Types of studies

If we want to put to good use everything we've learned so far about data, we'll need to know how to run studies in a way that gives us good, reliable data. In this section we'll talk about different kinds of studies we can use to collect data, including observational studies and experiments.

In the next section we'll talk about how to make sure these studies are producing reliable results.

The goal of collecting samples

The purpose of statistics is to gather information (data) about the world around us and analyze it in some way to help make sense of it.

Because collecting data for an entire population is usually difficult or impossible, we instead choose a smaller sample of the larger population, and then analyze the data for the sample, hoping that our results will translate to the larger population.

Characteristics like mean and standard deviation are called **statistics** when we calculate them for a sample. A **parameter** is the corresponding characteristic of the population that the statistic is trying to estimate. So we could choose a sample, calculate the sample mean (a statistic), and then use what we know about the sample mean to make inferences about the population mean (a parameter).

Observational study

In an **observational study**, we're just looking at the information that's already there, or measuring it in some way, but we're adding nothing to the population that will change it in any way. In statistics, something that changes a population is called a **treatment**, so for an observational study, no treatment is applied.

One-way tables

For example, let's say we want to know whether all the students at our school prefer peanut butter or jelly. We may choose to use the students in our classroom as a sample in order to estimate the preferences of the entire population (all the students at our school).

We ask every student in your classroom if they prefer peanut butter or jelly, and they give us an answer of "peanut butter" or an answer of "jelly." We now have data for a one-way table in which the individuals are the students in our classroom, and the variable is "Peanut butter or jelly?" If we find that 70% of our classmates prefer peanut butter, we might infer that 70% of all the students at our school also prefer peanut butter.

Notice that we didn't do anything here except ask a question and record the responses. We didn't do anything that would change anyone's mind in any way, because we just wanted to make an observation about what was already going on.

Keep in mind that it's only true that 70% of the students in our school prefer peanut butter if the students in our classroom make up a random, representative, unbiased sample of the whole school, which they may not.

In fact, we'd say that we introduced bias into our study by convenience sampling, which we'll talk about soon.

Two-way tables

Sometimes we might want to collect data in an observational study for a two-way table, (as opposed to just a one-way table in the above example) and understand how two parameters might move together in a population.

Maybe this time we want to know how height affects peanut butter and jelly preference. In other words, this time we'll survey all the students in our school, asking them whether they prefer peanut butter or jelly, and record this information along with their height.

We're looking to see how much height and peanut butter/jelly preference are correlated, if at all.

Keep in mind that even if we found that peanut butter/jelly preference and height were positively correlated, such that taller students were more likely to prefer peanut butter, and shorter students were the more likely to prefer jelly, we could only show correlation, not causation.

Two variables are **correlated** when they move together predictably. The variables are **positively correlated** when they increase together or decrease together. Variables are **negatively correlated** when they increase and decrease in opposite directions: one goes down while the other goes up, or one goes up while the other goes down.

On the other hand, **causation** means that one variable *causes* another variable to change. But just because we show correlation does not mean that we've proven causation.

For example, even if height and peanut butter/jelly preference are correlated, we don't know if being taller causes someone to like peanut butter more, or if liking peanut butter more causes someone to be tall. We don't know which variable causes which. Nor do we know if there's a **confounding variable**, which is a third variable that leads to both of the variables that were correlated. For example, being male might cause someone to be both taller and to prefer peanut butter.

Experimental studies

In an **experiment**, we're manipulating what's happening, and trying to establish causality, not just correlation.

To run an experiment, we assign people into at least two different groups, hopefully using good random sampling techniques, so that our groups aren't biased in some way.

One group acts as the **control group**, which is the group that does nothing, receives nothing, or isn't manipulated, and the other is the **treatment group** (also called the experimental group), which is the group that does something, receives something, or is treated in some way. The classic example of this is in medical studies, where the treatment group receives some kind of new drug, and the control group receives a **placebo**, or sugar pill.

In an experiment, we're looking to see whether one or more **explanatory variables** (the treatment) has an effect on the **response variable** (whatever is expected to be effected). If we're testing to see whether a new drug decreases blood pressure, the new drug would be the explanatory variable (the thing that explains the change), and blood pressure would be the response variable (the thing that might decrease as a result of the drug).

Even if our experiment shows a change in the response variable, we still may need to be skeptical of our conclusion. Did we run a good experiment? Could the results have been biased in some way? Was the effect we saw simply due to random chance or the placebo effect?

There are other things we can do to make our experiment more reliable. For example, we could make our experiment blind or double-blind. A **blind experiment** is when the participants don't know whether they're in the control group or the treatment group. A **double-blind experiment** is when neither the participants nor the people administering the experiment know which group anyone is in.

Matched pairs

When researchers separate participants into like groups, it's called **blocking**. For example, researchers might choose to block on gender by randomly selecting an equal number of men and women, instead of a truly random sample in which the number of men and women isn't controlled.

If they then treat half of the men and half of the women with the drug, and give a placebo to the other half of the men and the other half of the women, the blocking on gender helps them to see if the drug effects men and women differently.

A **matched pairs experiment** is a more specific kind of blocking where we make sure that the participants in our experimental group and control group are matched based on similar characteristics.

Maybe these researchers want to see how both gender and age change the effect of the blood pressure drug. They could match the ages and genders in the control group with the ages and genders in the experimental group. For example, they could put one 18-year-old woman in the treatment group, and put her matched pair (another 18-year-old woman) in the control group.

A matched pairs experiment design is an improvement over a completely randomized design, because participants are still randomly assigned into the treatment and control groups, but potentially confounding variables, like age and gender, are controlled for.

Replication

We also want to make sure that other people can replicate our experiment. If other people can run the same experiment in the same way, and they get the same results that we do, that provides more evidence that our results are legitimate.