Data Collection

Sampling

The goal of statistical work is to make rational conclusions or decisions based on the incomplete information we have in our data; this procedure is known as **statistical inference**. In inferential statistics we want to be able to answer the question: "If I see something in my data, say a difference between two groups or a relationship between two variables, could this be simply due to chance or is it an actual real difference in relationship?" If we obtain results that we think are not down to chance we would like to know what broader conclusions we can draw. For example, can we generalize them to a larger group or even to the whole world? Or perhaps we suspect our results support a novel theoretical model? And when we do see a relationship between two variables, we would like to know if we can say one variable *causes* the other to change.

Whether or not we can answer these questions, the methods we use to do so and the correctness of the conclusions that we can make all depend on how the data were collected. In this section we will focus in particular on how to collect data to ensure that we can make generalizations from our data to the group that we are interested in.

In statistical inference, we can imagine that we have a real world where we observe data, and a theoretical world where we have scientific models, where we think of plausible explanations for how our data arose, and statistical models that explain the variation in the data. The inferential process connects what we have observed in the real world to what we can say about the theoretical world.

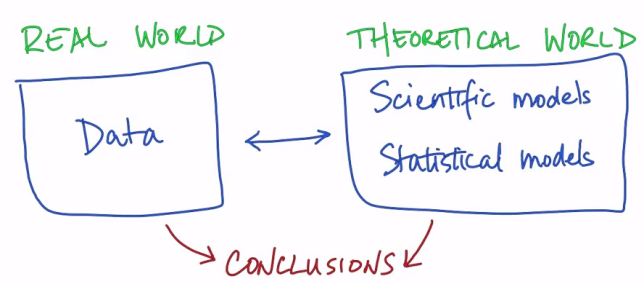

Figure 1: Process of statistical inference

Often we are interested in making conclusions about a particular group, for example, all humans, all Canadians, or women aged 15 to 25. We call this group that we are ultimately interested in the **population**. Data collected on the *entire* population is called a **census**. Many countries carry out censuses on its citizens every decade or so in order to inform future policies based on observed socio-economic trends. But in most situations, taking a census is

impractical because it is simply too time consuming and/or expensive. In some situations, such as testing a manufactured product, it might be necessary to destroy the item in order to study it and a census would then be inappropriate.

When we can't conduct a census we often collect data on a **sample**, or a subset, of our population. However, our goal remains unchanged: given the observed sample, what can we say about the population? As always, our population corresponds to the theoretical world that we are interested in.

Example 1
To have a better understanding of the concepts above, let's imagine a typical opinion poll on a politician (e.g., Barack Obama, the US President). Our population is people living in the relevant area (United States), and we are interested in what they think about an issue or how they would vote if an election were held. A poll would ask a sample of people a series of questions in order to draw conclusions about the opinions of the entire population.

Recall that a **statistic** is a value calculated from our observed data. Typically, a statistic is chosen to estimate a feature of the theoretical world or the population. The corresponding feature in the theoretical world or the population is called a **parameter**. We calculate a statistic in order to estimate an unknown parameter. When our data are measurements on a sample from a population, we need a **representative sample** in order to be able to generalize from the statistics calculated on the sample to the population parameters.

The key to selecting a sample that is representative is to use **randomization**. Using random selection ensures that we do not systematically exclude or over or under-represent any part of the population. There are multiple methods of collecting a random sample:

- **Simple Random Sample (SRS)**: each individual from the population is equally likely to be chosen.
- **Stratified Sampling**: first divide the population into non-overlapping subgroups (called stratas) and perform an SRS within each strata.
- **Cluster Sampling**: divide the population into non-overlapping subgroups (called clusters), select clusters at random, and include all individuals within the chosen clusters in the final sample.

The simplest is the Simple Random Sample. You can think of this as putting all members of the population into a (very large) hat, mixing it well, and randomly and blindly picking a sample from the hat.

For Stratified Sampling, typical subgroups include provinces or states when taking a sample of the people in a country. Stratified sampling is sometimes more convenient and ensures that the characteristics of each subgroup can be examined. It can even give more accurate estimates if there is less variation within a subgroup than there is in the whole population.

Cluster sampling is useful when it is easier to select groups instead of individuals from a population, and is often a less expensive way to get a random sample. As an example of cluster sampling, if the population is all students in an area, it may make more sense to randomly select schools than students and take our measurements on the students in the selected schools. Cluster sampling is also useful if each cluster is representative of the population. For the purposes of this course, we will stick just to simple random samples.

Many samples that are used in studies are not random samples. Some methods of non-random sample selection include:

- **Systematic Sampling**: select every kth individual from a list of the population, where the position of the first person is randomly selected from the first k individuals.
- **Convenience or Volunteer Sampling**: use the first n individuals who are available or the individuals who volunteer to participate

Systematic samples are fine if the ordering has no meaning and if we randomly pick which observation we start selecting on, but it can sometimes lead to non-representative samples if there is some structure to the list, such as hidden repeating patterns.

A convenience or volunteer sample almost always results in an unrepresentative sample. Suppose we wanted to learn about the views of the general population on the importance of statistics, and I asked the students in this course since I can conveniently reach you for your opinion. However, your view is most likely not representative of a population that also includes people who have never taken, or do not want to take, a statistics course.

If a sample is not representative, it can introduce bias into our results. We say that a sample is **biased** if it differs from its corresponding population in a systematic way. Bias can occur because of the way the sample was selected or because of the way the data were collected from the sample; this can result in statistics that are consistently too large or too small. Some types of bias are:

- **Selection Bias**: occurs when the sample is selected in such a way that it systematically excludes or under-represents part of the population. For example, we would miss everyone who only have cellular phone service in a phone survey that only calls landlines.
- **Measurement or Response Bias**: occurs when the sample is selected in such a way that it tends to result in observed values that are different from the actual vale in some systematic way. Response bias can occur because of improperly worded questions that may lead the respondent to give an opinion for or against a contentious issue, or because a question asks about a controversial issue for which people may hesitate at answering truthfully. Measurement bias occurs because the measuring instrument is faulty in some way.
- **Nonresponse Bias**: occurs when responses are not obtained from all individuals selected for inclusion in a sample. For example, if working parents tend to be too busy

to answer surveys then their views will be under-represented.

Example 2

There are many examples of non-representative sampling that have lead to unreliable data. A classic example is the Literary Digest poll of the 1936 U.S. Presidential Election. Literary Digest magazine sent surveys to 10 million people who were subscribed or owned cars or telephones. They received responses from 2.3 million people and based on these responses, they predicted a three to two ratio of votes in favor of the Republican candidate. However, his Democratic opponent won. So what went wrong?

The people surveyed either subscribed to the magazine or owned a car or a telephone. In 1936, this was only wealthy people. This is an example of selection bias. Moreover, this survey was voluntary response which typically represents angry people or people who want to change. Therefore nonresponse bias comes into play as well.

If our goal is to collect sample data that we can use to generalize about a population with no bias in our result, we need to select a random sample. In upcoming sections, we will discuss how the method we have used to collect our data can affect the strength of the conclusions we can make.