STATISTICS WORKSHEET 1

1.a

2.a

3.b

4.d

5.c

6.b

7.b

8.a

9.c

10.Normal Distribution:

The normal distribution is the most widely known and used of all distributions, Because the normal distribution approximates many natural phenomena so well, It has developed into a standard of reference for many probability problems.

Normal distribution is 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.

There is a very strong connection between the size of a sample N and the extent which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximately by the normal distribution even though the population distribution itself is definitely not normal.

11. First and foremost we have to check the missing data from the given dataset. If there is any missing data. Then, We have to apply the techniques to handle missing data.

- Use deletion methods to eliminate missing data. The deletion methods only work for certain datasets where participants have missing fields.
- Use regression analysis to systematically eliminate data.

Common methods are:

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.

Data can be missing in the following ways:

1.Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.

2.Missing At Random (MAR): The key difference between MACR and MAR is that under MAR the data is not missing randomly across all observations, but data is missing randomly only within the sub-samples of data.

3.Not Missing At Random (NMAR): When the missing data has a structure to it, We cannot treat it as missing data at random.

12.What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

A/B testing is also known as split testing. An AB test is an example of statistical hypothesis testing, A process where a hypothesis is made about the relationship between two data sets and those sets are the compared against each other to determine if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

It is a non-standard, it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e., continuous as well as categorical data and it makes advantageous over the other imputations.

Limitations:

- Imputing the mean preserves the mean of the observed data. If the data is missing completely at random, the estimate of the mean remains unbiased.
- Mean Imputation leads to an underestimate of standard errors.

14.What is linear regression in statistics?

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (Dependent) variable?
- Which variables in particular are significant predictors of the outcome variable, and in what way do they-indicated by the magnitude and sign of the beta estimates-impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c+b*x, where y=estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Types of Linear Regression:

- Simple Linear Regression
- Multiple Linear Regression
- Logistic Linear Regression
- Ordinal Regression
- Multinomial regression

15. What are the various branches of statistics?

The two main branches of statistics are:

- **Descriptive Statistics**
- **Inferential Statistics**.

Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics:

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

Inferential Statistics:

The branch of statistics that analyzes sample data to draw conclusions about a population.