

10.2: Digging deeper into t-test plus the Welch test

Introduction

We need to spend some more time with the **independent sample t-test**; by tearing it apart, we can learn about how **parametric tests** work in general.

Our assumptions for the independent sample t-test are like the one-sample t-test the data must be continuous and **normally distributed** (one of our standard assumptions for parametric tests, see [Chapter 13](#)). The formula is very similar to the one-sample t-test, except that now we have two sample means and the formula for the **standard error** (SE) has also changed.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

We see that the test statistic t is large if the numerator is large compared to the denominator. Large values of t will be evidence in favor of rejecting the null hypothesis.

The numerator is straight-forward: we subtract one sample mean from the other — if there is no difference between the samples, then this difference will be close to zero.

The denominator requires additional discussion. What is it called? It is the **pooled standard error of the mean** (pooled SEM). Provided the assumption of equal variances holds — an additional standard assumption for parametric tests, see [Chapter 13](#) — then sample variances estimate the population variance and we can use this information to our advantage to best test the hypothesis about the sample means. In other words, we don't have to lose a degree of freedom to account for differences in variability for the two groups (see Welch test). More degrees of freedom means more statistical power to test the null hypothesis and at the same time, more confidence that the test is performing to its best.

Let's break down the pooled standard error of the mean in order to see how the assumption of equal variances affects the t-test. We assume that $s_1^2 = s_2^2$.

Recall that $H_0 : \bar{X}_1 = \bar{X}_2$.

Now we want a “pooled SE” that is the pooled standard error for both samples. The variance of the difference between the means can be written as

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

First we need to calculate the pooled variance, where v_1 and v_2 are degrees of freedom for sample one and sample two, respectively. Note that this is just simply a combined formula of the sample variance.

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2}$$

What is SS ? It is the sum of squares, where SS_1 and SS_2 refer to sum of squares for each of the two groups.

$$SS = \sum_{i=1}^n (X_i - \bar{X})^2$$

You should recognize SS from your definition of the sample variance, [Chapter 3.2](#).

And the **standard error for the difference** between two means can now be written as

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

so the 2-sample t-test can be written to reflect the pooled sample standard error of the difference between two sample means. We can see how **unequal sample size** is accommodated by the t-test.

Conducting the test by hand follows the same form as the one-sample t-test. Find the degrees of freedom (DF), but now for each sample.

Finally we evaluate with the critical value in Table ([Appendix, Table of Student's t distribution](#)) and compare the **t test statistic** against the critical value with the appropriate degrees of freedom.

Because this is the t-test, again, we are assuming that the variances are the same between the two populations (**homoscedasticity**) and this allows us to pool the variance. As it turns out, the T-test is not overly sensitive to other deviations from the assumptions, but if the variances are in fact different, then the standard formula may yield incorrect Type I error rates compared to stated probability level (α).

However, it would be poor statistical choice to use a test where there are alternatives. This is why in part that R (`Rcmdr`) sets as the default for the t-test that variances are unequal! In fact, R does not do a t-test unless you change the default to assume equal variances, which, as we now know, is the t-test.

Welch test

What to do when assumptions for the t-test are not met? Many options have been proposed, and **Welch's approximate t** is a good alternative to the two-sample t-test — it would be appropriate if the normal assumption still held.

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

The degrees of freedom for the Welch's test are now

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot v_1} + \frac{s_2^4}{n_2^2 \cdot v_2}}$$

Note that all the Welch test does is remove the pooled estimate of the standard error, replaced with both variance estimates directly.

As a default option, R and Rcmdr uses a variation of Welch's test when you select to do the t-test without making the assumption of equal variance.

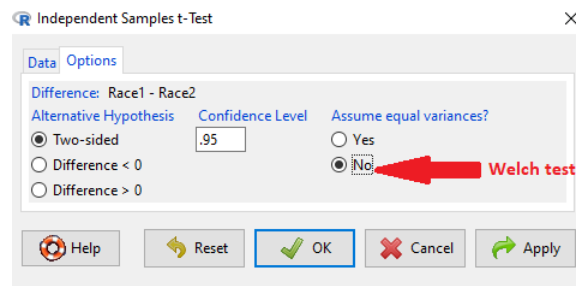


Figure 10.2.1: Screenshot Rcmdr t-test options. Default is “No” for Assume equal variances, i.e., the Welch test.

The Welch test is not a nonparametric test, it is a different formulation of the t-test.

Justification for beginning with t-test

It's unlikely that you will need the t-test in today's research climate. Data sets are large, experiments are complex with multiple variables and samples. Why do we have to consider the t-test, and then a separate test in the case for unequal sample size? I view it as a teaching moment. It makes the general point that ALL statistical tests make assumptions about how the calculations are done and as to the nature of the data set.

This is our first experience with what to do if there is a violation of an assumption of parametric tests (see [Chapter 13](#)). Here, the assumption is that the two groups have equal sample size. When they do not, the standard t-test tends to **biased estimates**. On the other hand, if the assumptions are met, the standard test is the best test because it has more power to do what we intended — that is, it is best at the actual test of the null hypothesis! Take heart — this point is not always appreciated even by scientists (cf. Fagerland 2012).

Let's us approach the problem of violation of assumptions in a couple of ways as an introduction to how, in general, to approach choice of statistical tests.

1. **Power of a test.** Much current statistical research focuses on learning about how a particular statistical test works when assumptions are violated. Thus, in addition to learning what tests are designed to do, we need to consider the effects of violations of assumptions on the performance of the test (namely, is the Type I error at the stated alpha level?). This is a matter of statistical power; power of a test reflects how well the test is able to get you the correct result even if assumptions are violated. Often tests perform well if sample sizes are large, despite violation of assumptions. We mention without proving that the two-sample t-test is robust to violations of normality assumption, to lack of equal sample sizes, and even to unequal variances. But good experimental design attempts to meet the assumptions because the test does better!
2. **Alternate forms of some tests are available to handle some aspects of test violations.** For example, the simple two-sample t-test can be modified to accommodate different variances (Welch's formula). Or you must find a different test (e.g. nonparametric tests).

In conclusion, all tests begin with consideration of the assumptions. In some cases we can test our assumptions. For example, we learned about testing the assumption of normal distributions of sample data. We can also test the assumption of equal variances.

Questions

1. Consider a clinical trial in which resting blood pressure is recorded on hypertensive subjects at the start of the trial, then 6 weeks after subjects have received daily supplements of flaxseed. Thus, for each subject there are two measures of blood pressure, BEFORE and AFTER.
 - Write the null hypothesis.
 - Write the alternative hypothesis.
 - Justify why or why not the hypothesis should be two-tailed. Explain why or why not an independent sample t-test may be used to compare.
2. Justify why the default t-test in R and therefore Rcmdr applies the Welch test, not the t-test?

This page titled [10.2: Digging deeper into t-test plus the Welch test](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by [Michael R Dohm](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.