PYTHON WORKSHEET 1

Q1- C

Q2- B

Q3- C

Q4- A

Q5- D

Q6- C

Q7-A

Q8- C

Q9- A, C

Q10- A, B

STATISTICS WORKSHEET 1

Q1- a

Q2- a

Q3- b

Q4- d

Q5- c

Q6- b

Q7- b

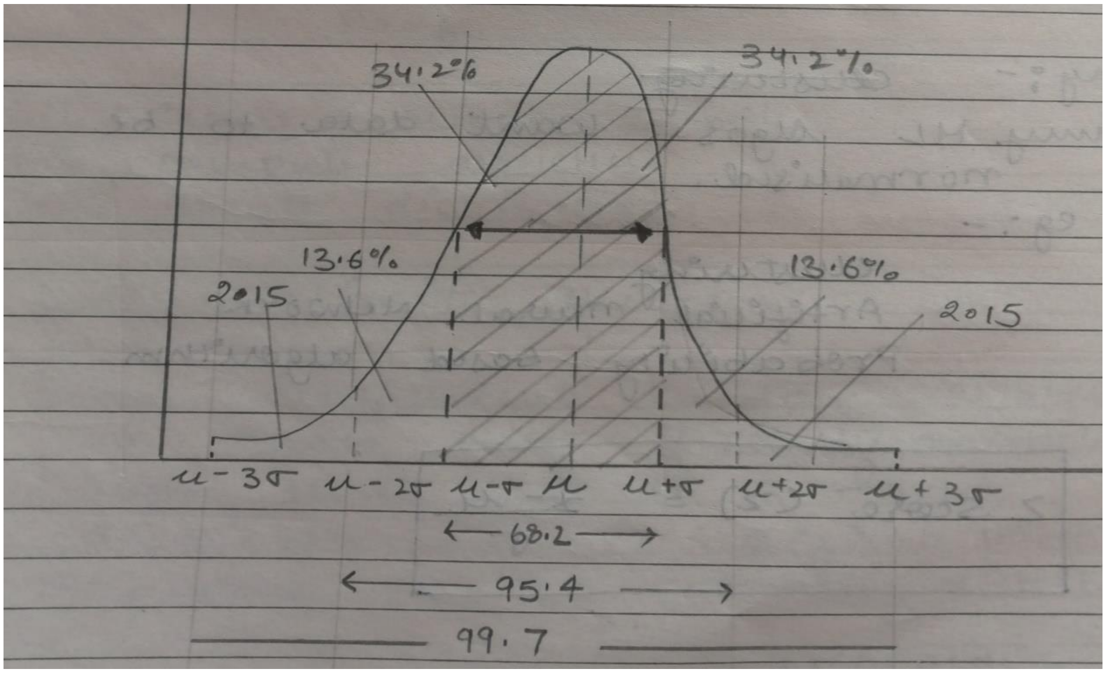
Q8- a

Q9- c

Q10- **Normal Distribution-** It is also called bell curve. This distribution is used when majority of data lies near mean. For example, average height of a human. A human can never be too short or too tall but majority of people are of average height so if the x axis has height starting from short to average to tall and y has number of people majority of data points will be at the center causing the graph to look like a bell.

In a normal distribution

- Mean defines the center of the graph
- Standard deviation defines the width of the graph
- Total are under the graph is 1
- 68.2% of the area is under the curve is within 1 sigma of the mean
- 95.4% of the area is under the curve is within 2 sigma of the mean
- 99.7% of the area is under the curve is within 4 sigma of the mean



Q11 **Missing Data-** It is the data which is unavailable for a particular variable for observation. It reduces the quality of analysis and reduces statistical power of the analysis. There are two methods of handling missing data

1. Imputation – It fills the unavailable data with the nearest guesses. It is most useful and efficient if the number of missing data is comparatively lower. If we use imputation when missing data is in large amount, then the result will inefficient result of analysis. Following are the imputation methods
    - Simple imputation
        - ➢ Mean, median, mode
        - ➢ Time series
        - ➢ Linear interpolation
    - Multiple imputation
        - K nearest neighbors
2. Deletion- When we are dealing with missing data then one method to handle the situation is removing or deletion of data. In this technique the related list, roe, column is permanently removed from the dataset. This method creates problem when there are not enough observations available.

In k nearest neighbor the missing data is filled by a value that is obtained by related records from the whole dataset. This results in more efficient analysis because filling values according to the dataset tend to be more true and natural. Therefore I recommend k nearest neighbor multiple imputation method.

Q12- **A/B testing** – It can be referred to as a statistical hypothesis testing in which we make predictions by analyzing two different datasets (for situations) A and B. Predictions are made on these two individually with the help of null and alternate hypothesis, taking sample from population and then after that comparing these two predictions for datasets for A and B. The one with better and efficient performance is chosen.

Q13- No mean imputation is not acceptable because it hinders with other statistical aspect of the data such as variance, correlation. Also due to mean imputation results are more likely to be biased.

Q14- Linear Regression in statistics refers to the process of defining a relationship between a dependent and an independent variable by fitting straight line through collected observations. Equation of the line is y=mx+c.

Q15- Two main branches of statistics are Descriptive statistics and Inferential statistics

1) Descriptive Statistics- It shows basic description of given data. It shows us about how dispersed the data is or how much different values are there. For example, in a study showing favorite subjects of students in 3rd grade, out of 90 students 25 like math, 13 like Science, 12 like Hindi, 20 like English and 20 like Social Science.

2) Inferential statistics- In this branch inferences are made out of the give observations. This means we can make predictions from the given data. This is done on very large population by taking samples from the whole population. For example analyzing average height of man in India .

MACHINE LEARNING WORKSHEET 1

Q1-A
Q2- A
Q3 -B
Q4-A
Q5- C
Q6-B
Q7-D
Q8-D
Q9-C
Q10-B
Q11-B
Q12-A, B

Q13- **Regularization-** It is a technique which prevents the data to over fit in the training model. Due to overfitting although the data may perform well on training model but does not give as good result in test dataset as compared to training model. Regularization prevents overfitting by adding extra factors and information.

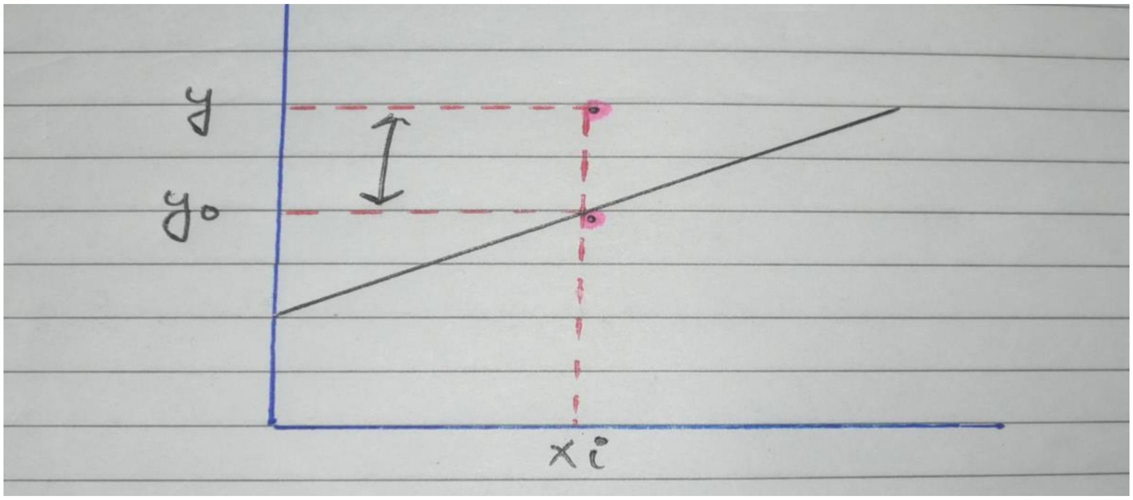
Q14- Following are the algorithm used in regularization-

Ridge regression (L2 regularization)- (sum of weight square + loss function)

Lasso regression (L1 regularization)

Q15- **Error in linear regression-** It is referred to as distance between the actual value and the corresponding point that lies on the regression line.

Y=mx + c + error



For x_i, y is the actual value and y_0 is the predicted value and error= $y-y_0$