Introduction to Statistics

Colm Mulcahy

Math 107-03, Spring 2020, Spelman College

27 Mar 2020

Overview

We'll briefly survey how data is collected, processed and reported.

We'll discuss "population parameters" versus "sample statistics".

We'll mention common issues with sampling.

We'll distinguish "observational studies" versus "experiments".

We'll highlight issues with surveys and opinion polls.

We'll provide advice about judging reliability of claimed statistics.

We'll overview common graphical representations such as bar charts, histograms and pie charts, and how they can be abused.

We will caution against allowing oneself to be bamboozled by mere numbers: if something sounds very surprising there needs to be hard evidence to back it up.

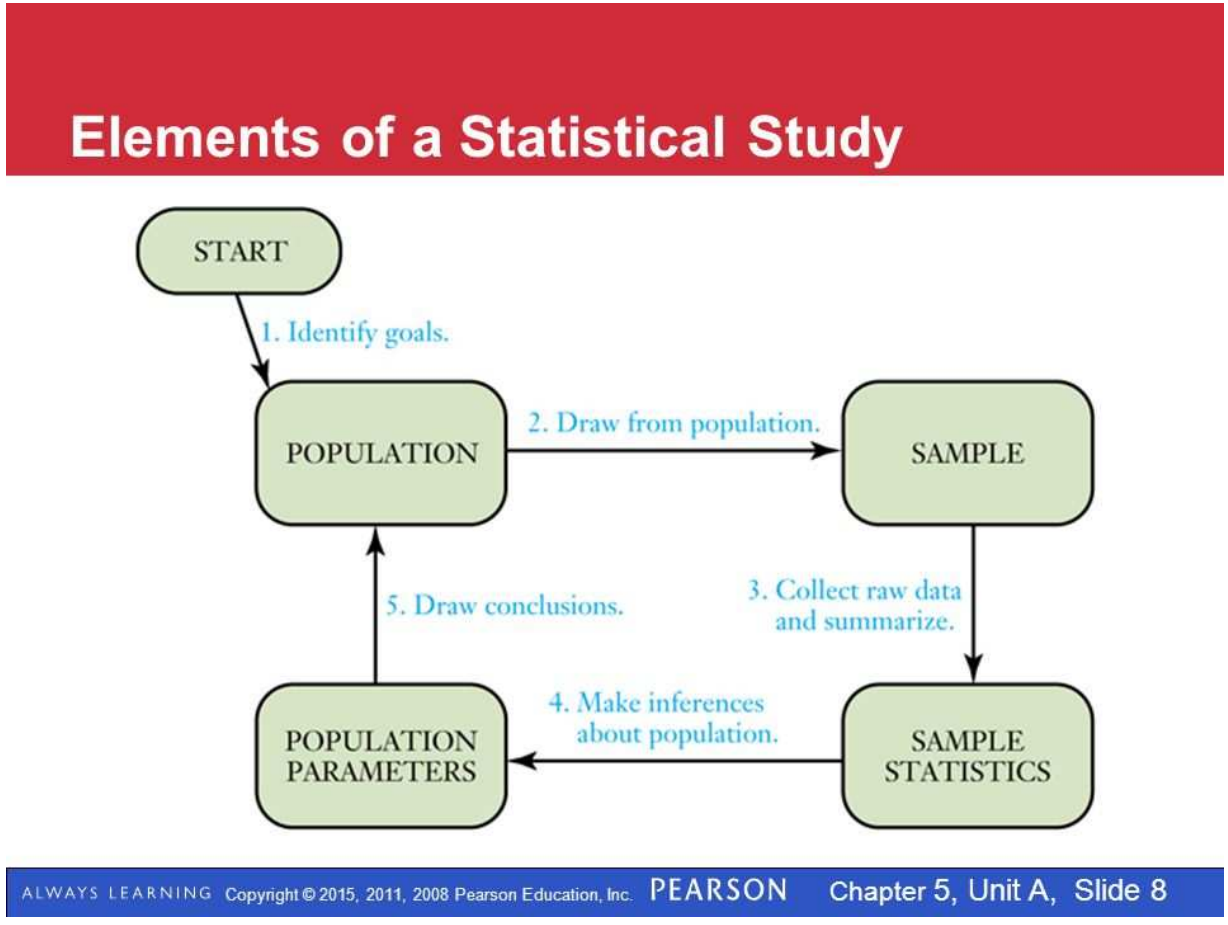
Statistics: how data is collected, processed and reported

Suppose we want to know the average height of all currently enrolled Spelman students: here the full student body is the "population" and the actual average height is the **population parameter**.

This "population parameter" could be determined exactly by measuring every student (data collection). We're assuming here that "average" refers to the mean: add up all the heights and divide by the total number of students. This does commit us to collecting over 2000 pieces of data!

If we want to repeat this for all students in the USA clearly we have a bigger problem. We need to "sample": maybe just collect the heights of *some* students nationwide and work out the average of those numbers; that's what's called a **sample statistic**.

We could have sampled in the Spelman case: just take heights of ten of your friends here and average those! Then claim this number is the average of all students' heights? (That's called **inference**).



What if your friends are all basketball players? What if they are all female? Or first year students? Or all have some Asian ancestry?

Statistics: how data is collected

Sampling techniques aim to avoid making choices that makes the data collected useless. We want "representative samples".

Now suppose we are studying the weights of Spelman students, and we have scales at the ready to weigh people we "sample".

We might sample 100 Spelman student, of whom 25 are first years, 25 are sophomores, 25 juniors and 25 seniors. But do you see a possible issue? (This relates to "stratefied sampling".)

We could sample 100 people as they leave the cafeteria. That's also a bad idea. (That would be called "convenience sampling".)

Suppose we emailed a survey to all students and used the first 100 responses received. (That would lead to "self-selection bias".)

As for offering $5 gift certificates to the first 100 people to participate, that too leads to biased results.

Observational Studies v Experiments, Treatment v Control

If we want to see if drinking milk impacts blood pressure, we can ask 50 people to drink milk regularly and check their blood pressure at 11am every day for a week, also checking the blood pressures of 50 people who don't drink milk.

That's an **observational study**: we're just checking blood pressure on two sample groups without seeking to change their behavior.

By contrast, if we want to see if taking aspirin impacts blood pressure, we might opt to ask a **treatment group** of 50 people to take aspirin regularly, and ask a corresponding **control group** of 50 people to avoid aspirin. That's an **experiment**.

However, if we want to see if taking heroin impacts blood pressure, we can hardly ask 50 people to take heroin! In that case, for moral and ethical reasons, we would have to settle for an observational study, acknowledging that data collection would be difficult.

Placebos and blinding issue

If we wanted to see if taking aspirin (or any medication) impacted how people responded to the statement "I have felt depressed in recent days" (say, making people chose on a scale from 0 for "not at all" to 5 for "very depressed") then there is another issue to consider.

Simply giving the medication to a treatment group and comparing the responses to those from an equal sized control group won't work well.

Why not? Human psychology!

The results will be far more reliable if the control group is given a **placebo**, such as fake aspirin, *and no participant knows whether they are in the treatment or the control group*.

This last condition is called a **single-blind experiment**.

Placebos and blinding issue

Now imagine that instead of using a questionaire, the effectiveness of the medication is judged by a team people who interview each person a month after "treatment" began.

If the people asking the questions and then making judgement calls (on how depressed the interviewees are) know which patient took the real medication and which one took the placebo, the results will naturally be tainted, again due to human psychology.

A **double-blind experiment** would be smarter: that's one in which neither the participants nor the people assessing the effectiveness know who is taking the real medication and who is only taking a placebo.