Chapter 1.1 - Samples, Populations, and Estimates

"Try not to have a good time–this is supposed to be educational." - Charles Schulz

The general goal of applied statistics is to develop methods for representing information in such a way that decisions can be made from it. These chapters will aim to put you into the shoes of an applied statistics. So to speak— the lessons are meant to lead you towards developing these (very well known and studied) methods as if you were the first to propose the idea.

It's important to practice this skill of creative thinking and self-discovery. Mathematics is a field where imagination and disruptive thinking separate the good from the great.

Statistics as a Science

In the sciences we often talk about the "Scientific Method", this idea of proposing a question to the universe then systematically working towards an answer.

Observe
Question
Hypothesize
Experiment
Analyze
Conclude

Many students tend to think of statistics as a process only involved in the "Analyze" step, but what if it wasn't? Let's consider the possibility that statistics, just like biology or chemistry, is a science.

Since statistics is a science we can begin from the same place, a question (or three):

- **How many students are on the third floor of any dorm on campus?**
- **How many students are enrolled at K-State?**
- **How many undergraduates are there in America? The world?**

These questions have different scales to them, we can't answer them right now but we can easily see how the first question is much less involved than the last.

Importantly, these aren't very scientific questions. They're more along the lines of "interest" questions— the type of thing we'd plug into Google and move on with our lives. We're here to do science, so let's ask a more scientific question.

When I was first shipped off the college I was warned about the insidious "Freshman 15" where the diet associated with dorm living is so unhealthy and calorie dense that almost everyone puts on 15 lbs. in their first year. As I went through the year that concept was showcased left and right, everyone seemed to be putting on weight (myself included).

But is it actually 15 lbs that everyone puts on? Or is it an urban legend that we've just never looked into?

I would naturally want to check the average caloric intake of undergraduate students in the US. This is troublesome since I live in Kansas, not the entirety of the US. How, then, can I find the answer to my question?

As one of my early mentors in data science would say: If you want a specific answer you have to ask a specific question. So we would be smart to specify our question a little before we go any further:

> What is the average caloric intake of undergraduate students in the US?

Then the scientific method would have us forming a hypothesis.

> The average caloric intake of undergraduate students in the US should be much higher than the caloric intake of non-students of similar ages.

The next step seems straight forward; I have to run an experiment. So I take my question and hand it over to a 5 different Universities across the US:

- Kansas State University
- UC Davis
- (The) Ohio State University
- UCONN (University of Connecticut)
- Texas A&M

I tell them to select 200 students at random and determine their caloric intake. All in all, I end up with 1000 students representing the totality of American Undergraduates.

What have I done? I've taken a **sample** from my **population**.

- **Population:** the *entire collection* of individuals about which information is sought.
- **Sample:** a *subset* of population, containing the individuals that are actually observed.

I calculate an average from my sample, and I end up finding out *some value* that I use to infer the caloric intake for undergraduates across the entire US.

What have I done now? I've made *inference* about a *population* from a *sample*; I've done **Statistics**.

- **Statistics**: is the study of procedures for collecting, describing, and drawing conclusions from information.

To put it plainly, statistics is the act of describing or making inferences about a population, from a sample.

If I wanted to move forward I would have to *test* my hypothesis somehow and form a conclusion. I encourage you to think about how you could do this, then dive further and ask how you could be *certain* that your hypothesis is correct. Just like any good TV show, you'll find out the plot twist in that question at the end of the "season".

Parameters and Statistics

Statistics is an entire language. Fortunately we can use English for most of our education in Statistics. We'll slowly discuss the special cases where we use the "Statistical Alphabet" but for now we can jump straight into learning the basic operating words we'll use every day in our "conversational" statistics.

Let's look at a different example

Raccoons get rabies, more than normal for most mammals. The Kansas Department of Wildlife & Parks (KDWP) decides to investigate how prevalent rabies is in the state.

KDWP estimates there are roughly 3.3 million raccoons in Kansas. They capture 10000 raccoons across the state and test them for rabies. They find 382 raccoons that test positive for rabies, with the rest being negative.

- In this study, what is considered the population?
- What is the sample?
- What does the study tell us about raccoons and rabies in Kansas?

In our study, we had a distinct population and sample, with a distinct quantity for each. This number can be very useful, but is generally *insufficiently informative*.

The Center for Disease Control (CDC) estimates roughly 10-14% of raccoons carry *Rabies lyssavirus*. From KDWP's study, they found that 3.82% of the raccoons in their sample had rabies, and extrpolated that to the entire population.

$$\frac{382}{10000} = 0.0382 \times 100 = 3.82\%$$

What we now have are two values that *describe* our population and our sample.

- **Parameter**: a value that describes an entire population.
- **Statistic**: a value that describes a sample.

a. What was our parameter in the above study?

b. Our statistic?

Sampling Techniques

If we want to conduct science we should know how to build a proper study. Statistics is a field that's rather reliant on gathering information, so naturally there's a lot of defined vocabulary for the different methods by which one can gather that information.

Let's say I want to know how many individuals in a (random) Division of Biology consider Cell Biology to be an enjoyable class.

I decide to assign every student who's taken the class and declared a major that falls under this Division of Biology's umbrella a number from 1 to 500.

I then generate 50 random numbers from 1 to N, and select those students to participate in my one question survey.

What I've done has resulted in a sample size of $n = 50$, where *every individual was equally likely to be selected*.

$$\text{Probability of being selected first} = \frac{1}{500}$$

> As a "fun" (i.e., optional) exercise: What would happen to this probability if I select the students one at a time and remove them from the pool every time they're selected?

What I've performed is called a **Simple Random Sample**.

- **Simple Random Sample (SRS)**: a sample chosen by a method in which collection of n population items is equally likely to make up the sample.

Let's say I divide the 500 students into two groups: Pre-Med and Not-Pre-Med

I end up with a split of $n_1 = 300$ Pre-Med students and $n_2 = 200$ Not-Pre-Med students, then I perform my SRS on each group.

This is a **Stratified Sample**.

- **Stratified Sample**: The population is divided into groups, called **strata**, where the members of each *stratum* are similar in some way. Then a SRS is drawn from each stratum.

Let's go back to our raccoon example:

KDWP samples those raccoons from pre-defined areas, subsections of Kansas, rather than going across the entirety of Kansas in a big fire line and snatching up suspicious racoons.

Each of those subsections of land are called **clusters**, and the technique we've used here is called **Cluster Sampling**.

- **Cluster Sampling**: Items are drawn from the population in groups, or clusters.

I drive a Honda Fit, which was built on an assembly line, and part of the process of building that car on an assembly line was something called *quality assurance* (QA). Considering my car has yet to blow up on me, it seems to have passed the QA check. Likely because they used a proper sampling technique:

The part of the assembly line that produces the muffler for Honda Fits decides that every day they'll draw a number, k, between 3 and 6. Whatever number they draw is now the first muffler that comes off of the line to be checked for defects. Then they'll check the k^{th} item moving forward.

They draw the number 4. So the 4^{th} muffler that comes off of the line will be evaluated for possible defects, then every 4 mufflers after that will also be checked for defects.

This is called **Systematic Sampling**.

- **Systematic Sampling**: a starting point is chosen randomly and then every k^{th} item in the population is selected.

Recall our example with Undergraduate caloric intake:

I instead instruct the 5 chosen Universities to send the survey to every student's email inbox. They are given the option on whether or not to participate.

This is naturally called a **Voluntary Response Sample**, due to the participants getting to *choose* whether they are involved or not.

Let's say I decide to calculate the caloric intake of Undergraduates in the US based off of the students in my morning lecture.

That's called a **Sample of Convenience**. It was easy, I could finalize it right now, and it'd be fairly incorrect.