Two sample t-tests

Introducing two sample tests

The one sample t-test isn't terribly useful in biology (or other fields). One of the problems is that often have no idea what μ should be. For example, why should μ have the following values?

> $\mu = 100$ for IQ.
>
> $\mu = 69$ inches for height in men.
>
> $\mu = 500$ grams for weight of a gray squirrel.

Where do these values come from (i.e., 100, 69, and 500)? Unless we already know something about our population, we often don't have any idea what μ should be. Because of this, if we only have one sample, Confidence Intervals are often a better approach as they bracket μ without having to guess what μ actually might be.

In biology, we find ourselves much more likely to take two samples and then compare them to see if we think the populations means are different. The advantage here is that we don't need to know anything about the actual value of μ for either of these *two* samples. For example, we might want be interested in the following:

> Does a new medicine work? We divide our volunteers into two groups, group *1* gets the new medication, group *2* gets a placebo (so we have *two samples*). We then measure to see if there is a statistically significant difference in the two groups.
>
> We want to know if men are the same average height as women. We measure a sample of men and a sample of women (*two samples!*) and see if there is a statistically significant difference in height.
>
> Do lions and tigers have the same average litter size? We take a sample of litter sizes from lions and tigers (again, *two samples*) and compare to see if there is a statistically significant difference in litter size.

This is different from our one sample test since now we're comparing two samples. One big advantage is that we need to know (or guess) nothing about the values of μ in any of these examples. We're comparing μ_1 with μ_2, and this doesn't require knowing or guessing anything about any value of μ. For example, our H_0 for the above examples would be the following:

H_0 : population mean for placebo group = population mean for medicated group.

H_0 : the population means for the heights in men and women are the same.

H_0 : the population mean litter sizes for lions and tigers are the same.

Or in symbols we would say:

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_{men} = \mu_{women}$$

$$H_0 : \mu_{lions} = \mu_{tigers}$$

This would give us the following alternative hypotheses (this time only in symbols):

$$H_1 : \mu_1 \neq \mu_2$$

$$H_1 : \mu_{men} \neq \mu_{women}$$

$$H_1 : \mu_{lions} \neq \mu_{tigers}$$

So we have our null and alternative hypotheses for two sample tests. Next we proceed as usual and pick our value for α (say, 0.05, or 0.01, or whatever value we want).

Then we need to calculate t^*. The problem is, what is our t^* now? We have two values for $\bar{y}$ and two values for s. How do we combine this and get a single value of t^*?

Introducing Welch's two sample t-test

As it turns out there are several ways of doing this, although only two are really commonly used. But even for just these two, we will need to make an assumption. We will need to assume that the two population variances (careful - *not* the sample variances) are not equal. We are assuming:

$$\sigma_1^2 \neq \sigma_2^2$$

Again, we are making this assumption about the population variances, not the sample variances. Obviously we could also use $\sigma_1 \neq \sigma_2$, but this particular assumption has always been referred to by the variances.

If we make this assumption, that let's us calculate t^* using something called *Welch's* test or *Welch's* t-test. This is to differentiate it from the regular t-test for two samples which

we'll learn about a little later:

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

But now we remember that the null hypothesis says the population means are equal ($H_0 : \mu_1 = \mu_2$). This implies that we can simplify our equation:

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The denominator of this expression ($\sqrt{s_1^2/n_1 + s_2^2/n_2}$) is actually the standard error of the numerator. We say:

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

Remember that the SE can take on different values - a concept we introduced when we first discussed the SE. In any case, hopefully you can see that this quantity is a type of average for the two estimates we have of our standard error (i.e., SE_1 and SE_2). It's not a true average, but you can think of it that way (a true average is what we have when we assume $\sigma_1 = \sigma_2$, which we'll talk about later).

Incidentally, notice that we can rewrite our expression for t^* in terms of the SE's of $\bar{y}_1$ and $\bar{y}_2$:

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{SE_{\bar{y}_1}^2 + SE_{\bar{y}_2}^2}}$$

Now we're almost ready to look up our tabulated t value and compare it to our value of t^*. But what are our degrees of freedom. What is $d.f.$ (or what is ν, which $= d.f.$)? It turns out this does not have an easy answer:

$$d.f. = \nu = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}} \qquad (OUCH!)$$

Fortunately most statistical software will automatically do this for us. Even some calculators will do this (you want to verify that your calculator does this correctly if you have one that does). This expression will almost never give you an integer, so we need to round down the degrees of freedom (always round down) before we can look up a value in the t-table.

Once we have all this in place, we can then compare our t^* to the t_{table} value and make our decision the same way we we did for a one sample t-test:

If $|t^*| \geq t_{\text{table}}$ we reject H_0.

If $|t^*| < t_{\text{table}}$ we fail to reject H_0.

Or, of course, we could use α and compare this to a *p-value* to make our decision.

So let's do some examples. First we'll use some of the data from snapping turtles that we used earlier this semester. We want to find out if there's a difference in length between male and female snapping turtles.

Let's set up our hypotheses:

H_0 : The true mean length of males is the same as the true mean length of females.

H_1 : The true mean length of males is not the same as the true mean length of females.

Or in symbols:

$H_0 : \mu_{♂} = \mu_{♀}$

$H_1 : \mu_{♂} \neq \mu_{♀}$

Let's pick $\alpha = 0.05$

We get the following results:

	Males	Females
n	111	53
$\bar{y}$	268	295
s	79.6	65.2
SE	7.56	8.95

So let's calculate t^* :

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{SE_{\bar{y}_1}^2 + SE_{\bar{y}_2}^2}} = \frac{268 - 295}{\sqrt{7.56^2 + 8.95^2}} = -2.30$$

$$\implies |t^*| = 2.30$$

Now we need to calculate our degrees of freedom:

$$d.f. = \nu = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}} = \frac{(7.56^2 + 8.95^2)^2}{\frac{7.56^4}{110} + \frac{8.95^4}{52}} = 123.06$$

And now we can do our comparison (we need to use $d.f. = \nu = 100$ since 123 is not in our t-tables):

Since $|t^*| = 2.30 \geq t_{100,0.05} = 1.984$, we *reject* the null hypothesis.

We conclude that we have enough evidence to show that male and female snapping turtles are not the same length.

Incidentally, R tells us that $t_{123,0.05} = 1.979$, which is pretty close to what we used from our tables. R also tells us that the *p-value* is 0.023, which is less than α as expected.

Our second example of Welch's t-test looks to see if there's a difference in fasting blood sugar levels between men and women. This time we have a sample of 10 men and 10 women, and we get the following results in mg/dL:

	Men	Women
$\bar{y}$	87	89
s	12.1	11.3

Again, let's set up our hypothesis

H_0 : The true mean blood sugar level in men is the same as in women.

H_1 : The true mean blood sugar level in men is not the same as in women.

Or in symbols:

$H_0 : \mu_{♂} = \mu_{♀}$

$H_1 : \mu_{♂} \neq \mu_{♀}$

Let's pick $\alpha = 0.10$.

And then calculate t^*:

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{87 - 89}{\sqrt{\frac{12.1^2}{10} + \frac{11.3^2}{10}}} = -0.3820$$

$$\implies |t^*| = 0.3820$$

Which is a small value for t^*, but let's do do it right and finish the problem.

We note that:

$$SE_{\male} = \frac{12.1}{\sqrt{10}} = 3.826$$

$$SE_{\female} = \frac{11.3}{\sqrt{10}} = 3.573$$

So we get:

$$d.f. = \nu = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}} = \frac{(3.826^2 + 3.573^2)^2}{\frac{3.826^4}{9} + \frac{3.573^4}{9}} = 17.9$$

And finally we compare t^* with t_{table} :

Since $|t^*| = 0.382 < t_{0.10,17} = 1.740$ we fail to reject H_0.

We conclude that we don't have enough evidence to show a difference in fasting blood sugar levels between men and women.

(And R shows our *p-value* to be 0.7072, which is larger than α.)

The classic sample t-test (assuming equal variances)

Above was assumed that the (population) variances are not equal. But suppose we somehow did know enough about our research to assume the variances are equal. In other words, we assume that $\sigma_1^2 = \sigma_2^2$ (which is obviously the same as $\sigma_1 = \sigma_2$).

This is termed the *equal variance assumption*, or the *pooled variance assumption*.

Notice that we're assuming the *population* variances are the same. It should be obvious that the sample variances (s_1^2 and s_2^2) may or may not be the same, regardless of this assumption (they will almost never be identical).

How does this change our two sample t-test? It turns out our denominator changes; we now used a pooled estimate for our variance:

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This is a *weighted* mean of s_1^2 and s_2^2. It adjusts the overall variance ($s^2_{poooled}$) for the size of each sample (the larger sample counts more). In fact, if $n_1 = n_2$, then:

$$s^2_{pooled} = \frac{s_1^2 + s_2^2}{2}$$

We then use s^2_{pooled} in the denominator of our calculation for t^*:

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Which isn't that different from Welch's formula for t^*.

It's not difficult to show that if $n_1 = n_2$ or $s_1^2 = s_2^2$ (or both), then the two formulas are actually the same (this is really easy to see in the extremely unlikely case that $s_1^2 = s_2^2$). But if neither of these conditions is true, then the two formulas for $|t^*|$ will give you different results.

What's important to realize is that the formula for the $d.f.$ also changes, and that in this case the two formulas will virtually never give you the same result for $d.f.$, so you have to be careful to use the correct formula for $d.f.$. If we assume equal variances, our formula for the degrees of freedom becomes:

$$d.f. = \nu = n_1 + n_2 - 2$$

This is obviously much easier than the other formula for $d.f.$!

Let's do an example. We'll use the fasting blood sugar example from above, but this time assume that the population variances are the same. Here are the data again (in mg/dL):

	Men	Women
$\bar{y}$	87	89
s	12.1	11.3

Our hypotheses don't change (just symbols this time):

$H_0 : \mu_{\male} = \mu_{\female}$

$H_1 : \mu_{\male} \neq \mu_{\female}$

Let's pick the same α as before: $\alpha = 0.10$. And now we calculate t^* using the formula for equal variances (first we need to calculate s^2_{pooled}):

$$s^2_{pooled} = \frac{(10-1)12.1^2 + (10-1)11.3^2}{10+10-2} = \frac{12.1^2 + 11.3^2}{2} \text{ (since } n_1 = n_2)$$

$$= 137.05$$

and now we do t^*:

$$t^* = \frac{87 - 89}{\sqrt{137.05\left(\frac{1}{10} + \frac{1}{10}\right)}} = -0.3820$$

$$\implies |t^*| = 0.3820$$

Which is the same as our t^* above since $n_1 = n_2$. Let's now get our degrees of freedom:

$$n_1 + n_2 - 2 = 10 + 10 - 2 = 18$$

And finally we compare t^* with t_{table} :

Since $|t^*| = 0.3814 < t_{0.10,18} = 1.734$ we fail to reject H_0.

And just as before *we conclude that we don't have enough evidence to show a difference in fasting blood sugar levels between men and women.* Incidentally, R shows our *p-value* to be 0.7072, which again is larger than α.)

So what does it all mean? If it is true that $\sigma_1^2 = \sigma_2^2$, then the equal variance t-test will have a little more power. It's not a big difference, but it does give a little more power.

So how much power does Welch's t-test have if the populations variances are actually equal? Almost as much. It isn't a big difference (actual power comparisons like this aren't really in the scope of an introductory class).

On the other hand, if $\sigma_1^2 \neq \sigma_2^2$ then what happens? The equal variance t-test can make bad mistakes:

It can reject when it shouldn't.

It can fail to reject when it shouldn't.

In other words, it can give you badly wrong answers (but see item (3.) below).

In other words, Welch's t-test often does *much* better.

So let's summarize:

If the variances are equal ($\sigma_1^2 = \sigma_2^2$) then the equal variance t-test has a little more power than Welch's test.

If the variances are not equal ($\sigma_1^2 \neq \sigma_2^2$), then Welch's test can perform much better (particularly if the sample sizes are also not equal).

So what do you do? Unless you're pretty certain about the assumption that the population variances are the same ($\sigma_1^2 = \sigma_2^2$), you should always use the unequal variance t-test (Welch's test). Welch's test will do almost as well even if the variances are actually equal. Some statisticians actually claim there is no difference in power.

A few final comments on all this:

1. How often do you really know enough about your data so you can assume that the *population* variances are the same?

   Not often. Some people might be able to make educated guesses about the means, but very few people know enough about their data to do the same for the variances.

2. It is possible to test to see if the population variances are the same:

   You could do $H_0 : \sigma_1^2 = \sigma_2^2$.

   But remember - just because you *fail to reject* H_0, doesn't mean that H_0 is true! You can *not* prove the null hypothesis is true. This test ($H_0 : \sigma_1^2 = \sigma_2^2$) also has awful power (it's not a very good test). For these reasons, this procedure isn't recommended, and we won't discuss it any further.

3. If $n_1 \approx n_2$ the equal variance t-test will at least not fail too badly. You should still use Welch's test unless you're certain about the population variances, but at least the equal variance t-test won't mess up too badly.