Chapter 2.1 - Observational Studies

> "As far as the laws of mathematics refer to reality, they are not certain, as far as they are certain, they do not refer to reality." - Albert Einstein

A surprising number of students take issue with applied mathematics. Their sole exposure prior to completely divorcing themselves of math education is usually in the form of an incoherent word question— one where Matthew has seemingly acquired 100 watermelons and is unsure how to divide them among his 3 pickup trucks of varying sizes. This is because mathematicians aren't very fond of going outside; any attempts at explaining real world scenarios are met with the same success as Jeff Bezos guessing the price of a gallon of milk. Statisticians are much better at describing real world phenomena because we work with real world *data*.

The benefit of data is that it inherently holds applied word problems within it. The downside of data is that it has to be collected (and cleaned) before it can be analyzed. This chapter covers the two pillars of data collection, their strengths, and their weaknesses. It should be noted that, as this is an introductory textbook, these methods of sourcing data won't be comprehensively discussed. Entire textbooks and fields of study have been developed on each method alone. Once again, if you turn the final page on this book with unanswered questions, it would benefit you *immensely* to perform some self-discovery.

Observational Studies

Scientific education puts a lot of stock into teaching students about experiments early on, only for students to graduate into researchers and discover that they'll never perform any experiments. The reason for this may not be obvious, but it *is* intuitive.

Imagine you've been asked to determine the effects of smoking cigarettes on heart health. If we were to follow the instructions laid out for us in secondary school we'd acquire some arbitrary number of participants then divide them into groups of smokers and non-smokers. The smokers would be prescribed a number of daily cigarettes and the non-smokers would be held as a "control" group. We'd check up on our participants regularly over many years until we had enough information to make our conclusion.

Hopefully it's clear why that's a problematic thing to do. We've violated *at least* one ethical principal, failed our course in moral philosophy, and likely blocked off the next 5 years of our lives for court appearances. But we **know** smoking causes heart problems. How could we without an experiment?

Imagine if we designed a different kind of experiment. One where we allowed the world to occur naturally; people will smoke of their own accord and others won't. Years down the line we seek out participants with similar characteristics, except our "smoking" and "non-smoking" groups were defined long before we, the researchers, even began rounding them up. We might not be able to get the same kind of data from this experiment as the more "traditional" one, but we could still pull a lot of science out of this group. We could survey both groups for any heart attack or stroke incidents in their lives, have them provide their general medical history, even screen them in the moment for heart health.

This "experiment" is a known and named method called an **observational study**.

Observational study: A study where the independent variable is not controlled by the researchers.

"*The independent variable is not controlled"*, is the key component to identify. Many observational studies could look very similar to experiments until we look deeper into where the researchers stop influencing the study.

Consider the concept of city-wide mask mandates to slow disease spread. The city government has enforced something that *should* have a quantifiable effect on the number of confirmed positive disease cases. This looks like controlling the independent variable, but it's not.

Even if everyone in the city follows the mask mandate there could still be positive cases. An epidemiologist observing the data from the duration of the mask mandate wouldn't have known prior to observing the results how many individuals were going to be positive. While the city *influenced* the independent variable, they didn't have any **control** over it.

Case-control Studies

1 in every 1000 male newborns are affected by Jacob's syndrome, a rare genetic condition where a male receives an extra Y chromosome. If you were to sample 10000 men using a simple random sampling technique, how many positive cases of Jacob's syndrome would you expect to find?

$$\frac{1}{1000} \times 10000 = 10$$

Eventually we'll discuss just how bad of a sample this is for statistics, but for now we can appreciate the complete lack of practicality going on. Even if the method of sampling used a compulsory survey we would have gone through a lot of effort to end up with 10 subjects. If we're interested in another unique phenomena within Jacob's syndrome subjects, possibly one that occurs in 1 out of every 10 individuals with the condition, we've *potentially* paid to survey 10000 men just to end up with 1 useful data point.

It'd be more reliable to put out an open notice for individuals **with Jacob's syndrome** to submit themselves to a study. All of those resources could be directed at confirming positive status instead of finding it to begin with. Even if we ended up with 100 subjects, we now have 10 useful data points for what's likely the same resource input. We call these observation studies **case-control studies**.

Case-control studies: Intentionally selecting confirmed case positive participants (cases) and pairing them with confirmed case negative participants (controls).

These studies fall into two categories: Prospective and retrospective.

Prospective studies gather subjects then observe outcomes as they observe, (i.e., checking for onset of heart disease in smokers).

Retrospective studies looks into subject's pasts, (i.e., asking smokers if they've ever experienced chest pain when seated on the couch).

Cohort Studies

Millions of adults in the US have Type 2 Diabetes. As such, research into treatment is a wildly lucrative industry. Developing higher quality medical treatments tends to involve looking into the various reactions different demographic groups have to treatments. Experiments are useful for looking at immediate or short-term reactions, but these are almost entirely confined to animal experimentation and very cost ineffective for long-term reactions. We need an **ethical** way to tease out the long-term effects of medical treatments, and when ethics are a problem observation studies are good at providing the solution.

Say we wanted to determine the long-term side-effects related to Ozempic on men. This drug is FDA approved so it should be safe and fully understood, right? Right. But we're paranoid researchers so we resolve to develop an observational study to determine these effects.

We gather 1000 men across four different age groups, all of whom currently use Ozempic. We check in with these men every 6 months for the next 5 years. As participants stop using the drug they're removed from the study without being replaced. We continue this until either no participants are left or the full 5 years have elapsed. This is another special type of observational study known as a **cohort study**.

Cohort study: Subjects sharing a common demographic characteristic are enrolled and observed at regular intervals over an extended period of time.

Cohort studies are generally prospective studies, although retrospective cohort studies do occur. The major advantage of these studies is that they're **highly** informative due to their incorporation of temporality (passage of time). Another important convenience is that they're restricted to demographically similar participants, hence the existence of retrospective cohort studies. Realistically any gathering of demographically similar participants that disregards positive or negative outcome status and looks over a long period of time, past or present, can be considered a cohort study.

Problems in Sampling

A recurring theme in this chapter is that all of our methods are slightly terrible. The sooner we accept that and understand their disadvantages the sooner we can get back to doing proper science. So there's a few ideas we need to nail down before we can proceed.

1. **Making up data is always bad**. Simulation studies are something statisticians engage in when developing new methods, but these are not replacements for proper data. If we're generating fake data and trying to draw real conclusions from it, we're generating fake science.
2. **Samples of convenience are easy, cheap, and easy to intentionally bias**. This makes them *incredibly popular* with media outlets. Whenever you're reading about a trivial study reported on by a news station you should be certain to check for any declaration of study participants. Rounding up 12 people from the streets of Los Angeles and asking them how they feel about GMOs doesn't provide the same inference as a properly designed survey.
3. **Voluntary response surveys can work well if designed well, but they're very easy to design poorly**. If you'd like to become exceedingly wealthy one of the ways you could do that is by fixing how political polling works. Currently, US pollsters call randomly selected registered voters. On the phone. It's fair to say that they're struggling to get young people to answer the phone, let alone their questions. On top of that the ones that answer their questions are *fundamentally different* from the ones who don't, meaning that the sample being analyzed is representing a **specific subset** of the population rather than casting a general net over it.

Surveys are particularly tricky since they make up a large fraction of the data collection done in observational studies. Below are some points of severe bias in survey design and questions you can ask yourself and your lab group to determine if the survey has been designed improperly:

Bias	Question
Undercoverage	Did we reach all of the intended groups in our population?
Nonresponse	Did we complete our survey with all respondants?
Response Bias	Were the respondants likely to avoid answering a question truthfully?
Question Wording	Do we invoke bias through the way we construct our questions?
Order of Questions	Do previous questions have an effect on the response to later questions?