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 - 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

Ans- d) All of the mentioned

5. _____ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans- c) 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. 4. Normalized data are cantered 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 is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions.

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

Ans - We either use the Bayesian approach of simply treating the missing data as latent variables and thus integrate them out, or use multiple imputation.

1. Mean, Median, Mode imputation - A simple guess of a missing value is the mean, median, or mode of that variable.

2. Regression imputation - If we know there is a correlation between the missing value and other variables, we can often get better guesses by regressing the missing variable on other variables.

3. K-nearest neighbour (KNN) imputation - For a discrete variable, KNN imputer uses the most frequent value among the k nearest neighbours and, for a continuous variable, use the mean or mode.

12. What is A/B testing?

Ans- An A/B test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

Ans- Yes, it is an acceptable practice. imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased. Also, by imputing the mean, you are able to keep your sample size up to the full sample size.

14. What is linear regression in statistics?

Ans- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

15. What are the various branches of statistics?

Ans- 1. Descriptive statistics.

2. Inferential statistics.