STATISTICS WORKSHEET-1

### OUESTION-1

### TRUE

### QUESTION-2

### (a)Central Limit Theorum

### QUESTION-3

### (b)Modelling bounded count data

### QUESTION-4

### (d)All of the mentioned

### QUESTION-5

### (c)poisson

### QUESTION-6

### (b)false

### QUESTION-7

### (b)Hypothesis

### QUESTION- 8

### 0

### QUESTION-9

### (c) Outliers cannot conform to the regression relationship

### OUESTION-10

ANSWER->**Normal distribution**, also called **Gaussian distribution**, the most common distribution function  for independent, randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation.

The graph of the normal distribution is characterized by two parameters: the mean,or average, which is the maximum  of the graph and about which the graph is always symmetric; and the,standard deviation which determines the amount of dispersion away from the mean. A small standard deviation (compared with the mean) produces a steep graph, whereas a large standard deviation (again compared with the mean) produces a flat graph.

### OUESTION-11

ANSWER->In real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. [Handling](https://analyticsindiamag.com/get-started-preparing-data-machine-learning/) the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models.

Here are a few imputation techniques to handle missing data:

### 1. Delete Rows

This method is commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is preferred only when there are enough samples in the data set. The analyst has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

## **2. Assign a unique category**

A categorical feature has a definite number of possibilities. Since they have a definite number of classes, we can assign another class for the missing values’. This strategy will add more information into the dataset which will result in the change of variance.

## **3. Replace with mean/median/mode**

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.

## **4. Use Algorithms which support missing values**

(1)KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values.

(2) RandomForest. This model produces a robust result because it works well on non-linear and the categorical data. It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

## **5. Predict the missing values**

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm.

Another algorithm which can be used here is RandomForest. This model produces a robust result because it works well on non-linear and the categorical data. It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

### QUESTION-12

### ANSWER->A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The**population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

### OUESTION-13

### ANSWER->Yes,the mean imputation of data is an acceptable but not an advisable practice.The estimation of the mean remains unbiased in this technique.

### However,it is not advisable due to the following reasons:

### It leads to an underestimation of standard errors.

### It reduces variance of the data.

### It does not preserve the relationships among the variables.

OUESTION-14

### ANSWER->Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) If a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable

### Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

OUESTION-15

### ANSWER->The two main branches of statistics are:

### Descriptive Statistics

### Infrerential Statistics

### (A)Descriptive Statistics

### It is associated with the colletion and presentation of data.It refers to the analysis of data that helps summarize data in a way that meaningful patterns can be derived from it.It makes interpretation a lot easier for the analyst

### This is usually done in two ways:

### Measuring central tendency

### Measuring the spread

### skewness

### (B)Inferential statistics

### It is associated with making inferences about the larger population by analysing the sample drawn from it.

### The analyst needs to make sure if the sample drawn accurately represents the larger population.

### The conclusions drawn from the sample are geneaelized to a population.