Metrics for Assessing Replication: Statistical Significance

STAT 461

Winter 2018

The Open Science Collaboration (2015), when evaluating if findings have been replicated argue, "There is no single standard for evaluating replication success." They analyzed replicate pairs in a variety of ways.

Correspondence of Significance

> Assuming a two-tailed test and significance or a level of 0.05, all test results of original and replication studies were classified as statistically significant ($p \leq 0.05$) and nonsignificant ($p > 0.05$). However, original studies that interpreted nonsignificant p values as significant were coded as significant (four cases, all with p values < 0.06).

The quote above suggests one metric of assessing replication is to determine if findings are statistically significant or not. The OSC is not alone in considering this metric (see Steiner and Wong, 2016; Valentine, 2012) By this standard, the OSC concluded that over 60% of their attempted replications failed.

However, it is not clear how exactly to interpret these results. For one, the difference in between statistically significant and nonsignificant findings may not reflect meaningful scientific differences in effects. As well, the properties of this as an inference procedure—e.g., error rates—are necessary to understand the OSC findings. How often, for instance, might this analysis construe negligible scientific differences between studies as nonreplication?

Using a statistical framework motivated by meta-analysis, this vignette attemts to clarify subjective ideas of replication as implied by comparing statistical significance. It describes the types of definitions of replication such analyses might correspond to, and delineates some of the statistical properties in those scenarios.

Statistical Model and Analysis

We can represent the true underlying scientific effect parameter in study i as θ_i, which we refer to as the study's results. This is the value that would be observed if study i had perfect precision (zero measurement and sampling variance). We do not actually observe θ_i, but instead observe an estimate T_i that has some variance v_i—due, for example, to sampling. Four simplifying assumptions used in this vignette are that T_i are independent, unbiased, normally distributed, with known sampling variance v_i:

$$T_i \overset{indep}{\sim} N(\theta_i, v_i)$$

In general, we may let i range from 1 (the original study) up to an arbitrary number $k > 1$ (to reflect $k - 1$ replication attempts), but to evaluate the analysis by the OSC, we set $k = 2$. Thus $i = 1$ corresponds to the original study and $i = 2$ refers to the (single) replicate study.

We would argue that replication can be defined in terms of underlying effect parameters θ_i. They may vary due to differences in sample compositions—studies i and j sample from different populations—or experimental contexts—minor variations in the way studies i and j carry out the experiment. However, if both are ignoreable (or nearly so), then the effect parameters ought to be (nearly) the same. Conversely, estimates T_i may vary due to random chance; that is, even if studies share the same underlying effect θ, estimates will differ due to variation from sampling. Thus, as a guiding principle, we would argue that coherent scientific and statistical definitions of replication should involve similarty between the θ_i.

Comparison of statistical significance involves computing p-values for the replicates and determining if both are either nonsignificant $p > 0.05$ or significant $p \leq 0.05$. As stated above, this is the totality of this analysis, and we can summarize this metric as

$$\mathbf{1}\{p_1 \leq 0.05 \wedge p_2 \leq 0.05\} + \mathbf{1}\{p_1 > 0.05 \wedge p_2 > 0.05\} \tag{1}$$

Under the model (and simplifying assumptions), we can rewrite this as

$$\mathbf{1}\left\{\frac{|T_1|}{\sqrt{v_1}} \geq 1.96 \wedge \frac{|T_2|}{\sqrt{v_2}} \geq 1.96\right\} + \mathbf{1}\left\{\frac{|T_1|}{\sqrt{v_1}} < 1.96 \wedge \frac{|T_2|}{\sqrt{v_2}} < 1.96\right\} \tag{2}$$

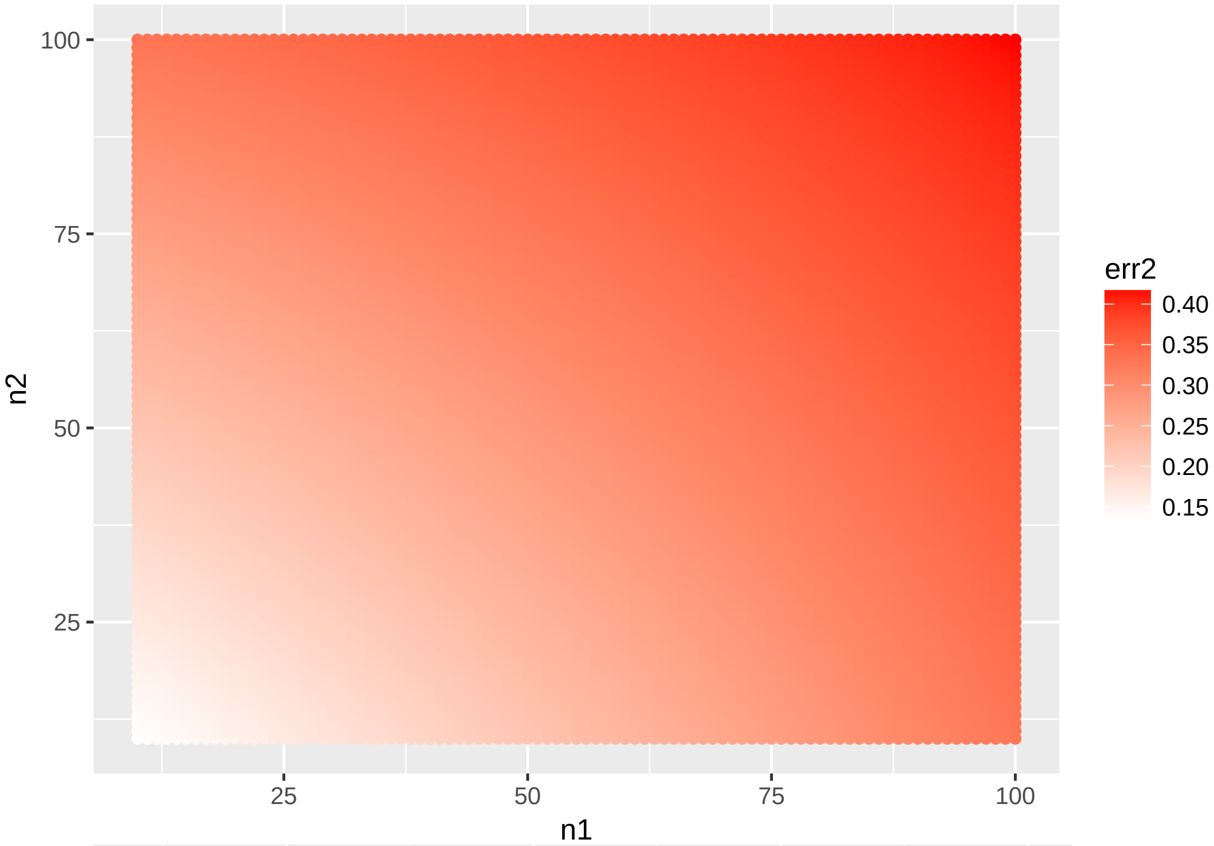
A type I error occurs when one study is significant, and the other is not (assuming that the studies replicate). Then, the probability of a false positive is given by

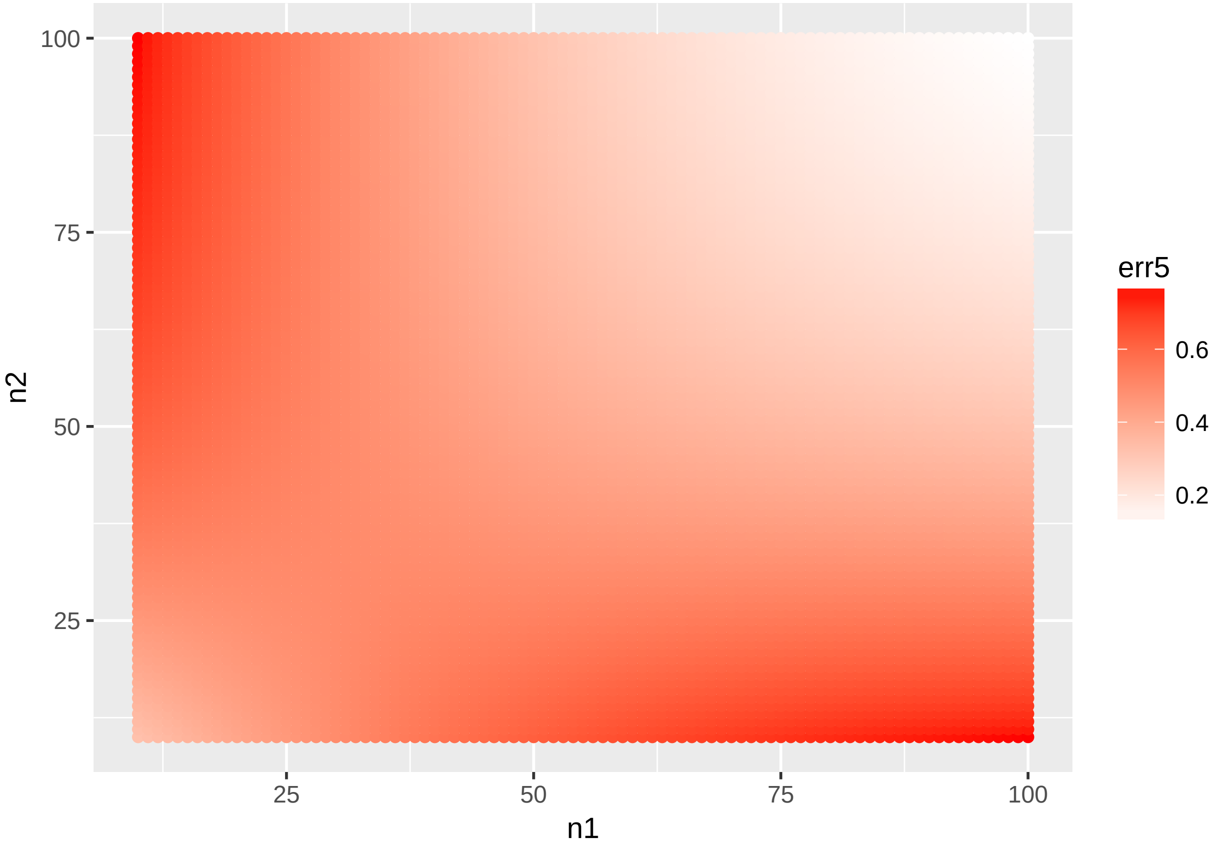
$$\begin{aligned}
P\left[\text{Type I Error}\right] &= P\left[\frac{|T_1|}{\sqrt{v_1}} \geq 1.96 \wedge \frac{|T_2|}{\sqrt{v_2}} < 1.96\right] + P\left[\frac{|T_1|}{\sqrt{v_1}} > 1.96 \wedge \frac{|T_2|}{\sqrt{v_2}} \geq 1.96\right] \\
&= P\left[\frac{|T_1|}{\sqrt{v_1}} \geq 1.96\right] P\left[\frac{|T_2|}{\sqrt{v_2}} < 1.96\right] + P\left[\frac{|T_1|}{\sqrt{v_1}} > 1.96\right] P\left[\frac{|T_2|}{\sqrt{v_2}} \geq 1.96\right] \\
&= \left[\Phi\left(-1.96 - \frac{\theta_1}{\sqrt{v_1}}\right) + 1 - \Phi\left(1.96 - \frac{\theta_1}{\sqrt{v_1}}\right)\right]\left[\Phi\left(1.96 - \frac{\theta_2}{\sqrt{v_2}}\right) - \Phi\left(-1.96 - \frac{\theta_2}{\sqrt{v_2}}\right)\right] + \\
&\quad \left[\Phi\left(1.96 - \frac{\theta_1}{\sqrt{v_1}}\right) - \Phi\left(-1.96 - \frac{\theta_1}{\sqrt{v_1}}\right)\right]\left[\Phi\left(-1.96 - \frac{\theta_2}{\sqrt{v_2}}\right) + 1 - \Phi\left(1.96 - \frac{\theta_2}{\sqrt{v_2}}\right)\right]
\end{aligned}$$

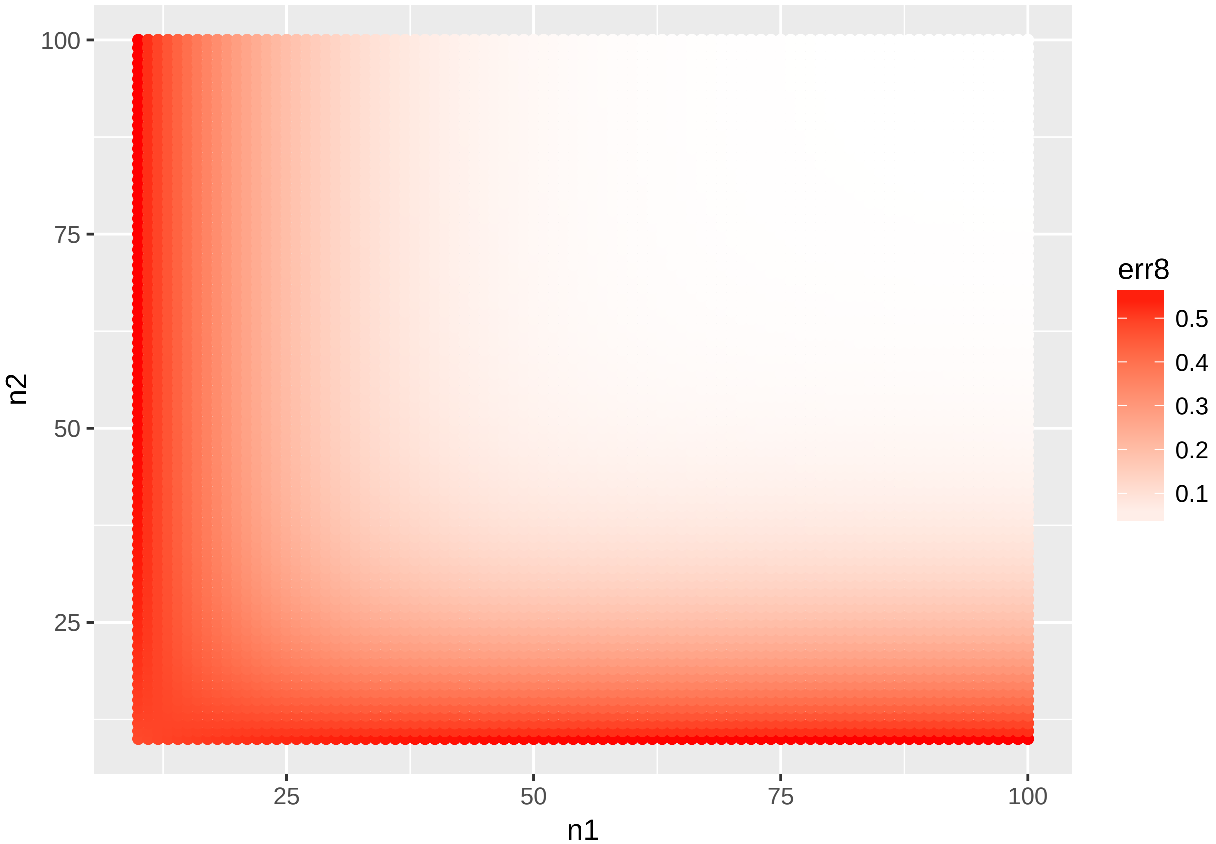
Note that when both studies replicate exactly, when $\theta_1 = \theta_2 = 0$, so that both studies have null effects, then the expression above reduces to

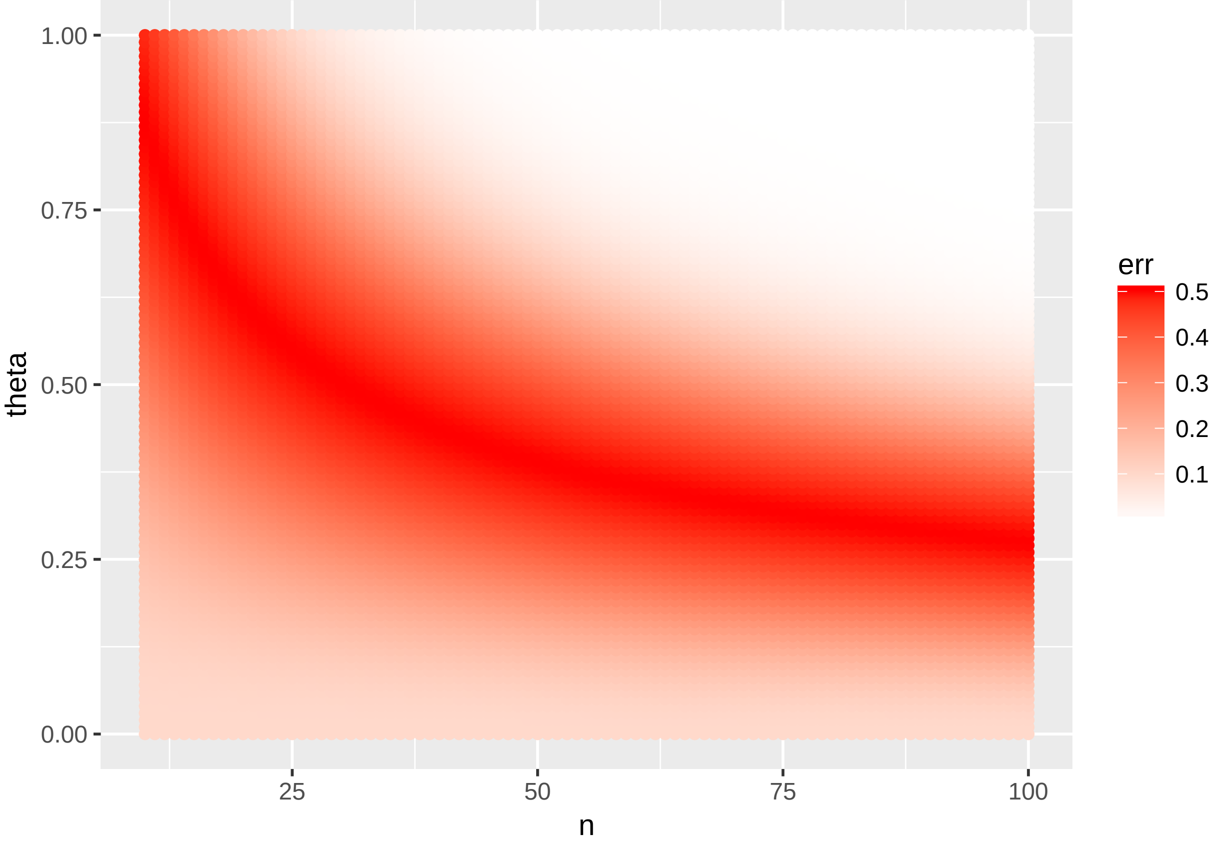
$$\left[\Phi\left(-c_\alpha\right) + 1 - \Phi\left(c_\alpha\right)\right]\left[\Phi\left(c_\alpha\right) - \Phi\left(-c_\alpha\right)\right] = 2\alpha - 4\alpha^2$$

Thus, if both significance tests are at the $\alpha = 0.05$ level, then the type I error rate in this case is 0.09.


100
75
50
25
n2
25
50
75
100
n1
err2
0.40
0.35
0.30
0.25
0.20
0.15


100
75
50
25
n2
25
50
75
100
n1
err5
0.6
0.4
0.2





However, others have noted that statistical significance may not correspond to replication if the significant estimates do not share the same sign. That is, if study 1 has a positive significant estimate and study 2 has a negative and significant estimate, one likely would not conclude they replicate. Thus, an alternative to (1) could be

$$\mathbf{1}\{p_1 \leq 0.05 \wedge p_2 \leq 0.05\}\mathbf{1}\{\text{sign}(T_1) = \text{sign}(T_2)\} + \mathbf{1}\{p_1 > 0.05 \wedge p_2 > 0.05\} \tag{3}$$

These are related analyses that have slightly different properties, which we illustrate below.