**Statistics Worksheet 1**

1.a

2.a

3.b

4.d

5.c

6.b

7.b

8.a

9.c

1. **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.

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

Analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

There are 2 primary ways of handling missing values:

1.Deleting the Missing values

2.Imputing the Missing Values

1.Deleting the Missing values:

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

The disadvantage of this method is one might end up deleting some useful data from the dataset.

There are 2 ways one can delete the missing values:

Deleting the entire row

If a row has many missing values then you can choose to drop the entire row.

If every row has some (column) value missing then you might end up deleting the whole data.

Deleting the entire column

If a certain column has many missing values then you can choose to drop the entire column.

2.Imputing the Missing Value

There are different ways of replacing the missing values. You can use the python libraries Pandas and Sci-kit learn as follows:

Replacing With Arbitrary Value

If you can make an educated guess about the missing value then you can replace it with some arbitrary value using the following code.

Replacing With Mean

This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate.

In such cases, outliers need to be treated first.

Replacing With Mode

Mode is the most frequently occurring value. It is used in the case of categorical features.

Replacing With Median

Median is the middlemost value. It’s better to use the median value for imputation in the case of outliers.

1. **What is A/B testing?**

A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies.

It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

1. **Is mean imputation of missing data acceptable practice?**

This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate.

In such cases, outliers need to be treated first.

1. **What is linear regression in statistics?**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.This form of analysis estimates the coefficients of the linear equation,

involving one or more independent variables that best predict the value of the dependent variable.

Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data.

You then estimate the value of X (dependent variable) from Y (independent variable).

1. **What are the various branches of statistics?**

Statistics have majorly categorised into two types:

Descriptive statistics

Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures.

For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

Measure of frequency

Measure of dispersion

Measure of central tendency

Measure of position

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics.

That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data.

Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population.

It grants us permission to give statements that goes beyond the available data or information.

For example, deriving estimates from hypothetical research.