1.04 Tests for two groups: Power

In this video we'll discuss the power of statistical tests and how to determine and increase power. Power refers to the probability of correctly rejecting the null hypothesis - the probability to detect a hypothesized effect if it truly exists in the population.

Before we discuss how power can be assessed and influenced, let's recall the type of correct decisions and errors we can make. We decide either not to reject or to reject the null hypothesis. Whether our decision is correct or incorrect depends on whether the null hypothesis is in fact true or false. We make correct decisions if a true null hypothesis is not rejected, or a false null hypothesis is rejected.
If we reject the null hypothesis when it's true we make a type I error, also called a false positive. If we don't reject the null hypothesis when it's false we make a type II error, also called a false negative.

The probability of a type I error - falsely rejecting the null hypothesis - is equivalent to the significance level alpha, usually set at 0.05. We identify the critical region of the test statistic distribution assuming the null is true. If the null is true, then in 5% of the cases we find a test statistic in the critical region and falsely reject the null.
The complement of alpha, (1 - alpha), is the probability of correctly *failing* to reject the null. It's referred to as the confidence level, which we use in confidence intervals.

The probability of a type II error - falsely failing to reject the null - is called beta. We can't visualize this probability as an area under the test statistic distribution that alpha falls under, because this is the distribution when the null is true and a type II error assumes the null is false.
We don't know where the test statistic distribution will be located if the null is false, since we don't know the true population value; but suppose it lies somewhere to the right. The probability of falsely failing to reject the null corresponds to an area under this second test statistic distribution. We fail to reject the null, so beta corresponds to the area in the *non*-critical region.

The complement of beta, (1 - beta), is the probability of correctly rejecting the null hypothesis. It's referred to as the power of a test and corresponds to the area in the critical region under the second test statistic distribution.
As you can see, the probability of making a type I error (alpha) and the probability of making a type II error (beta) are related. If we lower alpha the critical values move further into the tails, resulting in a larger

beta. Choosing a smaller significance level - or a larger confidence level in confidence intervals - will result in less power to detect a true effect. Reversely, if we're willing to accept a larger probability of making a type I error, we'll have more power.

We have other ways of increasing power. Consider these generic versions of the formulas for a test statistic and a confidence interval. More observations or less variance in the population - represented by a smaller standard deviation - will result in more power. In both cases the standard error will be smaller, which means a confidence interval will become narrower, with less chance of containing the null value; and a test statistic will become larger, with a bigger chance of falling in the critical region.
It's obvious how to increase the number of observations, but how do we lower the standard deviation? Well it helps if our measurement instruments are more reliable, eliminating some random variance. It also helps to select a sample that shows less variation, by selecting on - or controlling for - related variables. For example, if we compare the effect of diet on the health of cats it can help to select cats within a limited age range. Any variation in health related to age will be minimized.

Another way to increase power is to obtain a larger sample statistic value. This will result in a confidence interval located further away from the null value and a larger test statistic value. The sample statistic can be influenced by selecting a stronger manipulation or selecting a sample for which the effect is expected to be stronger.

Two final methods to increase power have to do with the type of test. One-sided tests have more power than two-sided tests. The critical value in a one-sided test is closer to mean of the test statistic distribution, so it's more likely that a test statistic in the expected direction will fall in this critical region. Finally, a parametric test, which involves assumptions about the shape and parameters of the population distribution, is usually more powerful than a non-parametric test. We'll discuss non-parametric tests later on.

We can estimate how large our power is post hoc, after the data have been collected. We use the sample statistic value as a proxy for the true population value so we can determine a location for the test statistic distribution when the null is false. Using software we can now determine the power to detect a true effect, assuming the population value is exactly the sample statistic value.
A better idea is to try and achieve a certain level of power a priori, *before* we collect any data. To do this, we first need to estimate how

large our sample statistic will be, using standard effect sizes - small, medium or large. Standard effect sizes have been published for the most commonly used statistics, based on typical findings in the social and behavioral sciences.

With an expected effect size and a chosen significance level we can calculate how large our sample needs to be to obtain a power of say 0.8. If the resulting sample size is impractically large we can choose to increase alpha or we can employ a stronger manipulation - or select a more homogeneous sample - so that we expect a larger effect size.
Performing a priori power calculations can help in the design of a research study and the selection of statistical tests. It also helps to determine whether a research study is worthwhile - whether the probability of detecting a true effect is large enough to justify performing a study in the first place.