Explorations in statistics: statistical facets of reproducibility

Douglas Curran-Everett

Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado; and Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado

Submitted 4 March 2016; accepted in final form 19 April 2016

Curran-Everett D. Explorations in statistics: statistical facets of reproducibility. Adv Physiol Educ 40: 248–252, 2016; doi:10.1152/advan.00042.2016.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This eleventh installment of Explorations in Statistics explores statistical facets of reproducibility. If we obtain an experimental result that is scientifically meaningful and statistically unusual, we would like to know that our result reflects a general biological phenomenon that another researcher could reproduce if she/he repeated our experiment. But more often than not, we may learn this researcher cannot replicate our result. The National Institutes of Health and the Federation of American Societies for Experimental Biology have created training modules and outlined strategies to help improve the reproducibility of research. These particular approaches are necessary, but they are not sufficient. The principles of hypothesis testing and estimation are inherent to the notion of reproducibility in science. If we want to improve the reproducibility of our research, then we need to rethink how we apply fundamental concepts of statistics to our science.

This eleventh paper in Explorations in Statistics (see Refs. 5–13 and 16) explores statistical facets of reproducibility. If we obtain an experimental result that is scientifically meaningful and statistically unusual (see Refs. 6 and 7), we would like to know that our result reflects a general biological phenomenon that other researchers could reproduce if they repeated our experiment. But if we consider our experimental \( P \) value to be an index of reproducibility, we may find that other researchers cannot replicate our experimental result. We should expect this: a \( P \) value is a notoriously weak indicator of evidence (1, 4, 7, 21, 22, 30, 32).

In 2005 Ioannidis (23) detailed his provocative position that most published research results are false—that most published research results cannot be reproduced. His paper spurred others to estimate theoretically (25) and empirically (28) the extent to which experimental results are reproducible. Needless to say, the estimates varied widely.

The National Institutes of Health (NIH) and the Federation of American Societies for Experimental Biology (FASEB) have acted to improve the reproducibility of research (3, 17). NIH has developed four training modules: 1) transparency, 2) blinding and randomization, 3) biological and technical replicates, and 4) sample size, outliers, and exclusion criteria. FASEB has identified three strategies: uniform definitions to improve the reproducibility of research. In this exploration, we remedy that omission.

A Brief History of Reproducibility in Science

The notion of reproducibility in science has a long history. In The Probable Error of a Mean (31), Gosset illustrated a procedure that would evolve into the one-sample \( t \) test:

First let us see what is the probability that \( [\text{drug} A] \) will on the average give increase of sleep. [Looking up the ratio of the sample mean to the sample standard deviation] in the table for ten experiments we find by interpolating...the odds are -887 to -113 that the mean is positive.

That is about 8 to 1 and would correspond to the normal curve to about 1-8 times the probable error. It is then very likely that \( [\text{drug} A] \) gives an increase of sleep, but would occasion no surprise if the results were reversed by further experiments.

Ref. 7 provides more background on this example.

In his pivotal Statistical Methods for Research Workers (18), Fisher wrote

Just as a single observation may be regarded as an individual, and its repetition as generating a population, so the entire result of an extensive experiment may be regarded as but one of a population of such experiments. The salutary habit of repeating important experiments...shows a tacit appreciation of the fact that the object of our study is not the individual result, but the population of possibilities of which we do our best to make our experiments representative.

A year later (19), Fisher wrote

A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give [a low] level of significance.

And as a final example, Yates (34) wrote

Research workers...have to accustom themselves to the fact that in many branches of research the really critical experiment is rare, and that it is frequently necessary to combine the results of numbers of experiments dealing with the same issue in order to form a satisfactory picture of the true situation. In such circumstances a number of experiments of moderate accuracy are of far greater value than a single experiment of very high accuracy.

With this brief history, we are almost ready to begin our exploration of reproducibility. First, we need to review the software we will use to help us learn about it.
R: Basic Operations

The first paper in this series (5) summarized R (29) and outlined its installation. For this exploration there are three more steps: download Advances_Statistics_Code_Reprod.R2 to your Advances folder, confirm you installed beeswarm in our previous explorations (12, 13, 16), and install the extra package ggplot2 (33).

To install ggplot2, open R and then click Packages | Install package(s) . . . Select a CRAN mirror close to your location and then click OK. Select ggplot2 and then click OK. When you have installed ggplot2, you will see 'ggplot2' successfully unpacked and MDS sums checked in the R Console.

To run R commands. If you use a Mac, highlight the commands you want to submit and then press (command key + enter). If you use a PC, highlight the commands you want to submit, right-click, and then click Run line or selection. Or, highlight the commands you want to submit and then press Ctrl + R.

When the Null Hypothesis Is True

As a prelude to our exploration of reproducibility, consider a situation for which reproducibility is likely not a big deal: when the null hypothesis is true. Suppose we want to learn if some intervention affects the biological thing we care about. If we use two groups—for example, a control group and a treated group—this is tantamount to asking if our two samples come from the same or different populations. This means we define the null and alternative hypotheses, H0 and H1, as

H0: The samples come from the same population.
H1: The samples come from different populations.

If we want to know whether the populations have the same mean, we write these as

H0: \( \mu_1 - \mu_0 = 0 \)
H1: \( \mu_1 - \mu_0 \neq 0 \)

where \( \mu_i \), the difference in population means, is \( \mu_1 - \mu_0 \), the difference between the means of the treated and control populations.

In our second exploration (7) we discovered we can reject a true null hypothesis by virtue of unusual sample observations. We also discovered we can control the chance we make this kind of error when we define the critical significance level \( \alpha \): when we define \( \alpha \), we declare we are willing to reject a true null hypothesis 100\( \alpha \)% of the time. For this prelude, suppose we define \( \alpha \) to be the traditional 0.05 (14). This means we expect to reject a true null hypothesis 5% of the time.

How can we know our null hypothesis, H0: \( \mu_1 - \mu_0 = 0 \), is true? By drawing observations for our two groups from the same population (Fig. 1). When we do this, regardless of the number of observations in each group, 5% of the observed \( P \) values are less than \( \alpha = 0.05 \).

\( \mu \) is the mean of the observed \( P \) values when the null hypothesis, \( H_0: \mu - \mu_0 = 0 \), is true. We drew at random two samples, each with \( n \) observations, from a standard normal distribution (top left; see Refs. 5 and 7), did a two-sample \( t \) test (using a critical significance level \( \alpha = 0.05 \)), and then repeated this process to generate a total of 100,000 replications. As expected (see Ref. 7), in 5% of the replications (gray area), \( P < 0.05 \): we reject a true null hypothesis. Advances_Statistics_Code_Reprod.R does not create this data graphic.

When the Null Hypothesis Is False

Now consider a situation for which reproducibility is a big deal: when the null hypothesis is false. To simplify our lives, suppose we use the same null and alternative hypotheses we just did:

H0: \( \mu_1 - \mu_0 = 0 \)
H1: \( \mu_1 - \mu_0 \neq 0 \)

But now define the populations to have means that differ by 0.5 units (Fig. 2). Only because we defined \( \mu_1 - \mu_0 = 0.5 \) do we know the null hypothesis \( H_0 \) is false.

In earlier explorations (7, 10) we recognized that we would like to reject a null hypothesis if it is false. We can boost our chances of doing that if we design our experiment so that power, the probability we reject our null hypothesis given it is false, is relatively high. Because we have defined our populations, power depends only on the number of observations in our two groups (10).

At this point, we need some data. Suppose we draw at random a sample of 10 observations from each population in Fig. 2. Then, to make an inference about our null hypothesis, we compute \( \bar{y} \), the difference in sample means, and do a two-sample \( t \) test (see Ref. 12). If we repeat this simulation, \( \bar{y} \) and the corresponding \( P \) value vary among the replications (Fig. 3).

At long last we are ready to explore the relationship of hypothesis testing—\( P \) values—and estimation to reproducibility. Hypothesis testing. Suppose we have an experimental result that is statistically unusual. If we had defined \( \alpha = 0.05 \), this means our initial observed \( P_1 < \alpha = 0.05 \). If our null hypothesis is true, our result is unusual. We reject our null hypothesis.

We have discovered something. We are pioneers! We assume...
our result reflects a general phenomenon that other researchers will reproduce if they repeat our experiment. But will they?

In our pioneering experiment, imagine we used the fundamental test statistic $z$ (see Ref. 5). Because $P_1 < \alpha$, we calculated an observed value of $z$ that was more extreme than the critical value of $z$: $z_1 > z^*$. If we assume the magnitude of the effect observed in our initial experiment equals the magnitude of the true effect, then the probability that the $P$ value from a second experiment, $P_2$, will be less than $\alpha = 0.05$ is

$$\Pr\{z > z^* - z_1\} = \Pr\{P_2 < \alpha = 0.05\} \, .$$

See Fig. 4. In this situation, if $P_1 = 0.05$, then the probability a duplicate experiment will achieve $P < 0.05$—the probability it will achieve ‘statistical significance’—is 50% (Table 1). If $P_1 = 0.01$, then the probability a duplicate experiment will achieve $P < 0.05$ is about 75%. Only when $P_1 = 0.001$ does the probability a duplicate experiment will achieve $P < 0.05$ exceed 90%. Power, experimental design, and the actual test statistic have little impact on this phenomenon (2).

Estimation. In contrast, power—with its inherent connection to the statistical threshold of $\alpha$—does impact the reproducibility of point and interval estimates of the magnitude of some biological effect (24). If an experiment of lower power happens to reject its null hypothesis—if the effect is statistically meaningful—then the estimate of the magnitude of that effect will be exaggerated (22). This phenomenon is sometimes referred to as the winner’s curse (22, 24).

We can see this if we draw at random $n$ observations from each population in Fig. 2, do a two-sample $t$ test, and then repeat this process. For each iteration, we obtain a $P$ value with which to assess our null hypothesis $H_0$: $\Delta\mu = 0$ and $\Delta\bar{y}$, an

### Table 1. Probability a duplicate experiment will achieve $P < 0.05$

<table>
<thead>
<tr>
<th>$P_1$</th>
<th>$\Pr{P_2 &lt; \alpha = 0.05}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
<td>0.13</td>
</tr>
<tr>
<td>0.30</td>
<td>0.18</td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>0.05</td>
<td>0.50</td>
</tr>
<tr>
<td>0.025</td>
<td>0.61</td>
</tr>
<tr>
<td>0.01</td>
<td>0.73</td>
</tr>
<tr>
<td>0.005</td>
<td>0.80</td>
</tr>
<tr>
<td>0.001</td>
<td>0.91</td>
</tr>
<tr>
<td>0.00031</td>
<td>0.95</td>
</tr>
<tr>
<td>0.00002</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Values represent $P_1$, the $P$ value from an initial experiment, and $\Pr\{P_2 < \alpha = 0.05\}$, the probability that the $P$ value from a duplicate experiment will be less than $\alpha = 0.05$. These results assume that the magnitude of the effect observed in the initial experiment equals the magnitude of the true effect, and they represent the best-case probabilities for the duplicate experiment (see Ref. 21). The commands in lines 327–358 of Advances_Statistics_Code_Reprod.R create the entries in this table. To generate this table, highlight and submit the lines of code from Table 1: first line to Table 1: last line.
estimate of the difference in means (Fig. 5). We know the true difference in means is $\Delta \mu = 0.5$.

Let us focus on just those simulations for which we reject our null hypothesis: for which $P < 0.05$. When power was 0.18 (Fig. 5A), the estimated difference $\bar{\Delta y}$ exceeded the true difference $\Delta \mu$ in all those simulations. In contrast, when power was 0.94 (Fig. 5D), the estimated difference $\bar{\Delta y}$ exceeded the true difference $\Delta \mu$ in roughly half (54%) those simulations. Experimental design and the actual test statistic have little impact on this phenomenon (24).

**Practical Considerations**

If we want to improve the chances a researcher could reproduce our result if (s)he repeats our experiment—and really, why wouldn’t we?!—what can we do from the perspective of statistics? There are three things. First, when we design our experiment, estimate sample size so that power approaches 0.90.5 Second, define the critical significance level $\alpha$—the benchmark for how statistically unusual our result needs to be before we reject our null hypothesis—to be 0.005 or even 0.001 (26, 30). And third, focus our attention away from a simple $P$ value and toward the potential scientific importance of our experimental result (6, 14, 15, 20, 22). Ioannidis (23, 24) and Sterne and Davey Smith (30) discuss other things we can do.

We have seen that power and $\alpha$ impact sample size (see Ref. 10). If we define $\alpha$ to be a more stringent 0.005, and if we design our experiment so that power approaches 0.90, might the sample size for our experiment be so large as to be practically impossible? Not necessarily (30). Moreover, it may help to remember what Yates wrote years ago (34):

---

**Fig. 5.** Contour plots of the estimated difference in means, $\Delta \bar{y}$, as a function of $\log_{10} P$. For each sample size (A–D), we drew at random $n$ observations from each population in Fig. 2, did a two-sample $t$ test (using a critical significance level $\alpha = 0.05$), and then repeated this process to generate a total of 1000 replications. In each contour plot, a red dot depicts the result of a single replication; the progressively smaller white contours depict greater numbers of replications. For example, in A, most of the replications cluster around $P = 0.70$ and $\Delta \bar{y} = 0.15$. A–D, bottom: the distribution of observed $P$ values. The number of replications in which $P$ was less than 0.001, 0.01, or 0.05 is listed. A–D, left: the distribution of the estimated difference $\bar{\Delta y}$ when the observed $P$ value was $<0.05$. In B, when theoretical power was 0.48, the estimated difference $\bar{\Delta y}$ exceeded the true difference of 0.50 in nearly all (94%) replications. In D, when theoretical power was 0.94, the estimated difference $\bar{\Delta y}$ exceeded the true difference in roughly half (54%) the replications. The commands in lines 157–306 of Advances_Statistics_Code_Reprod.R create the components for this data graphic. To replay these components, highlight and submit lines 316–322.

---

5 Although R can illustrate the concept of power (see Ref. 10), there are also online power calculators: see http://StatPages.org/, http://power.education.UConn.edu/OtherWebSites.htm, and http://PowerAndSampleSize.com/.
A number of experiments of moderate accuracy are of far greater value than a single experiment of very high accuracy.

Summary

NIH and FASEB have created training modules and outlined strategies to help improve the reproducibility of research. These particular approaches, which include transparency and the adequate reporting of experimental details such as the determination of sample size, an estimate of power, the process of randomization, and the handling of outliers, are necessary, but they are not sufficient. As this exploration has demonstrated, the principles of hypothesis testing and estimation are inherently related to the reproducibility of research. If we want to improve the reproducibility of our research, then we need to rethink how we apply fundamental concepts of statistics to our science.

ACKNOWLEDGMENTS

The author thanks Gerald DiBona (University of Iowa College of Medicine, Iowa City, IA) and Calvin Williams (Clemson University, Clemson, SC) for their helpful comments and suggestions.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

D.C.-E. conception and design of research; D.C.-E. performed experiments; D.C.-E. analyzed data; D.C.-E. interpreted results of experiments; D.C.-E. prepared figures; D.C.-E. drafted manuscript; D.C.-E. edited and revised manuscript; D.C.-E. approved final version of manuscript.

REFERENCES