Statistics Worksheet 1

1] Bernoulli random variables take (only) the values 1 and 0.

Answer :- a) True

2] Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Answer :- a) Central Limit Theorem

3] Which of the following is incorrect with respect to use of Poisson distribution?

Answer :- b) Modeling bounded count data

4] Point out the correct statement.

Answer :- d) All of the mentioned

5] _____ random variables are used to model rates

Answer :- c) Poisson

6] Usually replacing the standard error by its estimated value does change the CLT.

Answer :- b) False

7] Which of the following testing is concerned with making decisions using data?

Answer :- b) Hypothesis

8] Normalized data are centered at_____and have units equal to standard deviations of the original data.

Answer :- a) 0

9] Which of the following statement is incorrect with respect to outliers?

Answer :- c) Outliers cannot conform to the regression relationship

10] Normal Distribution is a distribution in which most data points lie in the middle of the range and

the rest taper off symmetrically towards the either end. In Normal Distribution mean, median & mode are all the same. The following are the important properties of Normal Distribution.

a) In a normal distribution, the mean, median & mode are equal

b) The total area under the curve should be equal to 1

c) The normally distributed curve should be symmetric at the centre

d) The normal distribution curve must have only one peak

e) The normal distribution must be defined by the mean & standard deviation.

11] Missing data can be handled in different ways.

a) If the entire row has null values, then we can delete the entire row from the dataset as it will not make any difference.

b) If the entire column has null values, then the entire column can be deleted as it will not make any difference

c) However, if some values are missing, then we will have to do imputation which is the process of substituting an estimate for the missing values.

d) If the missing data is of continuous data type, then we can substitute the missing data with mean or median of the data in that column.

e) If the missing data is of categorical data type, then we can substitute the missing data with mode of the data in that column.

f) Model Based Imputation :- We can consider the column having missing data as label and the other columns as feature and then use an appropriate model (say KNN model) to predict the missing values.

I would recommend using Model Based Imputation.

12] A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric.A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

13] Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Mean Imputation is not acceptable way of handling missing data because it does not preserve the relationships among variables.

Mean imputation will lead to underestimate of standard error.

14] Linear Regression Model predicts the dependent variable using a regression line based on independent variables.

Equation of Linear Regression is given by

y=m*x + c + e where m is the slope, x is the feature/independent variable, c is the intercept on Y axis and e is error.

This equation predicts the value of target variable based on given variables.

When we have one feature and one label, then it is called Simple Linear Regression. When we have multiple Features and one label, then it is called Multiple Linear Regression. Linear Regression is a supervised Machine Learning Algorithm because we train the model on some data to help it understand the relationship between feature and label and then use the same model to predict the output/label for some test data. For simple linear regression, the formula is given below.

y=mx + c

But for Multiple Linear Regression, the formula changes to the following

y=m1x1 + m2x2 + m3x3

The core idea in the regression model is to obtain a line equation that best fits the data. The best fit line is when the total prediction error for all the data points is considered as small as possible. The error is the distance between the point on the plane to the regression line.

15] The two main branches of statistics are :- a] Descriptive Statistics & b] Inferential Statistics

Descriptive Statistics deal with presentation & collection of data. Consider the example of elections in our constituency. We know our constituency very well and we know which candidate/party will win the elections. This is because the population is small.

Thus, when we are able to describe the outcome, then it is called descriptive statistics. Thus descriptive statistics is when we are

able to predict the outcome accurately because the population is less.

Descriptive Statistics have two parts.

a] Central tendency measures

b] Variability measures.

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

Mean

Mean is a conventional method used to describe the central tendency. Typically, calculate the average of values, count all values,

and then divide them with the number of available values.

Median

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals

and locate the result that is in the center of the distributed sample.

Mode

The mode is the frequently occurring value in the given data set.

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of

variability include quartiles, ranges, variances, and standard deviation.

Inferential Statistics on the other hand deal with understanding the data on a much bigger scale. It is clear that we can predict the elections

outcome for our constituency but we cannot predict the elections outcome for the entire nation. This is where inferential statistics comes

into play. In Inferential statistics, we collect sample data of the entire population and we try to make valid conclusions about the entire

population by understanding the sample data. Thus the process of understanding the data by sampling is called Inferential Statistics.