# Module 2
# Sampling Methods
# &
# Data Collection

Module Learning Outcomes

- Provide a description, in words, the population of interest.
- Provide a description, in words, the sample.
- Discuss if a given situation involves sampling error.
- Identify the most appropriate sampling method used in the given situation, from among the five sampling methods covered.
- Devise a sampling method (from the five methods), given the situation.
- Provide a description of the sampling frame.
- Discuss if a given situation involves selection bias using sampling frame.
- Understand the importance of having good or cleaned data set.
- Be aware of what data ethics in business is about.
- Discuss if a given situation involves response bias.
- Discuss if a given situation involves non-response bias.

## 2.1 Population

- In statistics, the term **population** is defined as the group of subjects we are interested in.
- First thing you want to know is that the meaning of the word population is quite different in English (refers to the number of residents in a country) as in Statistics (refers to a group of subjects).
- The term population is also sometimes called the **target population**.
- The number of subjects in the population is called the **population size** and its value is typically unknown. The letter or symbol “$N$” is used to denote population size.
- When providing a description population in this course, the word “all” has to be used. The reason is to use a more general term “all” to replace the unknown population size.
- MLO: Provide a description, in words, the population of interest.

### Why Don’t We Just Use the Population?

- There are two main issues with just using the population.
- 1) It is nearly impossible to get to all subjects in the population. In other words, if we cannot get to them, we cannot get information from them. To give you a perspective, Canada has about 35 million people and it is nearly impossible to reach out to all of them within a reasonable amount of time.
- 2) Sometimes, data collection involves a destructive process. For example, when testing battery life (in hours). It is simply not feasible to test all batteries for how long they last.
- Therefore, it is rather obvious we need a subset and yet representative of the population.

## 2.2 Sample

- A **sample** is defined as a group of subjects that we will get (at the planning stage) or have already collected (at the end of data collection).
- Note that it is always a subgroup of the population.
- The sample will eventually give you the information in the form of a data set.
- The number of subjects in the sample is called the **sample size** (denoted as “$n$”).
- As you can see, sample size is always smaller than the population size, i.e. $n < N$.
- When the number of subjects you plan to get (at the planning stage) is different from the number of subjects we get at the end of data collection, the term “sample size” refers to the former and the term “**effective sample size**” refers to the latter.
- For example, Suppose I have printed 120 surveys. I handed out 120 in the cafeteria but only get 105 back. Here the sample size is 120 and the effective sample size is only 105.
- There is only case when the sample is the population. This is called **census** – something only large-scale companies like Statistics Canada or U.S. Census Bureau can do it.
- In this course, the sample size (the value) is always part of the sample description. This is similar to the inclusion of the word “all” in the description of the population.
- MLO: Provide a description, in words, the sample.

## 2.3 Sampling Error

- **Sampling error** is defined as the error associated with getting samples.
- In statistics, we always have sampling error (because we always get samples in our studies).
- The only two situations where we do not have sampling error: (1) when we are doing a census, or (2) we do not get a sample at all (that means, not making use of statistics at all).
- Sampling error is a concept where some number we want to know from the population (called **parameter**; more later) is different from similar numbers we get from the sample (called **statistic**; more later).
- Important note: Sampling error generally goes down when the sample size gets bigger. This is intuitively true. We are generally more comfortable with bigger samples because we feel like that bigger samples have better resemblance to the population. Hence, it has less error (or smaller sampling error).
- MLO: Discuss if a given situation involves sampling error.

## 2.4 Sampling Methods

- There are two kinds of sampling methods: **Probability Sampling Methods** (ones making use of probability, and they are hence random) and **Non-Probability Sampling Methods** (non-random).
- Probability sampling methods are generally better than the non-probability counterpart is because subjects are selected without any subjectivity involved.

### A. Simple Random Sampling Method (or Simple Random SM)

- Subjects in a **simple random sample** (or simply **SRS**) are selected "randomly".
- Simple random sample is considered the "**gold standard**" in statistics. All subsequent methods will be compared to SRS.
- Note: In this course, the concept or idea of "randomness" can only be achieved by either using a random number generator (from Excel RAND() function, any scientific calculators etc.) or through old-school method like "drawing numbers from a hat". The latter would be too time-consuming to do when $N$ is big.
- A prime example of SRS is **Random Digit Dialing** (**RDD**). It was introduced back in the 1970's but is still in use even today.
- A quick test for SRS is to check if the sample you get satisfies the following three points.

1) Subjects are drawn randomly directly from the population (or sampling frame).
2) Each subject in the population has the same chance of being selected.
3) Each subject is only allowed to be selected once in the sample.

- Pros: SRS is the gold standard. In fact, most theories and formula we use in statistics are having an assumption that data are drawn from this method.
- Cons: It is nearly impossible to get a SRS because it requires you to know the population size or have a list of all subjects (called the **sampling frame**, more later) in advance.

- MLO: Identify the most appropriate sampling method used in the given situation, from among the five sampling methods covered.
- MLO: Devise a sampling method (from the five methods), given the situation.

## B. Convenience Sampling Method

- In this course, the term **convenience sample** is a collective terms of multiple non-probability sampling methods, which are not using any probability to get the subjects.
- Examples are voluntary sampling methods, judgment sampling methods, or web panel sampling methods etc.
- In other words, subjects in the convenience sample are either chosen based on convenience (nearby subjects) or judgment (at your discretion).
- Pros and Cons: Easy and fast to get sample but not very good for proper scientific studies.

## C. Sampling Frame

- The **sampling frame** or simply **frame** is the list (either a physical or an imaginary list) of subjects, from which the sample is drawn in practice.
- The main difference between the sampling frame and (target) population is that the frame is something tangible or some list of subjects that we could see or touch. In contrast, population is like some "imaginary world" and we rarely get to see or touch.
- But like population, the frame refers to a very big group of subjects and we typically do not know the size. Therefore, it also includes the word "all" in the description.
- Note that the frame is a group of subjects too, but not necessarily in a form of a list. Sometimes, a frame is a map or a floorplan. For example, the classroom floorplan can be considered as a frame because we could arbitrarily assign number to each seat, and there are 20 naturally occurring tables.
- For example, if the population is "all Canadians" and the sample is a "random sample of 100 expat Canadians living in US", then the sampling frame would be "all expat Canadians living in the US".
- MLO: Provide a description of the sampling frame.

## D. Selection Bias

- If the sampling frame matches well with the target population, then the sample (which is from the frame) is a good representative of the population. Hence, it is a good sample. Or, we do not have any issues with the **selection bias**.
- Otherwise, if the frame is rather different from the target population, we will have an issue with the selection bias. It basically means that we have chosen (or selected) the wrong group of subjects.
- In this course, we use selection bias to measure if a sample is good or not. A good sample would not have a huge issue with selection bias, while a bad sample would have an obvious issue.
- There are other terms that are used describe similar things – **undercoverage** or **overcoverage**. But they will not be used further in this course.

- For example, if the population is "all Canadians" and the sampling frame would be "all expat Canadians living in the US". Since they are not the sample, there exists selection bias.
- Note: The term "error" and "bias" are rather different in statistics. The former refers to some mistakes that could be reduced when sample size gets bigger (like sampling error) while the latter may or may not be reduced by only increasing the sample size (like selection bias).
- MLO: Discuss if a given situation involves selection bias using sampling frame.

Variations of Simple Random Sampling Method

- The following three methods are variations of simple random sampling method.
- Each of them has its advantages and disadvantages in certain situation, in reference to SRS.
- The key to success is to understand the basic features of each of them and know under what situation a particular method is preferred over SRS and why.

## E. Systematic Sampling Method (or Systematic SM)

- The **systematic sampling method** is typically used in situation where the sampling frame is readily available (either in a form of a list or a map). Or it is used when subjects are lined up nicely, like a list of names or a pile of application forms.
- The systematic sampling method has three main steps:
  1) A **stepper** is to be calculated by taking the ratio between $N$ and $n$, i.e. $k = N/n$ .
  When the calculated value of $k$ is not a whole number, the value will always be rounded down to the nearest whole number. It is because we would rather go over the minimum sample size requirement than go under.
  2) A random number (between 1 and $k$) is used as a starting point.
  3) Then, every $k^{th}$ subject on the frame will form part of the **systematic sample**.
- Pros: It is typically less tedious and less time-consuming compared to SRS. Hence, it is quicker.
- Cons: We need to know the population size ahead of time. This method is rather restrictive because subjects have to be in a list or in some natural order.
- MLO: Identify the most appropriate sampling method used in the given situation, from among the five sampling methods covered.
- MLO: Devise a sampling method (from the five methods), given the situation.

## F. Stratified Random Sampling Method (or Stratified Random SM)

- The **stratified random sampling method** is typically used when we want the sample and the population have a similar demographics.
- For example, if only 10% of 10,000 Langara students are left-handed (or 1,000 of them). Getting a random sample of 100 would not likely give you 10 left-handed students. But stratified random sampling method can do that!
- The stratified random sampling method has three main steps:
  1) Subjects in the population (or frame) are divided in to $m$ non-overlapping groups (by one or multiple demographic factors) called **strata** (singular **stratum**).
  2) A random selection of subjects is performed $m$ times, one for each stratum.

3) The collection of the $m$ mini random samples forms the **stratified random sample**.

- The division in step 1 is done primarily by demographic variable (or independent variables), such as gender, ethnic group, religion, or socio-economic class etc.)
- The general rule of thumb is to make use of as many demographic variables as you could, so as to get a **homogeneous** group of subjects for each stratum.
- Pros: We get a better (or more accurate) result, especially when the subjects within each stratum are more homogeneous.
- Cons: Demographics of subjects and the percentage in the population are typically unknown beforehand. Also, a lot of work is needed to make it a stratified random sampling method. In other words, it is almost impossible in practice.
- Note: In some cases, researchers use **post-stratification** – The criteria question used for stratification is included in the survey, then subjects are stratified according to their demographic response. However, this means that the original sample must be large enough to make sure it has enough subjects from all strata.
- **Proportionate Stratified Sample** is typically desired as the ratio of strata in the sample is the same (or very close to) the ratio of strata in the population.
- MLO: Identify the most appropriate sampling method used in the given situation, from among the five sampling methods covered.
- MLO: Devise a sampling method (from the five methods), given the situation.

## G. Cluster Sampling Method (or Cluster SM)

- The **cluster sampling method** is typically used when subjects in the population can be represented on a map or in a layout (i.e. geographic factors).
- For example, you do not need to know the name and phone number of your neighbours when you are canvassing donations around your neighbourhood because they can be laid out on a map.
- The cluster sampling method has three main steps:
  1) Subjects in the population are divided in to $q$ non-overlapping groups (by geographical factors) called **clusters**.
  2) A random selection of $p$ clusters is performed once, where $p < q$.
  3) The collection of all the subjects in the $p$ selected clusters forms the **cluster sample**.
- The division in step 1 is done primarily by geographical factors, such as province, city, community, or blocks in the neighbourhood, or by buildings in a city, or floors or rooms within a building etc.
- Pros: We get a better (or more accurate) result, especially when the subjects within each cluster are more **heterogeneous**. It is because each sampled cluster contains all sort of subjects with different demographic backgrounds.
- Cons: The number of subjects within each cluster is not known ahead of time. Therefore, we typically include "more" clusters to make sure the minimum sample size requirement is met.
- MLO: Identify the most appropriate sampling method used in the given situation, from among the five sampling methods covered.
- MLO: Devise a sampling method (from the five methods), given the situation.

Use of Simple Random Sampling Method in all Statistics Courses

- Although simple random sampling method is not easy to do in practice, it is still the basis of **ALL** statistical formula and theories in this course.
- Therefore, unless otherwise stated, simple random sampling method will be the only method used for the rest of the course.
- In other words, when "random sample" is used in this course, it always refers to simple random sample.

## 2.5 Data Collection

- Sampling plans or sampling in general allows us to know which subjects to get to, but we still have to perform the **data collection** complete the task!
- Data can be collected through survey or interview. And it typically has the following three steps: 1) **design of the questionnaire**, 2) **collect the data**, and 3) **enter and clean the data**.
- Believe it or not, statisticians spend about 60-70% of their time in the above steps, and the remaining time on data analysis and reporting.

### A. Design of Questionnaires

- Besides the aesthetic side of things (like how to make it more appealing etc.), a few technical aspects of **designing survey** are to determine the number of variables (or questions), the order of them, the wording of the questions, whether or not you provide options for respondents to choose from etc.
- The choice of variables (or the questions in the survey) should be consistent with the objective of the study.
- The length of the survey is also important. The general rule of thumb is to have a single page (one side) of questionnaire, with a reasonably bigger font size and plenty of white space. After 7-8 questions or about a page, the quality of the responses goes down significantly because respondents start to lose focus after a few minutes.
- Try not to go over 10 questions with online surveys.
- The choice of words in the questions are also important. Having chosen a wrong word or a set of words could change the meaning of the questions. Hence, respondents may get offended. So, extra care must be taken to prevent misleading and/or offending the respondents.
- Please google "common survey mistakes" and read one or two websites and you will get the idea.

### B. Collection of Data

- This refers to the part of actual "going out" and collecting the information
- There are a number of methods we could use: face-to-face interview, regular snail mail, electronic mail, text messages, social media networks etc.
- Either paper survey or web-based survey (like Survey Monkey or Google Form) could be used for the above methods.

- Face-to-face interviews are typically the most costly (need to hire personnel, train them to maintain the quality of the responses, and pay them to go around places etc.)
- On the other end of the spectrum, sending online survey is quick and cheap, but the quality of the responses would typically suffer (unless your questions are worded in such a way that answers you get match exactly what you are looking for, which is very rare).

## C. Data Entry and Cleaning

- **Entering data** and **cleaning data** are two iterative processes.
- While entering data is typically rather fast, cleaning may be very long, especially when the data is huge.
- There are typically two things to clean. One is to correct the improperly entered values. Like "Male" and "male" in the Gender column because statistical software applications, like Excel, will consider they are two different words or classes.
- The second is when a unit is entered into a numerical variable column, like entering "175cm" in the Height column. This not only gives you error messages (yes, computer is dumb!), but it also stops from considering that column as numerical variable because of that one non-number response in "175cm".
- Please also note that a proper data set should not have anything other than the data themselves. Other description of variables, annotation or documentation should be in a separate file or document. In other words, it is essential a documentation of the variable be provided, along side with the data set.
- MLO: Understand the importance of having good or cleaned data set.

## D. Issues Relating to Data Collection

- There are potentially many issues that we need to be aware of.
- 1) In an effort to collect "better" data, subjects might be coerced or forced to do certain unpleasant things just the sake of data collection. For example, to see how smoking affects the weight of newborn babies, you force one group of pregnant women to smoke for a period and observe the weight of the newborn afterwards.

Please note that it is a rather serious problem because it deals with **ethical issue**.
To learn more about the ethical conduct, you are suggested to complete an online tutorial. TCPS Certificate: As part of the group project, students are required to complete an on-line tutorial from the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* web site. This free, web-based course describes principles, standards and procedures for governing research involving human subjects. The URL is: http://tcps2core.ca/welcome

- 2) **Placebo effect** usually refers to the state of mind of the untreated subjects during the experiment. This could have a psychological effect on the results. To prevent this, a **blind experiment** would be used to make sure that untreated subjects would still have some sugary pills (called the **placebo**) which looks and feels the same as the real treatment.
- 3) A blind experiment is nice, but the person on duty could enter biased data along the line of what the researcher wants. So, a **double-blind experiment** is usually used these days. Not only do the untreated subjects get their placebo, but the person on duty also has no idea about the identity of the two groups (one treated and one untreated/control).

- The last two are mostly **moral issue** and they could typically be reduced with better education or training of the staff.

### E. Data Ethics

- First and foremost, make sure we can differentiate three things: 1) law, 2) morality, and 3) ethics.
- Keep in mind we are not looking at things from a legal perspective. There are many things that we legal but still unethical.
- There are many areas we could talk about, but I would put our focus in three main areas.
- 1) **Ownership**. Personal information is like personal belongings. Taking possession of other people's personal information without getting their consent is considered unethical.
- 2) **Intentions and Transparency**. How and why do you collect the data? Do you really need to collect such information without alienating respondents? Would there be other ways you could collect similar information? Would respondents know what you are going to do with the information (like for profit or not)?
- 3) **Privacy**. Even when respondents give you consent to use their information, and they are aware of what you are planning to do with it, how careful you are going to use it? Do you just want to leave it on Google Drive, where data are store in the US soil? Do you want to save it in a portable hard drive without password protection? https://www.scientificamerican.com/article/security-breach-lost-laptop/
- This course is not to tell you what you ought to do and what ought not. It merely lets you know that the ethical issues behind data collection. It is particularly important when data collection becomes automated, instead of manual work.
- One last word. We have learned a lot about Machine Learning and how machine can help us make decisions in business, from credit card applications in finance, to health care subsidy applications in health or social welfare, to screening applications in HR etc. Proponents have been saying that the machines are more objective than human because it does not create any unnecessary bias. Or does it? https://www.technologynetworks.com/informatics/news/widely-used-ai-model-makes-robots-racist-and-sexist-363042
- MLO: Be aware of what data ethics in business is about.

## 2.6 Biases in Data Collection

- The following two biases are a bit more systemic problem with collecting data.
- The root of the issue can be understood by thinking about kids' response: "Yes, No, Maybe so".
- As you can imagine, "Yes" would not cause any issue but the other two would.

### A. Response Bias

- **Response bias** is a problem that arises when respondents have provided a response to the question, but the response does not reflect the truth.

- Some possible reasons stem from the poor wording of the questions, like confusing meaning, using double negatives, etc.
- Another possible reason is because the questions are sensitive or offensive to respondents. When we get asked personal questions or about some touchy subjects, we typically would say something quite different from the truth to avoid getting into trouble.
- Please note that response bias is not something that we are aware of. So, the best practice is to avoid having this type of questions.
- MLO: Discuss if a given situation involves response bias.

## B. <u>Non-Response Bias</u>

- **Non-response bias** is defined as the problem that arises when respondents, while they are included as part of the sample, decide not to answer some or all questions in the questionnaire.
- In other words, for the non-response bias to materialize, the subjects or respondents must be first chosen in the sample. More importantly, the process of interviewing has already been started.
- For example, if you approach a person (who has been chosen to be the sample) and ask if he can spare five minutes to do a survey, but he politely refuses to do so. He does not cause any non-response bias. It is because from statistics point of view, the next (randomly selected) subject is as good as this lost one!
- Although all reasons from the response bias section can be applied here, but the main reason why someone does not want to give you an answer is because the questions are too sensitive, offensive or personal, or it may be due to time limitation as well.
- For example, if the question is about illicit drug use, sexual orientation etc., then the chances are great that you would not get an honest answer (response bias) or no answer at all (non-response bias).
- MLO: Discuss if a given situation involves non-response bias.