Collecting data

S520

August 23, 2016

Read this in addition to (not in place of) Trosset chapter 1.1 and 1.2

NOTE: This is an *R Markdown* file. To generate an HTML (or PDF) document from this file, open it in RStudio and click "Knit HTML" (or "Knit to PDF").

In this chapter:

- Why do scientists need statistics?
- What kinds of data are good for answering different kinds of questions?
- Why is probability so important to statisticians that we're devoting a third of the course to it?

Terminology

Trosset uses the term *experiment* quite generally, to mean any methodical collection of data. Read chapter 1.1 for his examples:

- Spinning a penny hundreds of times to see what percentage of the time it lands heads up.
- Measuring the speed of light (using a fixed mirror and a revolving mirror) a hundred times.
- Letting termites attack grids of toilet-paper rolls 30 times to see if there are patterns in termite foraging.

For now, the most important thing is that all these studies involve *repeating an experiment lots of times.* As we'll see in this course, we generally need lots of repetitions to work out what's going on in a given system. This means there's lots of data, which means we need statistics. (And as an aside, if we have lots of data, computers will make our lives a lot easier.)

Types of studies

Here are two general types of questions that good statistics can help us answer.

- *Describing a population:* How do American voters intend to vote in the next Presidential election? If you spin a penny, how often does it land heads up?
- *Cause-and-effect:* Does giving milk to children help them grow? Does a polio vaccine prevent polio? Does smoking increase your chance of getting cancer?

Notes:

- In the coin example, the population is hypothetical: There's an imaginary infinite set of coin spins. Fortunately, probability deals with infinite populations even more easily than finite populations.
- Another thing you can do with statistics is *prediction*, but this usually follows from a descriptive or a causal analysis. For example, if you want to predict whether someone's going to get lung cancer, you'd first want to study whether smoking causes lung cancer, and if so, how much it affects your chances.

When deciding how much trust to place in a study, we need to ask ourselves a few questions:

1. *What question was the study trying to answer? (Is the question descriptive or cause-and-effect?)*

2. *What data did the study collect to answer this question?*
3. *What analysis did they perform on the data?*
4. *Does this data analysis give a good answer to the question?*

The "analysis" part is the trickiest for both people doing studies and people reading them. In this course (and ones you'll take in the future), you'll learn a few analysis methods. We'll see that there are usually several valid ways to analyze a particular set of data, along with an infinite number of bad ways.

Let's now look at a few studies in detail.

The Lanarkshire Milk Experiment

From Trosset p. 8. The famous paper this example comes from, by the statistician "Student" (whom we'll hear more about during this course), is easily Googleable and quite readable.

"A 1930 experiment in the schools of Lanarkshire attempted to ascertain the effect of milk supplements on Scottish children. For four months, 5000 children received a daily supplement of 3/4 pint of raw milk, 5000 children received a adily supplement of 3/4 pint of pasteurized milk, and 10,000 children received no daily supplement. Each child was weighed (while wearing indoor clothing) and measured for height before the study commenced (in February) and after it ended (in June). The final observations of the control group exceeded the final observations of the treatment groups by average amounts equivalent to three months growth in weight and 4 months growth in weight, thereby suggesting that the milk supplements actually retarded growth! What went wrong?"

Let's see if we can get answers to our four questions above.

1. **What question was the study trying to answer?** Does milk help children grow? This is a cause-and-effect question: milk is the potential cause, growth in height and weight are the effects.
2. **What data did the study collect to answer this question?** The heights and weights of three groups of children – raw milk, pasteurized milk, and control – before and after the study.
3. **What analysis did they perform on the data?** A comparison of the final average heights and weights of the milk groups with the control group.
4. **Does this data analysis give a good answer to the question?** Apparently not!

Let's have a closer look at question 2 in particular. First, let's state this fundamental principle of cause-and-effect studies:

- *To make a valid estimate of cause-and-effect, we should compare groups that have no systematic differences except for the cause.* Then we can be sure that any differences are due to the cause, and not some other variable.

It follows that we should expand our second question:

2a. For studies of cause-and-effect: What groups are being compared? How are the groups assigned? Are there any systematic differences between the groups besides the cause under study?

We go back to Trosset and read the following important detail:

"An initial division into treatment versus control groups was made arbitrarily, e.g., using the alphabet. However, if the initial division appeared to produce groups with unbalanced numbers of well-fed or ill-nourished children, then teachers were allowed to swap children between the two groups. . . "

Letting the teachers have this discretion turned out to be fatal to the study. Bleeding heart teachers took malnourished (smaller) children out of the control group and put them into the milk group. This meant there was a systematic difference between the groups aside from milk: The control group (on average) was heavier than the milk groups to begin with, so it's unsurprising the control group was still heavier afterward.

The cliche is that "correlation is not causation" – that is, a relationship between two variables may not indicate cause-and-effect unless we can be sure the relationship isn't due to some other variable. We'll put this wording aside for now because to statisticians, correlation is a technical term that we won't get around to defining until chapter 14.

Student made two suggestions for a better experiment:

1. A **randomized controlled trial**, where the groups were assigned at random, without any discretion after the assignment.
2. A study of twins, where one (randomly chosen) twin gets milk and the other doesn't.

In both cases, **randomization** is the way to ensure there are no systematic differences between the groups. Then all the differences are either due to the milk or doe to chance, and one thing we'll get really good at during this course is working out how big a difference due to chance is likely to be. Student's experiment 1 is a bit better because it could be that twins are somehow different from the rest of the population in their reaction to milk, e.g. because they're smaller. On the other hand, experiment 2 could potentially be done with a much smaller sample size (i.e. much more cheaply.)

In addition to randomization, experiments involving people should ideally be **blind**: the subjects of the experiment should not know if they're in the treatment group or the control group. In medical experiments, this is done by giving **placebos** to the control group that appear the same as the actual treatment. This isn't really practical in the milk study, however – issues of cost aside, most people would be able to tell the difference between real milk and fake milk.

1948: Dewey vs. Truman

From Trosset p. 9:

"The 1948 presidential election pitted Harry Truman, the Democratic incumbent who had succeeded to the presidency with Franklin Roosevelt died in office, against Thomas Dewey, the Republican governor of New York. Each of the three major polling organizations that covered the campaign predicted that Dewey would win:"

- Crossley poll: Dewey 50%, Truman 45%
- Gallup poll: Dewey 50%, Truman 44%
- Roper poll: Dewey 53%, Truman 38%

"Dewey's election was considered a foregone conclusion until the votes were actually counted: in one of the great upsets in American politics, Truman received slightly less than 50% of the popular vote and Dewey received slightly more than 45%. What happened?"

1. **What question was the study trying to answer?** How will people vote in the 1948 presidential election? This is a descriptive question, not cause-and-effect: we're not directly interested in why someone supports a particular candidate.
2. **What data did the study collect to answer this question?** Some kinds of polls. How were they collected? We'll need to look into this further.
3. **What analysis did they perform on the data?** They probably counted how many people in their sample said they'd vote for each candidate, and turned those into percentages. (They might have made other adjustments as well, but we'll ignore these.)
4. **Does this data analysis give a good answer to the question?** Ask President Dewey.

Let's now state this fundamental principle of descriptive studies:

- *To draw conclusions about a population from a sample, the sample must be representative of the population.* With a real population, the simplest way to ensure this is to let every individual in the population have the same chance of being chosen.

Let's write a question 2b:

2b. In a descriptive study, did all members of the population have the same chance to be chosen? If not, were these differences adjusted for?

The 1948 polls used a method called **quota sampling**:

"For example, in a national quota sample survey conducted by the National Opinion Research Center in December 1947, an interviewer in St. Louis was assigned to interview 15 people."

- 7 men and 8 women
- 3 men under 40, 4 women under 40
- 1 black man, 1 black woman
- 6 people from the suburbs etc.

The idea is that once you put all the interviewers' samples together, you get the right proportions of men and women, of young and old, of black and white, etc. Unfortunately for 1948's pollsters, discretion rears its ugly head again. If interviewers tend to ask better-dressed people (subject to their quotas) and Republicans, for example, are more likely to be better dressed, then Republicans will be overrepresented in the sample. This leads to **bias**: a systematic difference between samples and the population as a whole. (Later on, we'll define bias in a mathematical way.)

The solution is to take **random samples**. Ideally there's a big list of everyone in the population, and you can use a chance mechanism (e.g. the random number generator in R) to select your sample. In practice there's no big list but nearly everyone has a phone, so polling firms can call numbers at random. (Even then there are additional complications, e.g. some people refuse to answer surveys, which require adjustments; these are beyond the scope of this lecture.)

Summary

- To study cause-and-effect, the best choice is a **randomized controlled trial**, where the randomization ensures that all differences between the groups being compared are either due to the treatment (cause) or due to chance.
- To do a descriptive study of a population, and assuming it's not practical to do a census of the whole population, the best choice is to take a **random sample** from the population.
- All other things being equal, large samples will give more accurate information than small samples. (We'll see why this is in chapter 8.)
- We rely on randomness because statisticians are really good at dealing with randomness. It follows that as wannabe statisticians, we need to get really good at dealing with randomness. So we'll spend the next part of the course learning probability.

To read for next time: Trosset ch. 2.2, 3.1-3.3