

# Write Up: OSF T-Test

*Mena Whalen*

=====

## OSF Article

In the OSF article the authors used a paired t-test to compare the p-values from the original study with its replicate study for the 100 studies. The authors omit one case where the p-values were unavailable for both original and replicate. The test statistic for the paired t-test looks like  $t = \frac{\bar{X}_{Diff}}{\frac{SD_{Diff}}{\sqrt{k}}}$  where  $\bar{X}_{Diff}$  is the mean of all the differences between original and replicate p-values,  $SD_{Diff}$  is the standard deviation of those differences from the mean, and  $k$  is the number of pairs. The t-test is testing the hypothesis that these two groups are different from one another, meaning that the p-value from the first experiment is the same p-value as the replicate study. A paired t-test is used in this case since each study is being compared with its replicate. They are assumed to be controlling the particular experiment being conducted within each of the different groups, original and replicate.

For the 99 pairs available, the dependent t-test was done, and found a mean difference of -0.2738438, showing that the replicate p-values were larger than the original. The test statistic was -8.2068 thus having a p-value less than 0.0001. This means that the pairs are not the same or that their difference is equal to 0, thus the original studies p-values are not the same and the replicate studies p-values. Leading back to that the original studys' estimate is not the same as the replicates' estimate.

## Theory

Using the set up provided with  $T_i \sim N(\theta_i, v_i)$ , if a study is significant, meaning p-value is less than or equal to 0.05, then  $\frac{|T_i|}{\sqrt{v_i}} \geq 1.96$  if testing  $H_0 : \theta_i = 0$ . We assume that the variance is known for each study and is  $\frac{4}{n}$  where  $n$  is the sample size. This means that the test statistic used for determining the p-value is dependent on the size of the estimate and the sample size. This p-value is a random variable coming from the given data, the distribution of which is dependent on the power of the test. The power to detect if  $\theta_i$  is statistically different from 0 depends on the estimate  $T_i$  and the sample size  $n$ . Given that each original study and its replicate study had possibly different estimates and sample sizes to calculate their p-value means that the groups could have different statistical power for each test.

When comparing the p-values of two paired groups using the t-test uses both cases' estimate and the sample size of each test. In a single example, study 1, there would be  $p_1$  the p-value from the original and  $p_2$  the p-value from the replicate study. This is equivalent to  $p_1 = \Phi^{-1}\left(\frac{|T_1|}{\sqrt{v_1}}\right)$  and similar can be shown for  $p_2$ , so subtracting the two leads to  $\Phi^{-1}\left(\frac{|T_1|}{\sqrt{v_1}}\right) - \Phi^{-1}\left(\frac{|T_2|}{\sqrt{v_2}}\right)$ . This is dependent on the power of each statistical test using the  $T$ 's and  $n$ 's for each group. Scaling this for  $k$  studies, the mean of the paired difference of p-values would be

$$\frac{1}{k} \sum_j^k \left[ \Phi^{-1}\left(\frac{|T_{1j}|}{\sqrt{v_{1j}}}\right) - \Phi^{-1}\left(\frac{|T_{2j}|}{\sqrt{v_{2j}}}\right) \right] \quad (1)$$

This mean takes into account both the estimate and the sample size, thus the power to detect if the p-values are different from one another depends upon the power of both the original and the replicate studies.

In the current use the t-test with p-values, it is testing if the group of original studies are similar to the group of replicate studies over all studies using p-values to assess similarity. When examining all studies together it takes away from a single study being replicated. Each study is dependent upon the power of both original and replicate being able to detect an effect. This is not the same thing as testing if an individual study did replicate which would relate back in hypothesis test form to

$$H_0 : \theta_1 = \theta_2 \quad H_A : \theta_1 \neq \theta_2 \quad (2)$$

where  $\theta_1$  is the parameter from the original and  $\theta_2$  from the replicate.

The t-test of p-values has two types of errors that can occur, one where the effect sizes from original to replicate are not equal but the t-test says that they are not significantly different meaning the p-values are the same. The other is when the effect sizes are equal but the t-test concludes that the p-values are significantly different from one another meaning to do not replicate. The first instance, which we will call Case 1, can occur when  $\theta_{1j} \neq \theta_{2j}$  but the powers are the same,  $(1 - \beta_{1j}) = (1 - \beta_{2j})$ . Power is determined by the effect size and the sample size, when the effect size is small the sample size needs to be bigger to have higher power. If two effect sizes are different, but they have the same power, their test statistic from the original study's work and the replicate experiment would look identical,  $\frac{T_1}{\sqrt{V_1}} = \frac{T_2}{\sqrt{V_2}}$ , since we know the  $T_i$ 's are different from one another then the denominator would have to be scaled to be the same. The smaller effect size would have a larger sample size making the variance become smaller and the larger effect size would have a smaller sample size resulting in a larger variance, meaning a scaled variance,  $cV_{small} = V_{big}$ . Since both test statistics are nearly equal then their p-values will be similar, even if each estimate is widely different from one another. Resulting in the paired t-test determining that p-values from original to replicate are not significantly different from one another. The power of this test goes to the  $\alpha$  level of the test, normally 0.05.

The second type of error that can occur from using t-test of p-values, Case 2, is from equal effect sizes with different power. In a simple thought experiment if we compared one study's original and replicate findings using the t-test of p-values the test statistic would look like  $\frac{p_1 - p_2}{\sqrt{V_{p_1 - p_2}}}$ . Let the original study be fixed, then the sample size of the replicate study is let to increase off to  $\infty$  then the variance of the replicate is 0 then making the p-value,  $p_2$  go to 0. This makes sense in a real life scenario to get a more accurate estimate of the first finding by increasing the sample size. This leaves the test statistic only having the p-value from the original study resulting  $\frac{p_1}{\sqrt{V_{p_1}}}$ , this makes the power of the test go to 1. The test will rarely say that the two p-values are significantly similar to one another. Another way to examine it is through the framework presented for Case 1 but where the test statistics from each case is not equal,  $\frac{T_1}{\sqrt{V_1}} \neq \frac{T_2}{\sqrt{V_2}}$ . Since the numerator is the same for each but the denominator is different since each study has different power, meaning different samples sizes and thus different variances. This makes the p-values different from one another so the test will highly reject that they are the same, resulting in the power of the test going to 0.80.

## Results

The simulations used to demonstrate Case 1 and Case 2 are presented in Table 1 and Table 2 below. In each simulation the effect sizes used were small 0.2, medium 0.5, and large 0.8. The powers used were 40,60,80 to range all possible powers that can occur. In Case 1 the rows represent the power of both original and replicate studies and the columns are the different effect sizes of the studies. For example, section (1,1) is if the original study had effect size 0.2 and the replicate had effect size 0.5 or vice versa, and both had 80% power. All results show are around 0.05 or below given the discussion made previously about Case 1 being the same as the  $\alpha$  level. Notice that as the power decreases, going down the rows, the power of the test increases being capped at 0.05. This is from a higher power being able to detect effects easier than a lower powers when comparing two high power studies with different effect but similar power the p-values will be closer together.

For Case 2, the rows are the effect sizes for both original and replicate study and the columns are different power of each study. The result in section (1,1) would be the power of the t-test of p-values from the effect

size of 0.2 with one study having power 40% and the second study having power 60%. All results in the table are over 80% power as discussed in the previous section since the estimates are the same size but with different powers resulting in different p-values for each result in the test, thus the t-test says that the two studies do not replicate even with the same effect size. In the second column with power 40% and power 80% the results are almost 1 which is similar to the thought experiment presented earlier since the sample size of the 80% powered test is much larger than the sample size of the 40% powered test, making the higher powered test's variance go to 0. The differences between each column is from the different between each power with the power between 60% and 80% being the smallest since both will have fairly large samples and better accuracy to detect an effect.

	Theta 0.2/0.5	Theta 0.2/0.8	Theta 0.5/0.8
Power 80	0.04821	0.04793	0.04753
Power 60	0.05006	0.04978	0.04889
Power 40	0.05002	0.05014	0.05080

	Power 40/60	Power 40/80	Power 60/80
Theta 0.2	0.84874	0.99996	0.82533
Theta 0.5	0.85056	0.99996	0.82641
Theta 0.8	0.85050	1.00000	0.82416