

## 1.06 Tests for two groups: Two independent means

In this video we'll discuss how to compare two independent groups on a quantitative variable, using a t-test for two independent means. We'll consider a version of this test that adds the assumption of equality of population variances. We'll also see how to calculate a confidence interval comparing two independent means.

We use a t-test or confidence interval for two independent means if we have a **quantitative response variable** and **binary independent variable** that distinguishes two independent samples. An example of a research question is: Does the average number of watched television shows differ between people with a full time job and people who are unemployed? Or: Is the mean score on a happiness scale lower for people who have children versus people who don't have children?

### Assumptions

The first requirement to test such questions is that the samples are **independent**; the cases should be assigned to the groups randomly or drawn randomly from the population - in a non-experimental design. Both samples should be distributed normally. However, the t-test is robust against violation of this assumption for large samples due to the Central Limit Theorem. It's even robust for small samples when using a two-sided test. Normality *is* important if the samples are small and the test is one-sided. There's no simple rule that says what sample size is large enough. It depends on the variation in the population and the true effect size. However, for the purpose of this course, let's say that for one-sided tests, both samples need to have more than thirty cases.

### Statistical hypotheses

The statistical hypotheses are expressed in terms of the difference between the population means  $\mu_1$  and  $\mu_2$ . If both groups form the same population the difference will be zero. This is the null hypothesis. Possible alternative hypotheses are that the difference is not zero, or that the difference will be greater, or smaller than zero.

### Test statistic

<sup>1</sup>The test statistic  $t$  equals the difference in sample means minus the expected value under the null hypothesis - zero, divided by the standard

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

error, which equals the square root of the sum of the group variance, divided by its sample size for each group.

### Test statistic distribution and p value

The test statistic follows a Student t-distribution,<sup>2</sup> with a rather complicated number of degrees of freedom. The formula looks horrific, but if you need to do this by hand, you already calculated the variances of each sample divided by their respective sizes when you determined the standard error. If you reuse that information the calculation is not so bad.

Once we know the degrees of freedom we can calculate or look up the one-sided or two-sided p-value and compare it to the predetermined significance level. Finally, we reject or fail to reject the null hypothesis.

### Example

Suppose I want to test whether a raw meat diet is healthier for cats than regular canned food. I randomly assign half of my entire sample to a raw meat diet and the other half to a canned food diet. Cat health is measured on a scale between zero and ten by a veterinarian.

To check whether the groups are normally distributed I look at the histograms of the health scores in each group. These look normal enough. The null hypothesis states that the difference in means between these groups will be zero, my one-side alternative hypothesis is that the difference will be larger than zero if I subtract the canned mean from the raw mean. I expect the raw diet to result in a higher mean health score. I'll set the significance level to 0.05.

<sup>3</sup>The test statistic value is  $5.12 - 4.86$  divided by the square root of  $(1.10/150 + 1.89/148)$ . This equals 1.87. As expected, the value is positive and falls in the right tail. <sup>4</sup>The degrees of freedom are 274.95.

If I use a table to look up the p-value I find that it lies somewhere between 0.025 and .050. If I calculate the p-value with statistical software, I find a value of 0.032.

$$^2 df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$$

$$^3 t = \frac{5.12 - 4.86}{\sqrt{\frac{1.10}{150} + \frac{1.89}{148}}} = 1.87$$

$$^4 df = \frac{\left(\frac{1.10}{150} + \frac{1.89}{148}\right)^2}{\frac{1}{150-1} \left(\frac{1.10}{150}\right)^2 + \frac{1}{148-1} \left(\frac{1.89}{148}\right)^2} = 274.95$$

This is smaller than the significance level of 0.05, so I can reject the null hypothesis. I conclude that the mean health score in the population is higher for cats on a raw meat diet, as compared to cats on a canned food diet.

### Additional assumption

There is an alternative version of the t-test for independent means that adds the assumption that the variances in the populations are the same:  $t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ . If we are willing to make this assumption the standard error and

degrees of freedom are calculated differently. The standard error is calculated by taking the pooled standard deviation times the square root of the sum of the reciprocals of the sample sizes:  $s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1) + (n_2-1)}}$ . You can think of the pooled standard deviation as the weighted average of the two standard deviations. Fortunately the calculation of the degrees of freedom is now much simpler:  $df = n_1 + n_2 - 2$ . They're equal to the total sample size minus two. The additional assumption of equal population variances can be useful because it results in larger degrees of freedom and possibly a smaller standard error, which results in a slightly larger chance of rejecting the null hypothesis. However, most people would recommend using the unequal variances version, to be on the safe side and because in almost all cases it's just as good.

Some people decide to make the extra assumption based on a test that determine whether the variances are equal, but these tests are not very robust against violation of normality and will be significant too often. If you want to make the extra assumption, then as a rule of thumb, only do so if the standard deviations differ by less than a factor of two.

### Confidence interval

We calculate the confidence interval using this formula:  $CI = \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \cdot se$  the difference in sample means plus and minus t times the standard error. Plus and minus t equal the t values associated with the required confidence level and the degrees of freedom that we just calculated. With 274.95 degrees of freedom and a confidence level of 95%, the values are minus and plus 1.96. The standard error is calculated the same way as before. Also, remember that we need to meet the same assumptions required for a two-sided t-test.

The confidence interval for our example data is 0.26 plus and minus 1.96 times 0.14. This results in a confidence interval that ranges from -0.01 to 0.54:  $CI = 5.12 - 4.86 \pm 1.96 \cdot 0.14 = (-0.01, 0.54)$ .

This corresponds to performing a two-sided test. Since the value zero - no difference in the means - lies inside the interval, zero is considered a plausible value. This means a two-sided, more conservative test than the one we performed - would have been non-significant.