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 - Predictive Analytics

Population

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

Sampling

(method)

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

Sample

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

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

In [ ]:

Merits & Demerits

Merits

- Less cost effective
- Time saving
- Higher accuracy

Demerits

- Chances of biasness
- Need of subject specific knowledge

In [ ]:

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 to the whole population.
   - Systematic Sampling
   - Cluster Sampling
   - Stratified Sampling
2. **Non-Probability Sampling**
   - Convenience Sampling
   - Judgmental Sampling
   - Snowball Sampling
   - Quota Sampling

In [ ]:

Make random data using pandas

In [11]:
```
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)

In [21]:
```
# rand_data (population)
rand_data = np.random.randint(low=5, high=100, size=(1000, 3))
```

In [3]:
```
rand_data
```

Out[3]:
```
array([[69, 50, 36],
       [51, 97, 23],
       [79, 84, 90],
       ...,
       [76, 58, 90],
       [96, 84, 79],
       [99, 12, 70]])
```

Create a dataframe with columns and data generated

In [4]:
```
# df
df = pd.DataFrame(data=rand_data, columns=['col_1', 'col_2', 'col_3'])
```

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

Out[6]:

	col_1	col_2	col_3
0	69	50	36
1	51	97	23
2	79	84	90
3	46	57	15
4	74	30	80

Population data (df) size is 1000

- N = 1000

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

Out[7]:
```
(1000, 3)
```

Simple random sample

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

In [8]:
```
# rand_sample_df
rand_sample_df = df.sample(n=100, random_state=2)
```

In [9]:
```
# shape
rand_sample_df.shape
```

Out[9]:
```
(100, 3)
```

In [10]:
```
# head
rand_sample_df.head()
```

Out[10]:

	col_1	col_2	col_3
37	14	73	87
726	93	45	59
846	48	7	10
295	96	74	70
924	53	76	49

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

In [11]:
```
# frac
rand_sample_df = df.sample(frac=0.5)
```

In [12]:
```
# shape
rand_sample_df.shape
```

Out[12]:
```
(500, 3)
```

In [13]:
```
# head
rand_sample_df.head()
```

Out[13]:

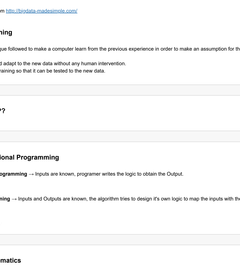
	col_1	col_2	col_3
825	63	8	49
203	27	17	57
732	9	45	74
588	22	76	66
46	45	38	88

In [15]:
```
# help(df.sample)
```

In [ ]:

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 https://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.

In [ ]:

What is this???

In [ ]:

ML and Traditional Programming

- **Traditional Programming** -> Inputs are known, programer 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

In [ ]:

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
- ...

In [ ]:

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 to 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 score.

In [ ]:

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.

More information -> https://www.sv-europe.com/blog/predictive-analytics-much-data-really-need/