Class 10
Machine Learning With Python

SAMPLING

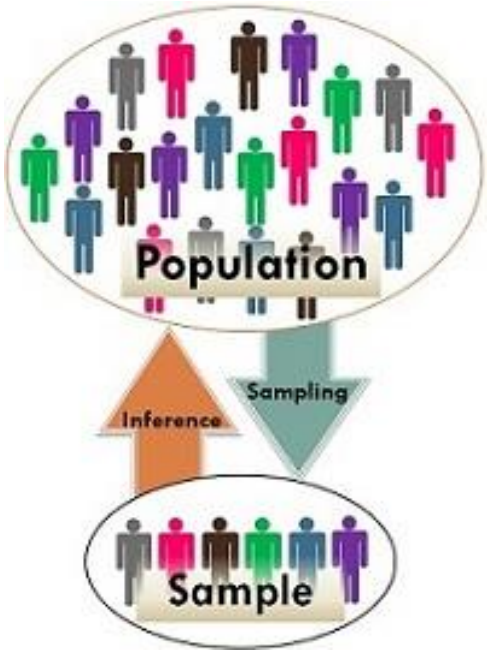
- The action or process of taking samples of something for analysis.

- Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed.

What is Sample.?

- A representation subset of the population is called as **sample**.
- A **sample** consists one or more observations drawn from the population

Note:- If we club Samples then we should get Population

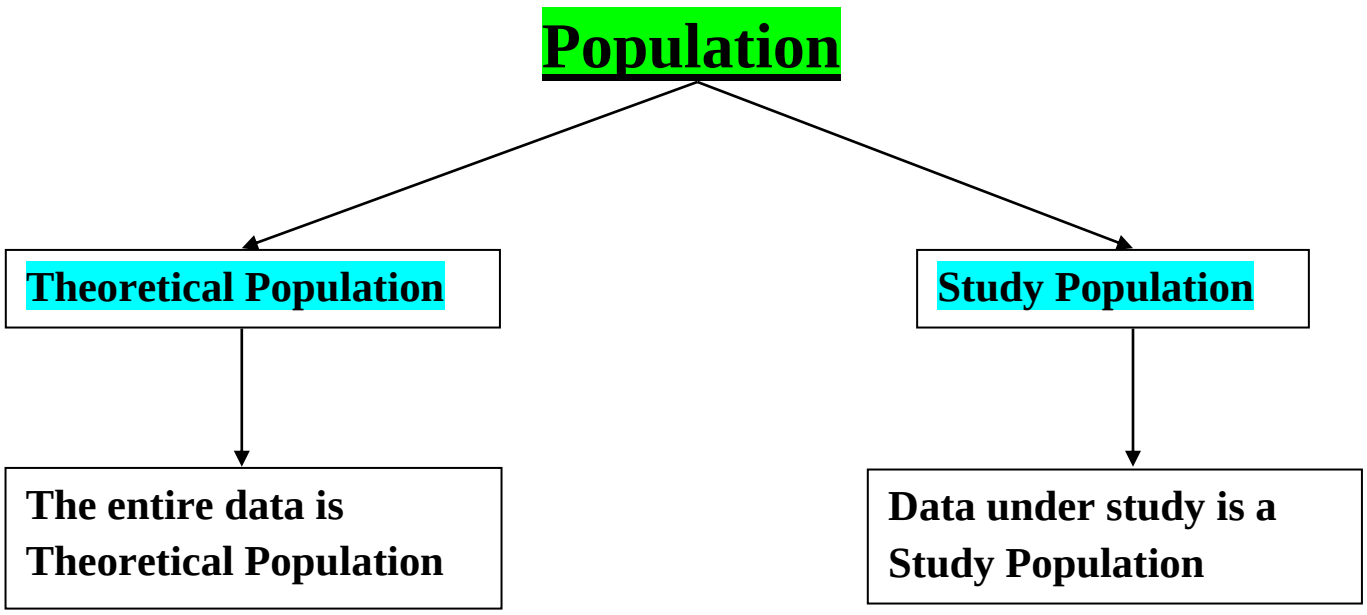


What is Population.?

The data that is collected overall from a studying experiments.

DIFFERENCE BETWEEN POPULATION AND SAMPLE

BASIS FOR COMPARISON	POPULATION	SAMPLE
Meaning	Population refers to the collection of all elements possessing common characteristics, that comprises universe.	Sample means a subgroup of the members of population chosen for participation in the study.
Includes	Each and every unit of the group.	Only a handful of units of population.
Characteristic	Parameter	Statistic
Data collection	Complete enumeration or census	Sample survey or sampling
Focus on	Identifying the characteristics.	Making inferences about population.



POPULATION DESCRIPTION :-

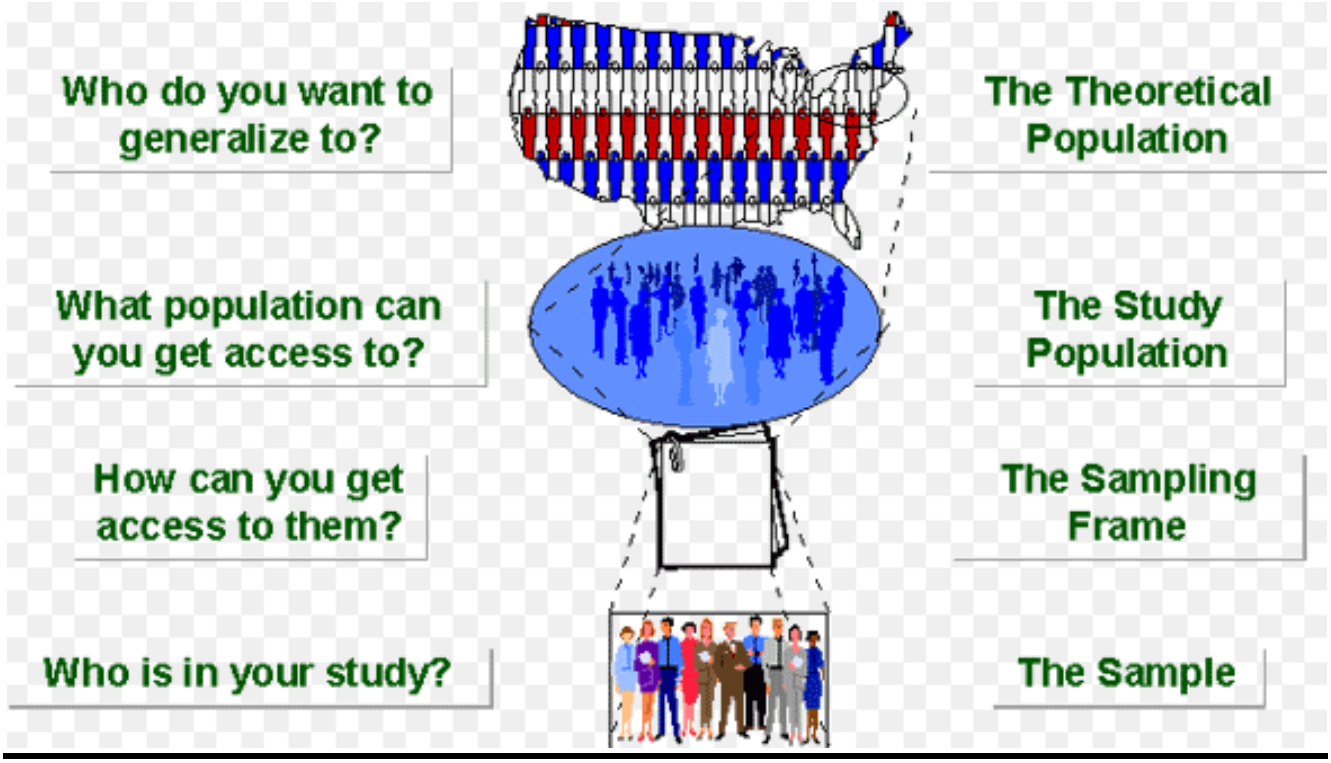
Population description are the **Filters** that retrieve Study population from Theoretical Population

for example:-

Theoretical - People working in TCS.

Study- How many debit card holder employees working in the organization who have active demat account in Noida branch.

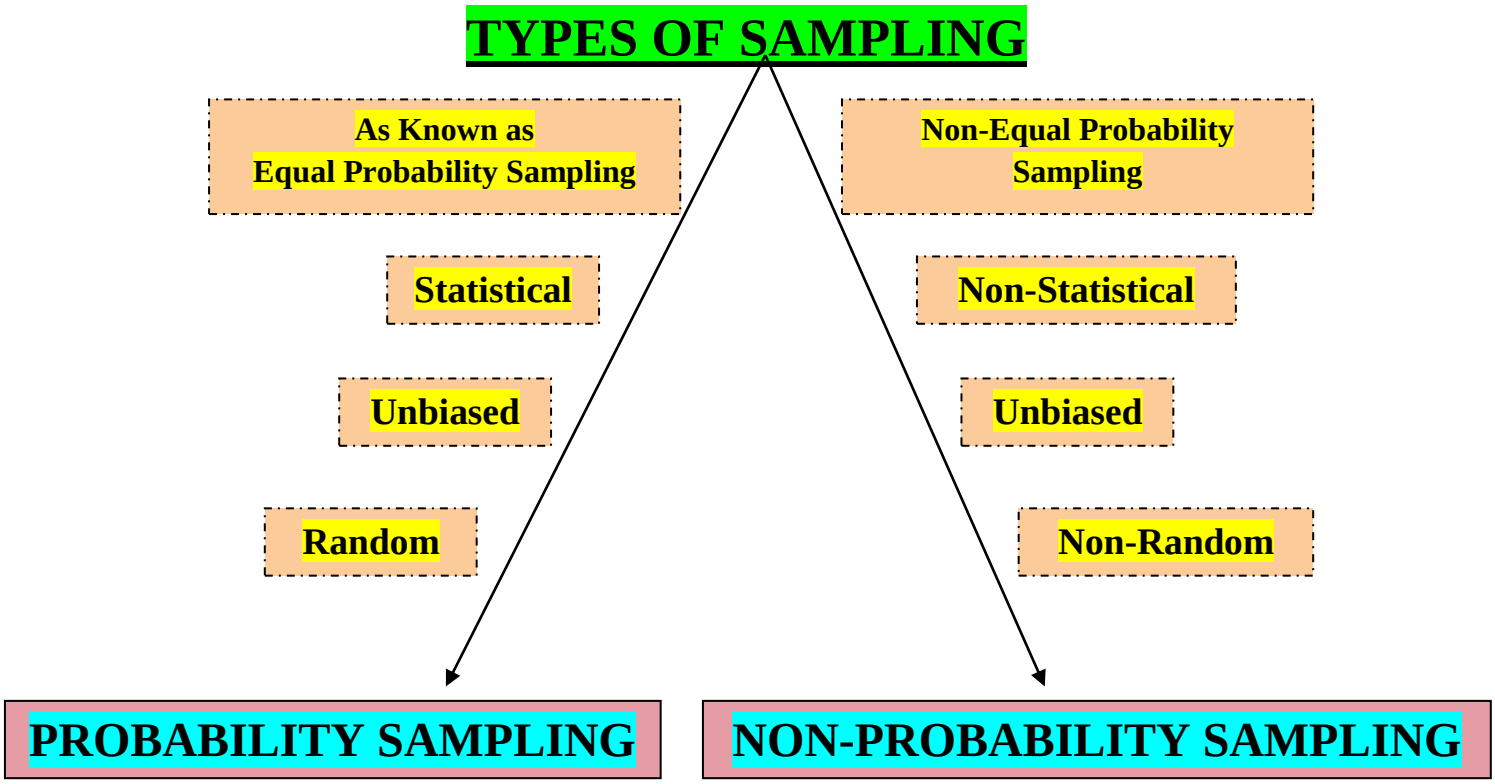
SAMPLING TERMINOLOGIES :-



Theoretical Population > **Study Population** > **Sample**

NEED OF SAMPLING

- Sampling provides a valid alternative than population
- It would be impracticable for you to survey the entire population.
- Too Expensive to gather information on the entire population
- your time constraints prevent you from surveying the entire population.
- you have collected all the data but need the result very quickly.
- Census requires enormous time, trained personnel, money etc. and slightest bias can get magnified when no. of observations are increased.
- Sometimes Sampling can be destructive, suppose a car company perform crash check quality, so if company pick 20 cars from a lot of 100 then at this time Sampling is destructive, in these type of scenario company pick only one car and perform crash test, if car passes all check list points, then company assumes that all cars will perform perfectly.

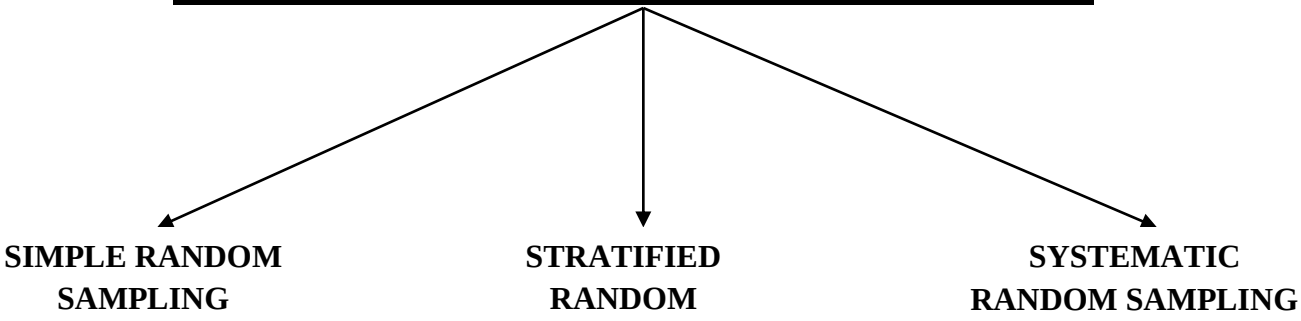


PROBABILITY SAMPLING (Equal Probability Sampling)

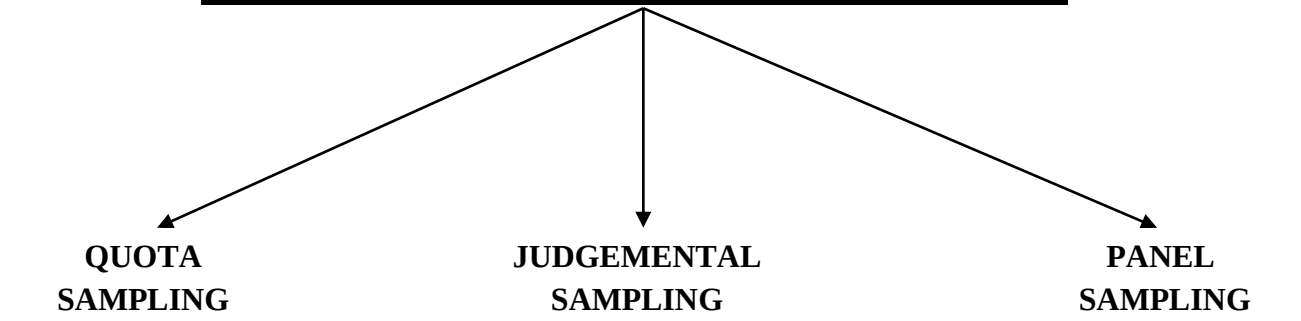
A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

When every element in the population does have the same probability of selection, this is known as an 'equal probability of sampling' (EPS) design. such designs are also referred to as ' self-weighting' because all sampled units are given the same weight.

TYPES OF EQUAL PROBABILITY SAMPLING



TYPES OF NON-PROBABILITY SAMPLING



SIMPLE RANDOM SAMPLING

The **Simple Random Sampling** is a sampling technique wherein every item of the population has an equal and likely chance of being selected in the sample. Here, the selection of the item solely depends on the chance and therefore, this method is also called as a **Method of Chance Selection**.
The selection of samples through simple random sampling is free from the personal bias as the investigator/researcher does not exercise his discretion of preference in choosing the items.

For example:- The most common method used to obtain the random samples from the population is a **Lottery Method.** Under this method, all the items of the population are numbered or named on the identical paper slips and then these slips are mixed up in a container. The investigators blindfold himself and select as many slips from the container that constitute the desired sample size.

Here, the selection is purely based on a chance and every item has an equal chance of getting selected. This method is very popular in the lottery draws. While using the lottery method, it is essential to check that the slips are of identical size, color, and the shape, otherwise, there is a possibility of personal bias and prejudice.

when data is bit homogenous then we do simple random sampling

Advantages

- There is an equal chance of selection.
- It requires less knowledge to complete the research.
- It is the simplest form of data collection.
- It is easier to form sample groups.
- Findings can be applied to the entire population base.
- Efficient.
- Homogenous.

Disadvantages

- No additional knowledge is taken into consideration.
- Researchers are required to have experience and a high skill level.
- No guarantee that the results will be universal is offered.
- It is easy to get the data wrong just as it is easy to get right.
- Heterogeneous.

STRATIFIED RANDOM SAMPLING

Stratified random sampling is a method of **sampling** that involves the division of a population into smaller groups known as strata. In **stratified random sampling**, or **stratification**, the strata are formed based on members' shared attributes or characteristics.

8 Steps to select a stratified random sample :-

1. Define the target audience.
2. Recognize the stratification variable or variables and figure out the number of strata to be used.
3. Use an already existent sampling frame or create a frame that's inclusive of all the information of the stratification variable for all the elements in the target audience.
4. Make changes after evaluating the sampling frame on the basis of lack of coverage, over-coverage, or grouping.
5. Considering the entire population, each stratum should be unique and should cover each and every member of the population. Within the stratum, the differences should be minimum whereas each stratum should be extremely different from one another. Each element of the population should belong to just one stratum.
6. Assign a random, unique number to each element.
7. Figure out the size of each stratum according to your requirement. The numerical distribution amongst all the elements in all the strata will determine the type of sampling to be implemented. It can either be proportional or disproportional stratified sampling.
8. The researcher can then select random elements from each stratum to form the sample. Minimum one element must be chosen from each stratum so that there's representation from every stratum

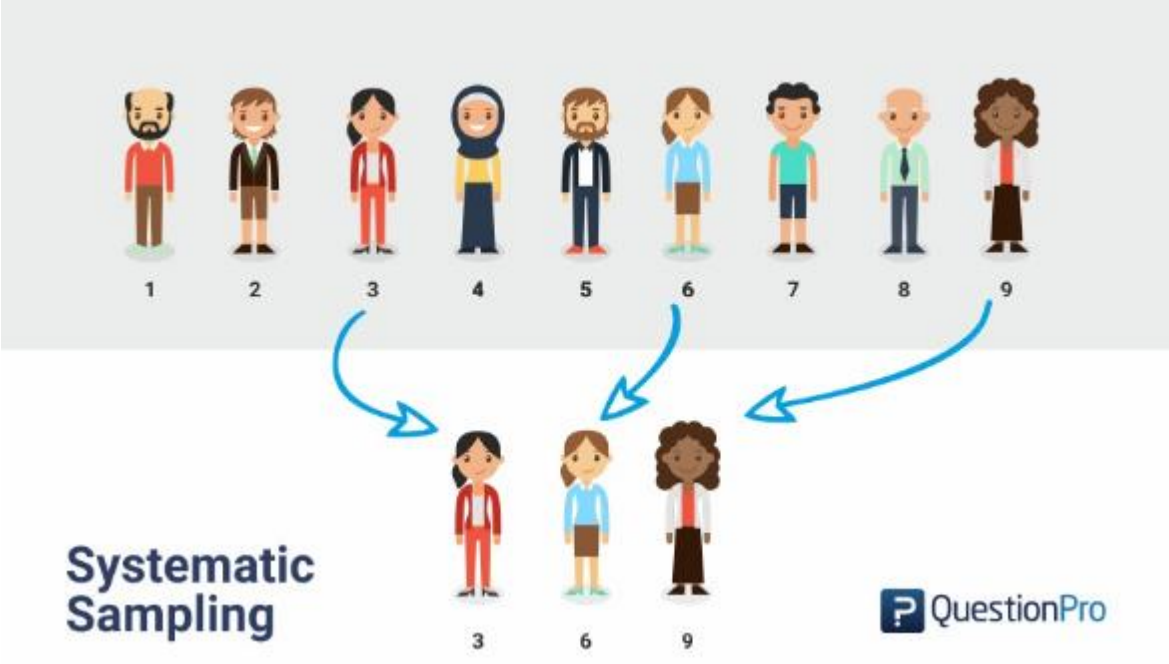
For Example:- Suppose there is a company has 10 process in it, Each process has 2 employee, and HR want 10 people randomly but 1 from each team.
So HR have to create Strata for each team and then pick one person from each **Strata.**

SYSTEMATIC RANDOM SAMPLING

- In **Systematic random Sampling** the start point is random, after that sampling happens at a fixed interval.
- There is a gap, or interval, between each selected unit in the sample.
- Selection of units is based on sample interval**..**

3 Steps to select a Systematic Random Sample :-

1. Number the units on your frame from 1 to N and the population are arranged in some way.
2. First sample drawn between **1** and **k** randomly(determine point / the random start).
3. Afterwards, every **K**th must be drawn until the total sample has been drawn.



For example - The researcher has a population total of 100 individuals and need 12 subjects. He first picks his starting number, 5.

Then the researcher picks his interval, 8. The members of his sample will be individuals 5, 13, 21, 29, 37, 45, 53, 61, 69, 77, 85, 93.

SAMPLE CALCULATOR

Key terms in Sample Calculator:-
1.Confidence Interval.
2.Margin of Error / Precision.
3. Historical Error Rate.

1.Confidence Interval:-

Express as percentage how much are you confident for repeating / obtaining the same results in an experiments / trial.

This measures how accurate or reliable you want your data to be. The percent indicates how confident you want to be that your results are correct. If you aren't sure what level to use, we recommend using 95%, which means that you can say with 95% certainty that your results are correct.

2.Margin of Error:-

Margin of Error is how much the sample would differ from population.

The margin of error is a statistic expressing the amount of random sampling error in a survey's results. The larger the margin of error, the less confidence one should have that the poll's reported results are close to the "true" figures; that is, the figures for the whole population.

3.Historical Error Rate:-

Error rates of Sample (Defects / Total).