

## Today's agenda

- Sample and Population
- Merits and Demerits of sampling
- Types of sampling
- Random sampling
  - Implementing the same using pandas
- Predictive analytics
- Importance of sampling in PA

# Sample Population & Sampling

## Population

- A set of similar items or events which is of interest for some question or experiment.
- We denote the population as  $N$ .

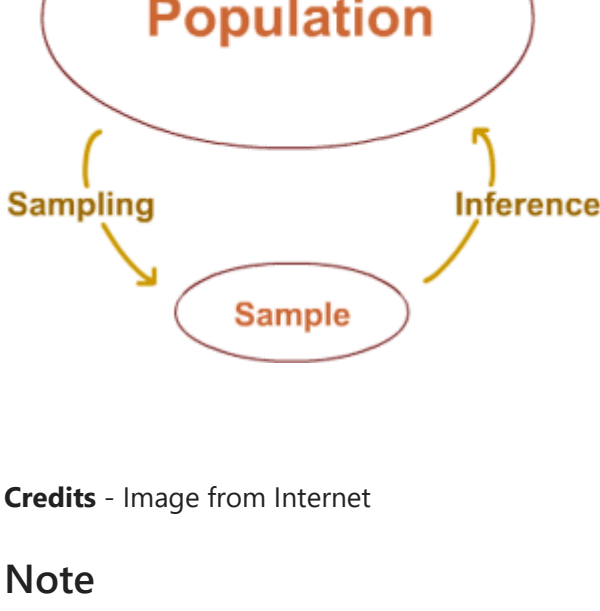
## Sample

- A subset of the population (a statistical sample) that is chosen to represent the population.
- We denote the sample as  $n$ .

## Sampling

(method)

- A selection of subset of individuals from statistical population to estimate the characteristics for the whole population.
- It is one such technique that is applied by everyone in our day to day activities.



Credits - Image from Internet

## Note

- By taking sample, statisticians tend to infer or conclude the characteristics/estimates to the whole population.

## Example

- Imagine you have a piece of land and you want to know if the land is fertile enough to grow plants.



- **Scenario 1**
  - Interpret the land's fertility by testing the whole land.
- **Scenario 2**
  - Interpret the land's fertility by just testing a sample (soil) in a container or jar.

Credits - Image from Internet

## Merits & Demerits

### Merits

- Less cost effective
- Time saving
- Higher accuracy

### Demerits

- Chances of biasness
- Need of subject specific knowledge

## Types of Sampling

1. **Probability Sampling**
  - Simple Random Sampling
    - It is a randomly selected subset where each member of the population has an exactly equal chance of being selected.
    - From the random sample that is selected, researcher tends to make statistical inferences on the whole population.
  - Systematic Sampling
  - Cluster Sampling
  - Stratified Sampling
2. **Non-Probability Sampling**
  - Convenience Sampling
  - Judgmental Sampling
  - Snowball Sampling
  - Quota Sampling

## What should be the size of the sample?

- Refer to → <https://www.tools4dev.org/resources/how-to-choose-a-sample-size/>

```
In [1]: import pandas as pd
import numpy as np
```

## Population data

Get random integers in the range of `low` and `high`

- size → (how\_many\_rows, how\_many\_columns) - (1000, 3)

Make random data using pandas

```
In [2]: # help(np.random.randint)
```

```
In [3]: # popn_data (population)
popn_data = np.random.randint(low=10, high=100, size=(10000, 3))
```

```
In [4]: # display popn_data
popn_data
```

```
Out[4]: array([[62, 97, 35],
               [99, 59, 71],
               [35, 39, 96],
               ...,
               [42, 18, 59],
               [37, 77, 70],
               [59, 36, 80]])
```

Create a dataframe with columns and data generated

```
In [5]: # df
df = pd.DataFrame(data=popn_data, columns=['col_1', 'col_2', 'col_3'])
```

```
In [6]: # head()
df.head()
```

```
Out[6]:   col_1  col_2  col_3
0      62     97     35
1      99     59     71
2      35     39     96
3      14     34     50
4      70     95     15
```

Population data (df) size is 10000

- N = 10000

```
In [7]: # shape
df.shape
```

```
Out[7]: (10000, 3)
```

## Simple random sample

- Select a sample dataframe from population (df) of size 100
- n = 100

```
In [8]: # dir(df)
```

```
In [9]: # ??df.sample
# help(df.sample)
```

```
In [10]: # rand_sample_df
# dir(df)
rand_sample_df = df.sample(n=100, random_state=42)
```

```
In [11]: # shape
rand_sample_df.shape
```

```
Out[11]: (100, 3)
```

```
In [12]: # head
rand_sample_df.head()
```

```
Out[12]:   col_1  col_2  col_3
6252     97     99     34
4684     28     95     54
1731     91     34     77
4742     17     48     99
4521     90     95     89
```

A descent way of sampling can be achieved by `frac`

```
In [13]: # frac
rand_sample_df_f = df.sample(frac=0.1, random_state=42)
```

```
In [14]: # head
rand_sample_df_f.shape
```

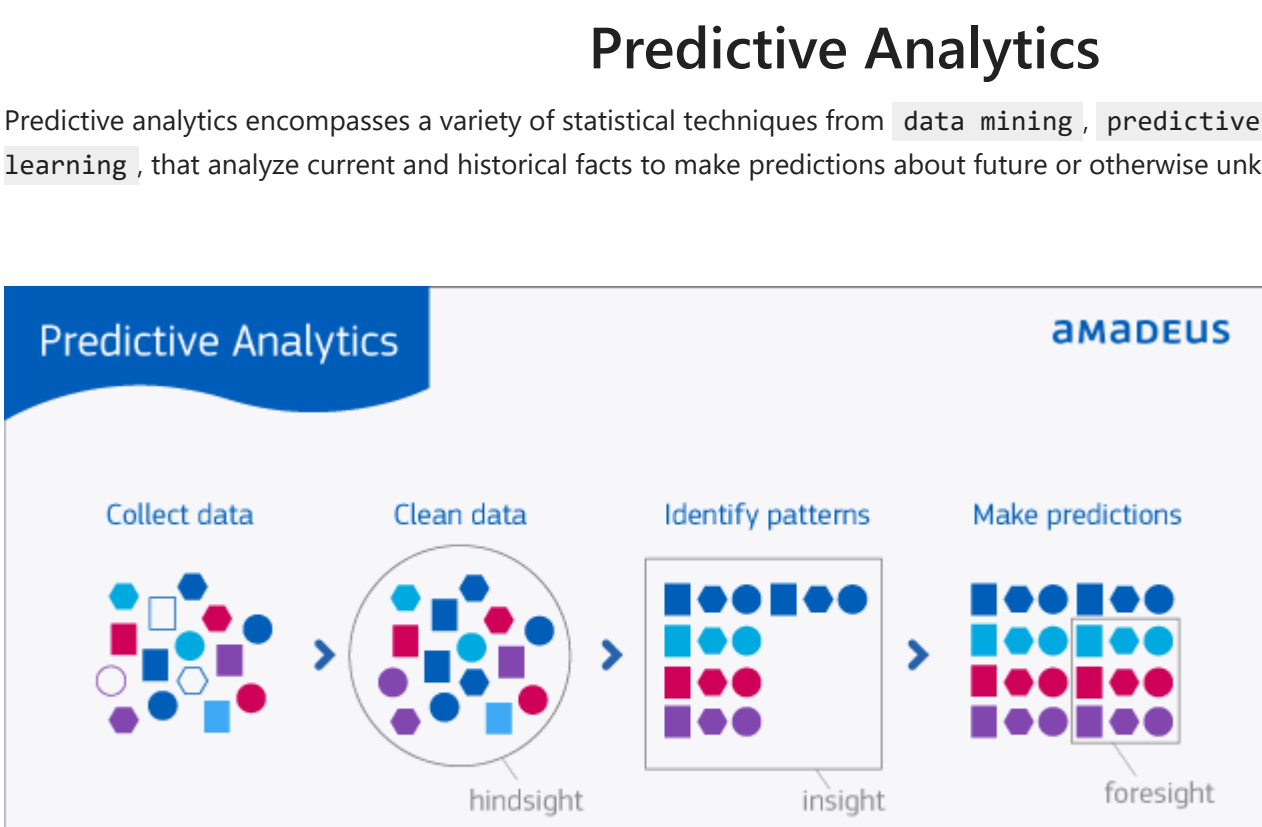
```
Out[14]: (1000, 3)
```

```
In [15]: # shape
rand_sample_df_f.head()
```

```
Out[15]:   col_1  col_2  col_3
6252     97     99     34
4684     28     95     54
1731     91     34     77
4742     17     48     99
4521     90     95     89
```

# Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from `data mining`, `predictive modelling`, and `machine learning`, that analyze current and historical facts to make predictions about future or otherwise unknown events.



Credits - Image from <http://bigdata-madesimple.com/>

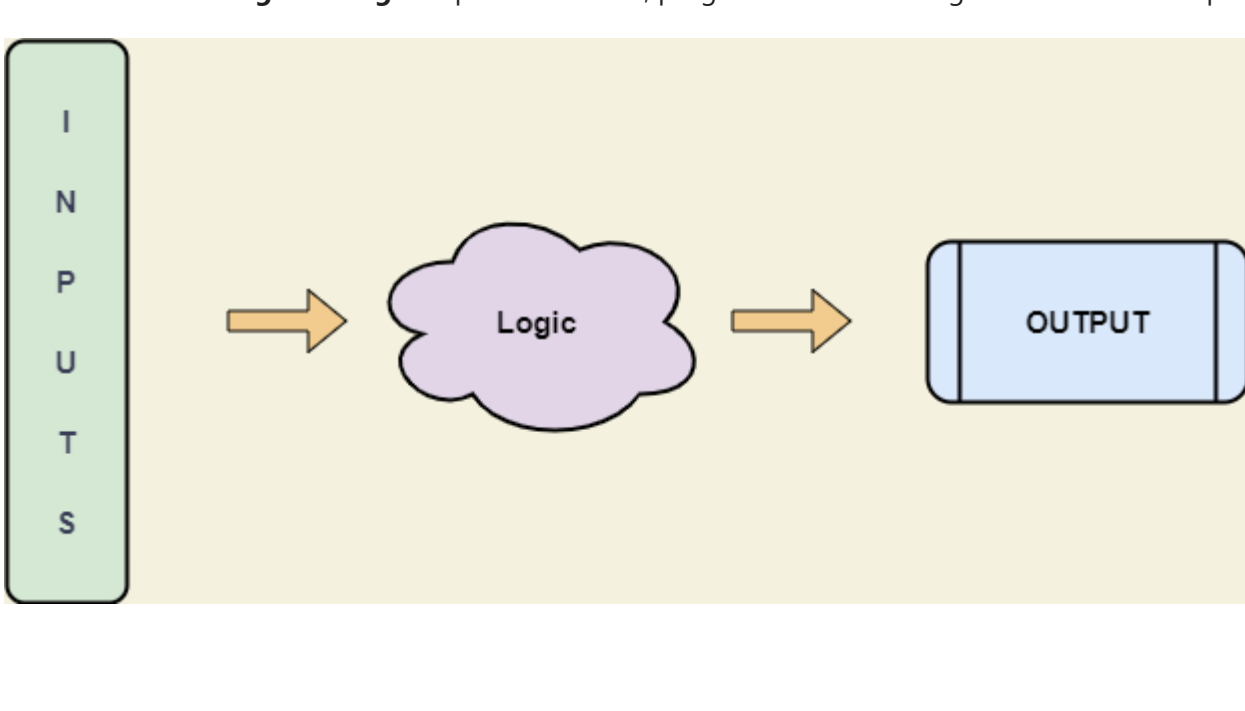
## Machine Learning

- ML is a technique followed to make a computer learn from the previous experience in order to make an assumption for the future outcome.
- It can learn and adapt to the new data without any human intervention.
- It needs prior training so that it can be tested to the new data.

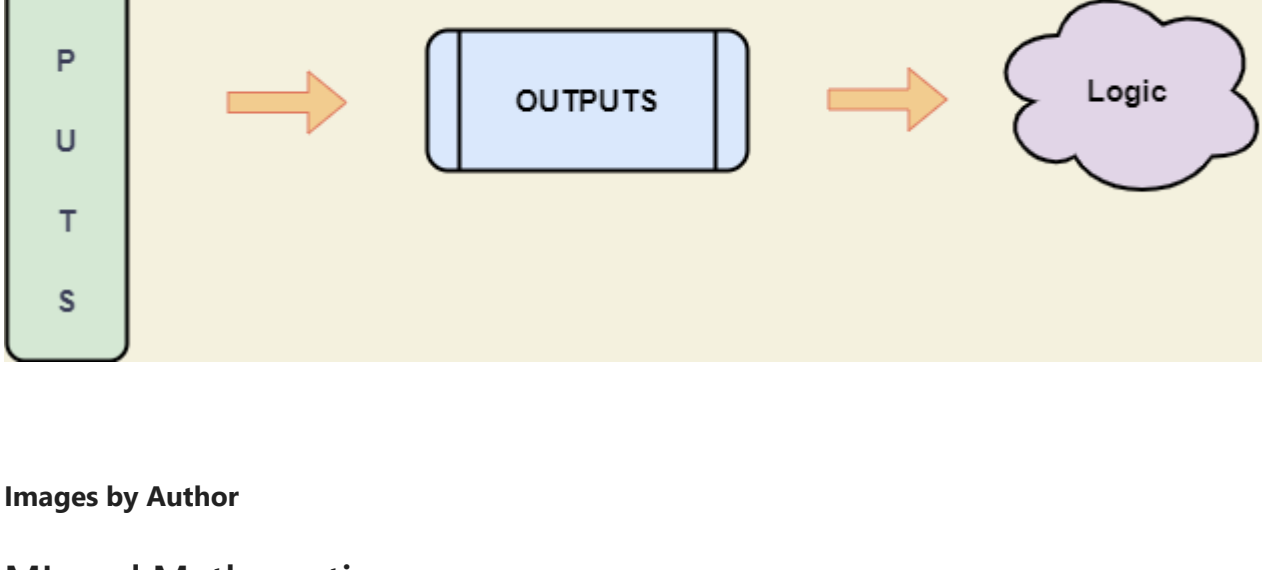
## What is this???

## ML and Traditional Programming

- **Traditional Programming** → Inputs are known, programmer writes the logic to obtain the Output.



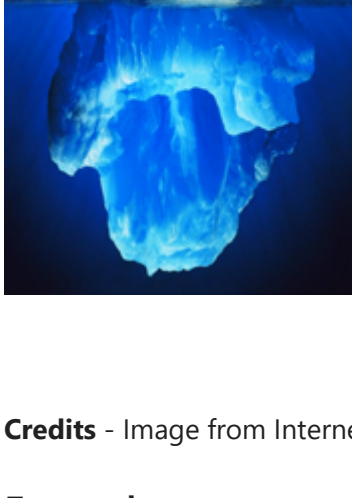
- **Machine Learning** → Inputs and Outputs are known, the algorithm tries to design it's own logic to map the inputs with the outputs.



Images by Author

## ML and Mathematics

- ML is just the tip of the iceberg.
- Math and Python code (algorithms) holding the iceberg is what we should be understanding.



Credits - Image from Internet

## Examples

- Email spam detector
- Auto-completion mode in the email
- Google photoes classification
- Weather forecasting - Time series prediction
- ...

## Types of ML

- **Supervised Learning**
  - The computer is presented with both example inputs and their respective outputs. The algorithm learns a general rule to map the inputs with the outputs.
- **Unsupervised Learning**
  - No outputs are given to the learning algorithm, instead the algorithm alone has to figure out the structure in the inputs and find the hidden patterns to get the final end.
- **Reinforcement Learning**
  - Works based on the reward system and the ultimate goal is to maximize the reward force.

## How much data do you really need for building a predictive model?

Often times, we have been told that - to build a machine learning predictive model, we need to have large amounts of data. Well that depends ultimately.

- Effective sampling is about maximizing the about (information) of the whole population from the sampling unit.
- A small random probability sample, as long as it is truly random and not biased in any way, can have very high predictive power.
- With less also, you can achieve more.