Slide 1 – Title Slide

Hello and welcome to Week 5, Part 3 of EGM101: Collecting Data. In this lesson, we'll talk about different sampling techniques, and how those can have an impact on the data that we gather.

Slide 2 – Need to Sample

In the first lesson this week, we introduced the terms population and sample, and discussed how we normally have to study a sample instead of an entire population. The reason for this is because it is often not feasible or even possible to study every member of a population, either because of limited resources such as money or time, or in some cases, the population may not even be finite.

When we study a sample, we use information about the sample to infer something about the population – a process known as inferential statistics.

So, in our salmon example, let's say we wanted to estimate the size of salmon that spawn in a particular river. Instead of measuring every single salmon, we would instead catch a number of salmon, measure the size of each of the fish in our sample, and use that information to determine the size of salmon in the whole population.

Slide 3 – Sampling Bias

To ensure that the information that we infer is not completely wrong, though, we want to make sure that the fish in our sample have similar characteristics as the population – that is, we want to make sure we have a representative sample.

So, let's say that instead of choosing these four fish as our sample, we instead chose four different fish. In the first example, we had fish that were kind of in the middle of the size range – they weren't all the biggest fish, they weren't all the smallest fish.

You can see how in this second example, we've chosen the four smallest fish. If we then measured these four fish, it would give us the wrong impression of the size of the population – what we have done is chosen a biased sample. That is, the smallest fish in the population are more likely to be included in our sample than the biggest fish.

To help make sure that we avoid sample bias and choose a representative sample, we need to consider how we choose our sample. How we sample has an impact on the sample, as well as what we are able to infer about the population.

In an ideal world, we would choose a truly random sample – that is, a sample where all members of the population have exactly the same chance of being part of the sample. In the rest of this lesson, we'll discuss different sampling strategies that can help make sure that we get as representative a sample as possible.

Slide 4 – Sampling Error

Remember from our definitions slide that a “statistic” describes a characteristic of a sample, while a “parameter” describes a characteristic of the population. The goal with calculating a sample statistic, then, is to approximate the population parameter.

The difference between the statistic and the parameter it approximates is what is known as the sampling error. Our goal with sampling is to reduce the sampling error as much as possible.

One way that we can minimize, or at least decrease, the sampling error is by taking larger random samples from the population – this is something that we’ll discuss in more detail when we cover probability during Week 7.

When we take random samples of the population, we reduce the sampling bias, as we will discuss over the rest of this lesson. We can also use the statistics that we derive from our random samples to estimate the sampling error, which is, again, something that we will come back to later on in the module.

Slide 5 – Simple Random Sampling

The most basic of these techniques is known as simple random sampling, which you might also hear referred to as the “lottery method.”

The steps of this method are as follows: first, we have to choose the target population, as well as the sample size. Then, we assign each member of the population a number, and then select numbers at random. The members of the population corresponding to the randomly-selected numbers make up our sample.

One of the main advantages of this method is that it helps minimize selection bias, as each member has an equally likely chance of being selected.

Unfortunately, this can also be time-consuming to implement, and it requires that we have access to all subjects/respondents/members of the population. One thing to keep in mind here is that while this method does help minimize selection bias, it doesn’t necessarily eliminate it – we might still end up with a non-representative sample.

Slide 6 – Systematic Random Sampling

The next method we’ll consider is systematic random sampling, which you might also see referred to as the “constant skip” method. Part of this method works the same as simple random sampling did: first, we need to choose the target population and a sample size, then we assign each member of the population a number. But, instead of randomly choosing every member of the sample as we did for simple random sampling, we only select the starting point at random, and the sample consists of every nth member after the starting point. So, using our salmon population as an example, we have randomly selected “1” as our starting point, and we’re taking every 4 members after that starting point to make up our sample.

Just as a side note here, there's nothing particularly special about the spacing between samples here – I chose "4" because of the constraint of fitting the illustration on the slide. You could just as easily choose every 5, or 10, or any other number you like here.

Some of the advantages to systematic random sampling are that it's very simple to do, although this depends on your application. If you're choosing house numbers for a door-to-door survey, it's not too difficult to do this; if you're trying to catch swimming upstream, it might be a little more difficult. Another benefit is that you only have to choose a single random number for your starting point.

One drawback to this approach is that your sampling frame, which is to say the way that you are selecting your samples, might have some periodicity. That is, you might accidentally build in some bias to your sample using this approach. As an example, let's say that we're doing a survey of temperature in a block of flats, and we choose to sample every 10^{th} flat. The way that many buildings are laid out, all of those flats might be directly on top of each other, which means that we wouldn't be getting a good spatial sample with this method. This is one of the reasons why we want to think carefully about what it is that we're sampling when we design our sampling strategy – again, we want to try to minimize these sort of biases as much as possible.

Slide 7 – Stratified Random Sampling

Sometimes, we have different characteristics that we want to make sure are represented in the sample in a similar way that they are present in the population – for example: age, spatial location, socioeconomic status, and so on. In this case, we want to take a stratified random sample. To do this, we first divide the population up into groups, called strata, based on similar characteristics. We want to be absolutely sure that all of our members belong to a group – that is, the groups must be "exhaustive". We also want to make sure that each member only belongs to a single group – that is, the groups must be "mutually exclusive."

Here, we have two groups of salmon – maybe these are different size or age ranges, or salmon that primarily eat different kinds of food. Next, we choose a simple random sample from each group – so, we have a random sample for this group, and one for this group. We then put these random samples together to form the sample that we analyze.

Stratified random sampling works well when there are obvious strata – that is, when we have obvious groups that we want to make sure are correctly represented in the sample. Most political or opinion polls, for example, use this technique to ensure that they get an appropriate representation of different age groups in the poll. Because we make sure that each group, or strata, is included, it means that we have a better chance of having a representative sample.

Unfortunately, not all datasets can be stratified efficiently, so we might not be able to divide our population into groups. Some strata might be extremely small, with only a few members – in that case, we would want to merge small strata into larger groups if possible.

Slide 8 – Cluster Sampling

The last technique we'll cover in this lesson is cluster sampling. Cluster sampling is similar to stratified random sampling in that we first divide the population into smaller groups, or clusters. Just like with stratified random sampling, we need to make sure that the clusters are both exhaustive and mutually exclusive – that is, each member of the population should belong to exactly one group.

Then, we choose a random sample of the clusters. We then measure/observe each member that belongs to one of the selected clusters, or we take a random sample from each cluster, to make up the sample for our study.

One of the main benefits of cluster sampling is that it's extremely efficient, especially if our study population is spread out over a large geographic area. we don't need a list of the entire population – we just need a list of the groups. This is especially useful for sampling that's done based on location – things like geographic area, streets/neighborhoods, and so on. For a survey of voters, for example, we only need to select a street, then survey everyone that lives on the street.

Unfortunately, the sample that we obtain may not be very representative of the population, which means that the sampling error might be quite a bit bigger when we do cluster sampling, compared to the other methods we've covered. It's important to think carefully about how you define the clusters to help minimize this problem. We also only end up getting information about the clusters that we've sampled, with no information at all about the unsampled clusters. Because of this, another strategy that is often used is to have many small clusters instead of a few large clusters.

Slide 9 – Summary

In this lesson, we've discussed how studying entire populations is often infeasible or impossible, which means that we have to sample the population, rather than study every single member of the population.

In order to do this effectively, however, we need to avoid sampling bias, or at least minimize it as much as possible.

The way that we can do this is by using one of the various forms of random sampling that we've discussed, such as simple random sampling, systematic random sampling, stratified random sampling, or cluster sampling. As we've covered, each of these methods has their own advantages or disadvantages, which means that which one we choose is going to depend mostly on the type of study that we're doing – there isn't necessarily a "one size fits all" approach that covers all applications.

Slide 10 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapter 1.2; Caswell, Chapter 2; and Weiss, Chapters 1.2 to 1.4.

For more about Sampling or Selection Bias, you can also read Chapter 6 of Bergstrom and West, which talks about many more examples of Selection Bias and how it can affect research. Finally, I've included a link to a video from Khan Academy, which goes into more detail about random sampling techniques.

That’s all for this lesson – I hope you found it interesting, and you have any questions, please don’t hesitate to e-mail me or post in the discussion forum on blackboard. Bye!