Properties of Methods for Assessing Replication

Hedges Working Group

February 3, 2018

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 and have zero sampling variance. Each pair of studies is further associate with estimates of these true effects: $(T_{1,i}, T_{2,i})$. 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, we assume that $T_{j,i}$ are independent.

For the purposes of exploring the proposed methods of assessing replication in the Open Science Collaboration article (2015), we assume a two-sided hypothesis test for each of the $2k$ studies with a significance level of 0.05. Thus, we add another dimension by which we can characterize the study pairs: their p-values. Let $p_{1,i}$ denote the p-value of the i^{th} original study and $p_{2,i}$ denote the p-value of the i$^{\text{th}}$ replication study.

1 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 as they pertain to a particular framework for assessing replication.

The authors assume a two-sided hypothesis test for each of the $2k$ studies with a significance level of 0.05. For some of the proposed methods, the resulting p-values are classified as either significant ($p_{j,i} \leq 0.05$) or non-significant ($p_{j,i} > 0.05$).

1.1 McNemar's test

McNemar's test tests the hypothesis that the proportion of statistically significant results among original studies is equal to the proportion of statistically significant results among the replication studies. Applying this test requires categorizing each $p_{1,i}$ and $p_{2,i}$ 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 values: $(0,0), (0,1), (1,0), (1,1)$. This information is generally summarized in a 2×2 contingency table containing counts of occurrences of these four pair values in the data.

	Rep Sig	Rep Non-Sig
Orig Sig	# (1,1)	#(1, 0)
Orig Non-Sig	# (0, 1)	# (0,0)

The null hypothesis for McNemar's test is $H_0 : P((1,1) \cap (1,0)) = P((1,1) \cap (0,1))$ which can conveniently be rewritten as $H_0 : P(p_{1,i} \leq 0.05) = P(p_{2,i} \leq 0.05)$ for any study i, randomly drawn from the k studies. Note that this null hypothesis can be translated or mapped onto our replication framework which does not include p-values. That is, by 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:

$$
\begin{aligned}
H_0 : P\Big(\frac{|\tilde{T}_{1,i}|}{\sqrt{\tilde{V}_{1,i}}} \geq 1.96\Big) &= P\Big(\frac{|\tilde{T}_{2,i}|}{\sqrt{\tilde{V}_{2,i}}} \geq 1.96\Big) \\
&\equiv \\
H_0 : \sum_{i=1}^{k} P\Big(\frac{T_{1i}}{V_{1i}} \geq 1.96|i\Big)P(i) &= \sum_{i=1}^{k} P\Big(\frac{T_{2i}}{V_{2i}} \geq 1.96|i\Big)P(i)
\end{aligned}
\tag{1}
$$

Thus, we can reformulate the null hypothesis for this particular test as a comparison of the average power of the original studies to the average power of the replicate studies.

2 Fisher's Method

The authors also propose assessing replication by applying Fisher's method on only the "nonsignificant P values of the replication studies." Using our notation, the hypotheses for this particular test can be defined as follows:

$$
\begin{aligned}
&H_0 : \theta_{2,1} = \theta_{2,2} = \cdots = \theta_{2,k} \\
&H_1 : \text{at least one of } \theta_{2,i} \text{ is not } = 0
\end{aligned}
\tag{2}
$$

When the null hypotheses are true, the test statistic $-2\sum_{i=1}^{k} ln(p_{2,i})$ follows a χ^2 distribution with $2k$ degrees of freedom. This result depends on the fact that $p_{2,i} \sim U(0,1) \forall i$.

Given the constraints presented by the authors (i.e. including only those p-values that are > 0.05), the proposed test statistic is not appropriate. In essence, $-ln(p_{2,i}) \sim TEXP(1,3)$ which cannot possibly have a chi-squared distribution once scaled since the support for this statistic has an upper bound. That being said, we can derive a test statistic that *does* follow a chi-squared distribution. If we let $p_{2,i} \sim U(0.05,1)$, then $-ln\Big(\frac{X-0.05}{0.95}\Big) \sim Exp(1)$. It follows that $-2\sum_{i=1}^{k} ln\Big(\frac{p_{2,i}-0.05}{0.95}\Big) \sim \chi^2_{2k}$ where k is the number of tests being combined. The method as proposed by the author's will always underestimate the size of the test statistic, making it more likely that using this method, you will fail to reject the null hypothesis than when using the corrected test statistic.