A prediction model for understanding statistical replication

Andrew Neath*

Abstract

There is growing concern over the number of scientific findings that fail when replication is attempted. Traditional statistical inference is designed as a look back to how data originates. Perhaps we also need to look ahead in anticipation of what data we will see next. Through the use of a Bayesian prediction model, this paper seeks to determine what can reasonably be expected to occur in a replication trial.

Key Words: Bayesian inference, hypothesis testing, statistical evidence, reproducible research, scientific method

1. Introduction

There is growing concern over the number of scientific findings that fail when replication is attempted. Indeed, when repeated experimentation leads to distinct versions of "truth", the scientific method as a whole becomes suspect. The question of how best to understand and analyze replication studies has drawn great interest within the statistical community. At the 2017 Joint Statistical Meetings, over 25 presentations featured replication or reproducibility in the title or as a key word.

In this paper, we take the position that the random behavior governing replicated experiments is misunderstood by scientists in general, and perhaps by some statisticians as well. Consider a quote from legendary statistician John Tukey (1991): "Truly solid knowledge does not come from analyzing a single experiment, but from a demonstrated ability to repeat experiments, each of which shows confident direction at a reasonable error rate." In other words, a statistically significant result alone is not enough to draw a final conclusion. It is only after we see statistical significance across repeated experiments that we are willing to take a finding as truth. For a scientific claim to be established as credible, we must have the same finding appear in a replication attempt. However, this rarely occurs in practice. Open Science Collaboration (2015) attempted to replicate 100 prominent studies in psychology, with only 39 repeating the outcome of statistical significance. Replication attempts in the health sciences do not fare any better (Begley and Ellis, 2012). Is the problem with the science, or with unrealistic expectations for replicated experiments?

So, what is realistic for a replicated study? To answer, we will pose the question as a problem in prediction. With the use of a simple Bayesian model, we will show how the information from a single study can be used to determine what can reasonably be expected to occur in a replication trial. Bayesian methods handle prediction problems more naturally than other frameworks; thus our choice of a Bayes model. It should be noted that the lessons learned in this inquiry can be seen to translate across statistical frameworks, including frequentist approaches, information theoretic methods, and more intricate Bayesian assumptions.

*Southern Illinois University Edwardsville, College of Arts and Sciences, aneath@siue.edu

2. Predicting a Replication Trial

Consider an experiment conducted to study some scientific phenomenon, followed by a replication study where the intention is to either verify of refute the original result. Suppose the experimental results can be summarized by the observed effect sizes and their standard errors. Let M_1, M_2 represent the sample effect sizes for the original trial, and the replication trial, respectively. Let σ_1, σ_2 represent the standard errors of estimation for the respective trials. For simplicity, assume σ_1, σ_2 are equal, and that the common σ is known. We are assuming the trials are performed independently, under identical conditions. So, the true underlying effect size θ is constant across experiments.

Our model and notation is borrowed from Senn (2002). Take

$$M_1, M_2 \,|\, \theta \sim N\left(\theta, \sigma^2\right).$$

We are considering the problem of predicting the results of the replication trial, based on the observed results of the original experiment. Take a "zero weight" noninformative prior on θ. The only information used in our model is from the experimental results. In this way, we are mimicking the thinking of an experimenter that is unaware (intentionally or not) of any outside evidence connected to their study. Let m_1 denote the observed effect size for the first trial. (Without loss of generality, take $m_1 > 0$).

Although our approach is Bayesian, we would like our focus to be on the p-value, the most widely used measure of evidence. We can define a Bayesian analog to the p-value through the posterior probability of a type S error. A type S, or sign, error occurs when the observed effect size m is in the opposite direction of the true effect size θ (Gelman and Tuerlinkx, 2000). Since $m_1 > 0$, the initial observed effect is in the positive direction. A type S error occurs if the true effect is actually in the negative direction. Define

$$p_1 = P\left(\theta < 0 \,|\, m_1\right)$$

as the posterior probability of a type S error. Under our modeling conditions, one can establish the posterior distribution on θ as

$$\theta \,|\, m_1 \sim N\left(m_1, \sigma^2\right).$$

Thus, the posterior probability of a type S error is

$$p_1 = \Phi\left(\frac{0 - m_1}{\sigma}\right) = \Phi\left(-z_1\right),$$

the (one-sided) p-value for testing $H_o : \theta \leq 0$, $H_1 : \theta > 0$ with $z_1 = m_1/\sigma$, the standardized test statistic using results from the original trial. Let

$$\begin{aligned} P_2 &= P\left(\theta < 0 \,|\, M_2\right) \\ &= \Phi\left(-Z_2\right) \end{aligned}$$

denote the corresponding p-value from the replication trial, where $Z_2 = M_2/\sigma$. As the replication trial is the focus of our prediction problem, M_2 is not yet observed. Thus, P_2 is a random variable. The predictive distribution for M_2 given m_1 is established to be

$$M_2 \,|\, m_1 \sim N\left(m_1, 2\sigma^2\right).$$

The predictive distribution for Z_2 given z_1 follows directly as

$$Z_2 \,|\, z_1 \sim N\left(z_1, 2\right). \tag{1}$$

Table 1: Prediction Intervals for Replication P-values

p_1	z_1	95% p.i. for P_2
.05	1.645	$[.000, .870]$
.025	1.960	$[.000, .792]$
.005	2.576	$[.000, .578]$
.0001	3.719	$[.000, .172]$

We can summarize the predictive distribution for Z_2 using a 95% prediction interval. From (1), we get the interval

$$z_1 \pm 1.96\sqrt{2}. \tag{2}$$

The effect size m_1, the test statistic z_1, and the p-value p_1 provide equivalent information regarding the outcome of the first trial. We can derive the posterior predictive distribution for P_2 given p_1 from (1). A 95% prediction interval for P_2 can be derived from (2) as

$$\left[\Phi\left(-z_1 - 1.96\sqrt{2}\right), \Phi\left(-z_1 + 1.96\sqrt{2}\right)\right].$$

It is our interest to see how closely we are able to predict the replication p-value P_2 based on the p-value from the original trial. Table 1 displays the 95% prediction interval for P_2, conditional on the result from the original trial.

What stands out is the sizable range of plausible outcomes for the replicated trial. Taking a statistically significant result as our lone source of information is not enough to accurately predict the outcome of the next trial. As the examples in Table 1 illustrate, it should not be surprising for an original trial and a replication trial to differ in their findings.

The fact that p-values exhibit greater variability than may be appreciated has been revealed in various other ways. See Boos and Stefanski (2010) for a mathematical demonstration and Cumming (2012) for a graphical demonstration. In the study of statistical replication, the variability inherent to a replication trial does provide us with an explanation for why scientific findings seem to suffer from a failure to replicate. But it also leaves us in an uncomfortable position. How does science move forward when a replication study, even under identical conditions, can be so different from an original study? In the next section, we explore a possible answer to this important question.

3. Combining Information Across Trials

Let's now consider a different way to think of a successful replication. Rather than treating the trials as separate entities, consider how the trials combine to provide information about the true underlying effect. We can define a p-value for the combined trials, under the Bayesian model developed in Section 2, as

$$P_{12} = P\left(\theta < 0 \,|\, m_1, M_2\right).$$

Recall that P_1 is the probability of a type S error, given the result from the original trial, where we observed a positive effect m_1. So, P_{12} is the probability of a type S error, given the results from both trials. Recall that M_2 is not yet observed, and is treated as a random variable. The direction of the sign error for the combined trials is determined by the direction of the effect observed initially.

Table 2: Predicted Probabilities of Replication Success

p_1	z_1	$P_{success}$
.05	1.645	.752
.025	1.960	.792
.005	2.576	.857
.0001	3.719	.938

Under the zero weight prior on effect size θ, we get a posterior conditional on both experimental results as

$$\theta \,|\, m_1, M_2 \sim N\left(\frac{m_1 + M_2}{2}, \frac{\sigma^2}{2}\right). \tag{3}$$

From (3), we can derive

$$P_{12} = \Phi\left(-Z_{12}\right)$$

where

$$Z_{12} = \frac{z_1 + Z_2}{\sqrt{2}}$$

is the standardized test statistic when experimental results are combined. If evidence in favor of the effect observed in the first trial has increased based on the result from the replication trial, it seems reasonable to call this a successful replication. We can say evidence has increased when the probability of a type S error has decreased. Then $P_{12} < p_1$ is the condition to be met for a success. We write an equivalent condition as $Z_{12} > z_1$, which reduces to the inequality

$$Z_2 > \left(\sqrt{2} - 1\right) z_1.$$

Since $\left(\sqrt{2} - 1\right) \approx 0.41$, we only need a z statistic in the second trial that is 41% of the original. The requirement that we need a repeat of statistical significance is too strong of a condition for success in replication. There is a greater opportunity to add to the existing evidence than may be appreciated.

Let's bring prediction back into the problem. Once an original experimental result is observed, we would like to know the probability that the replication attempt will be successful where success is defined as an increase in evidence toward the originally observed effect. Using the predictive distribution in (1), we get

$$\begin{aligned} P_{success} &= P\left(Z_{12} > z_1 \,|\, z_1\right) \\ &= P\left(Z_2 > \left(\sqrt{2} - 1\right) z_1 \,|\, z_1\right) \\ &= \Phi\left(\left(\sqrt{2} - 1\right) z_1\right). \end{aligned}$$

Table 2 displays the predictive probability of replication success, conditional on the result for the initial trial.

Even for results which are marginally significant in the traditional sense, the probability of successfully adding to the existing pool of evidence is moderately large. For highly significant initial results, there is high probability of replication success. Viewing replication success through the lens of combined information tells a different story than the uproar over the replication crisis in science would suggest.

The need for a suitable definition of replication success is vital for understanding the statistical issues when experiments are repeated. See Patil, Peng, and Leek (2016) for a more detailed exposition, but with a similar conclusion to the one reached from our simple Bayesian model.

4. Conclusion

There is a counterintuitive nature to the statistical replication problem. We would like to believe that an experiment that is well-designed and carefully performed will give similar results across repeated trials. However, the natural variability between experimental results can be very large. We should not necessarily expect experimental results to replicate in the traditional sense. Since we are trained to put our trust into the scientific method, this realization may be uncomfortable and disconcerting. But there is a balancing aspect to the replication problem. It turns out there is also great flexibility in combining seemingly disparate experimental outcomes. Science does work, it's just that the march of science in the discovery of truth is not as linear as we may have previously believed.

REFERENCES

Begley, C. and Ellis, L. (2012), "Drug development: Raise standards for preclinical research," *Nature*, 483, 531-533.

Boos, D. and Stefanski, L. (2011), "P-value precision and reproducibility," *The American Statistician*, 82, 112-122.

Cumming, G. (2012), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, New York: Taylor and Francis.

Gelman, A. and Tuerlinkx, F. (2000), "Type S error rates for classical and Bayesian single and multiple comparison procedures," *Computational Statistics*, 15, 373-390.

Open Science Collaboration (2015), "Estimating the reproducibility of psychological science," *Science*, 349, aac4716.

Patil, P., Peng, R., and Leek, J. (2016), "What should we expect when we replicate? A statistical view of reproducibility in psychological science," *Perspectives in Psychological Science*, 11, 539-544.

Senn, S. (2002), "A comment on replication, p-values and evidence," *Statistics in Medicine*, 21, 2437-2444.

Tukey, J. (1991), "The philosophy of multiple comparisons," *Statistical Science*, 6, 100-116.