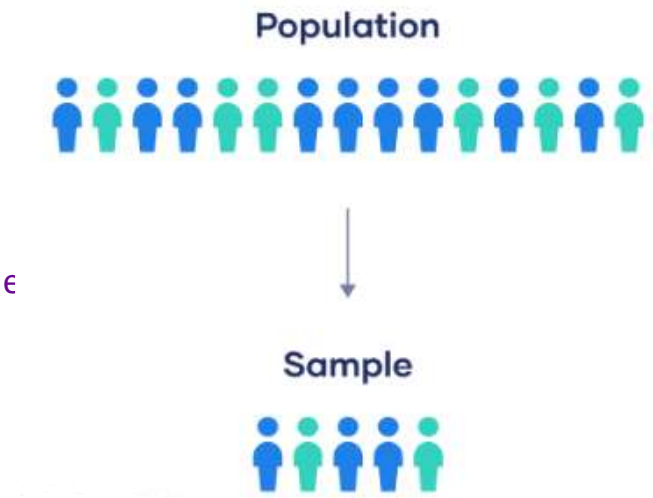
PROBABILITY SAMPLING IMPLEMENTED IN PYTHON

Github

POPULATION:

- A **population** is the entire group that you want to draw conclusions about.
- A **sample** is the specific group that we collect from population to represent the entire population.
- The size of the sample is always less than the total size of the population.
- **Sampling** is a method that allows that represents the information about the population based on results from sample.
- Sampling is categorized into 2 types:
    i. Probability Sampling
    ii. Non Probability Sampling



PROBABILITY SAMPLING:

- Probability sampling refers to the selection of a sample from a population, when this selection is based on the principle of randomization, that is, random selection or chance.
- When choosing a probability sample design, the goal is to minimize the sampling error of the estimates for the most important survey variables, while simultaneously minimizing the time and cost of conducting the survey.
- There are several different ways in which a probability sample can be selected. They are:
    i. Random Sampling
    ii. Systematic Sampling
    iii. Stratified Sampling
    iv. Cluster Sampling

RANDOM SAMPLING:

- In random sampling, every member of the population has an equal chance of being selected.
- The sampling frame should include the whole population.

```python
#Random Sampling
random_sample=data.sample(10)
```

✓ 0.0s Python

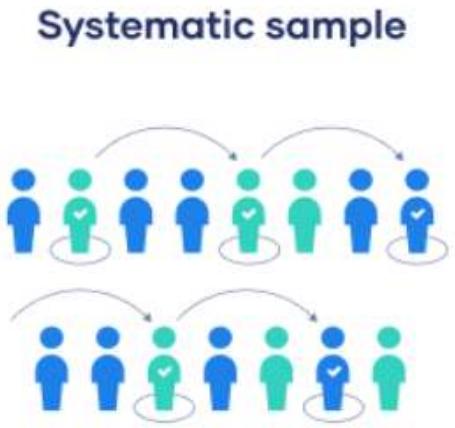
```python
random_sample
```

✓ 0.0s Python

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
103	104	6.3	2.9	5.6	1.8	Iris-virginica
91	92	6.1	3.0	4.6	1.4	Iris-versicolor
134	135	6.1	2.6	5.6	1.4	Iris-virginica
126	127	6.2	2.8	4.8	1.8	Iris-virginica
97	98	6.2	2.9	4.3	1.3	Iris-versicolor
83	84	6.0	2.7	5.1	1.6	Iris-versicolor
55	56	5.7	2.8	4.5	1.3	Iris-versicolor
24	25	4.8	3.4	1.9	0.2	Iris-setosa
125	126	7.2	3.2	6.0	1.8	Iris-virginica
26	27	5.0	3.4	1.6	0.4	Iris-setosa

SYSTEMATIC SAMPLING:

- Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct.
- Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.



```python
#Systematic Sampling
sys_sample=data.iloc[10:30:2]
```

✓ 0.0s Python

```python
sys_sample
```

✓ 0.0s Python

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
10	11	5.4	3.7	1.5	0.2	Iris-setosa
12	13	4.8	3.0	1.4	0.1	Iris-setosa
14	15	5.8	4.0	1.2	0.2	Iris-setosa
16	17	5.4	3.9	1.3	0.4	Iris-setosa
18	19	5.7	3.8	1.7	0.3	Iris-setosa
20	21	5.4	3.4	1.7	0.2	Iris-setosa
22	23	4.6	3.6	1.0	0.2	Iris-setosa
24	25	4.8	3.4	1.9	0.2	Iris-setosa
26	27	5.0	3.4	1.6	0.4	Iris-setosa
28	29	5.2	3.4	1.4	0.2	Iris-setosa

STRATIFIED SAMPLING:

- Stratified sampling involves dividing the population into subpopulations that may differ in important ways.
- It allows us to draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.
- To use this sampling method, we divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).
- Based on the overall proportions of the population, we calculate how many people should be sampled from each subgroup. Then we use random or systematic sampling to select a sample from each subgroup.

```python
#stratified Sampling
data.groupby('Species', group_keys=False).apply(lambda x: x.sample(2))
```

✓ 0.0s Python

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
41	42	4.5	2.3	1.3	0.3	Iris-setosa
34	35	4.9	3.1	1.5	0.1	Iris-setosa
59	60	5.2	2.7	3.9	1.4	Iris-versicolor
55	56	5.7	2.8	4.5	1.3	Iris-versicolor
112	113	6.8	3.0	5.5	2.1	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica

CLUSTER SAMPLING

- Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample.
- Instead of sampling individuals from each subgroup, we randomly select entire subgroups.
- If it is practically possible, we might include every individual from each sampled cluster.
- If the clusters themselves are large, we can also sample individuals from within each cluster using one of the techniques above.

```python
#Cluster Sampling
clusters = np.random.choice(np.arange(1,151), size=6, replace=False)
```

✓ 0.0s Python

```python
clusters
```

✓ 0.0s Python

```
array([105, 138,  55,  92,  18, 133])
```

```python
cluster_sample = data[data['Id'].isin(clusters)]
```

✓ 0.0s Python

```python
cluster_sample
```

✓ 0.0s Python

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
17	18	5.1	3.5	1.4	0.3	Iris-setosa
54	55	6.5	2.8	4.6	1.5	Iris-versicolor
91	92	6.1	3.0	4.6	1.4	Iris-versicolor
104	105	6.5	3.0	5.8	2.2	Iris-virginica
132	133	6.4	2.8	5.6	2.2	Iris-virginica
137	138	6.4	3.1	5.5	1.8	Iris-virginica