5 SAMPLING DISTRIBUTIONS

5.03 Sampling distribution

Researchers often use a sample to draw inferences about the population that sample is from. To do that, they make use of a probability distribution that is very important in the world of statistics: the **sampling distribution**. In this video I will explain what the sampling distribution is. Pay close attention, because the sampling distribution is the link that helps researchers to draw conclusions about a population on the basis of only one sample.

One note before we start. In this video we'll pretend that we know what a population looks like. In the actual research practice you'll never know that, but this step is necessary to understand inferential statistics later on.

Okay, here we go. Let's assume that a group of Scandinavian Hipsters organizes the Festival of the Beards on one of the small islands near Oslo, the capital of Norway. The audience of the festival is, as you might have expected, men with beards. The organization has sold about 5000 tickets and offers free transportation to the festival-island. All people with a ticket gather in the harbor of Oslo. The organization distributes all of them randomly over boats that carry the passengers to the island. Every boat carries 30 festival-goers.

Now, one of the boats gets lost in the maze of tiny Norwegian Islands. To make matters even worse, the mobile phone network has crashed, so it is impossible for the organization to reach the captain of the boat, or for the passengers to reach the organization. So, the Hipster-organization decides to send some employees to find the lost boat. *You* are one of those employees. After half an hour you see a crashed boat with about 30 people. Yes! You think, finally I've found them! You're about to send a message (by walkie-talkie) to the organization that the lost boat has been found. Then, however, you take a look at the people in the boat one more time. You see that they are comprised of families with young children. This is strange because you'd expect that a boat with randomly selected people who are going to the Festival of the Beards would be…. well…. men with beards…. Not young families. You decide that it is very unlikely that this is the boat you're looking for and decide to continue your search.

Later on, it turned out that this was a good decision. The boat you encountered was a boat carrying people to a Family Park on one of the other islands.

Why did I tell you this story? Well, if you understand this line of reasoning, you understand the basic idea behind the sampling distribution. The reasoning goes like this: if you draw a simple random sample from a population, it is very unlikely that this sample will strongly differ from the population from which it is drawn. In our case, the people going to the Festival of the Beards form our population. A boat with 30 randomly selected people from this population forms a simple random sample.

In fact, all the boats travelling between the harbor of Oslo and the festival-island could be seen as simple random samples. Of course, every boat is different from the next, but a large majority of them will contain a large proportion of bearded men. It is very unlikely that a boat will contain various young families. Of course, it's possible, there will be *some* families attending the Festival of the Beards, but finding a *randomly* composed boat *completely* consisting of young families is highly improbable.

Suppose you decide to measure the mean beard-length for every boat. In every boat there are 30 people going to the Festival of the Beards. Now, imagine you know that the mean beard-length in the population of 5000 festivalgoers is 10.3 millimeters. So μ is 10.3 millimeters. You also know that beard-length has a bell-shaped distribution in the population. In one boat you'll find a mean beard-length of, say, 9.4 millimeters, and in another one, say, 10.8 millimeters. However, it is very unlikely that you will encounter a boat with a mean beard-length as low as 3.4 millimeters. Or a boat with a mean beard-length as high as 19.2 millimeters. Because these boat-means could be seen as sample means we symbolize them by $\bar{x}$'s.

Now imagine you look at 3 boats. This is what the probability distribution could look like. In one boat the mean beard-length is 9.9 millimeters, in one 10.7, and in one 10.2. We have 3 boats, so every mean has a probability of 0.33. Now, imagine you look at 17 boats. That would look something like this. Now, you look at 40 boats. And now at 100 boats. You can see that the distribution of these beard-length means increasingly looks like a bell-shaped distribution. Moreover, you can see that the mean of the distribution is about 10.3 millimeters, exactly the population mean. This is not strange if you think about it. You would expect that in most cases the mean in a boat is close to the population mean. In one boat the mean will be a little higher, and in another a little lower. However, if you look at many boats you would expect that the mean of all these different boat-means would be the population mean.

Now imagine your population consists of all Norwegian men. You know that the mean beard-length in this population is 1.22 millimeter and that the variable has a bell-shaped distribution in the population. If you draw a simple random sample of 30 respondents you will find a mean close to this value. Say 1.34 millimeter. If you draw another random sample, the mean will be close to the population value of 1.22 as well. Say it is 1.19 millimeter. If you do that 5 times you will get 5 different values, which are all close to the population value. If we draw 20 samples the distribution of these samples will be approximately bell-shaped. If we do it 100 times it will be even more strongly bell-shaped. If we draw an *infinite* number of samples the distribution will be perfectly bell-shaped and the mean will be exactly 1.22 millimeter, which is the population value.

We call this distribution the sampling distribution of the sample mean. It is the distribution that you get if you draw an infinite number of samples from your population and compute the mean of all the collected sample means. For now you need to realize that in the actual research practice you will never collect an infinite number of samples from a certain population. However, it is important that you understand that if you would do that, the mean of that distribution would be equal to the mean in the population, and that we would call this distribution the sam*pling* distribution of the sample mean. Don't confuse this with the sam*ple* or data distribution, which is just the distribution of scores in the one sample that was actually drawn and for which data were actually collected.