5.1: Studies of a Population Paramete

Suppose we are studying some population, and in particular a variable defined on that population. We are typically interested in finding out the following kind of characteristic of our population:

DEFINITION 5.1.1. A [**population**] **parameter** is a number which is computed by knowing the values of a variable for every individual in the population.

EXAMPLE 5.1.2. If X is a quantitative variable on some population, the population mean μ_X of X is a population parameter – to compute this mean, you need to add together the values of X for *all* of individuals in the population. Likewise, the population standard deviation σ_X of X is another parameter.

For example, we asserted in Example 4.3.28 that the heights of adult American men are N (69, 2.8). Both the 69 and 2.8 are population parameters here.

EXAMPLE 5.1.3. If, instead, X were a categorical variable on some population, then the relative frequency (also called the **population proportion**) of some value A of X – the fraction of the population that has that value – is another population parameter. After all, to compute this fraction, you have to look at every single individual in the population, all N of them, say, and see how many of them, say NA, make the X take the value A, then compute the relative frequency NA/N.

Sometimes one doesn't have to look at the specific individuals and compute that fraction nA/N to find a population proportion. For example, in Example 4.3.28, we found that 14.1988% of adult American men are taller than 6 feet, assuming, as stated above, that adult American men's heights are distributed like N (69, 2.8) – using, notice, those parameters μ_X and σ_X of the height distribution, for which the entire population must have been examined. What this means is that the relative frequency of the value *"yes"* for the categorical variable *"is this person taller than* 6 *feet?"* is .141988. This relative frequency is also a parameter of the same population of adult American males.

Parameters must be thought of as fixed numbers, out there in the world, which have a single, specific value. However, they are very hard for researchers to get their hands on, since to compute a parameter, the variable values for the entire population must be measured. So while the parameter is a single, fixed value, usually that value is *unknown.*

What can (and does change) is a value coming from a sample.

DEFINITION 5.1.4. A [**sample**] **statistic** is a number which is computed by knowing the values of a variable for the individuals from only a sample.

EXAMPLE 5.1.5. Clearly, if we have a population and quantitative variable X, then any time we choose a sample out of that population, we get a sample mean and sample standard deviation S_x, both of which are statistics.

Similarly, if we instead have a categorical variable Y on some population, we take a sample of size n out of the population and count how many individuals in the sample – say nA – have some value A for their value of Y , then the nA/n is a statistic (which is also called the **sample proportion** and frequently denoted $\hat{p}$).

Two different researchers will choose different samples and so will almost certainly have different values for the statistics they compute, even if they are using the same formula for their statistic and are looking at the same population. Likewise, one researcher taking repeated samples from the same population will probably get different values each time for the statistics they compute. So we should think of a statistic as an easy, accessible number, changing with each sample we take, that is merely an estimate of the thing we want, the parameter, which is one, fixed number out in the world, but hidden from out knowledge.

So while getting sample statistics is practical, we need to be careful that they are good estimates of the corresponding parameters. Here are some ways to get better estimates of this kind:

1. *Pick a larger sample.* This seems quite obvious, because the larger is the sample, the closer it is to being the whole population and so the better its approximating statistics will estimate the parameters of interest. But in fact, things are not really quite so simple. In many very practical situations, it would be completely infeasible to collect sample data on a sample which was anything more than a miniscule part of the population of interest. For example, a national news organization might want to survey the American population, but it would be entirely prohibitive to get more than a few thousand sample data values, out of a population of hundreds of millions – so, on the order of tenths of a percent.

   Fortunately, there is a general theorem which tells us that, in the long run, one particular statistic is a good estimator of one particular parameter:

FACT 5.1.6. **The Law of Large Numbers**: Let X be a quantitative variable on some population. Then as the sizes of samples (each made up of individuals chosen randomly and *independently* from the population) get bigger and bigger, the corresponding sample means x get closer and closer to the population mean μ_X.

2. *Pick a better statistic.* It makes sense to use the sample mean as a statistic to estimate the population mean and the sample proportion to estimate the population proportion. But it is less clear where the somewhat odd formula for the sample standard deviation came from – remember, it differs from the population standard deviation by having an $n - 1$ in the denominator instead of an n. The reason, whose proof is too technical to be included here, is that the formula we gave for S_X is a better estimator for σ_X than would have be the version which simply had the same n in the denominator.

   In a larger sense, "picking a better statistic" is about getting higher quality estimates from your sample. Certainly using a statistic with a clever formula is one way to do that. Another is to make sure that your data is of the highest quality possible. For example, if you are surveying people for their opinions, the way you ask a question can have enormous consequences in how your subjects answer: *"Do you support a woman's right to control her own body and her reproduction?"* and *"Do you want to protect the lives of unborn children?"* are two heavy-handed approaches to asking a question about abortion. Collectively, the impacts of how a question is asked are called **wording effects**, and are an important topic social scientists must understand well.

3. *Pick a **better** sample.* Sample quality is, in many ways, the most important and hardest issue in this kind of statistical study. What we want, of course, is a sample for which the statistic(s) we can compute give good approximations for the parameters in which we are interested. There is a name for this kind of sample, and one technique which is best able to create these good samples: *randomness*.

DEFINITION 5.1.7. A sample is said to be **representative** of its population if the values of its sample means and sample proportions for all variables relevant to the subject of the research project are good approximations of the corresponding population means and proportions.

It follows almost by definition that a representative sample is a good one to use in the process of, as we have described above, using a sample statistic as an estimate of a population parameter in which you are interested. The question is, of course, *how to get a representative sample*.

The answer is that it is extremely hard to build a procedure for choosing samples which guarantees representative samples, but there is a method – using randomness – which at least can reduce as much as possible one specific kind of problem samples might have.

DEFINITION 5.1.8. Any process in a statistical study which tends to produce results which are *systematically different* from the true values of the population parameters under investigation is called **biased**. Such a systematic deviation from correct values is called **bias**.

The key word in this definition is *systematically:* a process which has a lot of variation might be annoying to use, it might require the researcher to collect a huge amount of data to average together, for example, in order for the estimate to settle down on something near the true value – but it might nevertheless not be *biased*. A biased process might have less variation, might seem to get close to some particular value very quickly, with little data, but would never give the correct answer, because of the systematic deviation it contained.

The hard part of finding bias is to figure out what might be causing that systematic deviation in the results. When presented with a sampling method for which we wish to think about sources of possible bias, we have to get creative.

EXAMPLE 5.1.9. In a democracy, the opinion of citizens about how good a job their elected officials are doing seems like an interesting measure of the health of that democracy. At the time of this writing, approximately two months after the inauguration of the 45th president of the United States, the widely respected Gallup polling organization reports [**Gal17**] that 56% of the population approve of the job the president is doing and 40% disapprove. [Presumably, 4% were neutral or had no opinion.]

According to the site from which these numbers are taken,

> *"Gallup tracks daily the percentage of Americans who approve or dis- approve of the job Donald Trump is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults...."*

Presumably, Gallup used the sample proportion as an estimator computed with the re- sponses from their sample of 1500 adults. So it was a good statistic for the job, and the sample size is quite respectable, even if not a very large fraction of the entire adult American population, which is presumably the target population of this study. Gallup has the reputation for being a quite neutral and careful organization, so we can also hope that the way they worded their questions did not introduce any bias.

A source of bias that does perhaps cause some concern here is that phrase "telephone interviews." It is impossible to do telephone interviews with people who don't have tele- phones, so there is one part of the population they will miss completely. Presumably, also, Gallup knew that if they called during normal working days and hours, they would not get working people at home or even on cell phones. So perhaps they called also, or only, in the evenings and on weekends – but this approach would tend systematically to miss people who had to work very long and/or late hours.

So we might worry that a strategy of telephone interviews only would be biased against those who work the longest hours, and those people might tend to have similar political views. In the end, that would result in a systematic error in this sampling method.

Another potential source of bias is that even when a person is able to answer their phone, it is their choice to do so: there is little reward in taking the time to answer an opinion survey, and it is easy simply not to answer or to hang up. It is likely, then, that only those who have quite strong feelings, either positive or negative, or some other strong personal or emotional reason to take the time, will have provided complete responses to this telephone survey. This is potentially distorting, even if we cannot be sure that the effects are systematically in one direction or the other.

[Of course, Gallup pollsters have an enormous amount of experience and have presum- ably thought the above issues through completely and figure out how to work around it – but we have no particular reason to be completely confident in their results other than our faith in their reputation, without more details about what work-arounds they used. In science, doubt is always appropriate.]

One of the issues we just mentioned about the Gallup polling of presidential approval ratings has its own name:

DEFINITION 5.1.10. A sample selection method that involves any substantial choice of whether to participate or not suffers from what is called **voluntary sample bias**.

Voluntary sample bias is incredibly common, and yet is such a strong source of bias that it should be taken as a reason to disregard completely the supposed results of any study that it affects. Volunteers tend to have strong feelings that drive them to participate, which can have entirely unpredictable but systematic distorting influence on the data they provide. Web-based opinion surveys, numbers of *thumbs-up* or *-down* or of positive or negative comments on a social media post, percentages of people who call in to vote for or against some public statement, *etc.*, *etc.* – such widely used polling methods produce nonsensical results which will be instantly rejected by anyone with even a modest statistical knowledge. Don't fall for them!

We did promise above one technique which can robustly combat bias: randomness. Since bias is based on a *systematic* distortion of data, any method which completely breaks all systematic processes in, for example, sample selection, will avoid bias. The strongest such sampling method is as follows.

DEFINITION 5.1.11. A **simple random sample** [**SRS**] is a sample of size n, say, chosen from a population by a method which produces all samples of size n from that population with equal probability.

It is oddly difficult to tell if a particular sample is an SRS. Given just a sample, in fact, there is no way to tell – one must ask to see the procedure that had been followed to make that sample and then check to see if that procedure would produce any subset of the population, of the same size as the sample, with equal probability. Often, it is easier to see that a sampling method *does not* make SRSs, by finding some subsets of the population which have the correct size but which the sampling method *would never choose*, meaning that they have probability zero of being chosen. That would mean some subsets of the correct size would have zero probability and others would have a positive probability, meaning that not all subsets of that size would have the same probability of being chosen.

Note also that in an SRS it is not that every *individual* has the same probability of being chosen, it must be that every *group of individuals of the size of the desired sample* has the same probability of being chosen. These are not the same thing!

EXAMPLE 5.1.12. Suppose that on Noah's Ark, the animals decide they will form an advisory council consisting of an SRS of 100 animals, to help Noah and his family run a tight ship. So a chimpanzee (because it has good hands) puts many small pieces of paper in a basket, one for each type of animal on the Ark, with the animal's name written on the paper. Then the chimpanzee shakes the basket well and picks fifty names from the basket. Both members of the breeding pair of that named type of animal are then put on the advisory council. Is this an SRS from the entire population of animals on the Ark?

First of all, each animal name has a chance of 50/N, where N is the total number of types of animals on the Ark, of being chosen. Then both the male and female of that type of animal are put on the council. In other words, every individual animal has the same probability – 50/N – of being on the council. And yet there are certainly collections of 100 animals from the Ark which do not consist of 50 breeding pairs: for example, take 50 female birds and 50 female mammals; that collection of 100 animals has no breeding pairs at all.

Therefore this is a selection method which picks each individual for the sample with equal probability, but *not* each collection of 100 animals with the same probability. So it is not an SRS.

With a computer, it is fairly quick and easy to generate an SRS:

FACT 5.1.13. Suppose we have a population of size N out of which we want to pick an SRS of size n, where n < N . Here is one way to do so: assign every individual in the popu- lation a unique ID number, with say d digits (maybe student IDs, Social Security numbers, new numbers from 1 to N chosen in any way you like – randomness not needed here, there is plenty of randomness in the next step). Have a computer generate completely random d-digit number, one after the other. Each time, pick the individual from the population with that ID number as a new member of the sample. If the next random number generated by the computer is a repeat of one seen before, or if it is a d-digit number that doesn't happen to be any individual's ID number, then simply skip to the next random number from the computer. Keep going until you have n individuals in your sample.

The sample created in this way will be an SRS.
