standard deviation, which determine its shape and spread. The shape of the normal distribution is often referred to as a "bell curve" due to its characteristic shape. Many real-world phenomena, such as heights, weights, test scores, and errors, can be approximated by a normal distribution. Understanding the properties and characteristics of the normal distribution is important for various statistical analyses and modeling techniques.

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

Ans : Handling missing data is an important task in data analysis and machine learning. There are several techniques for handling missing data, and the choice depends on the nature of the data and the specific problem you're working on. Here are a few common imputation techniques that you can consider:

Mean/Median Imputation: In this technique, you replace missing values with the mean or median of the available data for that feature. This approach is simple and can work well when the missing data is assumed to be missing at random.

Mode Imputation: Mode imputation involves replacing missing categorical data with the mode (most frequent value) of the available data for that feature. This technique is commonly used for categorical variables.

Forward/Backward Fill: In time-series data, you can use forward fill or backward fill to replace missing values with the nearest preceding or succeeding value, respectively. This method assumes that the missing values are likely to be similar to the surrounding values.

K-Nearest Neighbors (KNN) Imputation: KNN imputation involves replacing missing values with the average of the K nearest neighbors based on other features. This technique is useful when there is a strong relationship between the missing feature and other features.

Multiple Imputation: Multiple imputation involves creating multiple imputed datasets by estimating missing values based on observed values and their relationships. Statistical techniques are then used to combine the results from these multiple datasets. This approach can provide more accurate and robust results.

It's important to note that the choice of imputation technique depends on various factors such as the type of data, the amount of missing data, and the analysis or modeling task at hand. It's recommended to evaluate the impact of different imputation techniques on your specific dataset and choose the one that best suits your needs.

12.What is A/B testing?

Ans: A/B testing, also known as split testing, is a method used to compare two versions of a webpage, app, or any other user experience to determine which version performs better. It is a powerful technique commonly used in marketing, product development, and user experience optimization.

The process involves dividing your audience into two groups: Group A and Group B. Group A is shown the original version (control group), while Group B is shown a modified version (experimental group) with a specific change, such as a different layout, color scheme, or call-to-action button. The goal is to measure which version yields better results, such as higher click-through rates, conversion rates, or engagement metrics.

By comparing the performance of both versions, you can make data-driven decisions and determine the impact of the specific change on user behavior. A/B testing allows you to optimize your designs, content, and user experiences based on real user data, leading to improved performance and better decision-making.

13.Is mean imputation of missing data acceptable practice?

Ans: Mean imputation of missing data is a commonly used practice in data analysis and machine learning. It involves replacing missing values in a dataset with the mean value of the available data for that particular feature. However, while mean imputation is easy to implement and can provide a quick solution to handling missing data, it has certain limitations that you should be aware of.

One limitation is that mean imputation assumes that the missing values are missing completely at random (MCAR). In other words, it assumes that the reason for the missing values is unrelated to the values themselves or any other variables in the dataset. If this assumption does not hold true, mean imputation may introduce bias into the dataset and lead to inaccurate results.

Another limitation is that mean imputation does not account for any relationships or patterns in the data. It treats all missing values the same way, regardless of their context or the relationships they may have with other variables. This can result in a loss of valuable information and potentially impact the accuracy of your analysis or model.

There are alternative methods to handle missing data, such as multiple imputation, which generates multiple plausible values for missing data based on observed data and relationships. Other methods include regression imputation, where missing values are estimated based on relationships with other variables, or using machine learning algorithms to predict missing values.

In conclusion, mean imputation can be a quick and simple solution for handling missing data, but it has limitations. It's important to consider the assumptions and potential biases it introduces and to explore alternative methods if they are more appropriate for your specific use case.

14.What is linear regression in statistics?

Ans: Linear regression is a statistical modeling technique used to understand and analyze the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that the change in the dependent variable is directly proportional to the change in the independent variable(s).

In simpler terms, linear regression helps us determine how a change in one variable affects another variable. It allows us to make predictions based on the relationship between the variables. The goal of linear regression is to find the best-fit line that minimizes the distance between the observed data points and the predicted values.

Linear regression is widely used in various fields, including economics, finance, social sciences, and machine learning. It provides insights into the strength and direction of the relationship between variables, helps in forecasting or predicting outcomes, and can be used for hypothesis testing as well.

15.What are the various branches of statistics?

Ans: Statistics is a vast field with several branches that focus on different aspects of data analysis and interpretation. Here are some of the main branches of statistics:

Descriptive Statistics: This branch involves summarizing and describing data using measures such as mean, median, mode, standard deviation, and variance. It provides a snapshot of the data and helps in understanding its basic properties.

Inferential Statistics: Inferential statistics involves drawing conclusions or making predictions about a population based on a sample. It utilizes probability theory and hypothesis testing to make inferences about the larger population.