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
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
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
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
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   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
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
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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

10.Answer:-

Normal distribution is a type of probability distribution where a set of data is distributed along a bell-shaped curve. The normal distribution is also known as the Gaussian distribution and is the most commonly seen distribution in nature. It is symmetric about the mean, with its peak at the mean and its spread determined by the standard deviation. The normal distribution is used extensively in statistics, to model real-world phenomena, and to make predictions about future outcomes.

11.Answer:-

Missing data can be handled using a variety of techniques, depending on the type of data and the context in which it is being used. Generally, imputation is the process of replacing missing data with substituted values. Common imputation techniques include mean/median/mode imputation, k-nearest neighbor (KNN) imputation, and multiple imputation. Mean/median/mode imputation involves replacing missing values with the mean, median, or mode of the rest of the data. This is a simple and straightforward technique, but it may not be appropriate for all datasets as it can distort the data distribution and reduce the interpretability of the results. KNN imputation is a more sophisticated technique that uses the similarity between entries to fill in missing values. This technique is better at preserving the data distribution and can produce more accurate results. Multiple imputation is a technique that involves creating multiple sets of data with different imputed values for the missing data. This technique can be used to generate better estimates of the missing data and can also be used to calculate the uncertainty associated with the imputed values.

12.Answer:-

A/B testing is a type of test used to measure the performance of two versions of a web page or app against one another. It involves randomly assigning users to view one version of a page or app (A) or the other (B) and then analyzing the results to determine which version performs better. It is commonly used to test changes to the design, content, or functionality of a page.

13.Answer:-

Mean imputation of missing data can be an acceptable practice in certain cases, such as when the data is highly skewed and there are few outliers. However, if the data is more normally distributed, other methods of imputation may be more appropriate.

14.Answer:-

Linear regression is a statistical method for predicting a continuous response variable (e.g., a price or a quantity) based on one or more predictor variables (e.g., age, income, etc.). It is based on the assumption that the relationship between the predictor and the response variable is linear. Linear regression can be used to estimate the effects of one or more predictor variables on the response variable. It can also be used to identify potential outliers or influential points in the data.

15.Answer:-

The various branches of statistics include:
1. Descriptive Statistics
2. Inferential Statistics
3. Probability and Distribution Theory
4. Regression Analysis
5. Time Series Analysis
6. Multivariate Analysis
7. Nonparametric Statistics
8. Design of Experiments
9. Sampling Theory
10. Stochastic Processes
11. Data Mining
12. Computational Statistics
13. Bayesian Statistics
14. Survival Analysis
15. Statistical Quality Control