3.06 Simple Regression: Testing the model

In this video you'll learn how to perform a **statistical test** to see whether the predictor and the response variable are likely to be related in the **population**. You'll also learn how to calculate the corresponding **confidence interval**.

Consider the example where we predicted popularity of cat videos - measured as number of video views - using the cat's age as the predictor. We used regression *descriptively* to assess the relation between cat age and video popularity *in our sample*. But can we infer from our sample that the two variables are related *in general*?

Assumptions

To infer whether the variables are related at the population level, we use a statistical test. The inferential procedure only results in *valid* decisions if we meet a number of assumptions. I'll discuss these assumptions and how to check them elsewhere in detail.

Statistical hypotheses

For now, let's get back to the central question: How can we infer that cat age and video popularity are related in the population? Well, we need a null hypothesis that expresses *ab*sence of any relation - or *in*dependence - between the predictor and the response variable in the population.

If the predictor and response variable are independent, then the slope of the regression equation will be zero. So we test for *independence* by assuming the population regression coefficient beta is zero.
We can specify an alternative hypothesis, either two-sided - beta does not equal zero - or one-sided - beta is smaller or greater than zero if we expect a negative or a positive relation.
I don't have a well-founded theory or previous findings to support a directional hypothesis, but let's go crazy and assume the relation between cat age and video popularity is negative.

Test statistic

The test statistic is given by the formula t equals b minus the value of beta under the null hypothesis - beta-sub-zero, divided by the standard error of the regression coefficient b - s e-sub-b: $t = \frac{b-\beta_0}{se_b}$.
Because the value of the population regression coefficient under the null hypothesis is zero, this simplifies to the sample regression coefficient b

divided by its standard error. Remember, the standard error says something about the precision with which we estimate b.

I won't ask you to calculate the standard error manually, since it involves computing sums of squares for x and doesn't really help in understanding the procedure better. For cat age in our example the standard error is 1.76. To get the t-value we divide -3 by 1.76, which equals -1.70.

Test statistic distribution and p-value

To compute or look up the p-value we need to know the degrees of freedom; these equal n minus two. We lose two degrees of freedom because we have to estimate two population parameters, the intercept and the slope. We can now determine the p-value. If we find a p-value smaller than or equal to 0.05, we reject the null hypothesis in favor of the alternative hypothesis. We found a t-value of -1.7 with 5 - 2 = 3 degrees of freedom. Using tables we find the p-value is smaller than 0.10, but larger than 0.05. The exact value, calculated with software, is 0.09. This means we *can't* reject the null hypothesis. We *can't* infer that cat age and video popularity are negatively related.
Interestingly, the result of this test also applies to the correlation. If we can conclude that the regression coefficient is significantly different from zero, then so is Pearson's r. This makes sense because the regression coefficient is simply an unstandardized version of the correlation coefficient.

Confidence interval

We can also calculate a confidence interval for the regression coefficient. If the value of zero lies outside the interval we reject the null hypothesis in favor of the two-sided alternative hypothesis and assume the variables are related. We can also use the interval to evaluate the range of plausible values for the regression coefficient. If we find a statistically significant result, but the interval is very wide and close to zero, we might want to be more conservative in our interpretation of the practical relevance of this finding.

The boundaries of the confidence interval are calculated by taking the sample value of regression coefficient b and subtracting and adding the margin of error. The margin of error for a 95 percent confidence interval is the t-value associated with n minus two degrees of freedom and a p-value of 0.025. Remember, together with the two probabilities of 0.025 in the left and right tail, the 95 percent confidence level adds up to a hundred.

In our example the 95 percent confidence interval for the regression coefficient, which was -3, ranges from -8.6 to +2.6. We obtain

these values by taking -3 and subtracting and adding the margin of error: t - which equals 3.182 for this sample size - times the standard error 1.76, together equaling 5.600.
As you can see the interval is wide and contains zero, so based on this sample, we *can't* conclude that a 'true' relation between cat age and video popularity exists. Of course, in real research, we would never draw any conclusions based on only five observations.