STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
   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?
   Ans:
   a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   Ans:
   b) Modelling bounded count data

4. Point out the correct statement.
   Ans:
   d) All of the mentioned

5. ______ random variables are used to model rates.
   Ans:
   c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
   Ans:
   b) False

7. Which of the following testing is concerned with making decisions using data?
   Ans:
   b) Hypothesis

8. Normalized data are centered at______ and have units equal to standard deviations of the original data.
   Ans:
   a) 0

9. Which of the following statement is incorrect with respect to outliers?
   Ans:
   c) Outliers cannot conform to the regression relationship

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

Ans:
It is a probability distribution function for independent, randomly generated variables. That is symmetrical about the mean i.e. half the values fall below the mean and half above the mean and has no skewness. It shows that data near the mean are more frequent in occurrence than data far from the mean. When plotted on a graph, the data follows a bell shaped curved. It is also known as called Gaussian distribution. The mean, median and mode are exactly the same.

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

Ans:
Missing data is defined as the data that is not stored for some variables in the given dataset. It is a vital problem in real life datasets. It can bias the results of the machine learning models and reduce the accuracy of the model. There are two ways of handling missing values:
1. Deleting the Missing values (delete the corresponding rows and columns with missing values) and
2. Imputing the Missing values (replacing the missing data with some substitute value)

The basic technique of imputation is mean, median, mode and time series data imputation (back-fill, forward-fill). But it can skewed our histograms and also underestimates the variance in our data because we're making numerous values the exact same. But there are some advanced techniques like K Nearest Neighbours (KNN) and Multivariate Imputation by Chained Equations (MICE). Now we can choose the best imputation technique by considering the minimum root-mean-squared error (RMSE) value. Most of the cases MICE is the best method as the RMSE is minimum. So the recommended method is MICE.

12. What is A/B testing?

Ans:
A/B testing is a basic randomized control experiment. It is a popular way to compare the two versions of a variable to find out which performs better in a controlled environment. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. For instance, let's say there is a company XYZ. It wants to increase the sales of the product. Now let's divide the products into two parts – A and B. Here A will remain unchanged and there is some significant changes in B's packaging. We will use A/B testing and collect data to analyse which product performs better. On the basis of the response from customer groups who used A and B respectively, we try to decide which is performing better. The **population** refers to all the customers buying the product, while the **sample** refers to the number of customers that participated in

the test. Now we have to make two hypotheses i.e. Null hypothesis and the Alternative hypothesis and conclude the result. In this way A/B testing works.

13. Is mean imputation of missing data acceptable practice?

    Ans:
    Although imputing missing values by using the mean is a popular imputation technique, there are some problem with using mean imputation of missing data and they are:

    i. It reduces the variance of the imputed variables. The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable.
    ii. It does not preserve relationships between variables such as correlations.
    iii. This bias affects standard errors, confidence intervals and other inferential statistics.
    iv. Also, this is not acceptable if the variables have an odd distribution that makes the mean value meaningless.

    So, mean imputation of missing data is not acceptable for every dataset mainly when data set is small.

14. What is linear regression in statistics?

    Ans:
    These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The variable which we want to predict or explain is called the dependent variable and it is always continuous. Independent variables are variables that are used to predict or forecast the values of the dependent variable in the model and it may be continuous or categorical. It estimates the coefficients of the linear equation, involving one or more independent variables and predict the best value of the dependent variable. The simple linear regression calculators use a "least square errors" method to find the best-fit regression line for a set of paired data and the residuals of it follow normal distribution. It fits a straight line which minimizes the difference between predicted and actual output values.
    The equation of linear regression is: y = mx +c
    Example: It can be used to quantify the relative impacts of age, gender, and diet (the independent variables) on height (the dependent variable).

15. What are the various branches of statistics?

    Ans:
    The two main branches of statistics are descriptive statistics and inferential statistics.

    1. Descriptive statistics:-
       If data can be described without any statistical tools then it is called descriptive statistics. It is used to summarize the characteristics of a sample by utilizing certain

quantitative techniques and it is concerned with describing the characteristics of the known data. It can be classified into measures of central tendency and measures of dispersion.
Example- marks in class, height of student.

2. Inferential statistics:-
   If data is too big then then we use inferential statistics. It is a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. We take a few samples from different data and we find the average. This is called inferential statistics. The average is then applicable to all the data from where we have selected our samples.
   Example- Mean marks of 100 students in a particular country are known. Using this sample information the mean marks of students in the country can be approximated using inferential statistics.