13.4: The Independent Samples t-test (Welch Test)

The biggest problem with using the Student test in practice is the third assumption listed in the previous section: it assumes that both groups have the same standard deviation. This is rarely true in real life: if two samples don't have the same means, why should we expect them to have the same standard deviation? There's really no reason to expect this assumption to be true. We'll talk a little bit about how you can check this assumption later on because it does crop up in a few different places, not just the t-test. But right now I'll talk about a different form of the t-test (Welch 1947) that does not rely on this assumption. A graphical illustration of what the ***Welch t test*** assumes about the data is shown in Figure 13.10, to provide a contrast with the Student test version in Figure 13.9. I'll admit it's a bit odd to talk about the cure before talking about the diagnosis, but as it happens the Welch test is the default t-test in R, so this is probably the best place to discuss it.

The Welch test is very similar to the Student test. For example, the t-statistic that we use in the Welch test is calculated in much the same way as it is for the Student test. That is, we take the difference between the sample means, and then divide it by some estimate of the standard error of that difference:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}(\bar{X}_1 - \bar{X}_2)}$$

The main difference is that the standard error calculations are different. If the two populations have different standard deviations, then it's a complete nonsense to try to calculate a pooled standard deviation estimate, because you're averaging apples and oranges.[193] But you can still estimate the standard error of the difference between sample means; it just ends up looking different:

$$\mathrm{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

The reason why it's calculated this way is beyond the scope of this book. What matters for our purposes is that the t-statistic that comes out of the Welch test is actually somewhat different to the one that comes from the Student test.

The second difference between Welch and Student is that the degrees of freedom are calculated in a very different way. In the Welch test, the "degrees of freedom" doesn't have to be a whole number any more, and it doesn't correspond all that closely to the "number of data points minus the number of constraints" heuristic that I've been using up to this point. The degrees of freedom are, in fact…

$$\mathrm{df} = \frac{\left(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2\right)^2}{\left(\hat{\sigma}_1^2/N_1\right)^2/(N_1-1) + \left(\hat{\sigma}_2^2/N_2\right)^2/(N_2-1)}$$

… which is all pretty straightforward and obvious, right? Well, perhaps not. It doesn't really matter for out purposes. What matters is that you'll see that the "df" value that pops out of a Welch test tends to be a little bit smaller than the one used for the Student test, and it doesn't have to be a whole number.

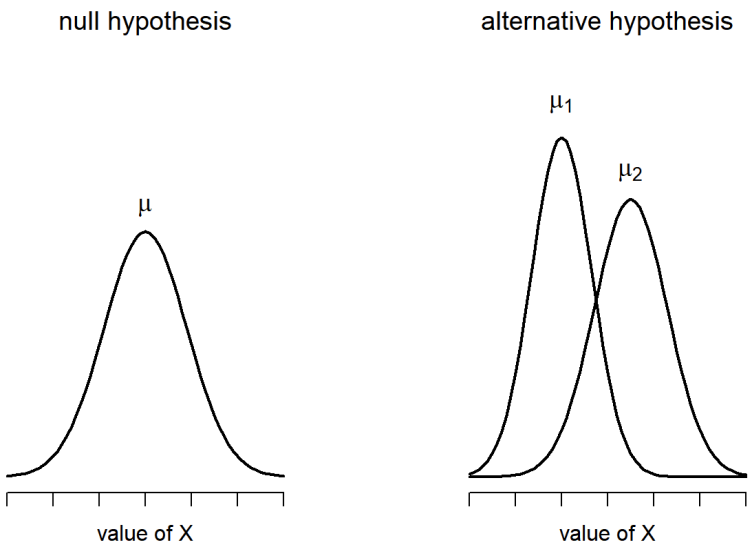

Figure 13.10: Graphical illustration of the null and alternative hypotheses assumed by the Welch t-test. Like the Student test we assume that both samples are drawn from a normal population; but the alternative hypothesis no longer requires the two populations to have equal variance.

13.4.1 Doing the test in R

To run a Welch test in R is pretty easy. All you have to do is not bother telling R to assume equal variances. That is, you take the command we used to run a Student's t-test and drop the `var.equal = TRUE` bit. So the command for a Welch test becomes:

```
independentSamplesTTest(
      formula = grade ~ tutor,  # formula specifying outcome and group variables
      data = harpo              # data frame that contains the variables
  )
```

```
##
##    Welch's independent samples t-test
##
## Outcome variable:   grade
## Grouping variable:  tutor
##
## Descriptive statistics:
##             Anastasia Bernadette
##    mean        74.533     69.056
##    std dev.     8.999      5.775
##
## Hypotheses:
##    null:        population means equal for both groups
##    alternative: different population means in each group
##
## Test results:
##    t-statistic:  2.034
##    degrees of freedom:  23.025
##    p-value:  0.054
##
## Other information:
##    two-sided 95% confidence interval:  [-0.092, 11.048]
##    estimated effect size (Cohen's d):  0.724
```

Not too difficult, right? Not surprisingly, the output has exactly the same format as it did last time too:

The very first line is different, because it's telling you that its run a Welch test rather than a Student test, and of course all the numbers are a bit different. But I hope that the interpretation of this output should be fairly obvious. You read the output in the same way that you would for the Student test. You've got your descriptive statistics, the hypotheses, the test results and some other information. So that's all pretty easy.

Except, except… our result isn't significant anymore. When we ran the Student test, we did get a significant effect; but the Welch test on the same data set is not ($t(23.03)=2.03$, $p=.054$). What does this mean? Should we panic? Is the sky burning? Probably not. The fact that one test is significant and the other isn't doesn't itself mean very much, especially since I kind of rigged the data so that this would happen. As a general rule, it's not a good idea to go out of your way to try to interpret or explain the difference between a p-value of .049 and a p-value of .051. If this sort of thing happens in real life, the *difference* in these p-values is almost certainly due to chance. What does matter is that you take a little bit of care in thinking about what test you use. The Student test and the Welch test have different strengths and weaknesses. If the two populations really do have equal variances, then the Student test is slightly more powerful (lower Type II error rate) than the Welch test. However, if they *don't* have the same variances, then the assumptions of the Student test are violated and you may not be able to trust it: you might end up with a higher Type I error rate. So it's a trade off. However, in real life, I tend to prefer the Welch test; because almost no-one *actually* believes that the population variances are identical.

13.4.2 Assumptions of the test

The assumptions of the Welch test are very similar to those made by the Student t-test (see Section 13.3.8), except that the Welch test does not assume homogeneity of variance. This leaves only the assumption of normality, and the assumption of independence. The specifics of these assumptions are the same for the Welch test as for the Student test.
