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 centered 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

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?
Ans - Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, 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.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans - There are several ways to handle missing values in the given data-

- **Dropping the values**
- **Deleting the observation (not always recommended)**
- **Replacing value with the mean, median and mode of the observation**
- **Predicting value with regression**
- **Finding appropriate value with clustering**

Mean or Median Imputation - When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.

Multivariate Imputation by Chained Equations (MICE) - MICE assumes that the missing data are Missing at Random (MAR). Random Forest

12. What is A/B testing?

Ans - It is a hypothesis testing for a randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for our business. It can be used to test everything from website copy to sales emails to search ads. An example of this could be identifying the click-through rate for a banner ad.

13. Is mean imputation of missing data acceptable practice?

Ans - This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, we can calculate the mean of the existing observations. However, when there are many missing variables, mean results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

14. What is linear regression in statistics?

Ans - Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression.

15. What are the various branches of statistics?

Ans - Statistics is a study of presentation, analysis, collection, interpretation and organization of data. There are two main branches of statistics - Inferential Statistic. - Descriptive Statistic. Inferential Statistics: Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population. Descriptive Statistics: Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.