Statistics Worksheet -1

1. a-True.

2. a-central limit theorem.

3. c-Modelling bounded count data.

4. d. All of the mentioned.

5. c. Poisson

6. b. False

7. b. Hypothesis.

8. a.0

9. c. Outliers cannot conform to the regression relationship.

10. Normal distribution:

Normal distribution is the distribution of data points which tends to have that perfect bell curve i.e the shape of the curve is symmetric at the peak. The peak or mean is the point where most of the observed data points are clustered and as the curve moves away from it, the frequency of data distribution also reduces. In other words the mean, median and mode are of same values.

The two main parameters that define normal distribution are: 1. Mean. 2. Standard Deviation.

1. Mean- This describes the distribution of variables, usually more points at the peak. Its an average of summation of total points to number of points.

2. Standard Deviation- This is in relative to the mean, this measures how far the data points are positioned away from the mean. Small deviation from mean gives a steep curve and large deviation gives a flatter curve.

properties:

1. It is symmetric.

2. the mean, median and mode are equal.

3.Empirical rule.

4.Skewness(no left or right skew).

11. Missing data and imputation techniques:

The data sets we observe to predict or make model through machine learning many times has missing data's and they are often represents by Nan's or felt blank. This missing data's cannot always be left missing as it hinders the performance mainly accuracy of the model we train as it lacks many information. So, it is a very necessary requirement to treat this null values.

Imputation techniques:

There are several data imputation techniques that's been used depending upon the requirement of the feature missing.

Imputation using Mean values:

The mean of non-missing data is taken to fill in for missing data. It can be only used with numeric data. Not efficient on categorical data.

Imputation using Median values:

This is same as like missing the mean values, but instead median is filled for the missing data independently.

Imputation using Mode:

This works well for categorical data, The most frequently occurring values are replaced for missing values.

Imputation using KNN method:

The missing values are found out by the k's closest neighbours using feature similarity. This technique is more accurate than the mean, median and mode. however its quite sensitive to the outliers in the data.

12. A/B Testing:

A/B testing is also known as split testing or bucket testing. It is a testing process where two versions of same web page or app is compared against each other to see which one performs better. In this testing process the main site or app is called control and the modified version is called variation.

The traffic of the website is routed both to control and variation to see which one records a better response. the user experience is measured and collected in the dashboard and is analysed through statistical engine. This process allows them construct hypotheses and know the best possible changes one can make to improvise there business.

13. Mean imputation is one of the practice followed to fill in the missing data in a data set. This practice fills in the missing values with the mean of the dataset.

pros:

1. This imputation technique can be used on small dataset.

2. It is one of the easiest method to fill in for missing values.

cons:

1. It can be used best only on numerical dataset.

2. Models built with mean imputation technique has considerably poor accuracy hence the performance.

3. This technique ignores feature corelation.

4. It decreases the variance of the data which bias the model.

Considering the pro's and con's mean imputation technique is not considered to be the best practice.

14. Linear Regression:

Linear Regression is of the supervised machine learning technique that models the relationship between two variables by fitting a linear equation to observed data. One of the variables would be dependent and other would be independent. A scatter plot would be a helpful tool to determine the strength of relationship between the proposed independent and dependent variable.

The equation of the line is y= a+bx+e

where: a= intercept

b=slope of the line

x=explanatory variable

e= error term.

15. Statistics in general comprises of two main branches

1. Descriptive Statistics:

As the name suggests, it is descriptive in nature that usually deals with presentation and collection of data. This is the first part of the analysis. It tells the distribution of data through visual representation in forms of graph, such as histogram or plots. this gives the general idea for the facts and figures to be considered for the statistical analysis and predict hypotheses for a problem statement. This is usually possible through mean, median, mode and variance calculation of the entire population of data.

2. Inferential Statistics:

An inference is drawn here based on the descriptive statistics. These two branches go hand in hand. The conclusions or predictions are made out of sample data's obtained from the population data to see if the hypotheses holds true or not.