

## 1.02 Tests for two groups: P-values

In the previous video we saw why we need the null hypothesis. Without it we can't determine the location of the test statistic distribution and calculate probabilities. Once we specify a null hypothesis and an alternative hypothesis we can calculate the **p-value**. This is the probability that we will find the same test statistic or a more extreme value, *assuming that the null hypothesis is true*. In this video we will see how we determine this probability and how we use it together with the significance level - to see if we can reject the null hypothesis.

Suppose I want to test whether a raw meat diet is healthier for cats than canned food. Cats are randomly assigned to a raw meat diet or a canned food diet. After two months a veterinarian rates each cat's health on a scale from zero to ten.

Our null hypothesis states that the difference in mean health rating between the raw meat and canned group is zero. We hope that our sample produces a test statistic that is very unlikely - with a very small p-value - so we can reject the null hypothesis. This p-value is determined - with statistical software - by calculating the area under the curve associated with the test statistic value and more extreme values. But what is more extreme? Should we take the area to the left or to the right of our test statistic value?

We need to decide beforehand, by specifying an alternative hypothesis. When we compare two proportions or means, this hypothesis can be unidirectional, for example if we expect raw meat to lead to better health. If we calculate the test statistic by taking the raw meat mean and subtracting the canned food mean, we expect a positive test statistic value. In this case we perform a one-sided test and calculate the probability by taking the area under the curve to the right of the observed test statistic value, with more extreme positive values.

If we had expected canned food to be healthier, we would expect a negative test statistic value and would calculate the probability by taking the area under the curve to the left of the observed test statistic value.

Our alternative hypothesis can also be bi-directional: We expect a difference between the diets but have no good reason to expect which diet will be healthier. In this case we expect either a positive or a negative test statistic. We perform a two-sided test and calculate the probability by seeing in what direction the test statistic falls, in this case positive, taking the area under the curve to the right - in a more extreme positive direction - and then doubling this p-value. The test statistic could also have been negative,



so we need to take the probability associated with a negative test statistic into account. If the observed test statistic had been negative, we would have taken the area under the curve to the left and doubled this, taking into account the possibility that the test statistic would have been positive.

How do we decide whether the p-value is small enough to reject the null hypothesis in favor of the alternative hypothesis? We decide by comparing the p-value to the significance level, denoted by  $\alpha$ . This is a value set beforehand, which represents the risk we are willing to run of mistakenly rejecting the null hypothesis, so rejecting it when it is in fact true. The most commonly used significance level value is 0.05, so a 5% chance of mistakenly rejecting the null hypothesis when it is in fact true. The choice for 5% is inherently arbitrary; it might as well have been 4%.

We decide if we can reject the null hypothesis by comparing the p-value to the significance level. If the p-value is smaller or equal we reject the null hypothesis. If the p-value is larger we cannot reject the null hypothesis. Please note that we never accept the null hypothesis, failing to reject it is not the same thing as showing it is true!

Nowadays we use statistical software to calculate the p-value, but you can also use tables. Back when computing power was limited or unavailable, calculating a test statistic value manually was relatively easy, but a p-value was a lot of work, especially for more complicated test statistic distributions like the t-, Chi-square and F-distribution. This is why tables were developed that list test statistic values and their associated p-values.

For complicated test statistic distributions with one or more degrees of freedom the p-values for commonly used significance level values are listed with their associated test statistic values. This way you can look up the test statistic value associated with the significance level that you set beforehand. This is called the critical boundary value.

If our observed test statistic is more extreme - in accordance with the alternative hypothesis - the observed test statistic value is said to fall in the critical region. You don't know the exact p-value, but you know that it is smaller than the significance level.

If I look up the value I look in the row with the appropriate number of degrees of freedom and starting at the left column, I check whether my t-value exceeds the listed value. If it does, I know my p-value is smaller than the p-value listed above that column. Once I encounter a value that is larger than my t-value I know my p-value is larger than the p-value



listed above that column but smaller than the p-value in the previous column.

If you're performing a two-sided test, remember to halve the significance level and consider two critical regions, one in the left tail and one in the right tail of the test statistic distribution.

One last thing to mention: Hypothesis testing always requires that certain assumptions are met, in parametric test like the z-, t- and F-test these assumptions concern the shape and parameters of the population distribution. If these assumptions are not met we can't be certain about the exact shape of the test statistic distribution. This means we might over- or underestimate the p-value and draw the wrong conclusion. So make sure you always know and check the assumptions of a statistical test!