WORKSHEET

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) Modelling 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: - d) All of the mentioned

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 (null)

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

WORKSHEET

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, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

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

Answer:- **Missing values can be handled in two ways:**

Replace the missing values with the mean value of that particular feature.

Arithmetic mean

It is the average of a group of numbers. Applicable for interval and ratio data. Not applicable for nominal and ordinal data

Weighted mean

In a dataset, we might want to assign more importance, or weight, to some of the numbers. Weighted average = Sum of the product of data & weight/Sum of weights

Drop the feature/ column is it is not significant while building the algorithm

Imputation techniques:- Generally used are,

1. **Mean or Median Imputation**

2. **Multivariate Imputation by Chained Equations (MICE).**

3. Random Forest.

12. What is A/B testing?

Answer:- 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?

Answer:- Yes, Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean. This method can lead into severely biased estimates even if data are MCAR.

Mean imputation **reduces the variance of the imputed variables**. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval. Mean imputation does not preserve relationships between variables such as correlations.

14. 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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What are the various branches of statistics?

Answer:- The various Branches of statistics are; There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics**

1. Data collection is the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques.

Furthermore, there are two types of data collection methods, namely, **primary data collection, and secondary data collection methods**.

2. Descriptive statistics: - Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

3. Inferential Statistics:- Inferential statistics **use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects**. There are two main types of inferential statistics that use different methods to draw conclusions about the population data. These are regression analysis and hypothesis testing specific research design and sample characteristics.