McNemar Test and T-Test of P-Values

Mena Whalen

OSF Article

In the OSF article a multitude of methods were used to attempt to access replication. Two of those methods, McNemar's and t-test for paired data, was used with the p-value of the original and replicate studies to determine if the results were replicated. The two methods are discussed together by the similar properties of each test, like the power and sample size of each study determining the error rates of the test. McNemar's test in the context of replication, is used to determine if there is a difference between the proportion of statistically significant results among original studies and the proportion of statistically significant results among replication studies. The t-test when used for determining replication is testing wether or not the original p-value and the replicate p-value are significantly different from on another.

Consider k pairs of studies, each containing an original study and a replication study. As described earlier, the original studies are characterized by an estimate $T_{1j} \sim N(\theta_{1j}, v_{1j})$ for $j = 1, \ldots, k$. Similarly, the replication studies are characterized by an estimate $T_{2j} \sim N(\theta_{1j}, v_{1j})$ for $j = 1, \ldots, k$. Further, let p_{1j} denote the p-value of the j^{th} original study and p_{2j} denote the p-value of the j^{th} replication study. Each p-value can be categorized as either significant or non-significant, creating a dichotomous variable suitable for the use of McNemar's test. This information is generally summarized in a 2×2 contingency table as shown below where n_{11} denotes the number of study pairs in which both the original and replication findings are significant, n_{10} denotes the number of study pairs in which the original finding is significant but the replication finding is not, n_{01} denotes the number of study pairs in which the original finding is not significant but the replication finding is, and n_{00} denotes the number of study pairs in which neither the original finding nor the replication finding is significant.

	Replication Finding	
Original Finding	Significant	Non-Significant
Significant	n_{11}	n_{10}
Non-Significant	n_{01}	n_{00}

The null hypothesis for NcNemar's test states that the marginal probabilities for each outcome are the same. That is, $H_0 : p_{11} + p_{10} = p_{11} + p_{01}$ where the p's are the probabilities of occurrence in cells with the corresponding labels. The test statistic is given by $\chi^2 = \frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}}$. Under the null hypothesis, the test statistic follows a chi-squared distribution with one degree of freedom. The null hypothesis is rejected for large values of the test statistic. The results of the OSF article are summarized in the contingency table below. 97% of original findings were significant compared to 36% of replication findings.The authors found a test statistic of 59.06, rejected the null hypothesis ($p = 1.53 \times 10^{-14}$), and concluded that the findings from the original studies did not replicate.

	Replication Finding	
Original Finding	Significant	Non-Significant
Significant	35	62
Non-Significant	1	2

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 study. They are assumed to be controlling the particular experiment being conducted within each of the different groups, original and replicate.

The paired t-test was used to compare the p-value from the original study with its replicate study for the 100 studies. The authors omit one case where the p-value were unavailable for both 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 of Test

McNemar's Test

Two properties you might use to assess the quality of a statistical test are its Type I and Type II error rates. To better understand these properties, we begin by rewriting the null hypothesis for McNemar's Test to reflect its relationship to the power of the original and replication studies. First note that the null hypothesis can be rewritten as

$$H_0 : P(p_{1j} \leq 0.05) = P(p_{2j} \leq 0.05)$$

In other words, under the null hypothesis, if we were to draw a study pair at random, the probability of obtaining a significant result in the original study is equal to the probability of obtaining a significant result in the replication study. Taking this observation into consideration and noting that $p_{ij} \leq 0.05 \equiv \frac{|T_{ij}|}{\sqrt{v_{ij}}} \geq 1.96$, we can reformulate the null hypothesis as follows.

$$H_0 : \sum_{j=1}^{k} P\Big(\frac{|T_{1j}|}{\sqrt{v_{1j}}} \geq 1.96|i\Big)P(i) = \sum_{j=1}^{k} P\Big(\frac{|T_{2j}|}{\sqrt{v_{2j}}} \geq 1.96|i\Big)P(i)$$

That is, we can think of McNemar's test as a test of difference in the average power rather than a test of difference in paired proportions. As such, a rejection of the null hypothesis is akin to concluding that there is a statistically significant difference between the average power of the original studies and the average power of the replication studies. However, as we will show, this does not necessarily imply that the original findings were not replicated in the replication studies, if we take replication to be the case in which $\theta_{1j} = \theta_{2j} \forall j$. Similarly, we will show that a failure to reject the null hypothesis for McNemar's test does not necessarily imply that the findings did, in fact, replicate.

Assuming the true null hypothesis for replication is $H_0 : \theta_{1,j} = \theta_{2,j}$, a Type I error occurs when the original and replications studies differ in their power to detect their common effect size. In this case, McNemar's test will almost always reject its null hyppothesis, leading to conclusions that original and replication findings do not replicate, despite having identical true effects. The power of each of the $2k$ original and replication studies is determined by the true effect size (θ_{ij}), the sample size (n_{ij})\$, and the population variance. If the original and replication studies have the same true effect size and population variance, we can let their powers differ by increasing the sample size of the replication studies. As n_{2j} grows large, $\frac{|T_{2j}|}{\sqrt{v_{2j}}}$ will also grow large. If we assume that original and replication studies are independent then,

$$n_{10} = \sum_{j=1}^{k} \mathbf{1}\Big\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} \geq 1.96\Big\}\mathbf{1}\Big\{\frac{|T_{2j}|}{\sqrt{v_{2j}}} < 1.96\Big\} \to 0$$

$$n_{01} = \sum_{j=1}^{k} \mathbf{1}\Big\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} < 1.96\Big\}\mathbf{1}\Big\{\frac{|T_{2j}|}{\sqrt{v_{2j}}} \geq 1.96\Big\} \to \sum_{j=1}^{k} \mathbf{1}\Big\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} < 1.96\Big\}$$

This implies that the McNemar's test statistic $X^2 \to \sum_{j=1}^{k} \mathbf{1}\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} < 1.96\}$ as the replication sample sizes grow large. Let's assume that all k original studies have 80% power to detect their true effect. Then, X^2 approaches $0.2k$ as the replication sample sizes grow large. Recall that McNemar's test rejects if $X^2 > \chi^2_{[1]}$. So, as long as $k > 2$, McNemar's test will always reject despite the true effect sizes being identical for all k study pairs.

A Type II error occurs when the original and replication studies have different true effects (i.e. $\theta_{1j} \neq \theta_{2j}$) but have the same power to detect those effects (i.e. $(1 - \beta_{1j}) = (1 - \beta_{2j})$). The impact of the power of the $2k$ original and replication studies on this test's Type II error rate is easiest understood through a consideration of the test power function. The power function of McNemar's test is given by

$$\Phi\Big\{\frac{(p_{10} - p_{01})\sqrt{k} - z_{1-\alpha/2}\sqrt{p_{10} + p_{01}}}{\sqrt{p_{10} + p_{01} - (p_{10} - p_{01})^2}}\Big\}$$

where p_{10} and p_{01} can be written as $\sum_{j=1}^{k}(1 - \beta_{1j})\beta_{2j}$ and $\sum_{j=1}^{k}\beta_{1j}(1 - \beta_{2j})$, respectively. When the power of the original study is equal to the power of the replication study for all k studies, $p_{10} = p_{01}$. The power of McNemar's test can thus be simplified to $\Phi\{-z_{1-\alpha/2}\}$. Assuming that $z = 1.96$, we know the power of McNemar's is 0.058. Thus, when the true effects of the original and replication studies are, in fact, different, McNemar's test will only reject the null hypothesis 5% of the time. In other words, the Type II error rate is 95%.

T-Test

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 θ_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 orginal and p_2 the p-value from the replicate study. This is equivalent to $p_1 = \Phi^{-1}\Big(\frac{|T_1|}{\sqrt{v_1}}\Big)$ and similar can be shown for p_2, so subtracting the two leads to $\Phi^{-1}\Big(\frac{|T_1|}{\sqrt{v_1}}\Big) - \Phi^{-1}\Big(\frac{|T_2|}{\sqrt{v_2}}\Big)$. 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

$$\frac{1}{k}\sum_{j}^{k}\Big[\Phi^{-1}\Big(\frac{|T_{1j}|}{\sqrt{v_{1j}}}\Big) - \Phi^{-1}\Big(\frac{|T_{2j}|}{\sqrt{v_{2j}}}\Big)\Big] \tag{1}$$

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

$$H_0 : \theta_1 = \theta_2 \qquad H_A : \theta_1 \neq \theta_2 \tag{2}$$

where θ_1 is the parameter from the original and θ_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 second type of error that can occur from using t-test of p-values, Type I, 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 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 ∞ 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 $\frac{p_1}{V_{p_1}}$, this makes the error rate 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 from the test statistics from each case being 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 error rate of the test going to 0.80.

The Type II error rate 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 error rate of this test goes to the α level of the test, normally 0.05.

Results

The simulations used to demonstrate Type I and Type II error of the t-test 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 each simulation three estimates were drawn one hundred times, given the stated effect size, and their p-values were calculated based on the effect size and stated power rate. Then a combination of different effect size's p-values were subtracted from one another since the test is for paired data. For example small effect size minus a large effect size, a small minus a medium effect size, and a medium minus a large effect size. Those three differences of p-values are then calculated into a t-test statistic, $t = \frac{\bar{X}_{Diff}}{\frac{SD_{Diff}}{\sqrt{100}}}$. These test statistics are then stored and the process is repeated a thousand times. Each test statistic's absolute value is then compared with a critical point, in this instance 1.98, to see if it is larger than the critical value. Those counts of the test statistic being larger are then divided by the number of times the t-test is repeated, this instance a thousand. The table error rates is bootstrapped 100 times to get standard errors for each combination of effect sizes and power rates.

For Type I, the rows are the effect sizes for both orginal 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 dicussed 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 eariler since the sample size of the 80% powered test is much larger then 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.

In Type II 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 vise 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 α 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.

	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