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 : 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
   Answer : 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
   Answer: 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
   Answer : c) The square of a standard normal random variable follow 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
   Answer : c) Poisson
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer : 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
   Answer : b) Hypothesis
8. 4. 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 : 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

Answer : 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?

Answer : Normal Distribution, is a continuous probability distribution that is commonly used to model random variables with a bell-shaped probability density function.

The normal distribution has several important properties, including the 68-95-99.7 rule, which states that approximately 68% of the observations fall within one standard deviation of the mean, 95% of the observations fall within two standard deviations of the mean, and 99.7% of the observations fall within three standard deviations of the mean.

Many natural phenomena, such as heights, weights, and IQ scores, tend to follow a normal distribution. The central limit theorem also states that the distribution of averages of independent and identically distributed random variables becomes approximately normal as the sample size increases.

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

Answer : Handling missing data is an important aspect of data analysis as it can affect the accuracy and validity of the results. Here are some techniques to handle missing data:

Deletion: One way to handle missing data is to simply remove any records with missing data. This method is called list-wise deletion or complete case analysis. However, this method can result in a loss of data and may not be feasible when the amount of missing data is substantial.

Imputation: Another approach is to impute or estimate the missing values. Imputation can be done using several techniques such as:

Mean/median imputation: This involves replacing the missing values with the mean or median of the available data.

Mode imputation: This involves replacing the missing values with the mode or most frequently occurring value in the available data.

The choice of imputation technique depends on the type and amount of missing data, the distribution of the data, and the analysis goals. It is important to carefully consider the implications of missing data and to choose a technique that is appropriate for the specific dataset and analysis.

12. What is A/B testing?

Answer:

A/B testing is a statistical hypothesis testing technique used to compare two versions of a product or service to determine which one performs better. In this technique, a sample group is divided into two groups randomly, where one group (called the control group) is shown the original version of the product or service, and the other group (called the treatment group) is shown the modified version. By comparing the performance of both groups, the experimenter can determine which version is better. A/B testing is commonly used in web design, marketing, and product development to optimize the user experience and maximize conversion rates.

13. Is mean imputation of missing data acceptable practice?

Answer:

Mean imputation of missing data is a common technique used to handle missing values. However, it has some limitations and may not always be the best approach.

One major limitation is that mean imputation assumes that the missing values are missing at random (MAR). This means that the probability of a value being missing depends only on observed data, and not on unobserved data. If data are missing not at random (MNAR), the mean imputation method can introduce bias and distort the analysis results.

Another limitation is that mean imputation reduces the variability of the data, which can affect statistical tests and confidence intervals. It can also lead to underestimation of standard errors and overestimation of statistical significance.

Therefore, while mean imputation can be a useful method for handling missing data, it is important to

carefully consider the underlying assumptions and limitations, and to explore alternative imputation techniques such as multiple imputation or regression imputation.

14. What is linear regression in statistics?

Answer: Linear regression is a statistical method that is used to model the relationship between two continuous variables, where one variable is considered as the dependent variable and the other is the independent variable. The relationship between these two variables is modeled as a linear equation with a slope and an intercept, where the slope represents the change in the dependent variable for every unit change in the independent variable.

The goal of linear regression is to estimate the coefficients of the linear equation that best describes the relationship between the two variables. This is done by minimizing the sum of the squared differences between the observed values of the dependent variable and the predicted values from the linear equation.

Linear regression can be used for both simple and multiple regression problems. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are multiple independent variables. Linear regression is a commonly used technique in various fields including finance, economics, social sciences, and engineering.

15. What are the various branches of statistics?

Answer:

Statistics can be broadly divided into two main branches:

Descriptive Statistics: It involves summarizing and describing the main features of a dataset. The main aim is to give a brief summary of the data in a meaningful way. Descriptive statistics includes measures like mean, median, mode, standard deviation, etc.

Inferential Statistics: It involves making inferences about a larger population based on the information obtained from a sample. The main aim is to draw conclusions about the population based on the sample data. Inferential statistics includes techniques like hypothesis testing, confidence intervals, regression analysis, etc.