5 SAMPLING DISTRIBUTIONS

5.04 The central limit theorem

If you draw an infinite number of samples from a bell-shaped population distribution, the distribution of means from this infinite number of samples will be bell-shaped, and the mean of this distribution of sample means will be exactly the same as the population mean. We call this distribution the sampling distribution of the sample mean. In this video I will discuss the sampling distribution of the sample mean and the **central limit theorem** – one of the most important theorems when it comes to inferential statistics.

The central limit theorem says that, provided that the sample size is sufficiently large, the sampling distribution of sample mean X-bar has an approximately normal distribution. *Even if the variable of interest is not normally distributed in the population!* Isn't that cool? No matter how a variable is distributed in the population, the sampling distribution of the sample mean is *always* approximately normal, as long as the sample size is large enough. As a guideline for 'large enough' a sample size of 30 or larger is often used.

Take a look at *these* possible shapes of population distributions. *This* is what the sampling distributions of the sample mean would look like if you drew samples of n equals 30. Remember, this means that you draw an infinite number of simple random samples of 30 respondents from the population and display all the resulting sample means in a distribution. You should realize that in practice it is impossible to draw an infinite number of samples. But then, the good news is that drawing multiple samples is not required at all to determine the shape of the sampling distribution. Because if it is normal, you can describe its shape with just two parameters: mean and standard deviation. So, it is sufficient to estimate these two parameters.

As I already told you earlier, the mean of the sampling distribution is equal to the mean of the population distribution. We can display that as follows: Mu-X-bar is equal to Mu. Mu stands for the population mean, and Mu-X-bar stands for the mean of the sampling distribution of all the sample mean. So, imagine you're interested in the average beard-length of Norwegian men. The population consists of all Norwegian men and Mu is the mean beard-length in the population. Let's assume that it's 1.22 millimeters. The mean of the sampling distribution of the sample mean, which is the mean of the distribution that we would get if we drew an infinite number of samples from the population, and wrote down the mean beard-length in each sample, is equal to the population mean 1.22. The X-bar is added to emphasize that the scores in the sampling distribution are sample means and not individual scores. In other words, the mean of the population distribution is the mean of the individual scores of *all* Norwegian men. The mean of the sampling distribution is the mean of the *sample means* of an infinite number of samples drawn from this population.

If we know what the distribution of the population looks like, we can also easily compute the standard deviation of the sampling distribution. The standard deviation of the sampling distribution is symbolized by Sigma-X-bar and is equal to Sigma divided by the square root of n.
The X-bar is added to make clear that we are talking about the standard deviation of the sampling distribution in which the scores are sample means, or in other words, X-bars. Sigma stands for the standard deviation in the population. And n stands for the sample size.

This formula shows that the standard deviation of the sampling distribution is affected by two characteristics. First, it is affected by the standard deviation in the population. Assume your n is 30 and your population standard deviation is 1, the standard deviation of your sampling distribution then is 1 divided by the square root of 30. That equals 0.18. If the standard deviation in your population increases to 2, the standard deviation of your sampling distribution becomes 2 divided by the square root of 30 makes 0.37. If Sigma becomes 3, Sigma-X-bar becomes 0.55 etcetera. So, if the standard deviation of the population distribution increases, the standard deviation of the sampling distribution increases as well. In other words: the larger the variability in the population, the larger the variability of the sample means. This makes sense intuitively, right? If you draw various samples of 30 subjects from a population in which men strongly differ from each other regarding the length of their beards, you can expect that the means of these samples differ more strongly from each other than if you draw various samples from a population in which men hardly differ from each other.

The standard deviation of the sampling distribution is also affected by the sample size. Look at the formula again. Assume that the population standard deviation equals 2. Now, if you have an n of 30, your Sigma-X-bar equals 2 divided by the square root of 30. That's 0.37. If you have an n of 100, the standard deviation of your sampling distribution becomes 2 divided by the square root of 100. That equals 0.2. If you have an n of 500, your Sigma X-bar becomes 0.09. This indicates that a larger sample size leads to a lower standard deviation of the sampling distribution.

This also makes sense intuitively. If the mean beard-length in the population of Norwegian men is 1.22 millimeters and you draw a sample of only 2 respondents, it wouldn't be strange if you find a mean that is much higher. If one of your two respondents has a long beard your mean will be pretty high. So, a mean of 5 or even 10 millimeter will not be exceptional if you have a sample of only 2 respondents. If you draw 5 samples, your sample means could look something like *this*. Now imagine that you draw a sample of 1000 subjects. It will be very unlikely that the mean of this sample will be 5 or 10 millimeter. After all, the long-bearded men will be counterbalanced by men with no facial hair at all. If you draw 5 samples, the sample means could look something like *this*. They will all be very close to the population mean of 1.22 millimeter. So, the larger the size of your sample, the closer the sample means will lie to the population mean, and the smaller the standard deviation of your sampling distribution will be.

To conclude: you now know that the central limit theorem tells you that no matter how a variable is distributed in the population, the sampling distribution of the sample mean is approximately normal as long as the sample size is at least 30. The mean of the sampling distribution, Mu-X-bar equals the population mean Mu and the standard deviation of the sampling distribution Sigma-X-bar equals the standard deviation of the population distribution (Sigma) divided by the square root of the sample size (n).