Objectives

- Sampling methods
- Simple random samples (used to avoid a bias in the sample)
- More reading (Section 1.3):
  - Chapters 1.3.2 and 1.3.3.
Topics: Sampling

- Learning objectives:
  - Be able to identify bad sampling methods
  - Know what a representative sample is.
  - Understand what a simple random sample is.
What is statistical inference?

- In statistics we rely on a sample – this a very small subset of a larger set, which is called a population.
- Example: Suppose you want to know the proportion of the US population who have attended university. It would be prohibitive to ask every American adult this question. Instead we query a relatively small number of Americans, and draw inference on the entire country based on this small sample.
- The procedure, whereby we convert information from the sample into intelligent guesses about the population, is called statistical inference.
- How to take a sample: A sample is a tiny subset of the population. Therefore it must be representative of the population, in the sense that no subpopulation is over or under-represented. If this happens it will induce a bias in the sample.
- On the next few slides we discuss bad sampling strategies and then the sampling method that gives a good representation of the population and under which statistical inference is based.
Bad sampling methods

Many people try to justify a point of view based on anecdotal evidence. **Anecdotal evidence** is based on individual cases, often selected because they are unusual in some way. They tend not to be representative of any larger group of cases and by definition tend to be biased. This is completely opposite to the statistical way of thinking.

Examples:

- Smoking is good for you “My great grandfather smoked like a chimney his entire life and live until he was 120”.
- There is no need to ban the use of cell phones. While driving “I always use a cell while I drive and I have yet to have an accident”.
- In many respects, anecdotal evidence is the opposite of statistical inferences. Whereas statistical inference looks at “typical” (representative) behavior, anecdotal evidence is notable because it is “atypical”.
Anecdotal evidence is one type of Haphazard sampling:

Examples:

- A substitute teacher wants to know how well students have done in an exam, so he asks the students in the front row their grades.
- Ask about gun control “on the street” in a large city in California or in a small town in Texas and you are likely to get very different answers. Even within the same area, answers would tend to differ if you did the survey on a university campus or at a ranchers’ cooperative.
- A study on high school SAT grades selects nearby schools because they are easy to reach.

- Haphazard samples are always biased: responses are limited to individuals who just happen to have been available.
- There is a very real danger of haphazard sampling in an observational study (which we define in a few slides).
Voluntary Response Sampling:

- Individuals choose to be involved. These samples are highly susceptible to being biased because people are motivated to respond (or not) for reasons often related to the response. Despite their common use, they are not considered to be valid or scientific.

- Examples:
  - To access the fitness of students, volunteers are asked to run 5K.
  - After buying a product you are requested to submit a review.
  - On a Have Your Say forum: 55% of the participants wrote in to say that immigration was bad for the country. But letters to forums are often written by disgruntled people. In a random sample of adults, only 30% thought that immigration was bad for the country.

- They are biased because the sampling method, can, systematically favor an outcome different from the truth.
Good sampling methods

A **Simple Random Sample (SRS)** is made of randomly selected individuals. This means each individual in the population has the same probability of being in the sample. *All possible samples* of size $n$ have the same chance of being the sample we select. A SRS is *completely unbiased*.

In practice this can be very difficult, so statisticians have many sophisticated methods for obtaining random samples. In this course, we will always assume we have an SRS.

A SRS is exactly what we are doing in class when selecting a sample of students to give their heights.
Size of population

- In a SRS there is always a chance that an individual will be selected more than once.
- In practice this usually cannot happen (a person does not like being interviewed twice).
- But if the population size is sufficient large as compared with the sample size, the chance that an individual can be sampled twice in a SRS is extremely small. Therefore, the impact of not using an exact SRS scheme is minimal.
- Since most data is not collected using an exact SRS scheme, a rule of thumb is that the sample size is at most 5% of the total population.
- If it is more than 5%, then most statistical analyses are not valid.
Collecting data through polls

- Polling is an easy and fun way to collect data.
- One can use web based applications such as easypolls.com to collect the data.
- However, one needs to be a little cautious when analyzing such data.
  - We recall that polling data is voluntary response, so people who respond often have a strong opinion on the subject.
  - If you send a poll out to friends and relatives, you need to be wary about the population you are analyzing. Since they are your friends or relatives they will be representative of a specific demographic such as young people, people of a particular race, people with a particular political view. Thus the population you are sampling from are not the citizens of the world but a far smaller subpopulation.
Objectives

Statistical studies and experiments

- Collecting data and problems with confounders
- Observational vs. experimental studies
- Sample vs. population

https://www.openintro.org/stat/textbook.php?stat_book=os Section 1.3.4, 1.4.1 and 1.5.1
Topics: Study Design

- Learning objectives
  - Understand what an observational study is
  - Understand how confounders (hidden variables) may lead to wrong conclusions
  - Understand what an experimental study is, and how can be used to remove the effect of hidden variables
Finding answers

To reliably answer specific questions set in a particular context, researchers must conduct an experimental or observational study. When the study is designed and conducted properly, statistical analysis can be used to formulate conclusions.

Data comes in many forms and from many sources.

Both the design of the study and the statistical analysis are required to have objective, reliable and justifiable conclusions.

Often data is publically available: collected in the past and made public for future purposes. This data is often produced by government organizations, and is not biased (as special interest groups are not involved). Check out websites such as the EPA, FDA, and data.gov (you can get data from these sites for your future project).
There are two kinds of studies

**Observational study:** Record data on individuals without attempting to influence the responses.

Example: Based on observations you make in nature, you suspect that female crickets choose their mates on the basis of their health. Method 1: Observe and record the health of male crickets that mated.

**Experimental study:** Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.

Method 2: Deliberately infect some males with intestinal parasites and see whether females tend to choose healthy rather than ill males.
Example 1: Observational studies

- Observational studies are essential sources of data on a variety of topics.
  - Example: To see whether child care had an influence on behavior, 1000 children were observed over a period of 12 years. It was `found’ that the more time children spent at day care between zero and 4.5 years of age, the more disruptive the children tend to be.

- This example, could be used to justify measures to encourage more mothers to stay at home. But extreme caution is required before we jump to such conclusions:
  - This example illustrates a problem with observational studies. It is not clear whether daycare had an influence on the child’s behavior or other factors/variables (such as low income families, stressed working parents, single parents) which may be more prevalent amongst parents who send their children to daycare.
Definition: Confounding

- The Daycare example illustrates the idea of confounding. It is not clear whether daycare or other hidden factors (low income, stress etc.) have increased the chance of disruptiveness in a child. This mix-up of variables between the factor (variable) we want to investigate and the other hidden factors (variables) is known as confounding.

- Example: An ice-cream company had a huge sales campaign in April. It was found that ice-cream sales had leapt in May, June and July. Does this mean the campaign worked?

  - Answer: May-July are typically the warm months in the year, where traditionally sales are high. Thus there are two factors which have could have contributed to the leap in sales (1) the sales campaign (2) the warm months. There is confounding between these two factors. Without additional information, it is impossible to disentangle them and know whether the campaign worked.
A study was done in the early 90s to see if the mortality rates between left and right handed people were the same. To do this, the psychologists collected the death records of 2000 people who had died in Southern California in May, 1990. They rang the families and asked whether the person was left or right handed. They also categorized the people into two groups: those who were above 60 when they died and those who were below 60. This is the data:

<table>
<thead>
<tr>
<th></th>
<th>Below 60</th>
<th>Above 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left handed</td>
<td>150</td>
<td>250</td>
</tr>
<tr>
<td>Right handed</td>
<td>300</td>
<td>1300</td>
</tr>
<tr>
<td></td>
<td>450</td>
<td>1550</td>
</tr>
</tbody>
</table>

Looking at the data we see that 37.5% of the left handed people died before 60, whereas 18.75% of the right handed people died before 60. This is a rather large difference. Does this mean if you are left handed you should be worried?
What the study did not take into account was that the proportion of people who are left handed has been **increasing over time**, due to shifts in social attitude.

60 years ago there was a policy to `change’ to the right side all children who showed signs of left handedness. Left handed people were turned into Right handed students.

In more recent times this measure has been phased out.

To understand how this would effect the data consider the following extreme case:

- Before 1930 no one was left handed. This means that all the left handed people who had died in 1990 must be below 60 (since here are no left handed people over 60). Without taking this change into account, the numbers would suggest that if you were left handed then you would die before 60! This of course is not the case, it is simply because there were no left handed people before 1930.

- The **hidden factor** is that there is an increase in the proportion of left handed people over time.

- This change in proportion is giving the perceived increase in mortality of left handed people at a younger age.
The Daily Eagle recently had the headline ‘20% of the people who died last year in the Bryan/College Station area have smoked marijuana (cannabis) in the past 10 years’. What conclusions can be drawn from this headline?

(A) Marijuana is bad for you, but only if smoked.  (B) Marijuana is bad for you.
(C) The data suggests that marijuana is bad for you when smoked.
(D) Nothing.  (E) Marijuana is good for you.
Experimental Studies

- These examples illustrate that it is often better (if possible) to design the experiment from which we collect the data.

- **In an experimental study** we deliberately impose some treatments on some individuals and observe their response.

- When our goal is to understand *cause and effect*, experiment studies are the only way to have fully convincing data.

- **Example 1**: Do smaller class sizes have an influence on exam results?
  
  An observational study will have too many confounders (hidden factors), because large class sizes tend to be in areas which are poorer and poorer areas have more social problems which can contribute to worse examination results.

  In the Tennessee experiment, researchers tried to prevent confounding by randomly allocating 6385 children to a normal class (approx 25), small class (approx 15) or normal class with extra teacher, and then follow these children over some years. This removes confounders such as income from the study.
Experimental Studies

- **Example 2:** To understand if a certain treatment has an influence on sick patients we need to randomly allocate a treatment or placebo to each patient. We should not give the sickest patients the treatment and the least sick patients the placebo, this would bias the results.

- **Example 3:** The CO2 obesity study discussed in the introduction is an experimental study. Different treatments were given to the rats and their weights/food consumption/hormone levels were observed.
Experimental Studies: problems

- Though ideal, sometimes it is impossible to conduct an experimental study.

- Example 4: To see whether daycare has an influence on a child’s behavior, at birth we randomly assign each child to either daycare or no daycare.
  - In practice this is impossible! This example illustrates situations where it is infeasible or indeed unethical to do an experimental study.

- Example 5: During the 1940s scientists observed that the incidence of lung cancer were higher amongst people who smoked.
  - This lead scientists to believe that a causality may be involved. That is smoking could directly be linked to an increase in cancer.
  - Ronald Fisher argued that such a claim could not be made on data collected in this way. He argued that an equally plausible explanation was that people who tended to smoke also tended to have a disposition for lung cancer. Though we now know that smoking does increase dramatically the chance of lung cancer, this line of argument is not wrong.
Question Time

What is the most effective method (may not be realistic) for establishing a causality between lactose consumption and the incidence of fraternal (non-identical) twins?

(A) Compare the rates of fraternal twins amongst vegans (people who don’t eat meat/dairy or eggs) and people who consume milk.

(B) Randomly place two groups of women (of child bearing age) into one of two groups (one group given dairy and the other not) and monitor over a few years.

(C) Compare rates of fraternal twins amongst women during the second world war (when dairy and meat was scarce) and now (when dairy is abundant).

(D) Ask a group of randomly selected women to place themselves into either the dairy and non-dairy groups and monitor them over a period of time.

(E) At a baby advise website ask for volunteers to give information about dairy consumption and whether or not they had fraternal twins.

What do you think may be a realistic method for looking for a relationship?
OkCupid, an online dating website, has spent a decade collecting and analyzing data from people who use the website.

The team randomly selects 5000 users and compares the average attractiveness scores they received (based on people who viewed them) with the number of messages they were sent in a month. This is an example of:

- (A) Experimental Study
- (B) Observational Study
Case study: antibiotics in farms


- Drug resistant bacteria (such as MRSA) is on the rise.
- The rise of this bacteria is due to the mass use of antibiotics (which causes the resistance).
- Antibiotics are widely used in farming
  - To increase the weight of animals (using low doses of antibiotics reduce the gut flora and increase its weight).
  - To reduce infection amongst animals in industrial farms (where conditions are cramped and infections easily caught and spread).
- An observational study
  - In a study conducted in Pennsylvania, people who were the most heavily exposed to crop fields treated with pig manure—for instance, because they lived near to them—had more than a 30% greater chance of developing MRSA infections compared with people who were the least exposed.
- This study (and many similar studies, read the above article) suggest that the wide spread use of antibiotics in farms is driving drug resistant infections in humans.
However, it has often been argued that farms use antibiotics that humans do not use and resistance that develops to these nonhuman drugs will not pose a risk to people.

Several studies have been done in labs to understand the influence feeding antibiotics have on drug resistance. They have shown that the above line of argument is incorrect.

However, it is still not fully understood how drug resistance spreads on large scale industrial farms and how this gets into the food chain.

The only way to understand this spread is to collect data using an experimental study on large scale farms.
Statistical Terminology

- The individuals in a study are the **experimental units**. We often call them **subjects**, especially if they are human. Collectively, the units are known as the **sample**.

- Each variable we measure on the units is called a **response**. It can be numerical or categorical.

- The purpose of the study is to compare the responses of different groups of subjects. The characteristic distinguishing the groups is called a **factor or a treatment**.

Example: To understand how exercise and diet effect blood pressure, one random group of people may be placed on a diet/exercise program for six months (treatment), and their blood pressure (response variable) would be compared with another group of people who did not diet or exercise.

- The larger (and often conceptual) group from which we select the sample is called the **population**. It usually is *much* larger.
Key point – we don’t see everyone

- The sample is usually far smaller than the population. But *the population is what we are interested in*, not just the individuals in the sample.

- The objective of statistics is to *extrapolate* from the evidence of the sample to the population (objectively and reliably). This is called *statistical inference*.

- Sometimes the sample is selected from a subpopulation of the population. If the subpopulation and population differ substantially then our results may be *biased*.

  Example: To understand how nutrition may influence the birth weight of a child, a simple random sample was taken from maternity hospitals in Texas. Problem: By excluding other states in the study certain parts of the population may not be represented in the sample. Thus leading to a biased sample.

- This illustrates that the sampling method is critical.
Accompanying problems associated with this Chapter

- Quiz 1 (Q1 and Q2)