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: 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

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

STATISTICS WORKSHEET-1

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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
Answer: A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer: Missing Data is handled by two techniques:

- Deletion - Pairwise Deletion, Listwise Deletion/ Dropping rows, Dropping complete columns
- Imputation - Imputation with a constant value, Imputation using the statistics (mean, median, mode) & K-Nearest Neighbor Imputation

12. What is A/B testing?
Answer: 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.

13. 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.

14. What is linear regression in statistics?
Answer: Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:
One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

15. What are the various branches of statistics?

Answer:
Branches Of Statistics

STATISTICS WORKSHEET-1

Descriptive Statistics: Descriptive statistics is the first part of statistics that deals with the collection of data. Descriptive statistics have two parts;

- Central tendency measures – Mean, Median, Mode
- Variability measures - The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

Inferential Statistics: Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. Different types of inferential statistics include:

Regression analysis: It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.

Analysis of variance (ANOVA): ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.

Analysis of covariance (ANCOVA): It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.

Statistical significance (t-test): It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.

Correlation analysis: It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.