The Center of Applied Data Science

python™

Sampling

Descriptive Statistics

In [ ]:

```
import numpy as np
import pandas as pd
import seaborn as sns
```

In [ ]:

```
import platform

sns.__version__ >= "0.9.0"
pd.__version__ >= "0.23.4"
np.__version__ >= "1.15.4"
platform.python_version() >= "3.6"
```

Content Outline

In [ ]:

```
import random
random.seed(42)
```

1. Introduction

Motivation

In a state of approximately 60,000 residents there are two candidates running for the post of Member of Parliament (MP) - Candidate A and Candidate B. A political analyst would like to determine whether the voting residents favor one candidate over the other. To accomplish this, he decides to run a poll for all voting residents of the state.

Based on the above scenario, take a minute to give your thoughts on the following:

1. Would it be possible to poll every single voting resident of the state?
2. What would be the cost of such an exercise?
3. Can we employ some strategy to obtain a reasonably similar result without expending as much resources?

Thankfully we can save quite a bit of time (and money!) by polling only a portion of the voters and with the magic of statistics, make reasonable inferences about the voting preferences of all voters in the state. This process is called **sampling.**

Populations

Before we dive deep into the process of sampling, let's first take a look at some preliminary concepts to start ourselves off on solid ground. In our motivational example above, the object of our interest is the voting preferences of the state's 60,000 residents, i.e. we are interested in knowing the vote of *all* voting residents in the town. In statistical terms, the set of all these voting residents is called a **population**. Simply put, a population is the set that contains **all** elements of interest for a particular study, and it exists to ensure that we don't draw observations from data unrelated to the problem at hand.

The definition of what constitutes a population is highly dependent on the context of the study, and to this extent a little bit of domain knowledge goes a long way.

Example

1. A researcher wishes to study the effect of steroid use in the National Football League (NFL). The population defined here would be the set of all active professional players in the NFL.
2. A grocery store owner would like to identify the most popular cereal he has on sale. The population defined here would be the set of all cereal products he currently sells at the store.
3. The Human Resource Department at KFRU, a large radio station, is interested in evaluating the effect of internal training programs on their staff. The population defined here would be the set of all current employees in the company.

In each of the above cases, the elements of interest are either people (NFL players, employees) or products (cereals) are said to be **members** of the population.

Guided Exercise

A. Suppose that you want to determine the movie preferences of students at a given university. Who/what are the members in your population?

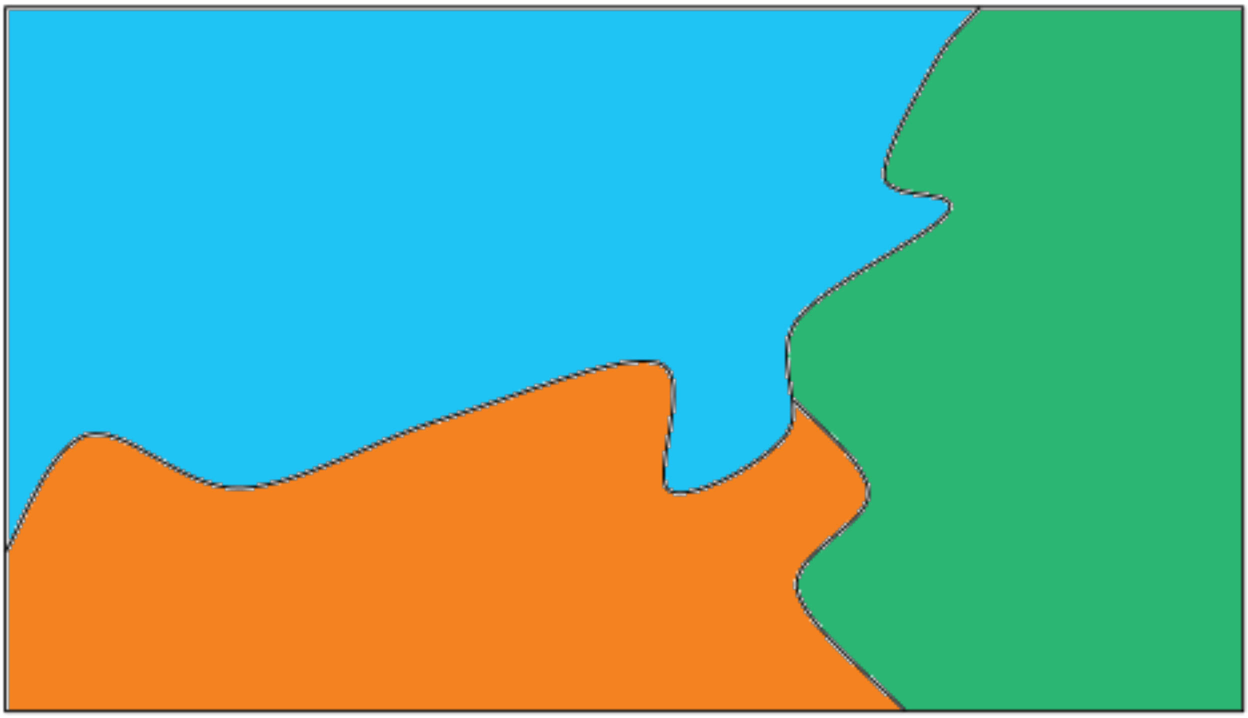
B. Suppose that you would like to study the job satisfaction of teachers in international schools in KL. Who/what are the members in your population?

```
# Type your answer here.
A. All students at the university.
B. All teachers working at international schools in KL.
```

Sub-populations

In many cases, a population can be divided into smaller groups called **sub-populations**, allowing us to narrow down our point of interest. For example, the population of all employees in a given company can be further split into all *male* employees in the company and all *female* employees in the company. Sub-populations can be defined by any chosen characteristic, but the division must obey two rules:

1. A member of one group cannot belong to another group. (**mutually exclusive**)
2. Each member in the population belongs to a group. (**collectively exhaustive**)



Example

Recall the Human Resource Department at KFRU from our previous example. We can sub-divide the population of all employees to the *employees from each department*.

There is no restriction on the number of sub-populations (as how we define a sub-population is subject to domain knowledge), and we can even go further to sub-divide the sub-populations itself!

Example

We can divide the employees in the Human Resource Department into *male* and *female* employees of the department.

Now, you may be wondering - do I really need to take every single sub-population into account?

Well, the answer is **no**. Sub-populations can safely be **ignored** if they aren't *perceived to have an impact* on the *object of interest* in our study.

Example

We wish to study the average income of a fresh graduate in KL but we are *not interested* in knowing if there is a significant difference in income between male and female fresh graduates. In such a case, we ignore the sub-populations of male and female fresh graduates and treat them as one whole set.

Guided Exercise

The `voter` data set for the state in our motivational example has been pre-loaded and summarized for you below. Examine the output and discuss how we can construct various sub-populations from the data.

In [ ]:

```
import pandas as pd
dtype = {'state_seat': str,
         'polling_district': str,
         'postcode': str}
voter = pd.read_csv('../data/voter.csv', dtype = dtype,
                    usecols = np.arange(1, 7))
```

In [ ]:

```
voter.head()
```

In [ ]:

```
voter.describe()
```

```
# Type your answer here.
There are 5 attributes that can be used to form sub-populations:
1. Gender
2. State Seat
3. Polling District
4. Ethnicity
5. Postcode
```

Exercise

The KFRU `HR` data set has been pre-loaded and summarized for you below. Examine the head of the DataFrame and discuss how we can construct various sub-populations from the data.

In [ ]:

```
dtype = {'Employee.Number': str,
         'Zip': str, }
hr = pd.read_csv('../data/hr.csv', dtype = dtype, parse_dates
hr.head().T
```

In [ ]:

```
 
```

In [ ]:

```
 
```

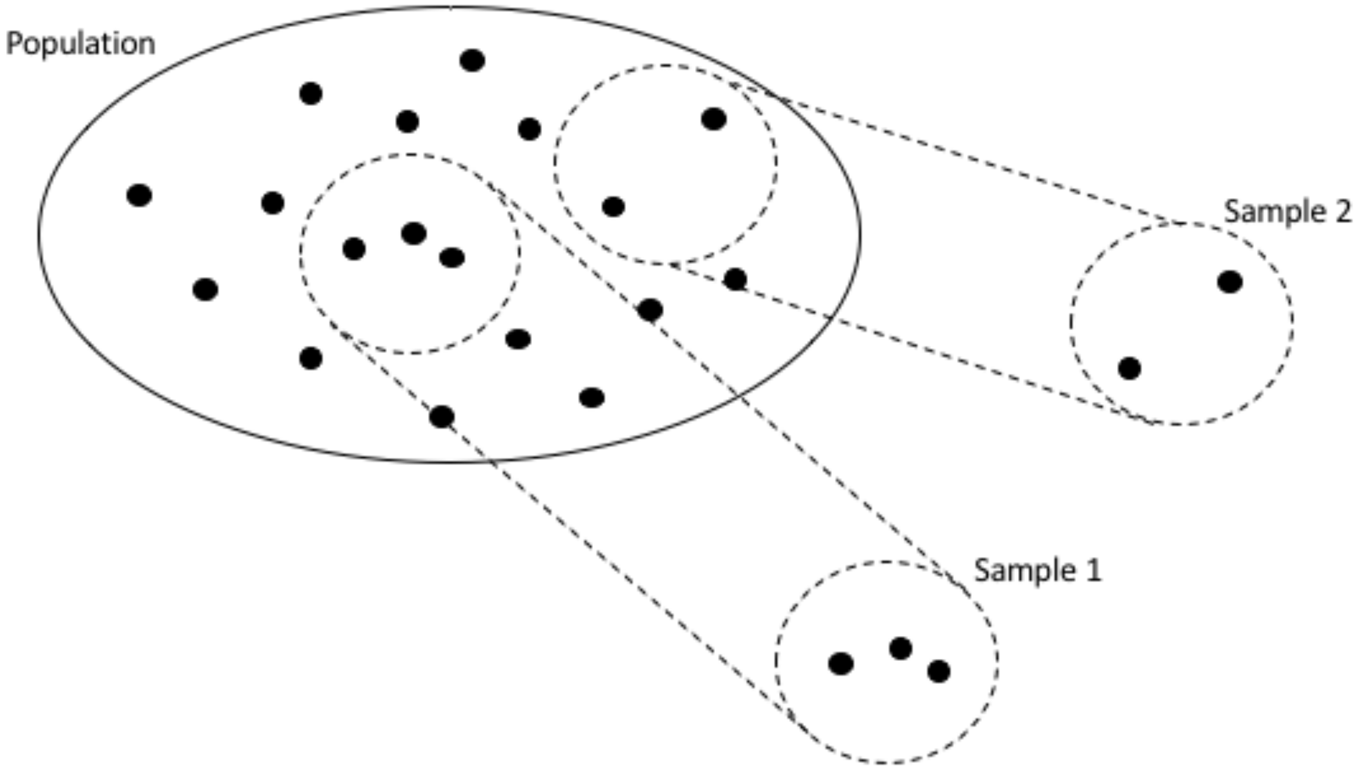
In [ ]:

```
 
```

```
# MC
# What attributes can be used to form sub-populations?

There are 11 attributes that can be used to form sub-populations:
1. Zip
2. Age
3. Sex
4. Marital Status
5. Citizenship Status
6. Racial Descent
7. Employment Status
8. Department
9. Position
10. Employee Source
11. Performance Score
```

Samples

As indicated in the beginning of this course, surveying an entire population can be a costly affair, and in most cases it is impossible to collate data from every single member of a population. By selecting/drawing a subset of the population, we obtain a **sample**, which can be used to estimate the properties of the population it is drawn from. We can draw as many samples as we want from a population, and we can even dictate the **size** of the samples, i.e. the *number of members in the sample*.



To avoid confusion, we commonly denote the size of a sample by $n n$, and the size of a population by $N N$. Sample sizes are restricted to the range $1 \leq n \leq N 1 \leq n \leq N$ as selecting none of the members yields nothing, and we cannot select more members than there are in the population. In the case of the diagram above, we have a population of size $N = 18 N = 18$, with two samples of size $n_1 = 3 n_1 = 3$ and $n_2 = 2 n_2 = 2$ respectively.

Example

Recall the example where we wished to study the average income of a fresh graduate in KL. Suppose that there are a total of N = 20,000 fresh graduates working in the city. To *draw* a sample of size n = 100, we would **select** 100 fresh graduates from this group of 20,000 to inquire about their income.

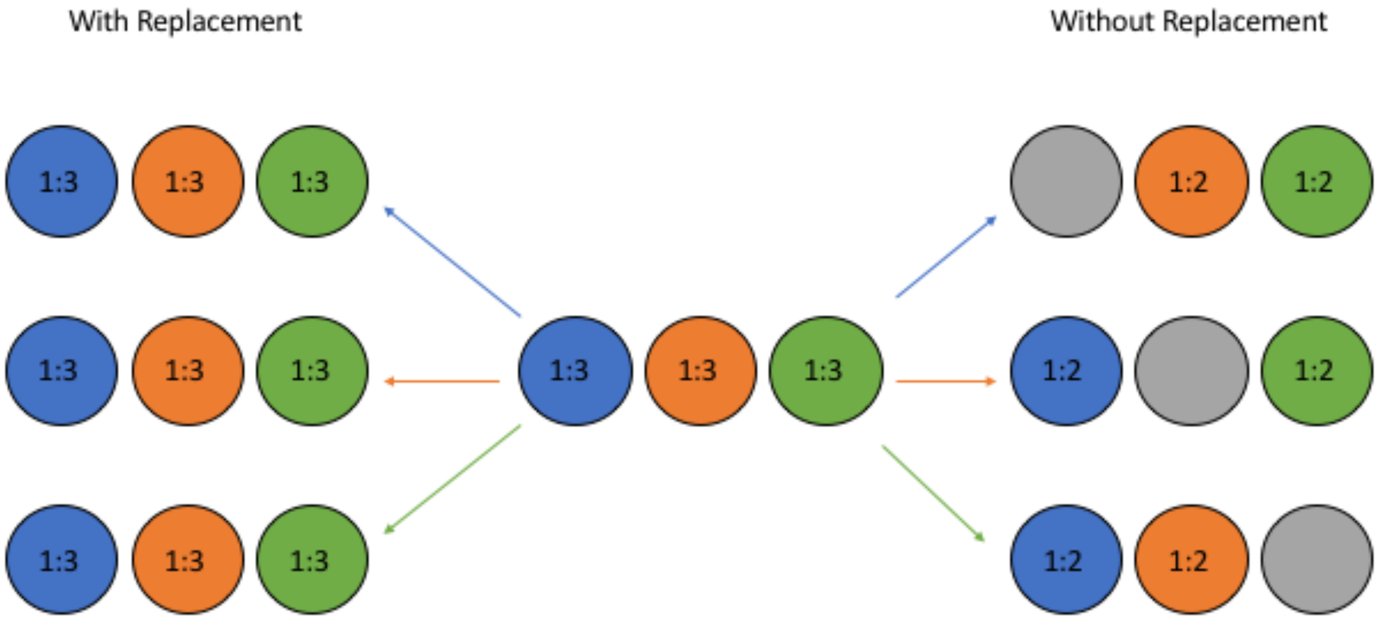
Samples can be drawn in **two** ways - *with* or *without* replacement. To illustrate the differences between these techniques, consider the following scenario:

Three colored balls are available to be drawn from a box - one blue, one orange, and one green. We would like to draw a sample of 2 balls and record the colors. The acts of drawing a ball are defined as follows:

Drawing *with* replacement: A ball is drawn and its color is recorded. The ball is then put back in the box before the second draw to be made.

Drawing *without* replacement: A ball is drawn and its color is recorded. The ball is **not** put back in the box before the second draw to be made.

The diagram below gives the odds of obtaining a ball of a specific color for the first and second draw.



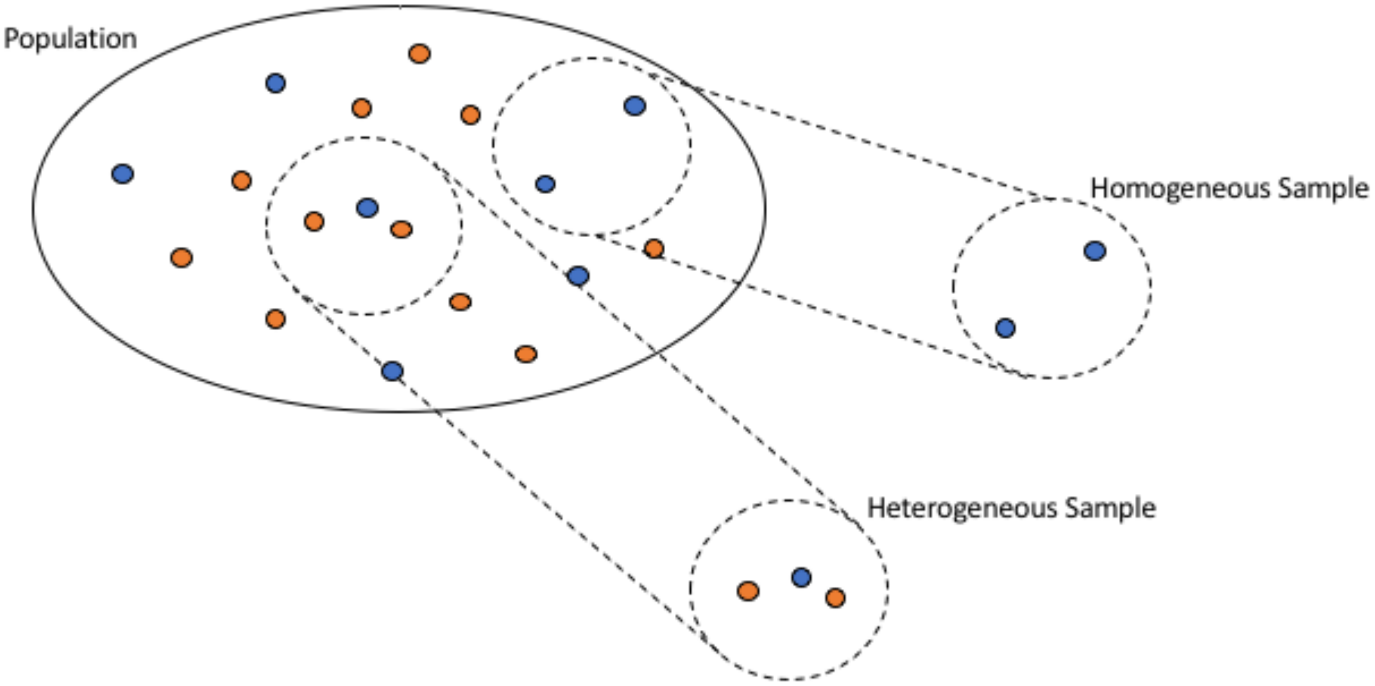
Odds of Selection for 3 Colored Balls

Note that on the first draw, both techniques have the same odds - 1 in 3 of obtaining a ball in each color. However, this no longer the case once the second ball is drawn. If the first ball drawn was not placed back in the box before the second draw, the odds of obtaining a ball of a different color than the first has now gone up to 1 in 2!

Sampling of real-world data is generally done **without replacement**. Sampling with replacement on the other hand is used in *oversampling* techniques, which are commonly employed in machine learning (and beyond the scope of this course).

Homogenous and Heterogeneous Samples

Samples are classified as **homogeneous** or **heterogenous** based on the members they contain, with the former having all members drawn from the *same sub-population* and the latter having members drawn from *multiple sub-populations*. The diagram below illustrates the difference between homogeneous and hetereogeneous samples:



Here we have a population that can be broken into two sub-populations. Samples that contain only members of *one* sub-population are deemed homogenous, and samples that contain members of *both* sub-populations are deemed heterogeneous.

Example

Suppose our population is defined as "All employees in KFRU."

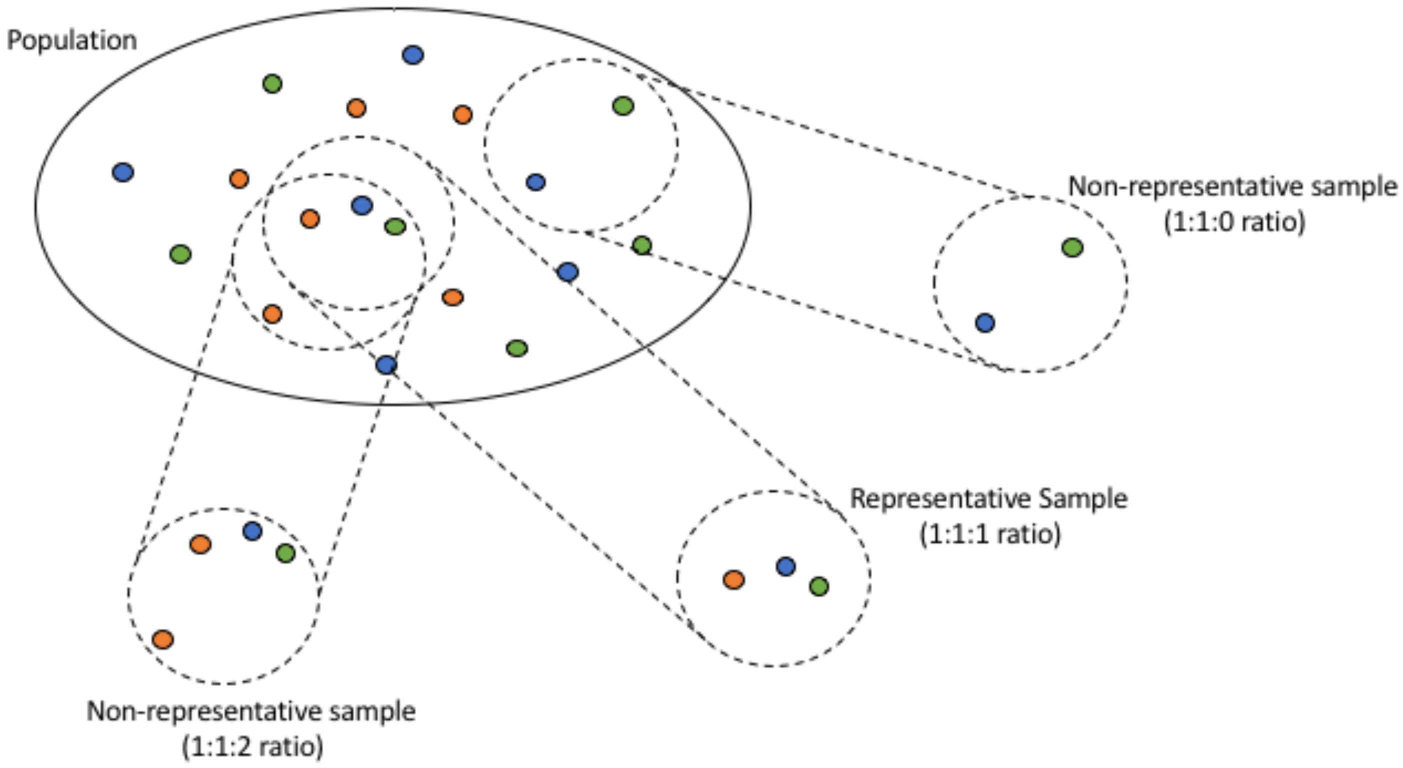
Homogeneous sample:

A sample of size $n = 25$ $n = 25$ where all sample members are from IT.

Heterogeneous sample:

A sample of size $n = 25$ $n = 25$, where 13 sample members are from IT, 2 are from Production, and the remainder are from Sales.

In the interest of ensuring that our samples accurately reflect the demographic proportions of the population, we often strive to draw samples with the same ratio of members across the various types to that of the population. As you can very well guess, obtaining an exact ratio isn't always possible if you fix the sample size so for cases such as these we favor sampling a **proportion** of the population.



In the above illustration, the population has a $1 : 1 : 1$ $1:1:1$ ratio between its 3 sub-populations. Taking a sample with the same ratio of members to that of the sub-population ratio results in what we call a **representative sample**. Conversely, if a sample is not representative we say it is **biased**. Numerical measures such as *mean* and *variance* for a biased sample do not accurately reflect that of the population and if used for business decision making can potentially lead to disastrous results.

Example

In a company there are 3 senior, 6 mid-level, and 9 junior employees.

Representative sample: A sample of size $n = 6$ $n = 6$ consisting of 1 senior, 2 mid-level, and 3 junior employees. This amounts to sampling 30% 30% of the population.

Biased sample: A sample of size $n = 6$ $n = 6$ consisting of 3 senior and 3 mid-level employees.

Guided Exercise

Using the `voter` data set, draw the following sample:

A sample of 1,000 voters (without replacement) from `polling_district` 16. Assign this to the variable `mysample1` and examine the resulting sample using the `describe` method.

In [ ]:

```
np.random.seed(420)

mysample1 = voter.loc[voter.polling_district == "16",].\
sample(n = 200, replace = False)
```

In [ ]:

```
mysample1.describe()
```

In [ ]:

```
voter.describe()
```

```
# MC
# Type your answer here.

1. The sample obtained is heterogeneous with respect to gender/ethnicity.
2. The sample is not representative.
```

Exercise

Using the `hr` data set, draw the following sample:

A sample of 20 employees (without replacement) from the `IT/IS` department. Assign this to a variable of your choice and examine if the resulting sample is representative using `describe` .

In [ ]:

```

```

In [ ]:

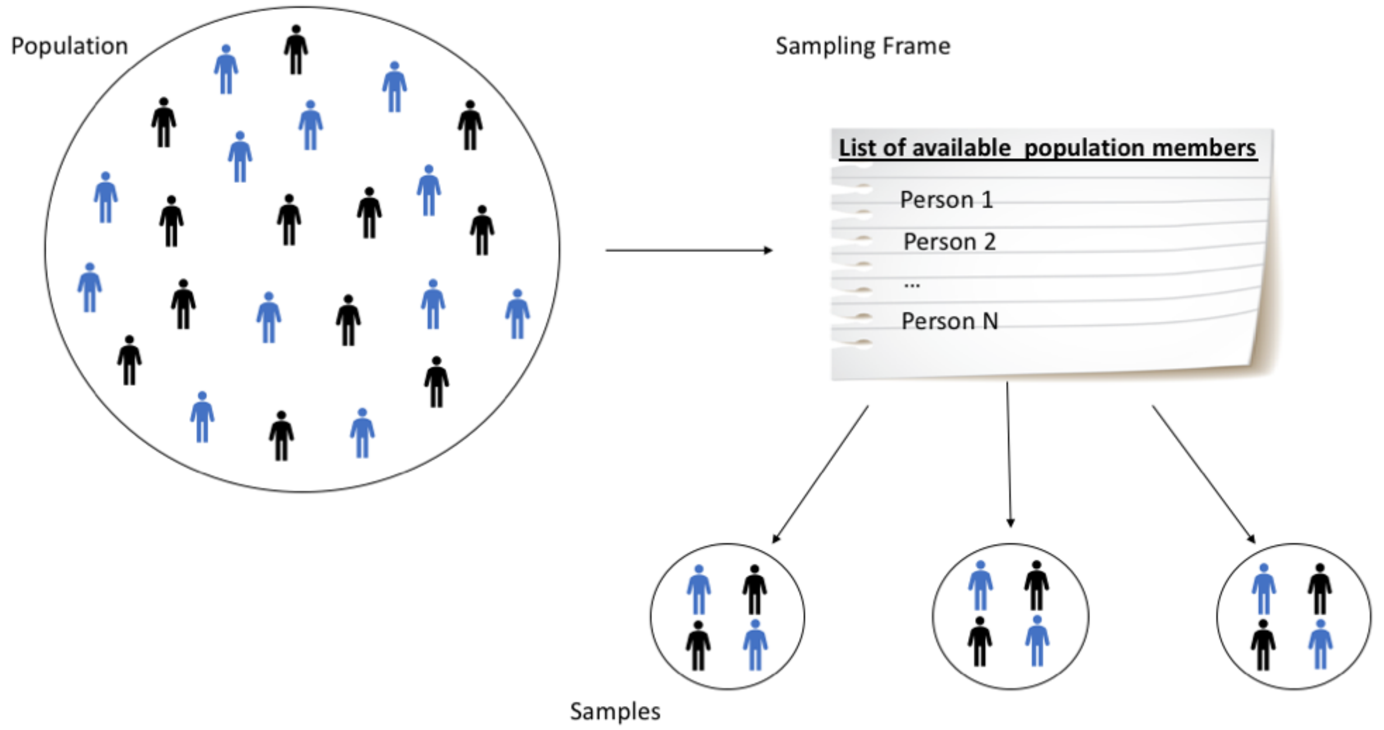
```

```

```
# MC
# Type your answer here.
# Examine the sample as in the previous example.

1. The sample obtained is heterogeneous with respect to zip/age/sex/marital status/citizenship status/racial descent/
employment status/position/employee source/performance score.
2. The sample is not representative.
```

Sampling Frames and Sampling Units

As we saw in the section on populations, a population needs to be defined such that it covers the entirety of data that is of interest in a given study. This raises the question of *how* we can identify each individual member of our population to facilitate drawing samples. To this extent, we would like to define a **sampling frame**, i.e. a *list* of all the members in the population for us to sample from. We will use this sampling frame as a basis to draw samples using one or more prescribed methods (sampling methodologies) to ensure consistency.

As you can very well deduce, it may not always be possible to get an entire list of the members in a population. In cases such as these, we strive to define a sampling frame that is **as close as possible** to the population in order to **minimize** bias. Looking back at our motivational example, the sampling frame here would be the *roster of registered voters* in the state!



The table below summarizes the concept of how we would draw samples from a population using a sampling frame:

Component	Example
Population	A company's entire customer base
Sampling Frame	Those customers the company has access to (contact details available)
Samples	Customers who you contact and actually respond to your survey

In addition to a sampling frame, we also wish to define the number of members we select at a single time when we draw our samples. This selection size is called the **sampling unit**. Sampling units come in handy when we wish to widen/narrow our scope without altering our entire sampling methodology.

Example

1. Recall the example where we wished to study the average income of a fresh graduate in KL. As we expect the income to vary between individuals, we would select the sampling unit here as *one person*.

2. Suppose we want to study the average household income across a country instead. Since we are not interested in the individual incomes of each family member, we can widen the sampling unit to *one household*.

2. Sampling Methodologies

There are a variety of *techniques* (methods) in which sampling can be carried out, each with its own strengths and weaknesses. These techniques can generally be broken down into two major categories - **probabilistic** and **non-probabilistic sampling**.

Probabilistic Sampling

Probabilistic sampling is a sampling technique in which the *likelihood* that a member of the population is selected as part of a sample is **known** (or can be calculated). There are **5** sampling techniques that fall under the umbrella of probabilistic sampling, namely:

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. Systematic Sampling
5. Multistage Sampling

To avoid confusion (and a ton of mathematical jargon), we will focus on the process of each sampling technique and skip the calculation of probabilities etc.

Simple Random Sampling

In simple random sampling, each member of the population is **equally likely** to be selected as part of the sample. This can be done by **randomly selecting** members from the population.

Example

Suppose we wish to study the the job satisfaction of teachers in international schools in KL. We would first need to obtain a sampling frame, i.e. compile a list of all teachers in international schools and their contact details. We would then randomly choose names on this list to contact for interviewing.

Pros

1. Samples can easily be drawn with the assistance of statistical software suites.
2. Selection bias is minimal as the process of drawing members of the population is random.

Cons

1. Requires a well-defined sampling frame, which in practice may not always be available.
2. May not yield a representative sample as there is no stipulation on the ratio of members drawn.

Guided Exercise

Draw a simple random sample of 2,500 voters from the `voter` data set. Assign this to the variable `mysample2` and use bar graphs to examine the demographics of the resulting sample with respect to `gender` , `ethnicity` , and `state_seat` .

In [ ]:

```python
mysample2 = voter.sample(n = 2500, replace = False)
```

In [ ]:

```python
def compare_pop_sample(pop_df, sample_df, col):
    pop = pop_df[col].value_counts(normalize = True)
    sample = sample_df[col].value_counts(normalize = True)
    pop_sample = pd.concat([pop, sample],
                           axis = 0,
                           keys = ["pop", "sample"]).reset_in
    sns.catplot(x = "level_1", y = col,
                hue = "level_0", kind = "bar",
                data = pop_sample)
```

In [ ]:

```python
compare_pop_sample(voter, mysample2, "gender")
```

In [ ]:

```python
compare_pop_sample(voter, mysample2, "ethnicity")
```

In [ ]:

```python
compare_pop_sample(voter, mysample2, "state_seat")
```

Exercise

Draw a simple random sample of 200 employees from the `hr` data set. Assign this to a variable of your choice and use bar graphs to examine the demographics of the resulting sample with respect to `MaritalDesc` and `Department`.

In [ ]:

```python

```

In [ ]:

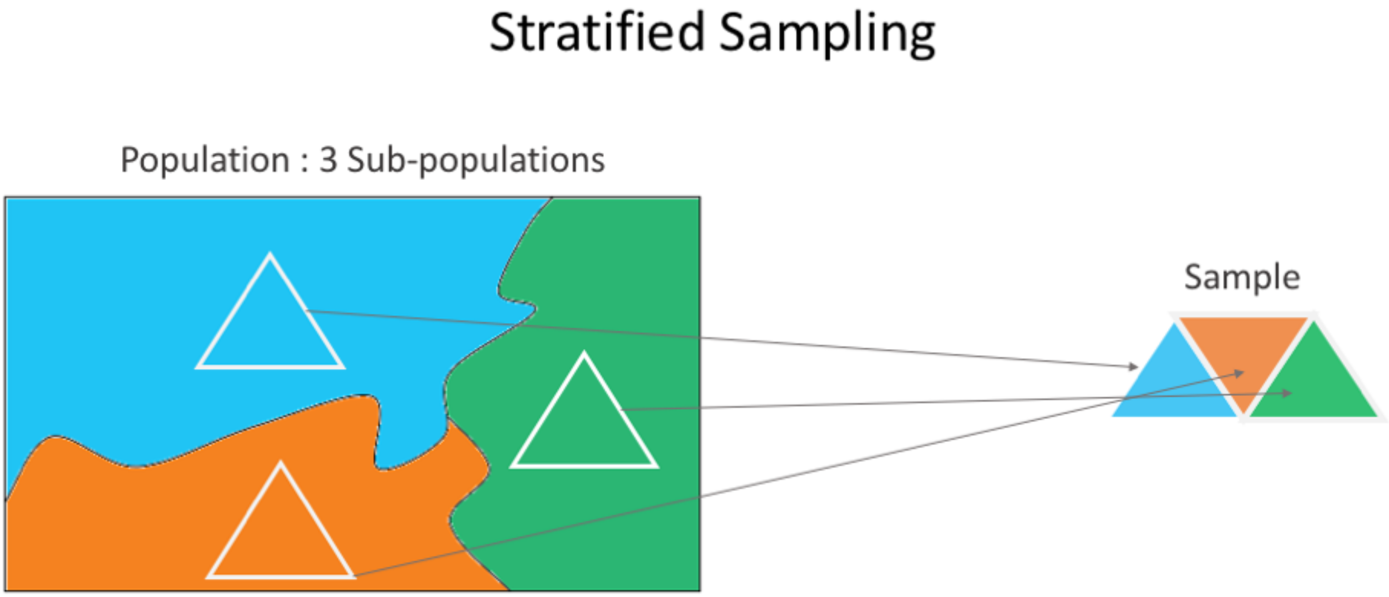
```python

```

In [ ]:

```python

```

Stratified Sampling

Stratified sampling is a sampling approach that is aimed at minimizing *representation bias*. It invovles grouping members of the population by their sub-populations and drawing samples from each one to preserve the ratio/balance from each group. For each sub-population $k k$, we draw a fixed percentage of $N_k N_k$ members to make up the sample.



Note that here we do not fix the sample size unlike simple random sampling. Instead, we specify a percentage of the population to allow for easier computation of how many members to draw from each sub-population.

Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. Suppose that we wanted to interview 30% of all teachers. We would then group the teachers by school and randomly select 30% from each school to make up our sample.

Stratified sampling is useful when we want to ensure that all the various sub-populations are well-represented. National census data is often collected in such a manner to ensure that minority groups etc. are taken into account when federal policies are developed.

Pros

1. Samples can easily be drawn from each sub-population with the assistance of statistical software suites.

2. Representation bias is minimal as stratification takes into account the ratio of sub-populations.

Cons

1. Requires a well-defined sampling frame, which in practice may not always be available.

2. If the variance within each sub-population is significantly different, stratified sampling may result in the sample variance being skewed.

3. If data collection involves travel to various geographical locations, stratification may incur greater cost as each sub-population needs to be accounted for.

Guided Exercise

Draw a sample of 10% of voters from the `voter` data set, stratified by `gender`. Assign this to the variable `mysample3` and examine the resulting sample using `describe`.

In [ ]:

```python
voter_stratified_sample = voter.groupby('state_seat').\
apply(lambda x: x.sample(frac = .1, random_state = 42))

voter_stratified_sample.describe()
```

In [ ]:

```python
round(voter_stratified_sample.state_seat.value_counts()/voter
```

In [ ]:

```python
compare_pop_sample(voter, voter_stratified_sample, "gender")
```

Exercise

Draw a sample of 20% of employees from the `hr` data set, stratified by `Sex`. Assign this to a variable of your choice and examine the resulting sample using `describe()`.

In [ ]:

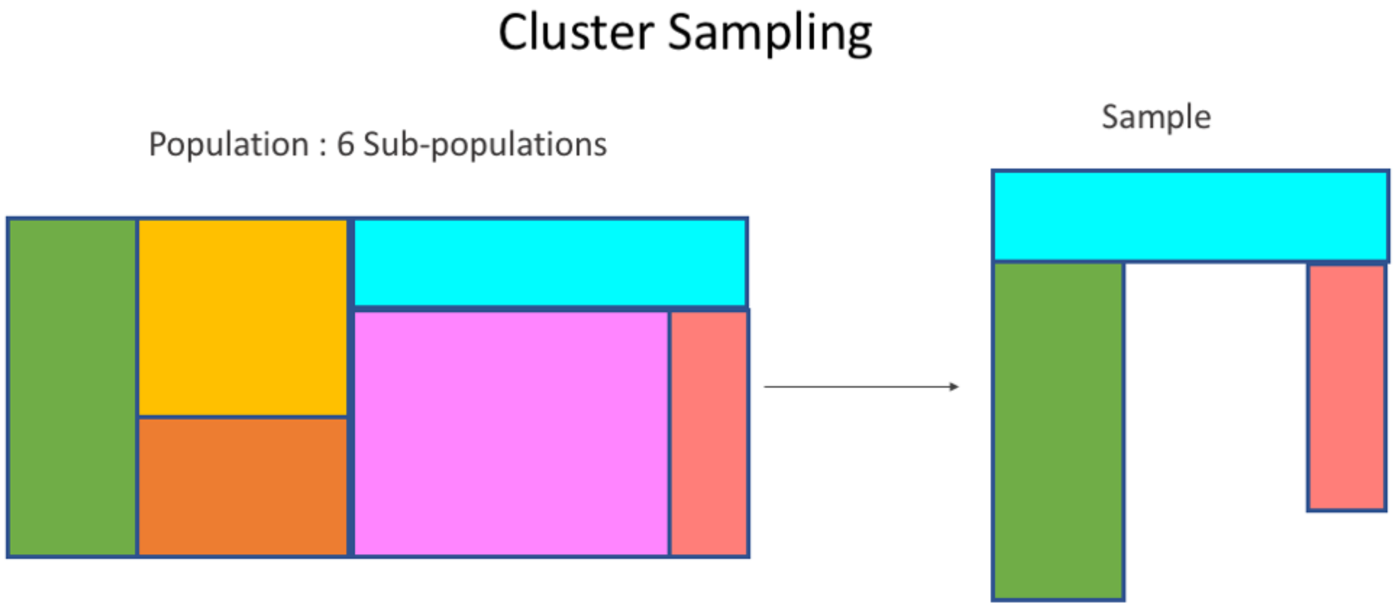
```

```

In [ ]:

```

```

Cluster Sampling

Cluster sampling is similar to stratified sampling in the sense that the sampling is done over sub-populations. The main difference however, is that instead of randomly sampling **within** each sub-population with the intention of preserving proportionality, we **randomly select** a **chosen number** of sub-populations and then proceed to sample from the selected sub-populations. If *all* elements in a chosen cluster are sampled, we call the process **one-stage cluster sampling**.



In the diagram above, we see that 3 random sub-populations (clusters) were chosen from the 7 that exist in the population. Though the number of clusters chosen is entirely up to the researcher, they are commonly dictated by resource availability (cost, manpower, etc.)

Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. To apply one-stage cluster sampling here, we would randomly select a number of international schools within KL (e.g. 20) and interview *all* teachers at these schools.

Cluster sampling is better suited towards large populations where sampling within clusters is easily done, but sampling **across** clusters is difficult. Geographical factors such as distance are often used as a measure of when cluster sampling is required.

Example

To ascertain a specific opinion from all Malaysians within Peninsula Malaysia, it is easier to divide the peninsula into its various states and perform cluster sampling, i.e. sampling only randomly selected states to cut down on traveling costs.

Pros

1. Clusters can easily be selected with the assistance of statistical software suites and/or random number generators.
2. Works well for large large populations.

Cons

1. Requires more than 2 sub-populations to be effective. Sampling from only 1 out of 2 sub-populations is biased.
2. Less control over sample size, poses a problem for populations where the number of members in each sub-population is not equal/similar.

Guided Exercise

Draw a one-stage cluster sample from the `voter` data set, clustered by `state_seat` by setting the number of clusters to `2` and using simple random sampling of clusters (`np.random.choice`). Assign this to the variable `mysample4` and examine the resulting sample using `describe` .

In [ ]:

```
# choose clusters with np.random.choice
clusters = np.random.choice(voter.state_seat.unique(),
                            size = 2,
                            replace = False)

cluster_mask = [i in clusters for i in voter.state_seat]

mysample4 = voter.loc[cluster_mask, :]

mysample4.describe()
```

Exercise

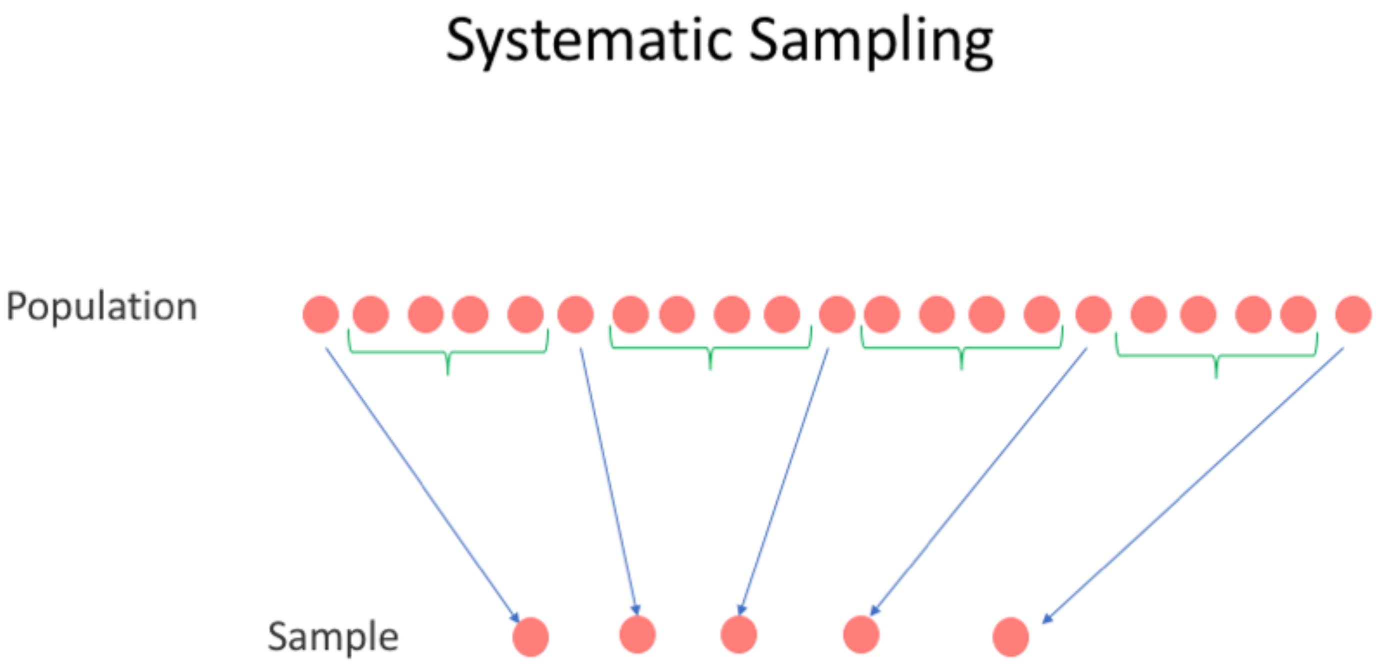
Draw a one-stage cluster sample from the `hr` data set, clustered by `Department` by setting the number of clusters to `4` and using simple random sampling of clusters. Assign this to a variable of your choice and examine the resulting sample using `describe` .

In [ ]:

In [ ]:

Systematic Sampling

Systematic sampling is carried out by randomly selecting an initial member of the population from the sampling frame and selecting additional members following a **pre-determined sequence**. This sequence is in the form of the index number for a member incremented/decremented by a fixed step size h h. For example, if we start with the 5th member of the population and set a step size of $h = 2$ $h = 2$, we will then select the 7th, 9th, 11th (so on and so forth) members of the population until we obtain a sample of desired size. Note that we can use modulo arithmetic to ensure that we don't overshoot the sample size when counting the indices.



Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. To obtain a systematic sample, we index the list of teachers and pick an random number (e.g. 45) as our starting index. If we want to select 100 members with the step size being $h = 10$ $h = 10$, then we select the 45th, 55th, 66th, ..., 445th members as our sample.

Systematic sampling is generally favored when the population is known to be large as it is easier to collect one large systematic sample than to draw multiple simple random samples. The conveniece however does come at the cost of bias if the data is ordered, e.g. picking 10 members from a set of 1,000 exam scores sorted in descending order will give a sample with a high average score if the step size is too small.

To deal with biases such as these, there is a variation of systematic sampling called *random systematic sampling*. This method is beyond the scope of this course as it requires in-depth knowledge of probability.

Pros

1. Only requires one instance of random selection, the remaining members are selected using a pre-determined sequence.
2. Easy to carry out, even with a loosely-defined sampling frame.

Cons

1. May introduce bias if the sampling frame is arranged to begin with.
2. If sub-populations exist, the resulting sample may not be representative.

Guided Exercise

Draw a systematic sample of size $n = 250$ $n = 250$ from the `voter` data set with step size `175` . Assign this to the variable `mysample5` and examine the resulting sample using `head` and `describe` .

In [ ]:

```
start = np.random.choice(len(voter), 1)
h, n = 175, 250
index = start + h * (np.arange(n)- 1)
mysample5 = voter.iloc[index % len(voter),]
```

In [ ]:

```
mysample5.head()
```

In [ ]:

```
mysample5.describe()
```

Exercise

Draw a systematic sample of size $n = 50$ $n = 50$ from the `hr` data set with step size `5` . Assign this to a variable of your choice and examine the resulting sample using `head` and `describe` .

In [ ]:

```

```

In [ ]:

```

```

Multistage Sampling

As the name suggests, multistage sampling is performed by **combining** two or more of the above sampling methodologies in **series**. The objective is to retain the strengths of each technique by leveraging the *order* in which we combine these techniques.

Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. Suppose we want to stratify our samples by gender of the teachers, but we would also like to conduct cluster sampling to reduce having to travel to all the different schools in KL. By combining both approaches, we can select a cluster of schools *then* pick stratified samples from each cluster.

As seen above, multistage sampling can introduce some complexity to the sampling process, but it does provide the best of both worlds of the combined methods when used appropriately. Ultimately, the decision to employ multistage sampling is subject to the researcher's scope of study.

Pros

1. A well-selected combination of sampling techniques may yield a significantly better sample than the use of only one sampling technique.

2. Combinations such as cluster and stratified sampling result in well-represented samples that are cost-effective to collect.

Cons

1. Poor selection of sampling techniques to combine can compound bias.

2. Each additional stage introduces added complexity to the overall sampling process.

Guided Exercise

Draw a multistage sample of size $n = 2000$ $n = 2000$ from the `voter` data set by selecting `2` random clusters from `state_seat` and stratifying each cluster by `ethnicity` . Assign this to the variable `mysample6` and examine the resulting sample using `head` and `describe` .

In [ ]:

```
# stage 1: simple random sample of clusters
clusters = np.random.choice(voter.state_seat, size = 2)
voter_stage1 = voter[[i in clusters for i in voter.state_seat
f=2000/voter_stage1.shape[0]

# stage 2: stratified sampling
voter_2stage = voter_stage1.groupby('ethnicity').\
apply(lambda x: x.sample(frac = f,random_state = 42))
voter_2stage.head()
```

In [ ]:

```
voter_2stage.describe()
```

In [ ]:

```
round(voter_2stage.ethnicity.value_counts()/voter_stage1.ethn
```

In [ ]:

```
pd.crosstab(voter.loc[:, "state_seat"],
            voter.loc[:, "ethnicity"],
            normalize = "index").round(1)
```

In [ ]:

```
voter_2stage = voter_2stage.reset_index(drop = True)
```

In [ ]:

```
pd.crosstab(voter_2stage.loc[:, "state_seat"],
            voter_2stage.loc[:, "ethnicity"],
            normalize = "index") .round(1)
```

Exercise

Draw a multistage sample of size $n = 20$ $n = 20$ from the `hr` data set by selecting `2` random clusters of size $N_k = 10$ $N_k = 10$ each from `Employee.Source` and stratifying each cluster by `Department` . Assign this to a variable of your choice and examine the resulting sample using `head` and `describe` .

In [ ]:

```

```

Non-probabilistic Sampling

Non-probabilistic sampling is a sampling technique in which the likelihood of a member of a population being selected for a sample is not entirely up to chance. Though easier to execute, we generally avoid non-probabilistic sampling as it yields significant **bias** compared to probabilistic methods. More often than not, non-probabilstic sampling involves sample selections that are based on the researcher's **subjective judgement**. The following sampling techniques fall under the umbrella of non-probabilistic sampling:

1. Convenience sampling
2. Volunteer sampling

There are other types of non-probabilistic sampling (*purposive*, *quota*, etc.), but we will focus on the above two as they are the most commonly used ones. We will aslo skip the pros and cons of these techniques as they are mostly negative, i.e. yield significant bias and should be avoided as far as possible.

Convenience Sampling

As the name suggests, convenience sampling is the selection of members from a population based on the researcher's **ease of access** to members of the population. The following are examples of convenience sampling:

1. A company decides to poll the first 100 customers who send them an email inquiring about their services.
2. A marketing firm conducts a survey at a shopping mall nearby their home office.

In both of the above cases, no care is given to *why* a specific person was polled/surveyed other than the fact that they were easy to contact. It goes without saying that such sampling practices will undoubtedly misrepresent the population being studied.

Example

A company decides to poll the first 100 customers who send them an email inquiring about their services. A marketing firm conducts a survey at a shopping mall nearby their home office.

Volunteer Sampling

Volunteer sampling is sampling in which the participants are all **willing volunteers**. This technique can yield significant bias because the sample is *chosen by the volunteers*, not the researcher. Volunteers often have a vested interest in the discussion/topic of study and are keen to share their point of view (this isn't always a good thing).

Here we see that the nature of the bias introduced differs from that of a convenience sample - the respondents here are already 'part of the system' being studied and we can expect their opinions to be skewed in favor of the researcher.

Example

1. During a show, a television station encourages their viewers to participate in an online poll on their Facebook page.
2. A band polls its fans on how they rate the band at a meet and greet event.

Exercise

A researcher wishes to investigate the movie preferences of all students at a university. Listed below are various ways in which he can do this. Identify the sampling method associated with each option.

1. He stands outside the library and asks students passing by questions on their movie preferences.
2. He obtains a student directory for the university and emails a questionnaire to 200 randomly selected students.
3. He obtains a student directory for the university and emails a questionnaire to every 15th name on the list.
4. He obtains a student directory and emails a questionnaire to 20% of the students from each department.
5. He obtains a student directory and emails a questionnaire to all students from the Engineering and Humanities departments.
6. He obtains a student directory and emails a questionnaire to 25% of the students from the Engineering and Business departments.

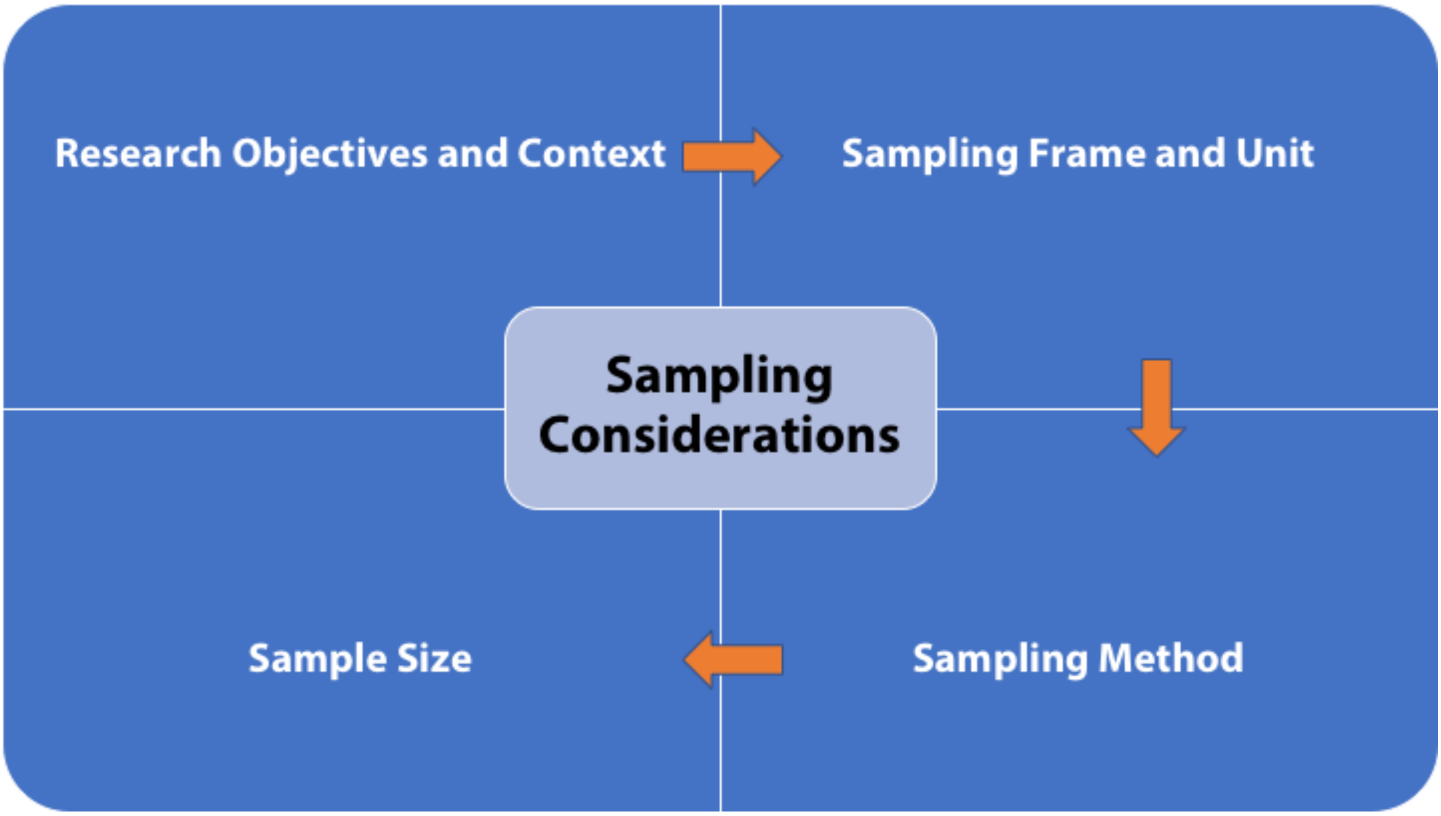
```
# MC
# Type your answer here.

1. Convenience sampling.
2. Simple random sampling.
3. Systematic sampling.
4. Stratified sampling.
5. One-stage cluster sampling.
6. Multistage sampling (cluster + stratified).
```

Summary

To summarize:

1. Sampling is the process of selecting a subset of the population. It is usually undertaken due to infeasibility of collecting data from the entire population.
2. There are various sampling methodologies, each with their own strengths/weaknesses. These methodologies can be combined to yield better results, but at the cost of complexity.
3. Sampling should always take the researcher's objectives into consideration. Assumptions about the impact of attributes should be noted upfront.

The chart below gives an overview of how we should approach the sampling process. Take care, and have fun sampling!



Exercise

With reference to our motivational example, draw an appropriate sample from the `voter` data set for the analyst to work on. You may take the following information into account:

1. Candidate A is a mother of 3, whereas candidate B is a bachelor.
2. Both candidates stay in the same housing estate.
3. The analyst's budget allows for a maximum sample size of 10,000 only.

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```

3. The Effect of Sample Size on Bias

Now that we are familiar with the various techniques involved in selecting samples, let us now discuss the impact of varying sample sizes on the accuracy of estimating our population. The code below simulates a population of $n = 1000$ $n = 1000$ members and draws samples of increasing sizes to compute the relative difference between the sample and population means.

In [ ]:

```python
np.random.seed(42)

N = 1000
mypop = np.random.choice(100, N, replace = True) + 1
popmean, popsd = np.mean(mypop), np.std(mypop, ddof = 1)
```

In [ ]:

```
relative_error = [abs(np.mean(np.random.choice(mypop, i)) - p
```

In [ ]:

```
rerror = pd.DataFrame({"n": np.arange(1000),
                       "Relative error": relative_error})

sns.relplot(x = "n", y = "Relative error", kind = "line", dat
```

As we can see from the graph, there is a clear trend for the relative difference to decrease as the sample size increases. In other words, the larger the sample size used, the better our estimates of the population will be.

In practice, the calculation of an optimal sample size is done using a technique called **power analysis** which is beyond the scope of this course. We can however take a quick glance at computing the minimum sample size using our motivational example.

Extra

Determining the Minimum Number of Voters to Sample

To compute the minimum sample size using power analysis, we need to have 4 pieces of information:

1. **The type of statistical test we would like to run**

   There are different types of statistical tests (t, chi-square, etc). Some tests are more complex and require larger samples to run effectively.

2. **Effect size**

   The effect size is a measure of how small/large a change we wish to be able to detect when using a statistical test. Larger effects are easier to detect as opposed to smaller ones, and this is reflected in the minimum sample size required. A larger effect size will require a smaller sample, whereas a smaller effect size requires a larger sample.

3. **Significance level**

   The significance level is a value between 0 and 1 that represents how likely we are to accidentally detect an effect that *isn't actually* present in our data.

4. **Power**

   Power is a value between 0 and 1 that represents how likely we are to detect an effect that is *actually* present in our data.

In Python, power analysis can be run using the `statsmodels` package. Alternatively, there are free tools available online such as *GPower* that can also be used to compute the required sample size. For the purpose of our motivational example, we will use the following parameters:

1. **The type of statistical test we would like to run**

   We will use a proportion test, `p.test`, to determine if either candidate garners more support than the other.

2. **Effect size**

   We will select an effect size of `h = 0.05`, i.e. we will only be able to detect differences of at least 0.05 in the proportion. Any differences smaller than this will go unnoticed.

3. **Significance level**

   We will select a significance level of `alpha = 0.05`. This means that we are only likely to accidentally detect a non-existent effect 5% of the time.

4. **Power**

   We will select a power of `power = 0.8`. This means that we will successfully detect an effect that is present in our data 80% of the time.

For the sake of brevity, we'll skip any discussion on the details of the above, including why the values given are appropriate and skip directly to the computation:

In [ ]:

```
import statsmodels.stats.power as smp
smp.NormalIndPower().solve_power(effect_size = 0.05, power=0.
# If ratio=0, then effect_size is the standardized mean in th
```

As we can see, the analyst would require a sample of at least 3140 voters to make a meaningful conclusion. Altering any of the 4 parameters (test type, effect size, significance level, power) will change the minimum sample size required to make a meaningful estimate of the population.

Sampling

Descriptive Statistics

```
In [ ]:
```

```
In [ ]:
```

Content Outline

In [ ]:

```

```

1. Introduction

Motivation

In a state of approximately 60,000 residents there are two candidates running for the post of Member of Parliament (MP) - Candidate A and Candidate B. A political analyst would like to determine whether the voting residents favor one candidate over the other. To accomplish this, he decides to run a poll for all voting residents of the state.

Based on the above scenario, take a minute to give your thoughts on the following:

1. Would it be possible to poll every single voting resident of the state?
2. What would be the cost of such an exercise?
3. Can we employ some strategy to obtain a reasonably similar result without expending as much resources?

Thankfully we can save quite a bit of time (and money!) by polling only a portion of the voters and with the magic of statistics, make reasonable inferences about the voting preferences of all voters in the state. This process is called **sampling.**

Populations

Before we dive deep into the process of sampling, let's first take a look at some preliminary concepts to start ourselves off on solid ground. In our motivational example above, the object of our interest is the voting preferences of the state's 60,000 residents, i.e. we are interested in knowing the vote of *all* voting residents in the town. In statistical terms, the set of all these voting residents is called a **population**. Simply put, a population is the set that contains **all** elements of interest for a particular study, and it exists to ensure that we don't draw observations from data unrelated to the problem at hand.

The definition of what constitutes a population is highly dependent on the context of the study, and to this extent a little bit of domain knowledge goes a long way.

Example

1. A researcher wishes to study the effect of steroid use in the National Football League (NFL). The population defined here would be the set of all active professional players in the NFL.
2. A grocery store owner would like to identify the most popular cereal he has on sale. The population defined here would be the set of all cereal products he currently sells at the store.
3. The Human Resource Department at KFRU, a large radio station, is interested in evaluating the effect of internal training programs on their staff. The population defined here would be the set of all current employees in the company.

In each of the above cases, the elements of interest are either people (NFL players, employees) or products (cereals) are said to be **members** of the population.

Guided Exercise

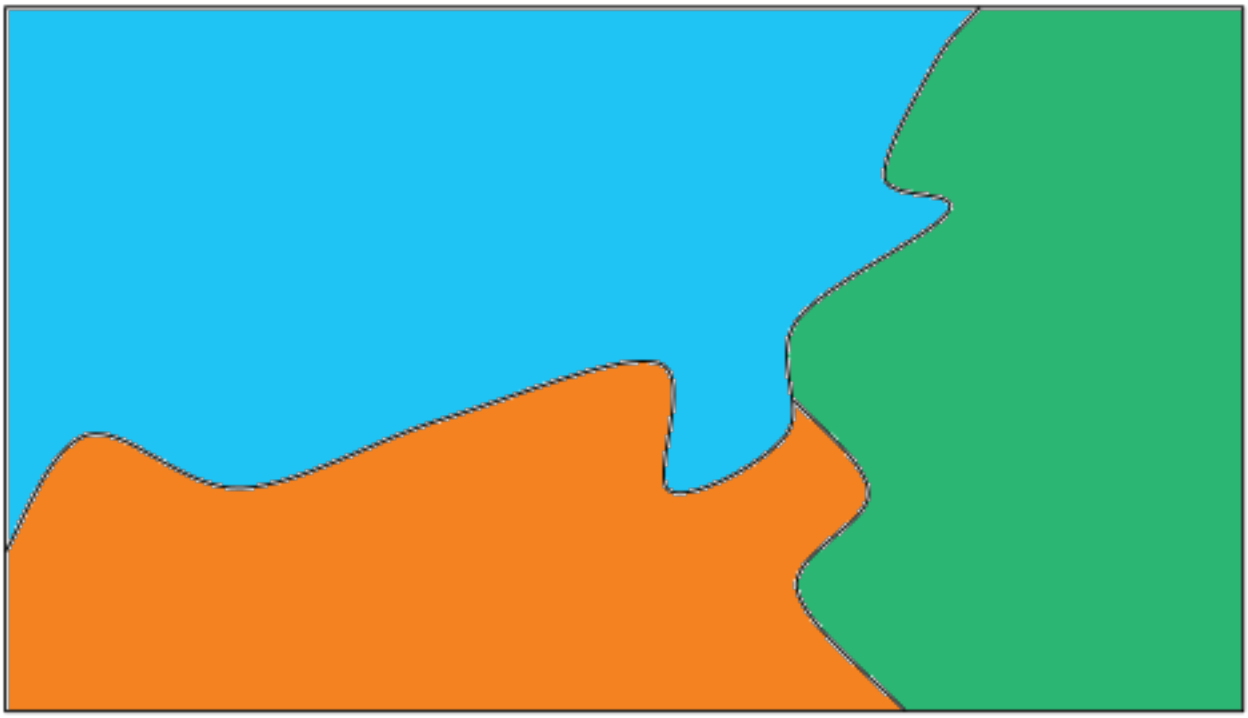
A. Suppose that you want to determine the movie preferences of students at a given university. Who/what are the members in your population?

B. Suppose that you would like to study the job satisfaction of teachers in international schools in KL. Who/what are the members in your population?

Sub-populations

In many cases, a population can be divided into smaller groups called **sub-populations**, allowing us to narrow down our point of interest. For example, the population of all employees in a given company can be further split into all *male* employees in the company and all *female* employees in the company. Sub-populations can be defined by any chosen characteristic, but the division must obey two rules:

1. A member of one group cannot belong to another group. (**mutually exclusive**)
2. Each member in the population belongs to a group. (**collectively exhaustive**)



Example

Recall the Human Resource Department at KFRU from our previous example. We can sub-divide the population of all employees to the *employees from each department*.

There is no restriction on the number of sub-populations (as how we define a sub-population is subject to domain knowledge), and we can even go further to sub-divide the sub-populations itself!

Example

We can divide the employees in the Human Resource Department into *male* and *female* employees of the department.

Now, you may be wondering - do I really need to take every single sub-population into account?

Well, the answer is **no**. Sub-populations can safely be **ignored** if they aren't *perceived to have an impact* on the *object of interest* in our study.

Example

We wish to study the average income of a fresh graduate in KL but we are *not interested* in knowing if there is a significant difference in income between male and female fresh graduates. In such a case, we ignore the sub-populations of male and female fresh graduates and treat them as one whole set.

Guided Exercise

The `voter` data set for the state in our motivational example has been pre-loaded and summarized for you below. Examine the output and discuss how we can construct various sub-populations from the data.

In [ ]:

In [ ]:

In [ ]:

Exercise

The KFRU `HR` data set has been pre-loaded and summarized for you below. Examine the head of the DataFrame and discuss how we can construct various sub-populations from the data.

In [ ]:

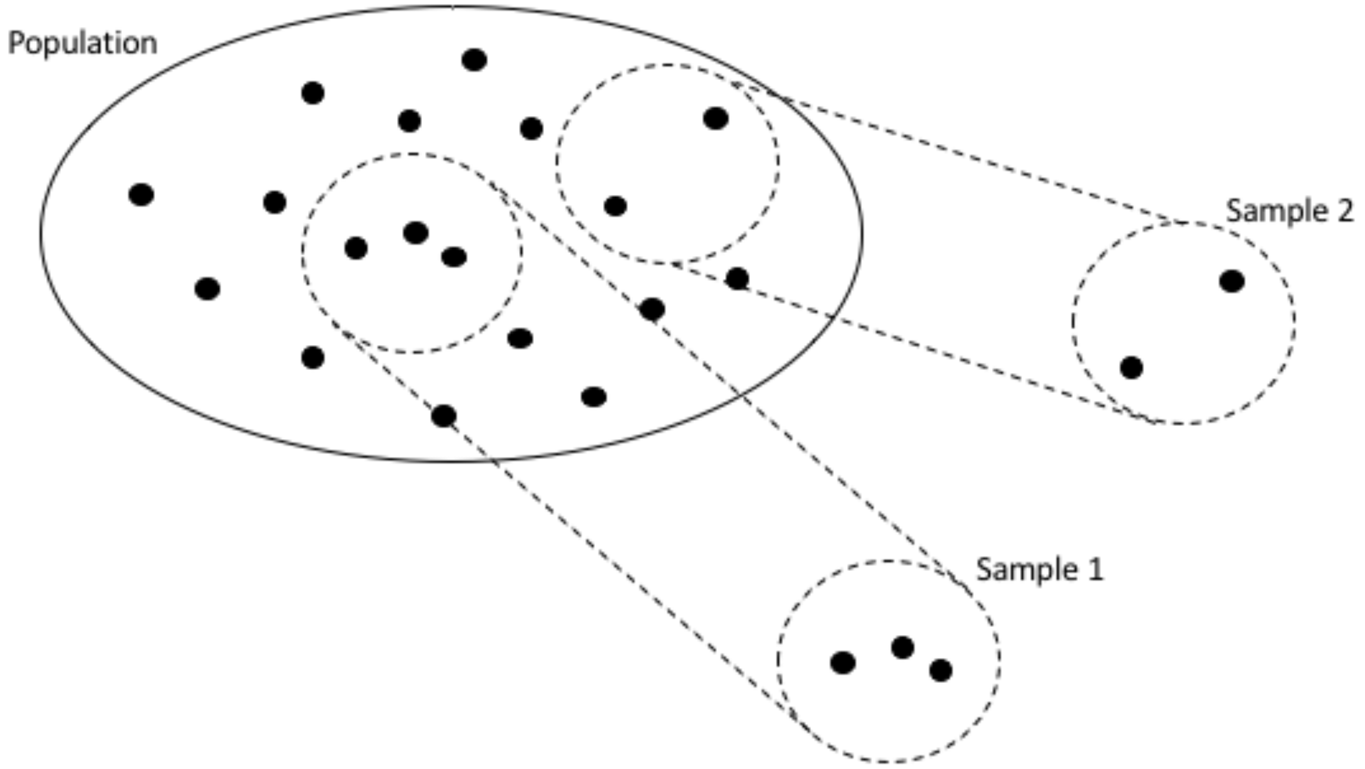
In [ ]:

In [ ]:

In [ ]:

Samples

As indicated in the beginning of this course, surveying an entire population can be a costly affair, and in most cases it is impossible to collate data from every single member of a population. By selecting/drawing a subset of the population, we obtain a **sample**, which can be used to estimate the properties of the population it is drawn from. We can draw as many samples as we want from a population, and we can even dictate the **size** of the samples, i.e. the *number of members in the sample*.



To avoid confusion, we commonly denote the size of a sample by n, and the size of a population by N. Sample sizes are restricted to the range $1 \leq n \leq N$ as selecting none of the members yields nothing, and we cannot select more members than there are in the population. In the case of the diagram above, we have a population of size $N = 18$, with two samples of size $n_1 = 3$ and $n_2 = 2$ respectively.

Example

Recall the example where we wished to study the average income of a fresh graduate in KL. Suppose that there are a total of N = 20,000 fresh graduates working in the city. To *draw* a sample of size n = 100, we would **select** 100 fresh graduates from this group of 20,000 to inquire about their income.

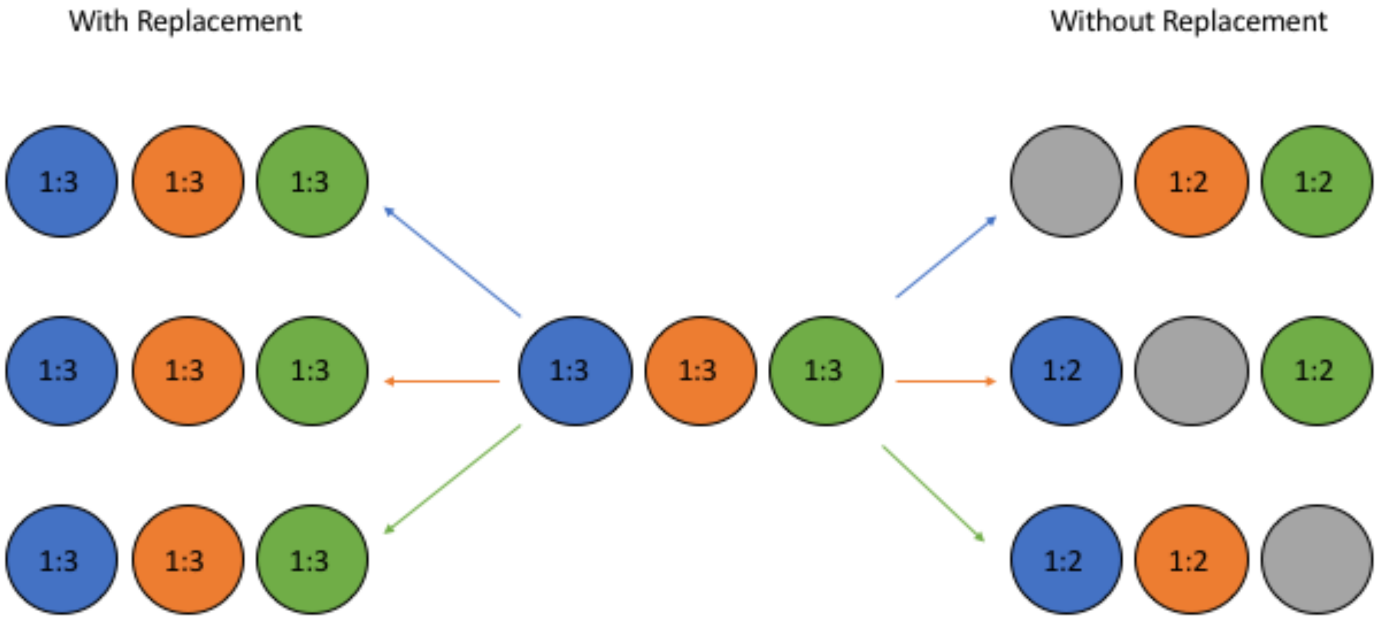
Samples can be drawn in **two** ways - *with* or *without* replacement. To illustrate the differences between these techniques, consider the following scenario:

Three colored balls are available to be drawn from a box - one blue, one orange, and one green. We would like to draw a sample of 2 balls and record the colors. The acts of drawing a ball are defined as follows:

Drawing *with* replacement: A ball is drawn and its color is recorded. The ball is then put back in the box before the second draw to be made.

Drawing *without* replacement: A ball is drawn and its color is recorded. The ball is **not** put back in the box before the second draw to be made.

The diagram below gives the odds of obtaining a ball of a specific color for the first and second draw.



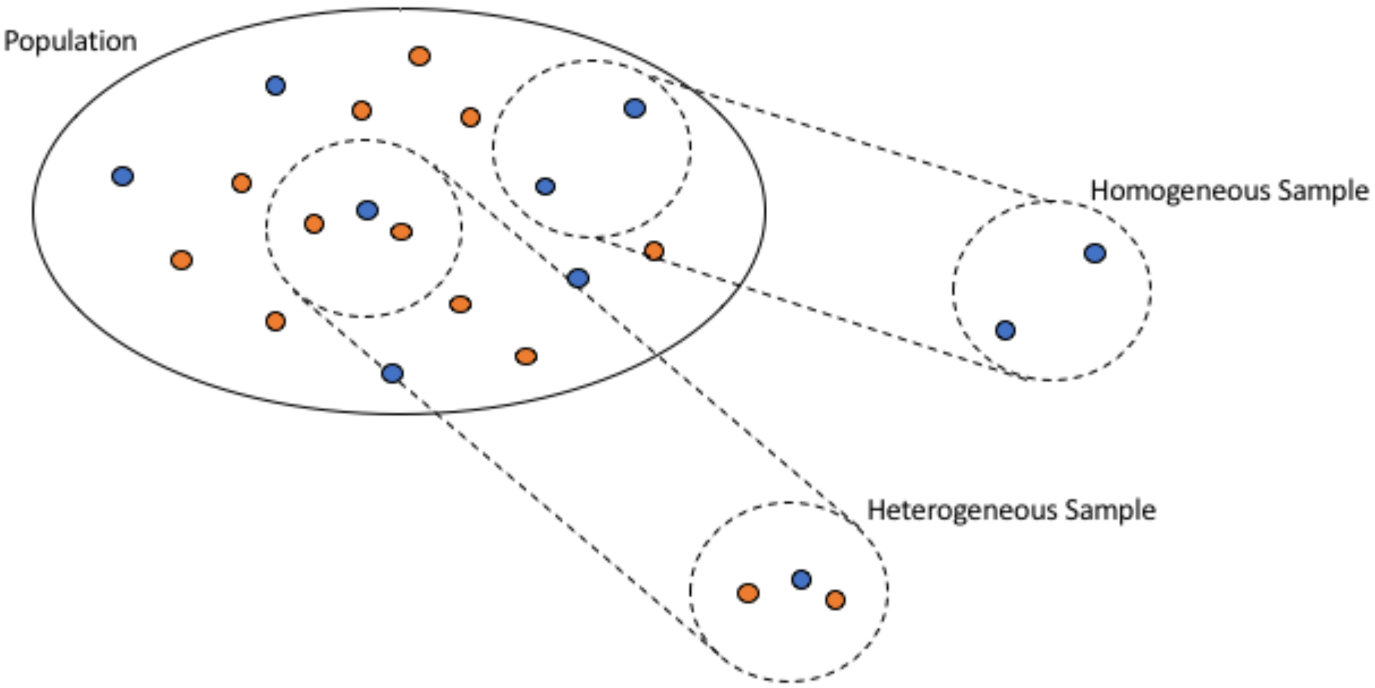
Odds of Selection for 3 Colored Balls

Note that on the first draw, both techniques have the same odds - 1 in 3 of obtaining a ball in each color. However, this no longer the case once the second ball is drawn. If the first ball drawn was not placed back in the box before the second draw, the odds of obtaining a ball of a different color than the first has now gone up to 1 in 2!

Sampling of real-world data is generally done **without replacement**. Sampling with replacement on the other hand is used in *oversampling* techniques, which are commonly employed in machine learning (and beyond the scope of this course).

Homogenous and Heterogeneous Samples

Samples are classified as **homogeneous** or **heterogenous** based on the members they contain, with the former having all members drawn from the *same sub-population* and the latter having members drawn from *multiple sub-populations*. The diagram below illustrates the difference between homogeneous and hetereogeneous samples:



Here we have a population that can be broken into two sub-populations. Samples that contain only members of *one* sub-population are deemed homogenous, and samples that contain members of *both* sub-populations are deemed heterogeneous.

Example

Suppose our population is defined as "All employees in KFRU."

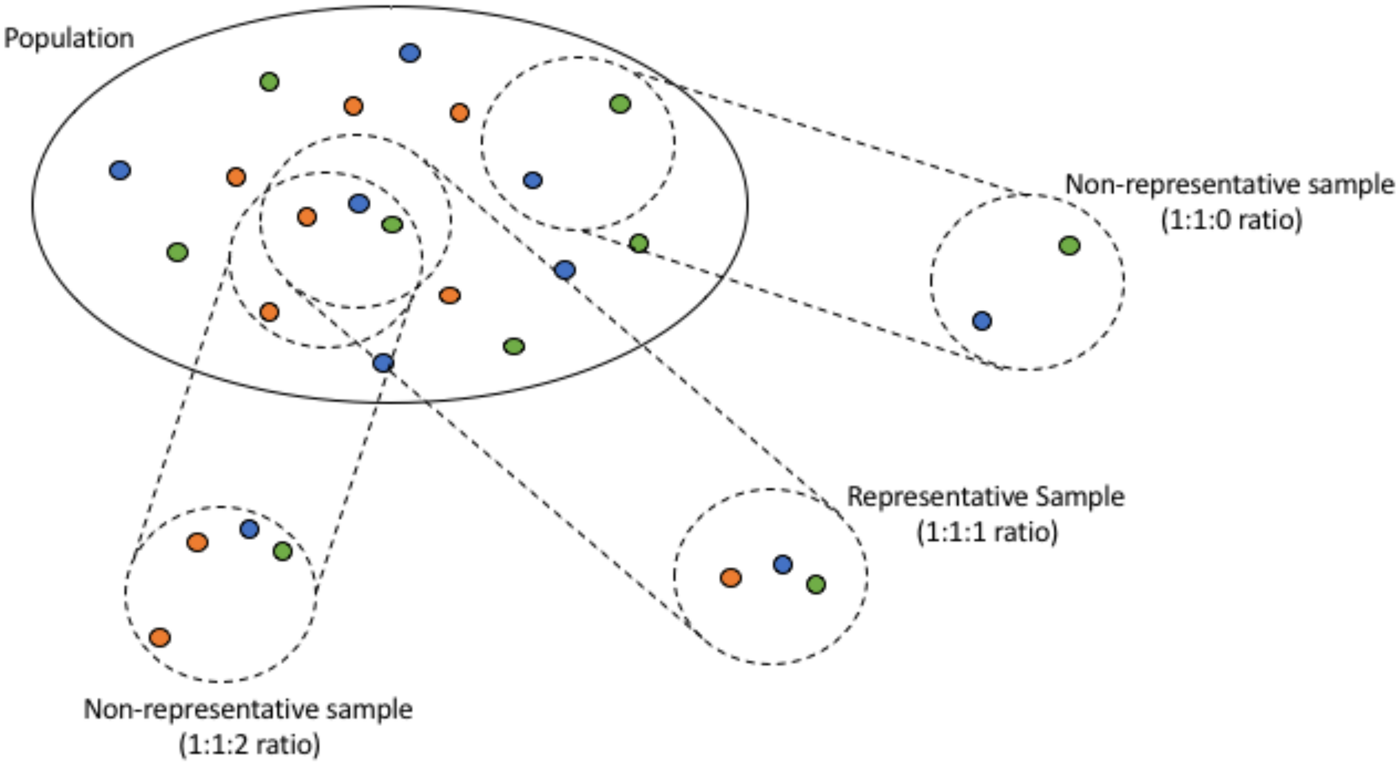
Homogeneous sample:

A sample of size $n = 25$ where all sample members are from IT.

Heterogeneous sample:

A sample of size $n = 25$, where 13 sample members are from IT, 2 are from Production, and the remainder are from Sales.

In the interest of ensuring that our samples accurately reflect the demographic proportions of the population, we often strive to draw samples with the same ratio of members across the various types to that of the population. As you can very well guess, obtaining an exact ratio isn't always possible if you fix the sample size so for cases such as these we favor sampling a **proportion** of the population.



In the above illustration, the population has a $1:1:1$ ratio between its 3 sub-populations. Taking a sample with the same ratio of members to that of the sub-population ratio results in what we call a **representative sample**. Conversely, if a sample is not representative we say it is **biased**. Numerical measures such as *mean* and *variance* for a biased sample do not accurately reflect that of the population and if used for business decision making can potentially lead to disastrous results.

Example

In a company there are 3 senior, 6 mid-level, and 9 junior employees.

Representative sample: A sample of size $n = 6$ consisting of 1 senior, 2 mid-level, and 3 junior employees. This amounts to sampling 30% of the population.

Biased sample: A sample of size $n = 6$ consisting of 3 senior and 3 mid-level employees.

Guided Exercise

Using the `voter` data set, draw the following sample:

A sample of 1,000 voters (without replacement) from `polling_district` 16. Assign this to the variable `mysample1` and examine the resulting sample using the `describe` method.

In [ ]:

In [ ]:

In [ ]:

Exercise

Using the `hr` data set, draw the following sample:

A sample of 20 employees (without replacement) from the `IT/IS` department. Assign this to a variable of your choice and examine if the resulting sample is representative using `describe` .

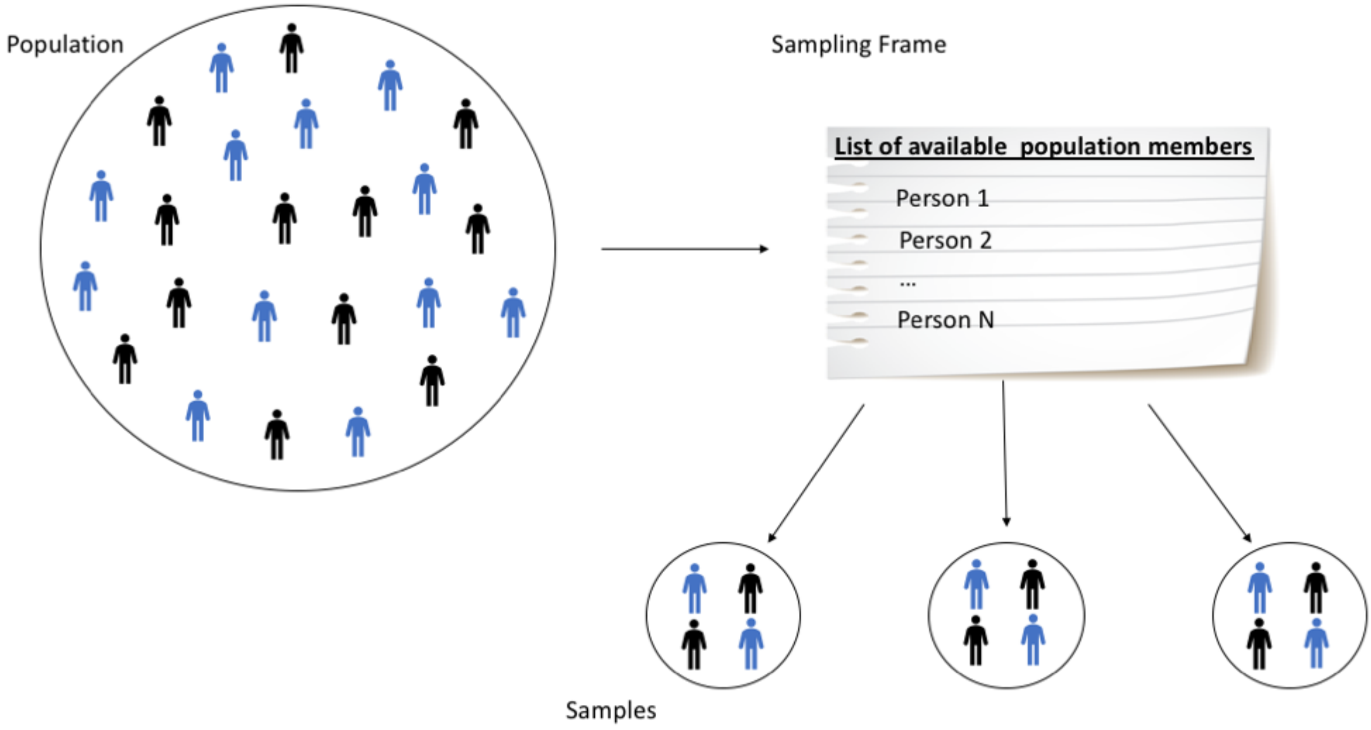
In [ ]:

In [ ]:

Sampling Frames and Sampling Units

As we saw in the section on populations, a population needs to be defined such that it covers the entirety of data that is of interest in a given study. This raises the question of *how* we can identify each individual member of our population to facilitate drawing samples. To this extent, we would like to define a **sampling frame**, i.e. a *list* of all the members in the population for us to sample from. We will use this sampling frame as a basis to draw samples using one or more prescribed methods (sampling methodologies) to ensure consistency.

As you can very well deduce, it may not always be possible to get an entire list of the members in a population. In cases such as these, we strive to define a sampling frame that is **as close as possible** to the population in order to **minimize** bias. Looking back at our motivational example, the sampling frame here would be the *roster of registered voters* in the state!



The table below summarizes the concept of how we would draw samples from a population using a sampling frame:

Component	Example
Population	A company's entire customer base
Sampling Frame	Those customers the company has access to (contact details available)
Samples	Customers who you contact and actually respond to your survey

In addition to a sampling frame, we also wish to define the number of members we select at a single time when we draw our samples. This selection size is called the **sampling unit**. Sampling units come in handy when we wish to widen/narrow our scope without altering our entire sampling methodology.

Example

1. Recall the example where we wished to study the average income of a fresh graduate in KL. As we expect the income to vary between individuals, we would select the sampling unit here as *one person*.

2. Suppose we want to study the average household income across a country instead. Since we are not interested in the individual incomes of each family member, we can widen the sampling unit to *one household*.

2. Sampling Methodologies

There are a variety of *techniques* (methods) in which sampling can be carried out, each with its own strengths and weaknesses. These techniques can generally be broken down into two major categories - **probabilistic** and **non-probabilistic sampling.**

Probabilistic Sampling

Probabilistic sampling is a sampling technique in which the *likelihood* that a member of the population is selected as part of a sample is **known** (or can be calculated). There are **5** sampling techniques that fall under the umbrella of probabilistic sampling, namely:

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. Systematic Sampling
5. Multistage Sampling

To avoid confusion (and a ton of mathematical jargon), we will focus on the process of each sampling technique and skip the calculation of probabilities etc.

Simple Random Sampling

In simple random sampling, each member of the population is **equally likely** to be selected as part of the sample. This can be done by **randomly selecting** members from the population.

Example

Suppose we wish to study the the job satisfaction of teachers in international schools in KL. We would first need to obtain a sampling frame, i.e. compile a list of all teachers in international schools and their contact details. We would then randomly choose names on this list to contact for interviewing.

Pros

1. Samples can easily be drawn with the assistance of statistical software suites.
2. Selection bias is minimal as the process of drawing members of the population is random.

Cons

1. Requires a well-defined sampling frame, which in practice may not always be available.
2. May not yield a representative sample as there is no stipulation on the ratio of members drawn.

Guided Exercise

Draw a simple random sample of 2,500 voters from the `voter` data set. Assign this to the variable `mysample2` and use bar graphs to examine the demographics of the resulting sample with respect to `gender` , `ethnicity` , and `state_seat` .

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Exercise

Draw a simple random sample of 200 employees from the `hr` data set. Assign this to a variable of your choice and use bar graphs to examine the demographics of the resulting sample with respect to `MaritalDesc` and `Department`.

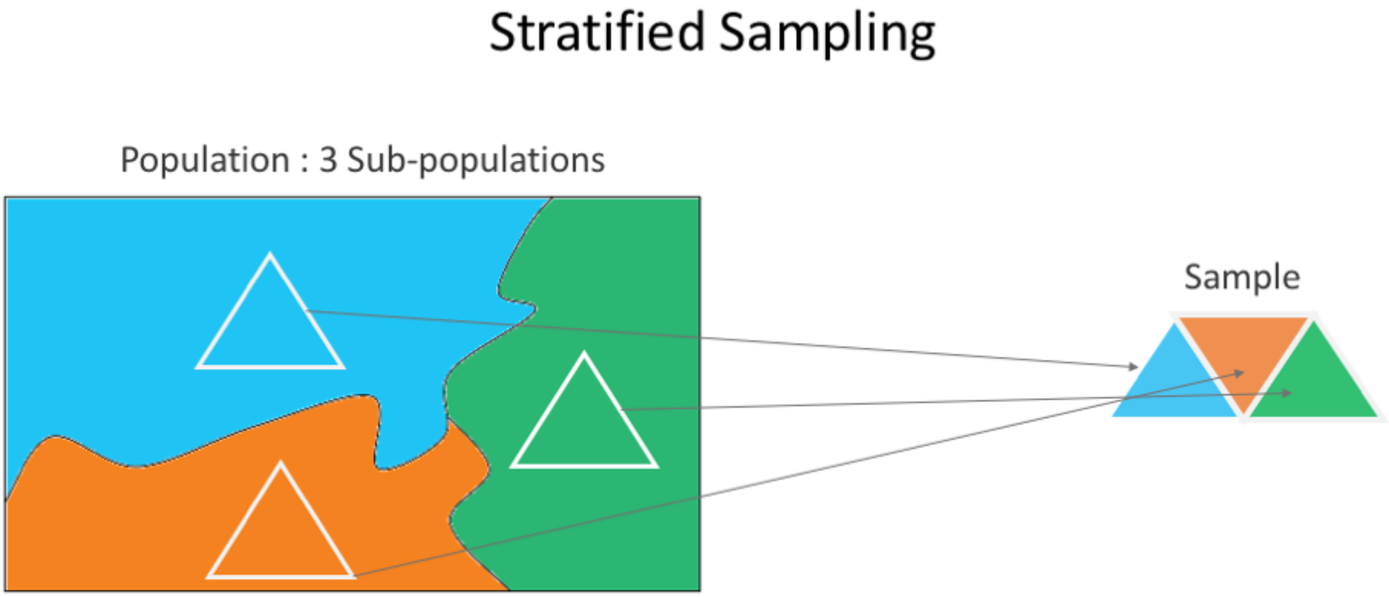
In [ ]:

In [ ]:

In [ ]:

Stratified Sampling

Stratified sampling is a sampling approach that is aimed at minimizing *representation bias*. It invovles grouping members of the population by their sub-populations and drawing samples from each one to preserve the ratio/balance from each group. For each sub-population k, we draw a fixed percentage of N_k members to make up the sample.



Note that here we do not fix the sample size unlike simple random sampling. Instead, we specify a percentage of the population to allow for easier computation of how many members to draw from each sub-population.

Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. Suppose that we wanted to interview 30% of all teachers. We would then group the teachers by school and randomly select 30% from each school to make up our sample.

Stratified sampling is useful when we want to ensure that all the various sub-populations are well-represented. National census data is often collected in such a manner to ensure that minority groups etc. are taken into account when federal policies are developed.

Pros

1. Samples can easily be drawn from each sub-population with the assistance of statistical software suites.

2. Representation bias is minimal as stratification takes into account the ratio of sub-populations.

Cons

1. Requires a well-defined sampling frame, which in practice may not always be available.

2. If the variance within each sub-population is significantly different, stratified sampling may result in the sample variance being skewed.

3. If data collection involves travel to various geographical locations, stratification may incur greater cost as each sub-population needs to be accounted for.

Guided Exercise

Draw a sample of 10% of voters from the `voter` data set, stratified by `gender`. Assign this to the variable `mysample3` and examine the resulting sample using `describe`.

In [ ]:

In [ ]:

In [ ]:

Exercise

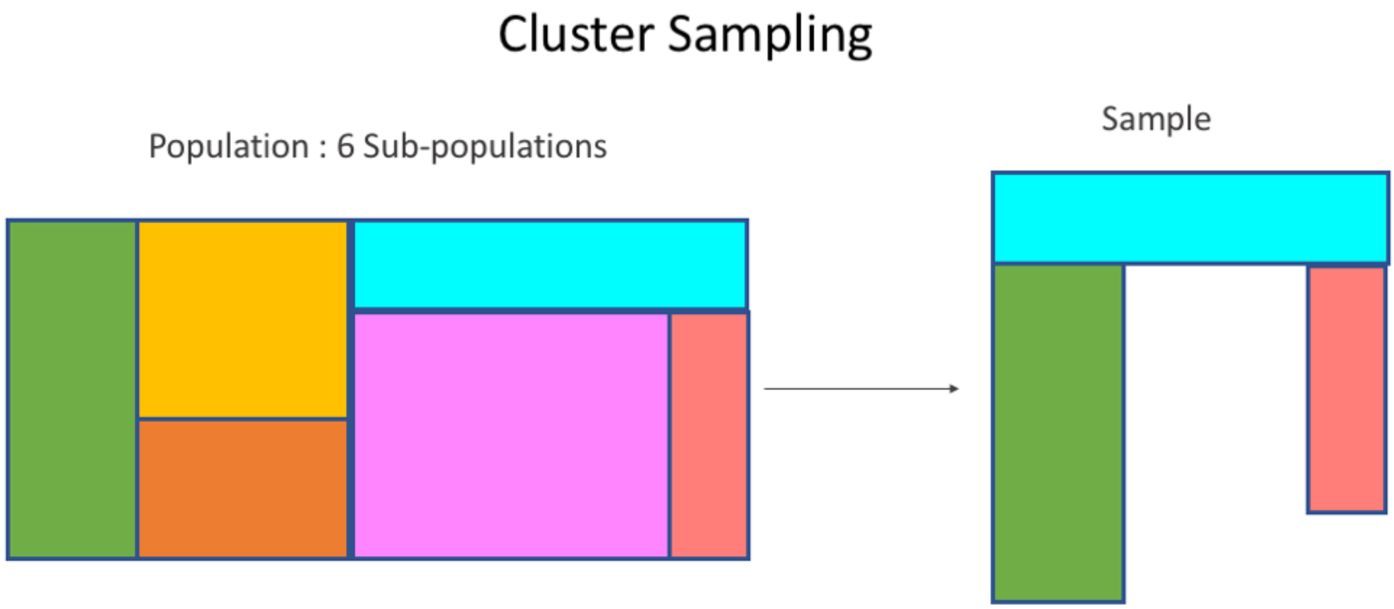
Draw a sample of 20% of employees from the `hr` data set, stratified by `Sex`. Assign this to a variable of your choice and examine the resulting sample using `describe()`.

```
In [ ]:
```

```
In [ ]:
```

Cluster Sampling

Cluster sampling is similar to stratified sampling in the sense that the sampling is done over sub-populations. The main difference however, is that instead of randomly sampling **within** each sub-population with the intention of preserving proportionality, we **randomly select** a **chosen number** of sub-populations and then proceed to sample from the selected sub-populations. If *all* elements in a chosen cluster are sampled, we call the process **one-stage cluster sampling**.



In the diagram above, we see that 3 random sub-populations (clusters) were chosen from the 7 that exist in the population. Though the number of clusters chosen is entirely up to the researcher, they are commonly dictated by resource availability (cost, manpower, etc.)

Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. To apply one-stage cluster sampling here, we would randomly select a number of international schools within KL (e.g. 20) and interview *all* teachers at these schools.

Cluster sampling is better suited towards large populations where sampling within clusters is easily done, but sampling **across** clusters is difficult. Geographical factors such as distance are often used as a measure of when cluster sampling is required.

Example

To ascertain a specific opinion from all Malaysians within Peninsula Malaysia, it is easier to divide the peninsula into its various states and perform cluster sampling, i.e. sampling only randomly selected states to cut down on traveling costs.

Pros

1. Clusters can easily be selected with the assistance of statistical software suites and/or random number generators.

2. Works well for large large populations.

Cons

1. Requires more than 2 sub-populations to be effective. Sampling from only 1 out of 2 sub-populations is biased.

2. Less control over sample size, poses a problem for populations where the number of members in each sub-population is not equal/similar.

Guided Exercise

Draw a one-stage cluster sample from the `voter` data set, clustered by `state_seat` by setting the number of clusters to `2` and using simple random sampling of clusters (`np.random.choice`). Assign this to the variable `mysample4` and examine the resulting sample using `describe` .

In [ ]:

```

```

Exercise

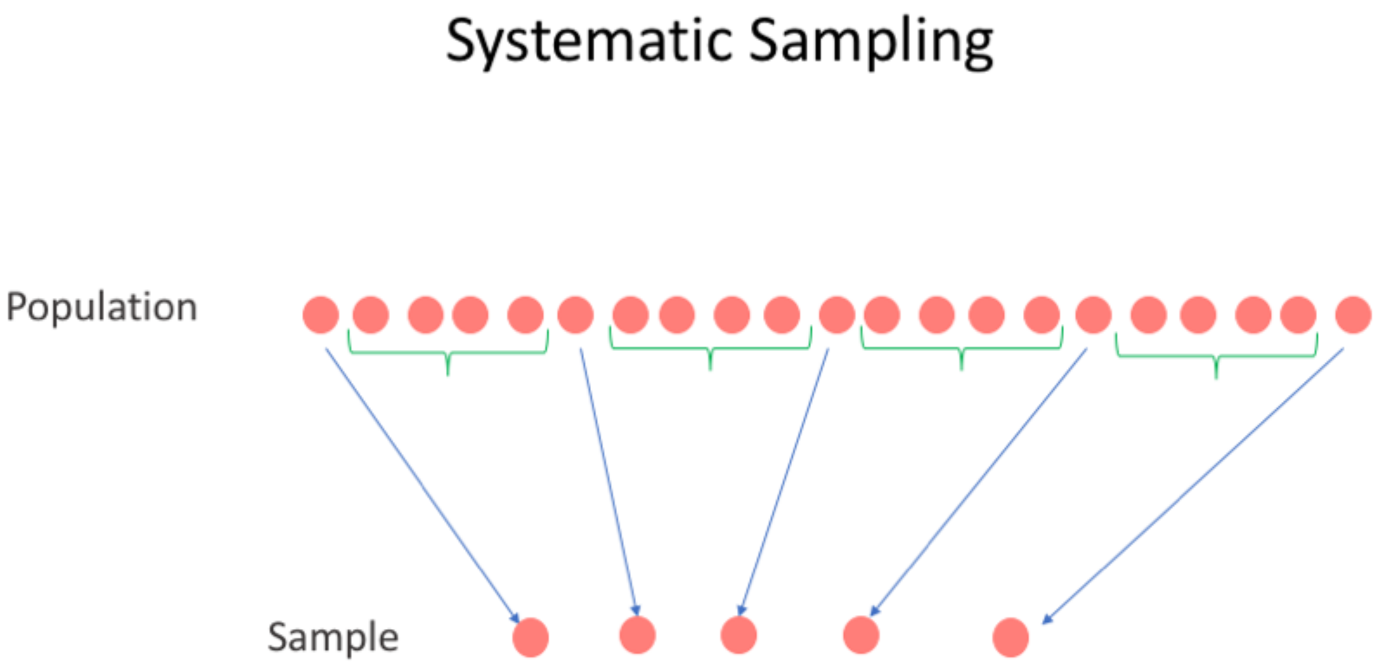
Draw a one-stage cluster sample from the `hr` data set, clustered by `Department` by setting the number of clusters to `4` and using simple random sampling of clusters. Assign this to a variable of your choice and examine the resulting sample using `describe` .

In [ ]:

In [ ]:

Systematic Sampling

Systematic sampling is carried out by randomly selecting an initial member of the population from the sampling frame and selecting additional members following a **pre-determined sequence**. This sequence is in the form of the index number for a member incremented/decremented by a fixed step size h. For example, if we start with the 5th member of the population and set a step size of $h = 2$, we will then select the 7th, 9th, 11th (so on and so forth) members of the population until we obtain a sample of desired size. Note that we can use modulo arithmetic to ensure that we don't overshoot the sample size when counting the indices.



Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. To obtain a systematic sample, we index the list of teachers and pick an random number (e.g. 45) as our starting index. If we want to select 100 members with the step size being $h = 10$, then we select the 45th, 55th, 66th, ..., 445th members as our sample.

Systematic sampling is generally favored when the population is known to be large as it is easier to collect one large systematic sample than to draw multiple simple random samples. The conveniece however does come at the cost of bias if the data is ordered, e.g. picking 10 members from a set of 1,000 exam scores sorted in descending order will give a sample with a high average score if the step size is too small.

To deal with biases such as these, there is a variation of systematic sampling called *random systematic sampling*. This method is beyond the scope of this course as it requires in-depth knowledge of probability.

Pros

1. Only requires one instance of random selection, the remaining members are selected using a pre-determined sequence.
2. Easy to carry out, even with a loosely-defined sampling frame.

Cons

1. May introduce bias if the sampling frame is arranged to begin with.
2. If sub-populations exist, the resulting sample may not be representative.

Guided Exercise

Draw a systematic sample of size $n = 250$ from the `voter` data set with step size `175`. Assign this to the variable `mysample5` and examine the resulting sample using `head` and `describe`.

In [ ]:

In [ ]:

In [ ]:

Exercise

Draw a systematic sample of size $n = 50$ from the `hr` data set with step size `5`. Assign this to a variable of your choice and examine the resulting sample using `head` and `describe`.

In [ ]:

In [ ]:

Multistage Sampling

As the name suggests, multistage sampling is performed by **combining** two or more of the above sampling methodologies in **series**. The objective is to retain the strengths of each technique by leveraging the *order* in which we combine these techniques.

Example

Recall the previous example where we wanted to study the the job satisfaction of teachers in international schools in KL. Suppose we want to stratify our samples by gender of the teachers, but we would also like to conduct cluster sampling to reduce having to travel to all the different schools in KL. By combining both approaches, we can select a cluster of schools *then* pick stratified samples from each cluster.

As seen above, multistage sampling can introduce some complexity to the sampling process, but it does provide the best of both worlds of the combined methods when used appropriately. Ultimately, the decision to employ multistage sampling is subject to the researcher's scope of study.

Pros

1. A well-selected combination of sampling techniques may yield a significantly better sample than the use of only one sampling technique.

2. Combinations such as cluster and stratified sampling result in well-represented samples that are cost-effective to collect.

Cons

1. Poor selection of sampling techniques to combine can compound bias.

2. Each additional stage introduces added complexity to the overall sampling process.

Guided Exercise

Draw a multistage sample of size $n = 2000$ from the `voter` data set by selecting `2` random clusters from `state_seat` and stratifying each cluster by `ethnicity`. Assign this to the variable `mysample6` and examine the resulting sample using `head` and `describe`.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Exercise

Draw a multistage sample of size $n = 20$ from the `hr` data set by selecting `2` random clusters of size $N_k = 10$ each from `Employee.Source` and stratifying each cluster by `Department` . Assign this to a variable of your choice and examine the resulting sample using `head` and `describe` .

```
In [ ]:
```

Non-probabilistic Sampling

Non-probabilistic sampling is a sampling technique in which the likelihood of a member of a population being selected for a sample is not entirely up to chance. Though easier to execute, we generally avoid non-probabilistic sampling as it yields significant **bias** compared to probabilistic methods. More often than not, non-probabilstic sampling involves sample selections that are based on the researcher's **subjective judgement**. The following sampling techniques fall under the umbrella of non-probabilistic sampling:

1. Convenience sampling
2. Volunteer sampling

There are other types of non-probabilistic sampling (*purposive*, *quota*, etc.), but we will focus on the above two as they are the most commonly used ones. We will aslo skip the pros and cons of these techniques as they are mostly negative, i.e. yield significant bias and should be avoided as far as possible.

Convenience Sampling

As the name suggests, convenience sampling is the selection of members from a population based on the researcher's **ease of access** to members of the population. The following are examples of convenience sampling:

1. A company decides to poll the first 100 customers who send them an email inquiring about their services.
2. A marketing firm conducts a survey at a shopping mall nearby their home office.

In both of the above cases, no care is given to *why* a specific person was polled/surveyed other than the fact that they were easy to contact. It goes without saying that such sampling practices will undoubtedly misrepresent the population being studied.

Example

A company decides to poll the first 100 customers who send them an email inquiring about their services. A marketing firm conducts a survey at a shopping mall nearby their home office.

Volunteer Sampling

Volunteer sampling is sampling in which the participants are all **willing volunteers**. This technique can yield significant bias because the sample is *chosen by the volunteers*, not the researcher. Volunteers often have a vested interest in the discussion/topic of study and are keen to share their point of view (this isn't always a good thing).

Here we see that the nature of the bias introduced differs from that of a convenience sample - the respondents here are already 'part of the system' being studied and we can expect their opinions to be skewed in favor of the researcher.

Example

1. During a show, a television station encourages their viewers to participate in an online poll on their Facebook page.

2. A band polls its fans on how they rate the band at a meet and greet event.

Exercise

A researcher wishes to investigate the movie preferences of all students at a university. Listed below are various ways in which he can do this. Identify the sampling method associated with each option.

1. He stands outside the library and asks students passing by questions on their movie preferences.

2. He obtains a student directory for the university and emails a questionnaire to 200 randomly selected students.

3. He obtains a student directory for the university and emails a questionnaire to every 15th name on the list.

4. He obtains a student directory and emails a questionnaire to 20% of the students from each department.

5. He obtains a student directory and emails a questionnaire to all students from the Engineering and Humanities departments.

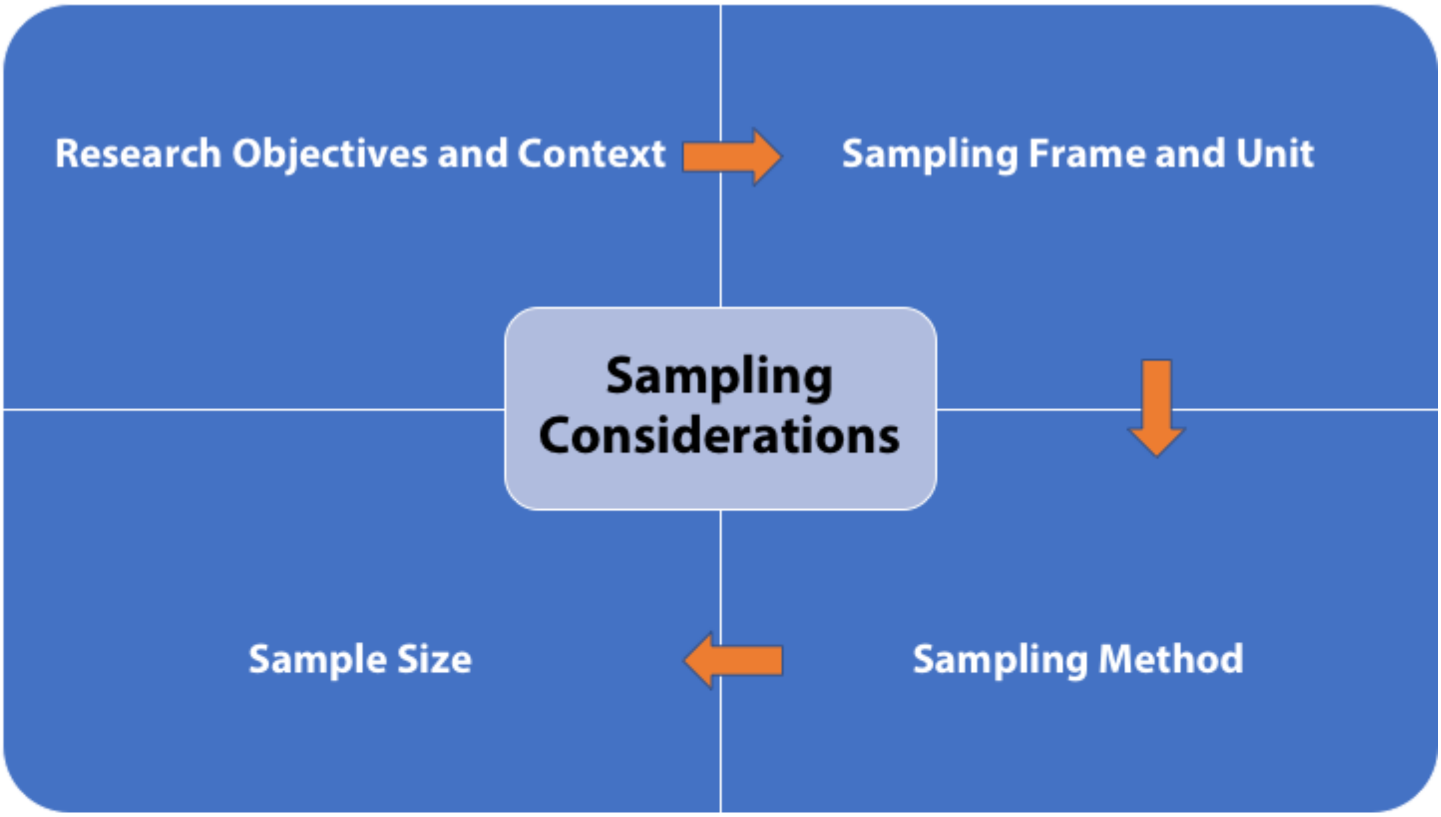
6. He obtains a student directory and emails a questionnaire to 25% of the students from the Engineering and Business departments.

Summary

To summarize:

1. Sampling is the process of selecting a subset of the population. It is usually undertaken due to infeasibility of collecting data from the entire population.
2. There are various sampling methodologies, each with their own strengths/weaknesses. These methodologies can be combined to yield better results, but at the cost of complexity.
3. Sampling should always take the researcher's objectives into consideration. Assumptions about the impact of attributes should be noted upfront.

The chart below gives an overview of how we should approach the sampling process. Take care, and have fun sampling!



Exercise

With reference to our motivational example, draw an appropriate sample from the `voter` data set for the analyst to work on. You may take the following information into account:

1. Candidate A is a mother of 3, whereas candidate B is a bachelor.
2. Both candidates stay in the same housing estate.
3. The analyst's budget allows for a maximum sample size of 10,000 only.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

3. The Effect of Sample Size on Bias

Now that we are familiar with the various techniques involved in selecting samples, let us now discuss the impact of varying sample sizes on the accuracy of estimating our population. The code below simulates a population of $n = 1000$ members and draws samples of increasing sizes to compute the relative difference between the sample and population means.

In [ ]:

In [ ]:

In [ ]:

As we can see from the graph, there is a clear trend for the relative difference to decrease as the sample size increases. In other words, the larger the sample size used, the better our estimates of the population will be.

In practice, the calculation of an optimal sample size is done using a technique called **power analysis** which is beyond the scope of this course. We can however take a quick glance at computing the minimum sample size using our motivational example.

Extra

Determining the Minimum Number of Voters to Sample

To compute the minimum sample size using power analysis, we need to have 4 pieces of information:

1. **The type of statistical test we would like to run**

   There are different types of statistical tests (t, chi-square, etc). Some tests are more complex and require larger samples to run effectively.

2. **Effect size**

   The effect size is a measure of how small/large a change we wish to be able to detect when using a statistical test. Larger effects are easier to detect as opposed to smaller ones, and this is reflected in the minimum sample size required. A larger effect size will require a smaller sample, whereas a smaller effect size requires a larger sample.

3. **Significance level**

   The significance level is a value between 0 and 1 that represents how likely we are to accidentally detect an effect that *isn't actually* present in our data.

4. **Power**

   Power is a value between 0 and 1 that represents how likely we are to detect an effect that is *actually* present in our data.

In Python, power analysis can be run using the `statsmodels` package. Alternatively, there are free tools available online such as *GPower* that can also be used to compute the required sample size. For the purpose of our motivational example, we will use the following parameters:

1. **The type of statistical test we would like to run**

   We will use a proportion test, `p.test` , to determine if either candidate garners more support than the other.

2. **Effect size**

   We will select an effect size of `h = 0.05` , i.e. we will only be able to detect differences of at least 0.05 in the proportion. Any differences smaller than

   this will go unnoticed.

3. **Significance level**

   We will select a significance level of `alpha = 0.05` . This means that we are only likely to accidentally detect a non-existent effect 5% of the

   time.

4. **Power**

   We will select a power of `power = 0.8` . This means that we will successfully detect an effect that is present in our data 80% of the time.

For the sake of brevity, we'll skip any discussion on the details of the above, including why the values given are appropriate and skip directly to the computation:

In [ ]:

As we can see, the analyst would require a sample of at least 3140 voters to make a meaningful conclusion. Altering any of the 4 parameters (test type, effect size, significance level, power) will change the minimum sample size required to make a meaningful estimate of the population.