**Basics of** **Statistics**

**Definitions -** It is a branch of science which involve analyzing and exploring the data in a large quantities, so that user can come up with solving various use casesand conclusions.

**Types:** Descriptive Statistics and Infrantial Statistics

**Descriptive Statistics:** it means analyzing the data, exploring, visualizing to understand the relation of the data. (Organizing and summarizing the data using Histogram, Bar graph, Pie chart, Scatter plot). Tools and Libraries – Power Bi, Tablue, Matplotlib, plotly and seaborn.

**Infrantial Statistics:** In the infrantial statistics user takes sample of the data and do some operation on the data and make conclusion using some tests such as Hypothesis testing, t-test, f-test, chi-square test, z-test and p value.

**Population and Sample:** In simple way population means whole data and it is denoted by (N). Sample means small amount of data of the whole data and it is denoted by (n). Based on the sample information user can make conclusion.

**Sampling Techniques:**

1.Random Sample – User select sample randomly. However, the random sampling has disadvantage such as sample overlapping. For some of the use cases it do not work, for an instance if user want to make survey of data scientist then user can not pick sample randomly as not each random sample would be a data scientist.

2. Stratified Sampling – Population data is divided into subgroups. E.g. If user wants to take survey of cosmetic products the he will only take survey of females which is the subgroup of Gender.

3. Systematic Sampling – sample takes from population data using N^th value, such as 6th sample per population data or 10th sample per population data.

4. Clustering Sampling – In cluster sampling, researchers divide a [population](https://www.scribbr.com/methodology/population-vs-sample/) into smaller groups known as **clusters**.  They then randomly select among these clusters to form a sample.

**Use of sampling data-**

Sampling data come up with train and test split. Machine learning model train on training data and test on testing data. The split of the data also has types like Random split, Stratified split, cluster split, cross validation split. E.g. 1Lrecords splits and 70k go in training dataset and 30k go in testing data set. And then model will evaluate on it.

**Measure of Central Tendency:**

**Topics –** Mean, Median and Mode

**Mean** – Average of the specific column in dataset is the mean of the column.

Age of 5th class student= [10, 12, 9, 10, 11, 10, 8, 9]

Mean of age = [10, 12, 9, 10, 11, 10, 8, 9] /9

Whenever dataset haa outlier do not use mean.

Age = [10, 12, 9, 10, 11, 10, 80, 93] --- here 80 and 93 is an outlier as person cannot live more than 100 years.

**Median** – Whenever outlier is present in dataset use median. Because median does not get impacted by outlier.

Steps🡪 sort data, pick up the middle element and that will be median of the dataset.

Age = [10, 12, 9, 10, 11, 10, 80, 93] = middle elements are 11 and 10 hence 10+11/2 =10.5 ……. Hence median is 10.5.

**Mode** – Most frequent occurring element in dataset is nothing but mode of the dataset.

Age = [10, 12, 9, 10, 11, 10, 80, 93] Mode is - 10 because it has occurred 3 times.

**Random Variable:** from random phenomena user can get some outputs which is nothing but Random Variable. For e.g. tossing coin, rolling dice.

Random variable has two types:

1. Categorical/Qualitative Random Variable
2. Quantitative Random Variable

**Categorical Random Variable** – e.g. Gender (Male, Female), Weekdays (sun, mon, sat)

1. Nominal Categorical Random Variable – Ranking is not important. E.g. Gender, Colors
2. Ordinal Categorical Random Variable -- Ranking is important in Ordinal. E.g. Weekdays, Months, Grades

**Quantitative Random Variable** – e.g. Height, Age,

Discreet Quantitative Variable and Normal Quantitative Variable.

1. Discreet Quantitative Variable: Whole number like Age, No. of people. (1,2,3,4,5)
2. Continuous Quantitative Variable: Fraction values like Height, Weight. (23.20,1.2)

**Independent and Dependent Features:**

In machine learning model user takes independent features and make some conclusion and produce output. That output is nothing but dependent feature. Dependent feature (output) is totally depends on independent features (input) as if independent features changes it will effect on dependent feature.

E.g. – user train model to predict the result of 10th grade students by dataset which is containing features like, 9th grade marks of student, no. of hour’s student study per day. Here 9th grade marks of student and no. of hour’s student study per day these both are independent features and by using these features user can make prediction about result of 10th grade students which is the output of model (dependent features).