1.03 Tests for two groups: Confidence intervals and two-sided tests

In this video we'll discuss the relation between null hypothesis testing and confidence intervals. Test statistic values and p-values are very useful to make a decision about the null hypothesis - to determine statistical significance. Unfortunately, even very small effects will become statistically significant if the sample size is very large. In most cases - at least in the social and behavioral sciences - such small effects are of little importance. The practical significance is very hard to tell just by looking at the test statistic and p-value.

Another way to draw statistical inferences is to use interval estimates that provide an indication of a range of plausible values. A confidence interval with a 95% confidence level answers the following question: If I were to repeatedly draw a sample and calculate a statistic, then what is the range of values around this statistic if I want the true population parameter to lie within this interval in 95% of my samples?

So if I take a hundred samples, calculate a sample statistic and a 95% confidence interval around this statistic each time, then I expect 95 of these intervals to contain the population parameter value. Note that this is not the same as saying that there's a 95% chance that the interval based on our sample contains the population value! In practice we have only one sample; it either contains the true population value or it doesn't, we don't know. But since it's likely that it does, we consider the values in the interval to be plausible.

Practical significance is more easily interpreted using a confidence interval: By considering the width of the interval and by seeing whether the parameter value under the null hypothesis lies inside the interval, or very close to it. The narrower, the more precise our interval estimate is. The closer to the null value, the less plausible it is that the effect has real practical significance.

Two-sided tests and confidence intervals are closely related. We'll look at how a confidence interval is constructed and then note the parallels with two-sided test. I'll use a capital T to denote test statistic values in general; depending on whether we're considering means or proportions this could be a t-value or a z-value.

Confidence intervals are based on the sampling distribution. We set the confidence level, say at 95%, and determine the margin left and right of the sample statistic. The boundaries are given by the *sample statistic* - $T_{alpha/2}$ * *se* and the *sample statistic* + $T_{alpha/2}$ * *se*. These boundaries

represent the number of sample statistic units above and below the observed sample statistic value, associated with an area of 95% under the sampling distribution.
With repeated sampling, 95% of the confidence intervals will contain the true population parameter. The true value will lie beyond these boundaries only in 5% of the samples. Now suppose for a minute that the null hypothesis is true and the population value is the null value. Then this value will lie outside the interval only in 5% of the samples.
If - in our sample - we find the value under the null *outside* the interval, then it's unlikely that the null value is the true population parameter value. So finding the null value outside the confidence interval corresponds to rejecting the null in a two-sided test.

In both cases a decision is based on whether boundary values are exceeded. The boundaries are determined by the significance level or the confidence level, which form each other's complement. With a two-sided test we hope to find a test statistic value outside the boundaries - in the critical region. With a confidence interval we hope to find the value under the null hypothesis outside the boundaries.
With a test, the interval is centered around the expected test statistic value under the null; with a confidence interval, it's centered around the sample statistic value. The margins around these centers are the same for two-sided tests and confidence intervals; they're just expressed in different units. For tests the margin equals the critical test statistic value $T_{alpha/2}$ and is expressed in standard errors. For confidence intervals the margin equals the same critical test statistic value $T_{alpha/2}$, but *times the standard error*. Because we multiply by the standard error the scale is returned to the original units of the sample statistic.