Study Design

Sampling

Individuals are the objects being measured or described. These may also be referred to as **observational units** or **experimental units** or **subjects**.

Variables are the characteristics of the individuals that we are recording or measuring.

When analyzing data, we need to be aware of the distinction between a population and a sample. A **population** is the set of all subjects of interest. A **sample** is the subset of the population for which we have data. We are often interested in using the data from a sample to draw conclusions about the overall population.

For example, in some recent work I was involved in, we surveyed students at Seattle University to ask about their experiences with bias and discrimination here. In this study, the individuals are the students, and the variables would be each of the questions asked in the survey. The population of interest is all current a near-term future students at Seattle U, while the sample is the specific students who completed the survey.

As another example, I have been working with a nonprofit called KiloWatts for Humanity to develop plans to install a solar powered energy kiosk in the town of Kanchomba, Zambia. As part of these plans, we needed to understand current energy usage and expenses in the community. We administered a survey, randomly selecting households in the community and asking about the energy sources they currently use and what they pay for these. In this case, the individuals are the households, and the variables are each of the questions asked. The population of interest is all households in Kanchomba, and the sample is the households that we surveyed.

As a third example, I worked on a project with the U.S. Forest Service to study the spread of an outbreak of Jeffrey Pine Beetles in the Lake Tahoe area. Each year, data was recorded for every tree in the region, including elevation, diameter, tree species, and whether the tree was infested with Jeffrey Pine Beetles. In this study, the individuals are the trees, and the variables were each of the attributes: elevation, diameter, species, infestation status. In this case, the sample was all of the trees in this region for the time period studied, and the population of interest would be trees in similar regions or in this region in the future.

We tend to store data in spreadsheets or tables with each row representing an individual, and each column representing a variable. So the U.S. Forest Service data described above might look something like this:

Tree ID	elevation	diameter	species	infested
1	2113	7	Jeffrey Pine	1
2	2119	15	Ponderosa	0
3	2110	11	Jeffrey Pine	0

Different types of variables will need to be analyzed in different ways.

There are two primary types of variables: **categorical** and **numerical**.

Categorical variables record attributes that are non-numerical. They are sometimes also referred to as **qualitative** variables or **factor variables**. In the USFS example, the tree species would be an example of a categorical variable. The infestation status would also be considered a categorical variable - even though it is being recorded as a 1 or a 0, these are not numerical measurements, but indications of what category a tree falls into: infested or not infested.

Numerical variables, also sometimes called **quantitative** variables, record numerical attributes. In the USFS example, both elevation and diameter would be numerical variables.

We can further divide numerical variables into two sub-types: **discrete** and **continuous**.

A **continuous** variable is a numerical variable which could take on any value within some range. The tree diameter, for example, would not be restricted to only whole numbers of inches. A tree could have a diameter of 6.38 inches, or 6.383, or any value out to however many significant figures we have instruments capable of measuring.

A **discrete** variable is a numerical variable which can only take on specific individual values. Most often, this takes the form of a count. In the USFS example above, we do not have any discrete variables. But suppose we had additionally recorded a variable that counted the number of other trees growing within 25 feet of each tree. For tree #1, this might be 7 trees, or 8 trees. But it couldn't be 7.62 trees. Only certain types of values, in this case whole numbers, are possible for this variable.

We can also subdivide categorical variables into two types: **nominal** and **ordinal**.

An **ordinal** variable is a categorical variable for which there is some natural ordering. For example, if we were recording t-shirt sizes, they could be small, medium, large, etc. These are categorical, not numerical, but we can still order them from smallest to largest.

A **nominal** variable is a categorical variable for which there is not a natural ordering. Tree species is an example of a nominal variable.

In much of the material we look at in this course, we will be interested in drawing a **random sample** from a population, and recording and analyzing the values of one or more variables. While the term **random sample** can be used to refer to any sample collected through a process that includes some sort of randomness, within the context of our course we will use it more specifically to refer to a sample of independent and identically distributed (i.i.d.) observations from the population.

We will use capital letters to denote random variables (such as X, Y, Z), and lower case letters to denote actual observed values (such as x_1, x_2, x_3, where the subscripts can indicate which specific observation we are referring to).

If a population is infinite (such as flips of a coin) then it is straightforward to consider an i.i.d. random sample. Each flip of the coin will be independent of each other flip, and each will follow the same distribution (in this case, a Bernoulli distribution with the probability of success being equal to 0.5).

If a population is finite, then we need to be a bit more careful. Suppose we have a class with twenty students, and want to sample five students and record their heights. One approach would be to sample with replacement. We select one student at random from the group of twenty and record their height as x_1, then again select one student at random from the group of twenty and record their height as x_2, and so on. Since we are picking out of all twenty students each time, each observation will be independent of all of the others.

But suppose instead we pick without replacement. We first select one student at random from the group of twenty and record their height as x_1, then select one student at random from the remaining nineteen students who have not been selected yet, and record their height as x_2, then select one student at random from the remaining eighteen students who have not been selected yet, and record their height as x_3, and so on. Now, the observations are no longer independent, because the population being sampled from is changing each time.

Many of the methods we will be learning in this course are based on the assumption that we have an i.i.d. random sample. But in the real world, sampling is most often done without replacement. How much of a problem is this? It depends on how large the population is relative to the sample. If we have a very large population relative to our sample, then sampling with replacement and without replacement will behave very similarly.

For example, suppose we have an urn filled with 40% red balls and 60% white balls. We will take a sample of four balls from the urn and count how many of them are red. If we are picking with replacement, then X = the number of red balls will follow a binomial distribution. If we are picking without replacement, then X will follow a hypergeometric distribution.

Suppose the urn has a total of 10 balls, 4 red and 6 white. If we pick with replacement, we can find P(X = 2) using the binomial distribution in R:

```
dbinom(2,4,.4)
```

Giving us a probability of 0.3456.

On the other hand, if we pick without replacement, we can find P(X=2) using the hypergeometric distribution in R:

```
dhyper(2,4,6,4)
```

Giving us a probability of 0.4285714.

These two results come out rather differently.

What if we were instead drawing from an urn with 40 red balls and 60 white balls?

Drawing with replacement, the binomial result remains the same:

```
dbinom(2,4,.4)
```

Giving us a probability of 0.3456.

But drawing without replacement, the hypergeometric result is now:

```
dhyper(2,40,60,4)
```

Giving us a probability of 0.3520839, very similar to the binomial result.

Parameters & Statistics

When we consider numerical summaries of populations versus samples, we use different terminology and notation.

In general, a numerical measurement describing a population is referred to as a **parameter**, and a numerical measurement describing a sample is referred to as a **statistic**.

You have likely seen examples of parameters and their corresponding statistics. When describing a probability distribution for a population, we could be interested in parameters such as the population mean, μ, or the population standard deviation, σ. With a set of sample data, we could be interested in statistics such as the sample mean, $\bar{x}$, or the sample standard deviation, s.

One key distinction between parameters and statistics is that parameters are fixed, while statistics are random. Suppose our population of interest was all students currently enrolled at Seattle University, and we were measuring students' heights. There is some mean height, μ, for this population. If we were to randomly sample twenty students, we could calculate the mean height, $\bar{x}$, for our sample. We could then think about taking a new sample, and calculating the mean height again for this new sample. Every time we take a new sample, the sample statistic can change, but the population parameter remains the same.

Sample Surveys

Suppose we are interested in studying fish in a particular lake. Our population is all fish living in that lake. But we cannot realistically capture and record data for every fish. So we capture and record data for a sample of the fish, and use that data to generalize to the population of all fish in the lake.

In an observational study, we need to be especially careful about how we select our sample, to ensure that it is representative of the entire population. The set of subjects from which we will draw our sample is known as the **sampling frame**. For example, with our lake, perhaps one end of the lake is shallower than the other. If we only sampled fish from this shallower end, that is, if only fish in the shallower end made up our sampling frame, we might find different sorts of fish than we would find at the deeper end.

When a study is conducted such that some types of outcomes are systematically more likely to occur, and others are systematically less likely to occur, we call this **bias**. We want to design studies to minimize bias. The specific type of bias in the example above, where some types of individuals are disproportionately excluded from the sampling frame, is known as **undercoverage**.

A key component to minimizing bias is using **probability sampling**. In a probability sample, the ultimate decision about which subjects are and are not sampled is determined through some random process, rather than being left to the researcher to decide.

The most straightforward version of a probability sample is a **simple random sample**. This is like picking names out of a hat - a sample is chosen such that each subject has the same chance of being chosen, and any one subject's chances of being chosen do not depend on what other subjects have been chosen.

A disadvantage to a simple random sample is that there is a possibility that, due to random chance, you might inadvertently pick more subjects of one type and fewer of another type.

Suppose you were working with several hospitals in the area to conduct a study. If you had a list of patients from all of these hospitals, and randomly selected 200 patients to recruit for your study, it's possible that you could, through random chance, end up picking more patients from some hospitals and fewer from others. The sorts of people who go to each hospital might be different, due to differing socioeconomic levels in different neighborhoods. So you might end up biasing your study.

This concern can be addressed through a process known as **stratified sampling**. Rather than taking a single random sample from the entire population, the population is divided into groups (referred to as strata), and then a random sample is taken from each stratum. This ensures that the strata are all appropriately represented in the sample. In our example, each hospital could be considered a stratum. We could choose separate random samples from each hospital.

Another issue that can arise is that samples might be costly to obtain, in terms of money, time, and resources. This arises most often in situations that require visiting subjects in person to collect data. This can be addressed by **cluster sampling**. In this sampling approach, we randomly select clusters of subjects, and collect data for all subjects of in a cluster. For example, if Seattle U wanted to survey students via a paper survey distributed in person, they might randomly select some classes, and then have the surveys distributed to all students in the selected classes.

In many studies, a number of different stages of stratification and/or clustering are used in selecting a sample. This sort of approach is known as **multistage sampling**. For example, the National Youth Tobacco Survey selects first a sample of counties nationwide; then for each chosen county, a sample of schools is chosen; lastly, within each chosen school, a sample of classrooms is drawn, and all of the students in the selected classrooms are surveyed.

Studies

There are two general ways of collecting data: **observational studies** and **experiments**.

In an observational study, a researcher observes and records data for some variables of interest, without influencing or controlling the values of any of those variables.

In an experiment, a researcher controls the values of one or more variables, and then examines the impact of this on other variables of interest.

Suppose that you suspect that eating spinach reduces cholesterol. If you were to conduct a study, you might find a group of subjects to sample, and for each person, ask them how much spinach they had eaten in the past week (or some other time period, or ask how often they eat spinach), and then measure their cholesterol.

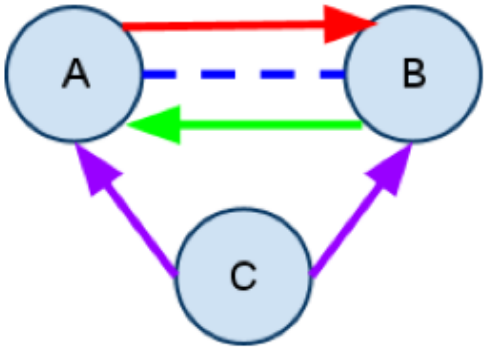
This would be an example of an observational study. The researcher is not influencing or controlling the values of any variables of interest, only recording them and analyzing them.

We could consider an alternate approach to studying the same issue: rather than just asking subjects how much spinach they eat, we would tell them how much to eat. We could divide subjects into a no-spinach (control) group, and a spinach (treatment) group. After a month of eating either their spinach-full or spinach-free diet, we could then look at how cholesterol levels compare for the two groups.

This would be an example of an experiment. The researcher is controlling the variable "spinach", and looking at how this impacts the variable "cholesterol."

There are advantages and disadvantages to both approaches. It can be difficult to talk about causation when we have an observational study.

Any time you observe an association between two variables, there are three general types of explanations possible.



Suppose that we observe an association between two variables, A and B. It could be that changes in variable A cause changes in variable B. Or it could be that changes in variable B cause changes in variable A. Or it could be that changes to some third variable C cause changes to both A and B. This third variable is then said to be confounded with the other variable.

Any of these causal explanations would result in the same observation of an association between variables A and B, so when you see an association, you need to think about what explanations seem plausible, and whether you can rule any out.

One of the most common mistakes in interpreting results from studies is in assuming a causal explanation based on having only observed some sort of association. Hence the saying "correlation does not imply causation."

In our observational study example, if you found that people who ate more spinach tended to have lower cholesterol, could you conclude that spinach lowers cholesterol?

That is one possible explanation. What other explanations are possible? Could cholesterol levels be impacting spinach consumption? Maybe something chemical happens in the body to cause people with lower cholesterol to crave leafy green vegetables.

Could some other variable be impacting both spinach consumption and cholesterol? In what ways might people who eat more spinach differ from people who eat less spinach? Could those factors also be related to cholesterol level? Maybe people who eat more vegetables also tend to exercise more often. It could be that the exercise is what's really impacting the cholesterol, not the spinach - if people changed their spinach consumption but did not change their level of exercise, maybe we wouldn't see any change in cholesterol.

In a controlled experiment, since the researcher is controlling the values of one variable, we can rule out most of these explanations. If we are controlling whether or not subjects are eating spinach, and still see an association between spinach consumption and cholesterol levels, we know that it couldn't be that differences in cholesterol were

causing people to choose to eat different amounts of spinach. Similarly, it couldn't be that some other factor, like level of personal health awareness, was impacting spinach consumption.

So an experiment has a major advantage in that it can allow us to talk about causation. Rather than just noting that people who eat more spinach tend to have lower cholesterol, it could allow us to conclude that eating more spinach causes a lowering of cholesterol (note that this is a hypothetical example - I do not know if there is actually evidence for this).

Observational studies have several advantages as well, though. They are often much less costly, in terms of money, time, and resources. Often data can be collected via a survey, or by examining existing records.

Observational studies also more closely reflect real life. If we recruit subjects for an experiment about cholesterol, they may behave differently than they normally would, since they know their cholesterol is being monitored. In an observational study, we can just ask about how they eat normally.

When studying the effect of some treatment in an experiment, we need a baseline for comparison. We can establish a baseline by separating our subjects into two (or more) groups: a **control group** who do not receive a treatment, and one or more **experimental groups** or **treatment groups** who do receive a treatment.

This allows us to adjust for any changes we would have expected to see in our subjects even without the treatment. For example, if we are considering the effectiveness of offering supplemental tutoring in math for 2nd graders, we know that they will gain some skills in math over the course of the school year even without the supplemental tutoring. So we can't just compare their performance at the end of the school year with their performance at the start of the school year. We need to compare them to other students who did not receive the supplemental tutoring.

In some situations, subjects will respond differently just because they think they are receiving a treatment. In order to adjust for this, studies make use of what is known as a **placebo** - something that appears to be a treatment, but does not actually do anything. This often takes the form of a sugar pill that contains no medicine, for example.

When a study is designed so that the subjects do not know if they are receiving the treatment or not, we refer to it as a **blinded** study.

But just keeping subjects from knowing is not always enough. If the researcher who is interacting with the subjects knows whether they received the treatment or not, they might unintentionally influence the outcomes. They might act differently towards subjects who receive the treatment versus those who do not. Or when recording their assessments, especially if any part of it is subjective, they might interpret things differently for patients receiving the treatment versus those in the control group.

Because of this, studies will often be designed so that the researchers interacting with the patients also do not know which patients are in which group. This is referred to as a **double-blind** experiment.

These issues of blinding do not arise in studies where the subjects are not people.

Additionally, we want to ensure that the groups receiving the treatment and not receiving the treatment are comparable. To avoid bias in determining who does and does not receive treatment, we generally use some sort of **random assignment** to the groups.