6 CONFIDENCE INTERVALS

6.03 CI for mean with unknown population standard deviation

A 95 percent confidence interval to estimate a population mean tells us that we have 95 percent confidence that this interval contains the actual population mean. With *this* formula you can compute the endpoints of the interval. There is one big problem with this formula, however. To compute the confidence interval, you need to know the population standard deviation, and usually, we don't know this value. After all, we use the sample to draw inferences about our population parameters. In this video I show you how you can solve this problem. The solution is that we *estimate* the population standard deviation, and therefore have to employ another distribution than the standard normal distribution: the **t-distribution**. Let me tell you how that works.

Imagine we've asked a sample of 60 new parents in Amsterdam how much sleeping hours they have lost after their first children were born. The mean is 2.6 hours and the standard deviation is 0.9 hours. To construct a 95% confidence interval we need this formula: X-bar plus and minus 1.96 times the standard deviation of the sampling distribution of the sample mean, which equals the population standard deviation divided by the square root of the sample size. We can also write that as: X-bar plus and minus the z-score for the 95% confidence level (which is 1.96) times the standard deviation of the sampling distribution. As you can see, it is impossible to compute the confidence interval, because we don't know the value of the population standard deviation. Therefore, we can't compute the standard deviation of the sampling distribution of the sample mean… To solve that problem we estimate the population standard deviation with the sample standard deviation.

This leads to the following formula: X-bar plus and minus the z-score for the 95% confidence level times the *estimated* standard deviation of the sampling distribution of the sample mean, which equals the *sample* standard deviation divided by the square root of the sample size. We call this estimated standard deviation of the sampling distribution the **standard error**. But because we now estimate the standard deviation we add extra error in our computation. For that reason we employ another distribution than the standard normal distribution (also called the z-distribution) we employed previously. Because of the extra error we now use the T-distribution. That leads to this formula: X-bar plus and minus the **t-score** for the 95% confidence level times the estimated standard deviation of the sampling distribution of the sample mean.

Let me now tell you a little more about t-distributions and t-scores. The t-distribution strongly resembles the standard normal distribution. It is bell-shaped, symmetric and has a mean of zero. Still, however, it is slightly different. Because we now estimate the standard deviation of the sampling distribution we introduce extra error. That error can be substantial when we have a small sample. The t-distribution takes into account that extra error for small samples. The T-distribution therefore has slightly thicker tails than the normal distribution and a larger standard deviation. You can see that here. The black distribution is the standard normal distribution. The blue one is the t-distribution.

The exact shape of the t-distribution depends on the size of the sample. The larger the sample, the more the t-distribution looks like the normal distribution. More precisely the shape of the t-distribution is dependent on a single parameter, the so-called **degrees of freedom**, symbolized by *df*. The degrees of freedom parameter in the t-distribution equals the sample size n minus one. This means that we actually have many different t-distributions: one separate distribution for every df. The blue t-distribution you see here has two degrees of freedom. A t-distribution with 5 degrees of freedom looks like *this*, and a t-distribution with 30 degrees of freedom looks like *this*. This means that when we have 30 or more degrees of freedom, the t-distribution is almost identical to the

standard normal distribution. More precisely, the standard normal distribution is the t-distribution with df equals infinity.

Just like with the standard normal distribution and z-scores, we can find cumulative probabilities pertaining to particular t-scores. An important difference with the standard normal distribution, however, is that these probabilities depend on the degrees of freedom. When you're computing a 95 percent confidence interval you can find the t-scores pertaining to the 95 percent confidence level for all the different possible degrees of freedom in a so-called t-table. This table is similar to the z-table.

Let me show you how that works by means of our "new parents study". The mean lost sleeping hours in our sample was 2.6 and the standard deviation was 0.9. The sample size was 60. *This* is the formula to compute the endpoints of the 95% confidence interval. Let's start with computing the standard error. It equals the sample standard deviation *s* divided by the square root of *n*. That is 0.9 divided by the square root of 60. That makes 0.116. So, our standard error, or, in other words, the estimated standard deviation of the sampling distribution of the sample mean equals 0.116. To compute the margin of error we have to multiply this value with the t-score for the 95% confidence level. As you know, the t-score depends on the degrees of freedom. Df equals n minus one. We have a sample of 60, so 60 minus one equals 59. In our T-table we look in the column of the 95% confidence level and in the row of 59 degrees of freedom. Because the table does not report 59 degrees of freedom, we settle for the closest *lower* value. That's 50 degrees of freedom. The relevant t-score is 2.009. So, we multiply 0.116 with 2.009. That's about 0.23. So we subtract this value from our sample mean 2.6 and also add it to the mean. That gives us the interval endpoints 2.37 and 2.83. The 95% confidence interval goes from 2.37 to 2.83. We have 95% confidence that this interval contains the actual population mean.

To compute a confidence interval for a population mean, two assumptions need to be satisfied. First, your data should be obtained by randomization. In other words, the sample should be a random sample. Otherwise your findings will not be valid. Second, your population should be approximately normally distributed. This might seem to be problematic, because many variables are not normally distributed in the population. However, things are not as bad as they seem. Using the t-distribution to construct a confidence interval for a mean is robust against violations of this second assumption. That a statistical method is robust means that it performs well even if the assumption is violated. Finally, you should also be wary of extreme outliers when constructing a confidence interval based on the T-distribution. When your data have extreme outliers, the method doesn't work well. So always check your data for outliers before you start!