t-Tests – Part 2

Part 1 covered the univariate and paired-samples *t*-tests, which are used to evaluate H_0: $\mu = V$ and H_0: $\mu_1 - \mu_2 = V$, where, in the second case, the data for Conditions 1 and 2 come from the same subjects or matched pairs of subjects. One point that was emphasized was how these two "versions" of the *t*-test are actually exactly the same, because both involve using a single set (or column) of data to test the idea that the population mean is equal to a particular value, *V*. This is obviously true in the univariate case, because you started with one set of data. In the paired-samples case, it's done by first calculating a difference score for each subject and then using this one set of data to do the same test (with $V = 0$).

We now turn to cases when we are testing H_0: $\mu_1 = \mu_2$ where the data for Conditions 1 and 2 come from different subjects. Here we cannot create a single set of difference scores because (1) there is no way to pair specific pieces of data from Condition 1 with specific pieces of data from Condition 2, and (2) you might not have the same number of subjects in each of the two conditions.

The Independent-samples *t*-test

As suggested above, these tests are used to compare population means when separate samples were taken from the two populations. To keep this situation clearly distinct from paired-samples *t*-tests, I suggest that you write the null hypothesis (that the two population means are the same) as H_0: $\mu_1 = \mu_2$, instead of H_0: $\mu_1 - \mu_2 = 0$. By having the two means on opposite sides of the equals-sign, the separateness of the two samples is made much clearer.

Before getting to the details of the independent-samples *t*-test, let me say a few more things about *t*-tests in general. In every case, the calculated *t*-statistic is a ratio. The numerator is always the mean violation of what was predicted by the null hypothesis and the denominator is always some form of standard error. In the case of the univariate *t*-test, the numerator is $\bar{X}-\mu$ and the associated error is $s_{\bar{X}}$. The paired-samples *t*-test is exactly the same after you swap the X for D, which stands for "difference score," and set μ to zero. In the case of the independent-samples *t*-test, however, the numerator is $\bar{X}1-\bar{X}2$ (because the null hypothesis predicts that these will be the same) and the associated error term is $s_{\bar{X}-\bar{X}}$, which is called the "standard error of the difference."

Second, note that *t*-statistics do not have units because the units of the numerator (which are the same as the units of the original data) are exactly canceled by the units of the denominator (which are also the same as the original data because standard errors are based on standard deviations, and standard deviations have the same units as the original data).

Finally, note that *t*-stats also have a measure of quality: the degrees of freedom (which is used to determine the final *p*-value). The *t*-stat inherited its value of df from the error term; it doesn't appear out of nowhere and it doesn't come from the numerator; it comes from the denominator.

The Standard Error of the Differences

The only new thing in the above brief review was the standard error of the difference, $s_{\bar{X}-\bar{X}}$, which we need for the denominator of the independent-samples *t*-test. Given the formula for the

standard error in the univariate case -- i.e., $s_{\bar{x}} = s / \sqrt{N}$ -- it probably (hopefully?) won't surprise you to learn that the standard error of the difference is also based on the standard deviations and samples sizes. The trick is in how to combine the values of s and N from the two different samples. And this is where a new assumption comes in.

The standard form of the independent-samples *t*-test assumes that the variances of the two populations are equal. Under this assumption, all of the data -- i.e., both samples combined -- can be used to estimate the (common across conditions) population variance and then this one value can be converted into an estimated standard deviation of the mean difference, which is just another name for the standard error of the difference. The degrees of freedom associated with the standard error of the difference is $(N_1 - 1) + (N_2 - 1)$; i.e., each of the samples contributes $N - 1$ df to the total. You can also think of this as $N_1 + N_2 - 2$ -- i.e., the total number of pieces of data minus the number of estimated parameters (which is two: the means for each of the two conditions).

How is the above actually done? In the first step you get the variances (using the $N - 1$ formula) for each of the two conditions (separately). We'll refer to these as s_1^2 and s_2^2. *Why the variance, instead of the standard deviation?* Because the two samples are here being treated as independent events (hence the name of the test) and it is the variances of independent events that are additive, not their standard deviations.

In the next step, we combine these two values into a single estimate of the (common) population variance -- this is the "pooled estimated variance" or s_p^2. When the two samples have equal Ns (and, therefore are of the same "quality"), we weight them equally, so just get the plain mean of the two separate values. When the two samples have unequal Ns, we get a mean that is weighted by the degrees of freedom:

$$s_p^2 = (((s_1^2 (N_1 - 1)) + (s_2^2 (N_2 - 1))) / (N_1 + N_2 - 2)$$

Note how the above reduces to $s_p^2 = (s_1^2 + s_2^2) / 2$ when $N_1 = N_2$. It's not that we use a different formula when $N_1 = N_2$; rather, there's simply a short-cut to the same answer.

In the last step, we switch back to unsquared units and then get the standard error of the difference by applying the equivalent of the "square-root of N" rule. This is another place where the additivity of variances, not standard deviations, plays a role. Recall, first, the "square-root of N" rule for single distributions:

$$s_{\bar{x}} = s / \sqrt{N}$$

By the distributed property of square-roots, this is the same as:

$$s_{\bar{x}} = \sqrt{ (s^2 / N)}$$

where, for lack of a better name, the quantity s^2 / N would be something like "standard variance" (which you square-root to get standard error). This is probably the better way to think about it in

general, even if it doesn't match the "square-root of N" rule label for converting s to $s_{\bar{x}}$.

With the second version of the formula for (plain) standard error in mind, along with the general idea that it's variances that are additive, we're now ready to expand the formula to the difference between two (independent) means, instead of the error associated with a single mean:

$$s_{\bar{x}-\bar{x}} = \sqrt{ (s_p^2 / N_1) + (s_p^2 / N_2) }$$

Note that you do not convert to two standard errors (one for each sample) and then add them together; you add the two "standard variances" together and then take the square-root in the very last sub-step.

As mentioned above, the degrees of freedom associated with $s_{\bar{x}-\bar{x}}$ is $N_1 + N_2 - 2$. So now you have everything that you need to calculate the *t*-stat and then go look up the *p*-value.

What if we don't assume that the two populations have the same variance? When the two population variances are ***not*** assumed to be equal (either because we have some prior reason to doubt this or because we conducted a preliminary test of the assumption and found evidence against it), then two changes are made to the typical analysis. First, rather than creating a single estimate of the (common) variance, s_p^2, you keep the two separate. When you get to the formula for the standard error of the difference, $s_{\bar{x}-\bar{x}}$, you don't use s_p^2 twice; you use s_1^2 and s_2^2. Second (and much more important), the degrees of freedom associated with the standard error is shrunk more and more as the two variances become more and more different. This almost always results in a value for df that is not a whole number, which is a clear sign (to the reviewers and readers) that the assumption of equal variance was not being used.

☞ Shrinking the degrees of freedom associated with the standard error is a "kluge." We are not doing this because the degrees of freedom really are smaller than $N_1 + N_2 - 2$; we are doing this to counter-act the problems that arise when the assumption of equal variance is violated. We didn't fix the problem; we added a second problem on top that counter-acts the first problem. (That's what the word "kluge" means; it's German for "two wrongs *can* make a right.")

A common name for the version of the independent-samples *t*-test that does not rely on the assumption of equal variance is the "clinical trials *t*-test." It is called this because clinical and control groups rarely have the same variance, so people who conduct tests that involve comparisons of clinical populations with controls use this version a lot. (The technical name is Brown-Forsythe, after the guys who came up with the particular method that we use to deal with violations of equal variance.)

If you have an *a priori* reason to suspect that the population variances are not equal or Levene's Test (see below) tells you that they are significantly different, then you should not use the standard, equal-variances-assumed version of the independent-samples *t*-test; you should use the clinical-trials or equal-variances-not-assumed version, instead. The latter is more conservative (mostly because it has fewer degrees of freedom), which is another way of saying that it has less power. This might sound bad (and, in some ways, I guess that it is), but using the equal-variances-assumed version of the test when the population variances are highly unequal can result in a falsely-low

p-value, which will raise the rate of Type I errors ***without any warning***. In the analysis of psychological data, few things are considered as bad as having a true rate of Type I errors that is above the agreed-upon value of 5%.

Testing for Equal Variances

A variety of tests have been proposed and used to evaluate the assumption of equal variances. Originally, we used an *F*-test, but this required some assumptions that were probably not warranted. These days, we use Levene's Test. This test is somewhat lacking in power, but isn't as bad as the other options. Fortunately for us, Levene's Test is automatically run whenever you ask SPSS to do an independent-samples *t*-test. In other words, you don't have to ask for the test, nor run it separately; it's all automatic. If the reported "sig" value is less than .05, then the assumption of equal variances must be rejected and the equal-variance-not-assumed version should be used. If Levene's Test is not significant, then you can use the more-powerful, equal-variance-assumed *t*-test (unless your sub-field or target journal requires the more conservative test even when Levene's Test is not significant).

Implications of (un)Equal Variance for the Test(s) of Normality

Like the univariate (or paired-samples) *t*-test, both forms of the independent-samples *t*-test assume that the data are normal. But now we have two separate sets of data, so we have two general options: either conduct separate tests of normality on each set of data or conduct a single test on the combined set of data. As a preview: we need to match the test(s) of normality to way in which the standard error of the difference was calculated. If we pool the two samples when calculating the standard error together (as is the default), then we conduct one test of normality on the combined data; if keep the two samples separate when getting the standard error (maybe because Levene's Test was significant), then we conduct separate tests of normality on each of the sample. This is why I discussed the issue of (un)equal variance before the question of whether the data are normal: the decision with regard to test(s) of normality is driven by the decision with regard to (un)equal variance.

(Deep breath.) Because the correct way of testing normality depends on whether we will be assuming equal variance, and because whether we will be assuming equal variance is often not known until we have conducted Levene's Test, and because Levene's Test is built into the independent-sample *t*-test procedure, we will actually start the analysis before we have conducted the test(s) of normality. (In an ideal world, we would be able to test all assumptions before running the analysis, but we often can't do this and this is your first example of such.)

How-to*: Independent-samples t-test using SPSS*

Ask for **Analyze... Compare Means... Independent-Samples T Test...** then highlight the data variable and "push" it over to the Test Variable(s) window. Then highlight the control variable that denotes group (or condition) and push it into Grouping Variable, and specify the value associated with each group by clicking Define Groups. The null hypothesis is that the two means are the same and you cannot change this. Click OK to finish.

Look first at the results from the automatic application of Levene's Test (and try to avoid looking at anything else). It's on the left side of the main output table. What you want is the *p*-value (you can ignore the *F*-stat). Everything that follows depends (in most cases) on whether Levene's is significant, starting with how you will be testing normality.

If you will be using the equal-variance-not-assumed output, either because Levene's was sig or your advisor or sub-field or the journal requires it, then you will be conducting separate tests of normality for each of the groups (or conditions). The way to do this is relatively simple. When you run **Analyze... Descriptive Stats... Explore** with **Normality plots with tests** clicked in the **Plots** sub-menu, be sure to push the control variable that defines the groups (or conditions) into the **Factor List** box (as well as pushing the data variable into the **Dependent List** box). This will get you separate tests of normality for each set of data. Report the Shapiro-Wilk *p*-value for each group (or condition) separately, maybe on the same line in your notes as the mean and standard error for a plot of the data. Then scroll up in the output to get the results from the *t*-test, using what is given on the row labeled "Equal variances not assumed." Remember to report the df of the *t*-test to two decimal places (even if -- by freakish accident -- it came out as a whole number), as a second reminder that equal variance was not being assumed.

Note: when equal variance is not being assumed, you can use the means and standard errors (for the two groups or conditions) from either **Analyze... Descriptive Stats... Explore** or from the "Group Statistics" table in the *t*-test output. They will be the same and both are correct.

In contrast, if you will be using the equal-variance-assumed output, because you're allowed to do this (by your advisor and target journal) and Levene's was not (close to) significant, then a combined test for normality is needed. This isn't simple, because you can't just run Shapiro-Wilk on the data variable (ignoring groups), because the groups might have different means, which can make the data bimodal, even if the spread of values within each group is perfectly normal. You need to remove the overall difference between the groups before testing normality. (You also want standard errors that are based on the combined or pooled error term and [annoyingly] these aren't provided by the *t*-test procedure, nor are they provided by **Analyze... Descriptive Stats... Explore**. In any event, you need to use something else to complete the analysis.)

The best way to do this is to re-run the *t*-test as an ANOVA, instead, because the ANOVA procedures have many more options. This is done using **Analyze... General Linear Model... Univariate**. Push the data variable into the **Dependent Variable** box and the control variable that defines the groups (or conditions) into the **Fixed Factor(s)** box. Then go into the **Save** sub-menu and click **Unstandardized** under **Residuals**. Then click **Continue** to exit the sub-menu. Then, for reasons that we'll get to in a moment, go into the **EM Means** sub-menu and push the control variable from left to right and then click **Continue**. Then run the ANOVA via **OK** but don't look at the output. Instead, immediately run a test of normality on the variable that SPSS just created, which will be called "RES_1" with **Factor List** empty (so only one test is conducted). This will get you the correct Shapiro-Wilk *p*-value for the pooled error term. Then scroll up to the table under "Estimated Marginal Means" for the means and standard errors for the plot. These standard errors are based on the pooled error term, which is what you want. (You don't want the standard errors from the *t*-test table, because those aren't based on the pooled error term.) Then scroll up some more to get the results from the *t*-test. Use the values for "Equal variances assumed."