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 : True

2. Which of the following theorem states that the distribution of averages of iid variables, properlynormalized, 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 : 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 : 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- normaldistribution
   b) Sums of normally distributed random variables are again normally distributed even if the variablesare dependent
   c) The square of a standard normal random variable follows what is called chi-squareddistribution
   d) All of the mentioned

Answer : All of the mentioned

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

Answer : Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Answer : 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 : Hypothesis

8. 4. Normalized data are centered at and have units equal to standard deviations of theoriginal data.
   a) 0
   b) 5
   c) 1
   d) 10

Answer : 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

Answer : Outliers cannot conform to the regression relationship

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

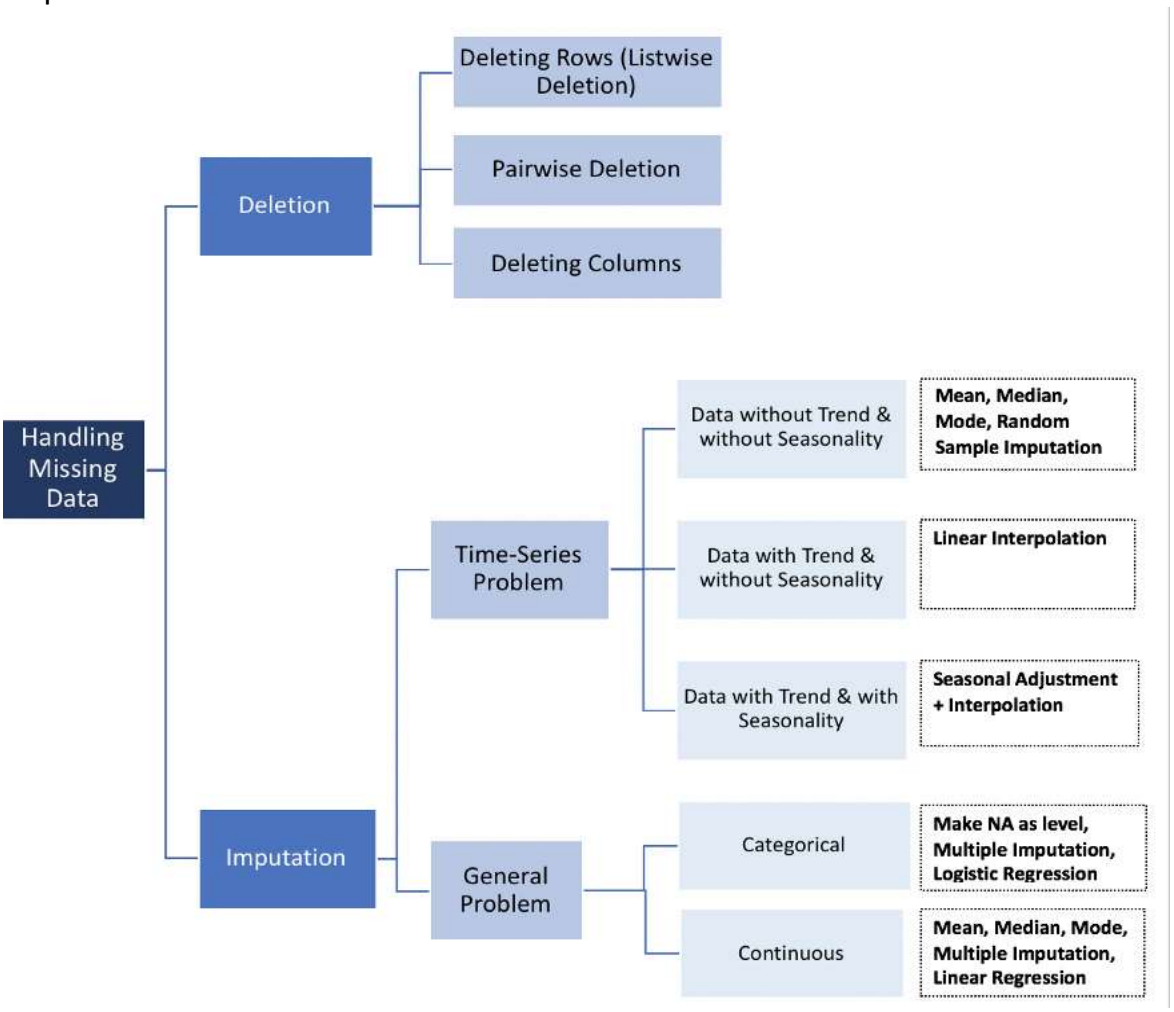
10. What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the

mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

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

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.
The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.
The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.



Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.
The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

12. What is A/B testing?

A/B testing refers to the experiments where two or more variations of the same webpage are compared against each other by displaying them to real-time visitors to determine which one performs better for a given goal. A/B testing is not limited by web pages only, you can A/B test your emails, popups, sign up forms, apps and more.
Like any type of scientific testing, A/B testing is basically *statistical hypothesis testing*, or, in other words, statistical inference. It is an analytical method for making decisions.

13. Is mean imputation of missing data acceptable practice?

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?

In statistics, linear regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with. Building blocks of a linear regression model are

- Discreet/continuous independent variables
- A best-fit regression line
- Continuous dependent variable i.e., A linear model regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the linear regression is :

$$Y=a+b*X + e$$

Where a is the intercept, b is the slope of the line, and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

15. What are the various branches of statistics?

Statistics:
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 are used to make inferences and describe the population. These stats are more useful when it's 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 graphycal form.