Properties of Methods for Assessing Replication

Sarah Peko-Spicer

February 7, 2018

A framework for replication

Let there be k pairs of studies each containing an original study and a replication study. Each pair of studies can be characterized by a vector of true effect parameters $(\theta_{1,i}, \theta_{2,i})$ that are unobserved with zero sampling variance. Each pair of studies is further associated with estimates $(T_{1,i}, T_{2,i})$ of these true effects. These estimates have sampling variances $V_{1,i}$ and $V_{2,i}$ which are assumed to be known. Let $T_{j,i} \sim N(\theta_{j,i}, V_{j,i})$ for $j = 1, 2$ and $i = 1, 2, \ldots, k$. Further, assume that $T_{j,i}$ are independent. Generally speaking, two studies replicate if $\theta_{1,i} \approx \theta_{2,i}$.

Assessing replication via significance and p-values

The authors propose four different tests for assessing replication based on significance and p-values. These methods include Fisher's method, McNemar's test, the Wilcoxon signed-rank test, and the t test for dependent samples. In what follows, we derive properties of each of these tests and discuss how they function within our proposed framework for replication.

The authors assume a two-sided hypothesis test for each of the $2k$ studies with a significance level of 0.05. Let $p_{1,i}$ denote the p-value of the i$^{\text{th}}$ original study and let $p_{2,i}$ denote the p-value of the i$^{\text{th}}$ replication study.

McNemar's Test

McNemar's test is typicaly used to test the difference in proportions in paired data. In this case, the test is used to determine if there is a statistically significant difference between the proportion of statistically significant results among original studies and the proportion of statistically significant results among replication studies. Applying this test requires categorizing each $p_{j,i}$ as either significant (1) or non-significant (0). That is, each pair of studies is associated with a vector of p-values $(p_{1,i}, p_{2,i})$ that can take one of the following pairs of vlaues: (0,0), (0,1), (1,0), (1,1). This information is generally summarized in a 2 × 2 contingency table containing counts of occurrences in the data of these four pairs of values.

The null hypothesis for this test is one of marginal homogeneity. That is, under the null hypothesis, the marginal probabilities for each outcome are the same. In our case, this can be stated as

$$\begin{aligned} H_0 &: P(p_{1,i} \leq 0.05 \cap p_{2,i} \leq 0.05) + P(p_{1,i} \leq 0.05 \cap p_{2,i} > 0.05) \\ &= P(p_{1,i} \leq 0.05 \cap p_{2,i} \leq 0.05) + P(p_{1,i} > 0.05 \cap p_{2,i} \leq 0.05) \end{aligned}$$

This can be be further simplified to

$$\begin{aligned} H_0 &: P(p_{1,i} \leq 0.05) = P(p_{2,i} \leq 0.05) \\ H_1 &: P(p_{1,i} \leq 0.05) \neq P(p_{2,i} \leq 0.05) \end{aligned}$$

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 $p_{j,i} \leq 0.05 \equiv \frac{|T_{j,i}|}{\sqrt{V_{j,i}}} \geq 1.96$, we can rewrite the null hypothesis as follows.

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

So, we can reformulate this test of difference in paired proportions into a test of difference in average power. A rejection of the null hypothesis leads to the conclusion that there is a statistically significant difference between the average power of the original studies and the average power of the replication studies. Given that the power of any of the $2k$ studies depends on a variety of factors including sample size, true effect size, and population variance, it should be clear that a rejection of the McNemar null does not immediately imply that, on average, the studies do not replicate. Conceivably, the true effect sizes can be quite similar within study pairs but differences in sample size or variance between original and replication studies would lead to a rejection of the McNemar null. This suggests that equality in average power is perhaps not the best metric for assessing replication, or even rates of replication.

It is also worth exploring the power of McNemar's test, particularly in terms of its relationship to the power of the $2k$ studies. The power for McNemar's test is given by

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

Noting that $p_{10} - p_{01} = (p_{10} + p_{11}) - (p_{01} + p_{11})$ and that $p_{01} + p_{10} = (p_{10} + p_{11}) + (p_{01} + p_{11}) - 2p_{11}$, we can rewrite the power function as

$$\Phi\Big(\frac{(p_{1+} - p_{+1})\sqrt{k} - z_{1-\alpha/2}\sqrt{p_{+1} + p_{1+} - 2p_{11}}}{\sqrt{p_{+1} + p_{1+} - 2p_{11} - (p_{1+} - p_{+1})^2}}\Big)$$

where $p_{1+} = P(p_{1,i} \leq 0.05)$ and $p_{2+} = P(p_{2,i} \leq 0.05)$ for a randomly drawn i.