STATISTICS WORKSHEET-1

1. (a)
2. (a)
3. (b)
4. (d)
5. (c)
6. (b)
7. (b)
8. (a)
9. (c)
10. Normal Distribution:

    A Normal distribution is the proper term for a" probability bell curve" .In a normal Distribution, the mean is zero and standard deviation is 1 and it is 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.

A normal distribution also known as the Gaussian distribution and normal distributions are symmetrical, but not all symmetrical distributions are normal.

11 .Missing Data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. This indicates that statistical programme will make the decision. Another way is imputation.

Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values .Some of the most prevalent methods are:

Mean Imputation:

Calculate the mean of the observed values for that variable for all non –missing people .It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks .Almost all the methods described below are superior to mean imputation Substitution:

Assume the value from a new person who was not there in the given samples .To put it another way, pick a new sample and employ their worth instead.

Hot Deck Imputation:

A Value picked at random from a sample member who has comparable values on other variables. Otherwise select all the sample members who are comparable on other factors, then choose one of their missing values at random.

Cold Deck Imputation:

A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck, but without the random variance.

Regression Imputation:

The result of regressing the missing variable on other factors to get a predicted value .As a result, instead of utilizing the mean, you're relying on the anticipated value which is influenced by other factors.

Stochastic Regression Imputation:

The predicted value of a regression plus a random residual value.

When list wise deletion eliminates data on the dataset, single imputation appears to be tempting option.

Multiple imputation eliminates many difficulties with missing data and when done correctly, leads to unbiased parameter estimations and accurate standard errors.

Otherwise we can use the algorithm which supports handling of missing values like XGBOOST.

12)A/B Testing:

This is a process of showing two versions of the same web page to different segments of web site visitors at the same time to determine which performs well.

A/B Testing, also known as split testing or bucket testing, the experiment With two or more variants of an ad, marketing email ,or web page are shown to users at random, and then different statistical analysis methods are used to determine which variant drives more conversions.

Typically in A/B testing, the variant that gives higher Conversions is the winning one, and that variant can help you optimize the site for better result.

13.) 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. For example, if we have a table with columns age and fitness scores and a sixty year old has a missing fitness score. If we average the fitness scores of remaining people between the ages of 15 and 60, the sixty year old person fitness score appears as greater fitness level than the actual fitness level he have.

And also, mean imputation decreases the variance of our data with increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14.) 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 predicts 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).

15.) Statistics is the branch of mathematics that deals with data .Data is a collection of values. A collection of data is often referred to as a data set or set of data.

There are three real branches of statistics.

- Data Collection
- Descriptive Statistics
- Inferential Statistics

Data Collection:

This is all about how the actual data is collected. For the most part, this need not concern us too much in terms of the mathematics, but there are significant issues to consider when actually collecting data.

For data such as marks in a class test, this is straight forward. The marks are simply collected together to make the data set.

Sometimes, data is harder to collect. Counting the number of bees in a colony is not easy, because they move and fly around; you have to approximate in such cases.

Also, if you are collecting data, you need to be careful where you get it from. So there are issues in the collection of data. You need to make sure that the data has been collected fairly before you go on to deal with it, and try to present it and make conclusions.

The words population and sample are used in general in statistics. The population is the entire set of data, and a sample is subset of the population.

Descriptive Statistics:

This is the part of statistics that deals with presenting the data we have. This can take two basic forms presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

Common visual techniques include graphs, bar charts and we shall focus mainly on numerical techniques such as averages and spreads.

The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little confuse to see, it needs to be summarised. Instead, you are presented with visual charts.

Inferential Statistics:

This is the aspect that deals with making conclusions about the data. This is quite a wide area, essentially you are asking “what is this data telling us, and what should we do”?

Take the data you have and make an ‘inference’ or ‘conclusion’ for it.