Statistics WorkSheet-1

Q1. A)

Q2. A)

Q3. B)

Q4. D)

Q5. C)

Q6. A)

Q7. B)

Q8. C)

Q9. C)

Q10. **Normal Distribution**- A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same.

Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

Q11. In Machine Learning the missing values are generally represented by Nan in datasets.

We generally replace Nan by mean or mode of that particular column depending on the type of the data is present in that particular feature column that is either continuous or categorical.

On the other hand we use some of the built in techniques like imputers to treat missing values. These are

- **KNN Imputer**- This technique tries to find relation with other columns and impute other columns and impute the data according with the other columns.
- **Iterative imputer**- This method treats other columns which does not have null values as feature and train on them and treat null columns as label. Finally, it will predict Nan data and impute.

Q12. **A/B testing**-: A/B testing is a marketing strategy that pits two different versions of a website, advert, email, popup, or landing page against each other to see which is most effective.

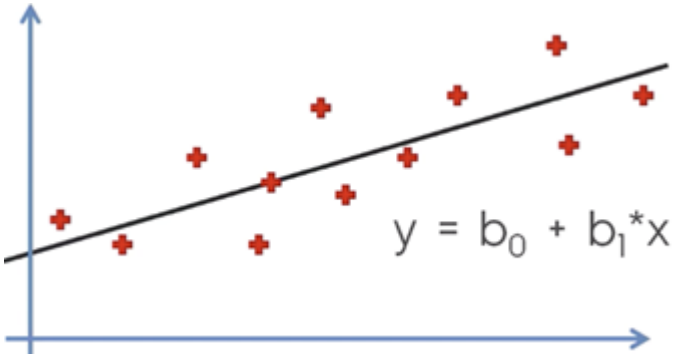
For example, you might test two different popups (to see which drives more webinar sign-ups) or two different Google Ads (to see which drives more purchases). This provides key insights on where and how to invest your marketing budget and gives you the courage to take potentially risky moves.

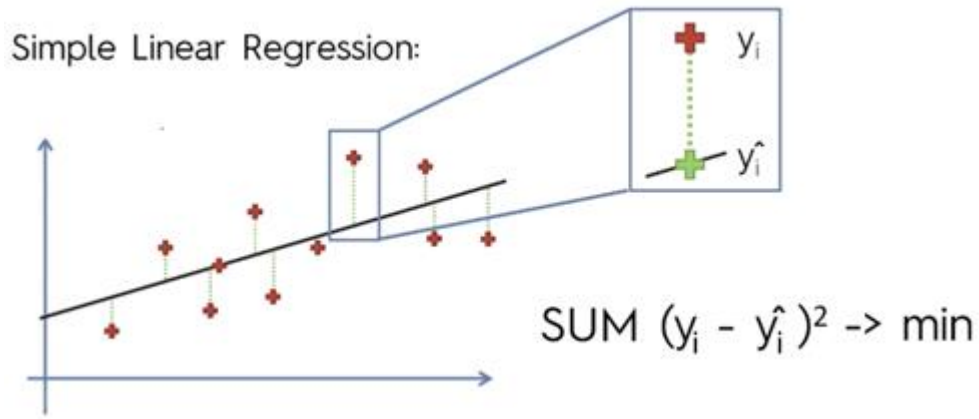
Q13. Generally in very few cases we use mean method to fill null values .We mostly use imputers to fill Nan values .The imputer is fit on a dataset to calculate the statistic for each column. The fit imputer is then applied to a dataset to create a copy of the dataset with all missing values for each column replaced with a statistic value.

Q14. **Linear regression** is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

There are two types of linear regression - **Simple** and **Multiple**.





Q15. There are two branches of statistics: - a) Descriptive Statistics

b) Inferential Statistics

Descriptive Statistics

The first aspect of statistics is descriptive statistics, which deals with the presentation and collection of data. It is not as simple as it appears. The statistician must know how to design and experiment, select the appropriate focus group, and prevent biases that are too easy to introduce into the experiment.

Generally, descriptive statistics can be categorized into

- Measures of central tendency
- Measures of variability

To understand both measures of tendency and variability, easily use graphs, tables, and general discussions.

Measures of Central Tendency

Measures of central tendency are used by statisticians to examine the value distribution center. These are the measures of tendency:

Mean

A mean is a common approach for describing the central tendency. To calculate the average of several values, count them all and divide them by the number of possible values.

Median

It is an outcome found in the middle of a set of values. In numerical journals, edit the results, and the result that is in the center of the distributed sample finds that one is an easy technique to get the median.

Mode

In the given data set, the value which occurs most frequently is the mode.

Measures of Variability

The measure of variability helps the statisticians analyze the distribution from a particular data set. Quartiles, ranges, variances, and standard deviations are the variability variables.

Inferential Statistics

Inference statistics (statistics branch) are statistical techniques that allow statisticians to utilize data from a sample to conclude, predict the behavior of a given population, and make judgments or decisions.

Using descriptive statistics, inference statistics frequently talk in terms of probability. Furthermore, a statistician uses these techniques mainly for data analysis, writing, and drawing conclusions from the limited data. This is accomplished by taking samples and determining their reliability.

Most future predictions and generalizations based on a population study of a smaller specimen are covered by inference statistics. Furthermore, the majority of social science experiments involve the investigation of a small sample population, and that helps in determining community behavior.

The researchers can bring the study-related conclusions by designing a practical experiment. When drawing conclusions, it is important to avoid drawing incorrect or biased conclusions.

There are some different types of inferential statistics, which include the following, which are shown below:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis