5 SAMPLING DISTRIBUTIONS

5.02 Sampling

Inferential statistics refers to methods used to draw conclusions about a population based on data coming from a sample. You can imagine that in order to understand methods of inferential statistics, it is of essential importance that you know how you should draw samples. In this video I will tell you how you can do that. I will pay attention to good sampling methods as well as some poor practices. I will also deal with different forms of bias you could encounter along the way.

A sample is nothing more than a subset of a population. Yet, for methods of inferential statistics, not every sample is appropriate. What you want is a representative sample. What you want, in other words, is that your sample is a micro-version of the entire population. A good way to achieve that goal is to draw a **simple random sample**. That means that you make sure that each subject in a population has the same chance of being selected.

Imagine you want to know to what extent students in London identify themselves as hipsters. You decide to draw a sample of 200 respondents. The mean 'Hipsterness' score, which could range from 0 to 10, is 3.12. The population here consists of all students in London. The parameter you're interested in is the population mean μ. The sample consists of the 200 selected students. The statistic you're going to use to draw conclusions about the population mean μ is the sample statistic $\bar{x}$. In order to be able to draw conclusions about the population, we want the sample to be a simple random sample. That means that every student in London must have the same chance of being selected. How can you make sure that's the case? Well, first you need to make clear what the population looks like. We already know that: all London students.

The second step is to compile a list of all subjects. This is what we call the **sampling frame**. Imagine that there is an organization in London that has an overview of all students, including their contact details. Moreover, this organization is willing to share this list with you. You ask a computer to randomly select 200 students out of this list. That's it. There you have your simple random sample.

The next step is to decide how you're going to approach your 200 respondents. In a personal or **face-to-face interview** you're in the same room as your respondent, and you ask your questions face-to-face. An advantage is that respondents are likely to participate, a disadvantage is that this way of collecting data is fairly expensive. Another option is the **telephone interview**. This is less expensive. But generally respondents are less patient on the phone, so the interview has to be short. You can also ask respondents to fill out the questionnaire themselves. This is what we call a **self-administered questionnaire**. Because respondents could complete the survey online, this is a much cheaper option. A disadvantage is that it is more common that people don't participate.

Along the way you will encounter various possible forms of **bias**. The first one is **undercoverage**. This means that not everyone in the population is included in the sampling frame. In our London students example this is what happens if the list of students is incomplete. Some students will then have zero chance of being included in the sample. There can also be **sampling bias**. This means that not every person in your sampling frame is equally likely to be included in the sample. This is what happens if you fail to draw a random sample. This is the case if, for instance, you randomly approach people on the street. This is what we call a **convenience sample**. It is not random, because some people will be less likely to be found on the street than others. These people will have a smaller chance to be included in the sample. Thirdly, once you have your sample, another possible form of bias is **nonresponse bias**. Some selected subjects might refuse to participate, or they might simply be unreachable. Some respondents who have agreed to participate might not be willing to respond to particular questions. The problem is that those who don't participate might be different from those in the overall sample.

Whether it is undercoverage, sampling bias or nonresponse bias, if the group that has a zero or different chance to be sampled, or if the group that refuses to respond differs from the rest of the population in terms of Hipsterness, we might over- or underestimate Hipsterness in the population. Our estimation will be biased due to systematic under- or overrepresentation of certain groups.

Finally, there can be **response bias**. In this case the actual given responses are biased. This could for instance be due to an interviewer asking leading questions or because respondents think that some answers are socially unacceptable. For instance, a student might identify as a Hipster but tell an interviewer that she doesn't because she thinks that the interviewer doesn't like Hipsters. In this case our estimation can become biased due to systematic misrepresentations of certain responses.

So, when drawing a sample, you should make sure that your sample is a simple random sample and that you keep these forms of bias to a minimum. However, many times it will be almost impossible to draw a simple random sample. Luckily, there are two other ways of random sampling that are almost as good. Let me first briefly repeat how simple random sampling works. If your population consists of all London students, and you want to draw a sample of 200 students, you write the names of all students on a piece of paper and you throw all these pieces of paper in a huge bin. Then you randomly select 200 pieces of paper from this bin. That's simple random sampling.

The first alternative is a **random multi-stage cluster sample**. It works as follows. First you identify a large number of clusters within your population; for instance, the various educational programs in London in which the students are enrolled. Every program is represented by a bucket and you put the pieces of paper with the names of students in the buckets of the programs in which they are enrolled. Next, you randomly pick a number of buckets. Say… 10. Then, you select all pieces of paper within these buckets. *That's* your sample. A random multi-stage cluster sample is a good choice if you don't have a good sampling frame or if drawing a simple random sample would be very expensive.

The second alternative is a **stratified random sample**. Now you divide the population in separate groups which you call "strata". For instance, the various universities in London. Every university is represented by a box and you put the pieces of papers with the names of the students in the box of the university where they are registered. Next, you select a simple random sample of pieces of paper from *each* box. Al these pieces of paper together form your sample. An advantage of this method is that you can make sure that you have enough subjects from every stratum in your sample. A disadvantage is that you need a sampling frame and that you need to know to which stratum each respondent belongs.

One final thing you need to know about samples: bigger is better. However, there's one important caveat: a bigger sample can *never* make up for a bad sampling procedure. If your sample is not random, you can increase your sample size as much as you want, your sample will never be good. However, *provided that* your sample is random, a bigger sample is technically always better than a smaller sample. However, once you pass a certain point, an increase in your sample only results in a very small increase in the precision of your estimation of the population parameter.