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
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?

Answer :- Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. There are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

Figure: Normal distribution in a bell curve

The random variables are distributed in the form of a symmetrical, bell-shaped curve. Properties of Normal Distribution are as follows;

1. Unimodal -one mode

2. Symmetrical -left and right halves are mirror images

3. Bell-shaped -maximum height (mode) at the mean

4. Mean, Mode, and Median are all located in the center

5. Asymptotic

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

Answer :- Handle missing data :-

1.) **Missing At Random (MAR) :-** Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of data. It is not known if the data should be there. The missing data can be predicted based on the complete observed data.

2.) **Missing Completely At Random (MCAR) :-** In the MCAR situation, the data is missing across all observation regardless of the expected value or other variables. Data scientists can compare two sets of data, one with missing observations and one without. Using a t-test, if there is no difference between the two data sets, the data is characterized as MCAR.

3.) **Missing Not At Random (MNAR) :-** The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing. In a survey, perhaps a specific group of people – say women ages 45 to 55 – did not answer a question. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown. Data scientists must model the missing data to develop an unbiased estimate. Simply removing observations with missing data could result in a model with bias.

Imputation Techniques :-

1.) **Complete Case Analysis (CCA) :-** This is quite straight forward method of handling the Missing Data, which directly removes the rows that have missing data. We consider only those rows where we have complete data. This method also popularly known as “Listwise deletion”.

2.) **Arbitrary Value Imputation :-** This is an important technique used in imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column.

3.) **Frequent Category Imputation :-** This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the mode of that column. This technique is also referred to as Mode Imputation.

12. What is A/B testing?

Answer :- A/B testing is the act of running a simultaneous experiment between two or more variants of a page to see which one performs the best . A/B testing is an online experiment conducted on a website, mobile application, to test potential improvements in comparison to a control, or original, version. A/B testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or “two-sample hypothesis testing”.

13. Is mean imputation of missing data acceptable practice?

Answer :- Two, imputation 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. Studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

Answer :- In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression. The process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

15. What are the various branches of statistics?

Answer :- There are three real branches of statistics :

1.) Data collection
2.) Descriptive statistics
3.) Inferential statistics

MCQ’S ANSWER’S

1.) (A) True
2.) (A) Central Limit Theorem
3.) (B) Modeling bounded count data
4.) (D) All of the mentioned
5.) (C) Poisson
6.) (B) False
7.) (B) Hypothesis
8.) (A) 0
9.) (C) Outliers cannot conform to the regression relationship