Experimental Power (and why it matters)

Dr Andrew J. Stewart

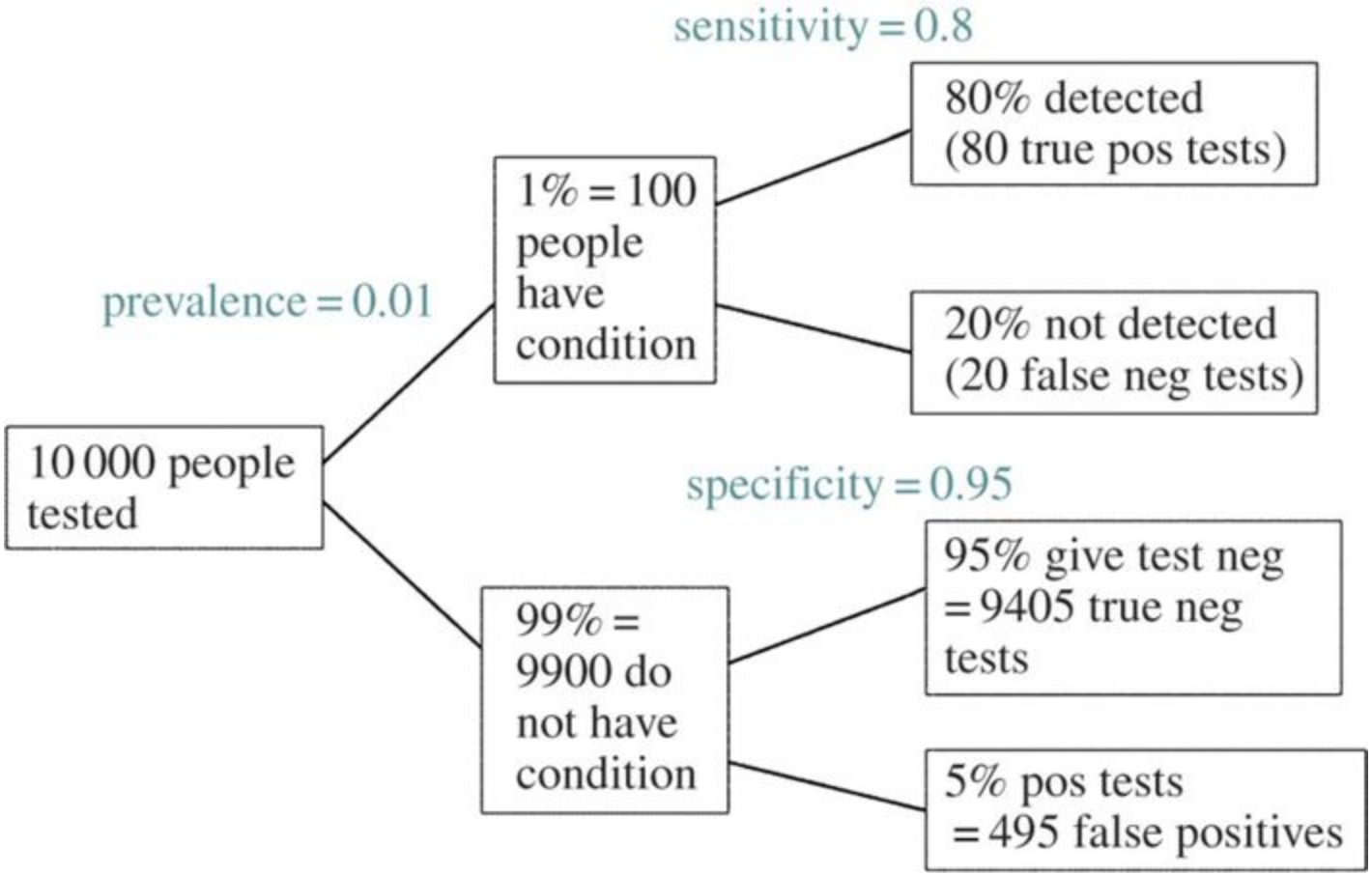
Fellow of the Software Sustainability Institute

E: drandrewjstewart@gmail.com
T: @ajstewart_lang
G: ajstewartlang

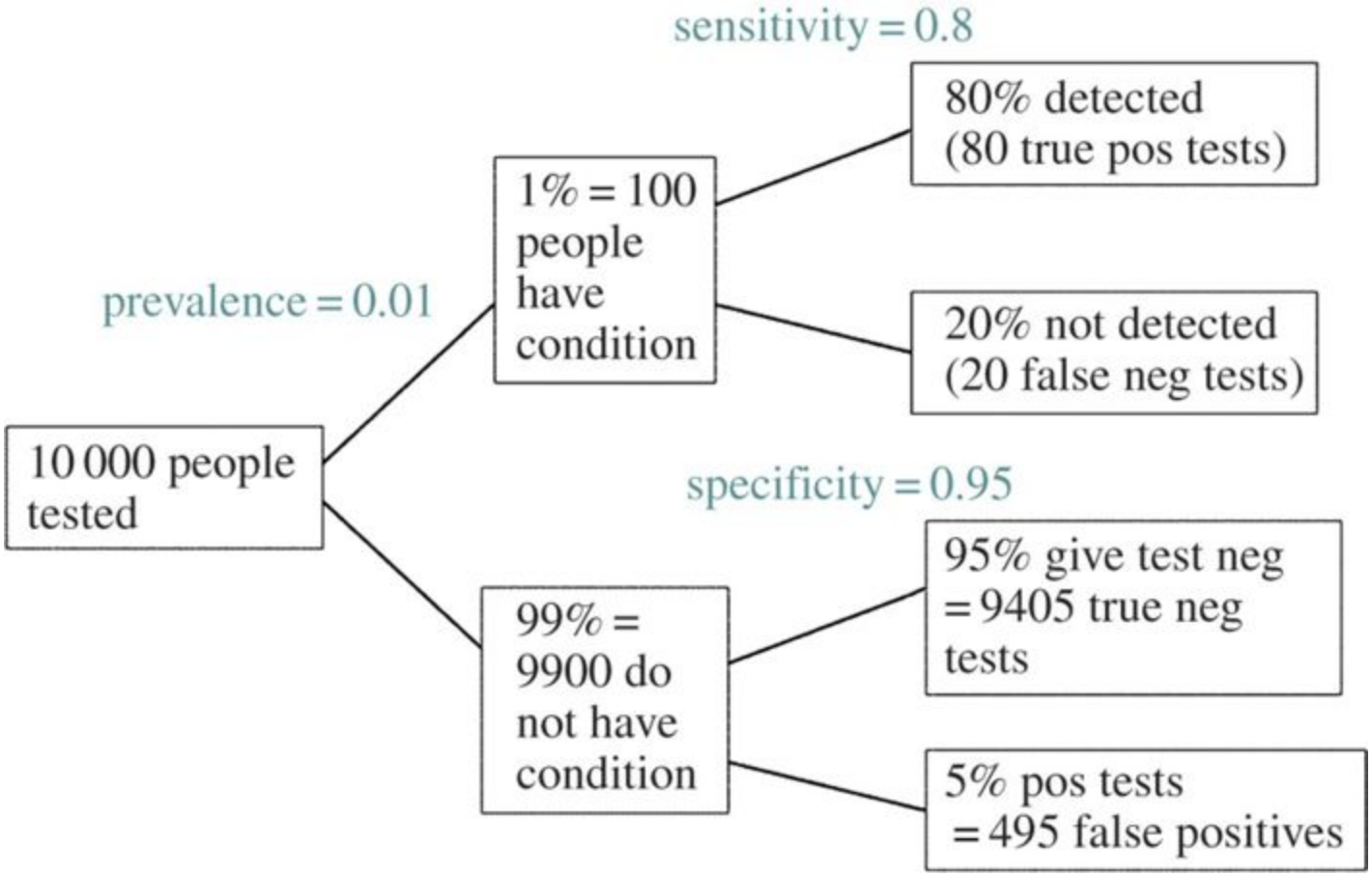
Understanding Statistics

- Imagine a test in which 95% of people without a medical condition will be correctly diagnosed as not having it (specificity = 0.95).
- Imagine the test is able to correctly diagnose 4 out of the 5 people who do have the medical condition (sensitivity = 0.8).
- Imagine the prevalence of the medical condition in the population is 1%.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. DOI: 10.1098/rsos.140216



Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. DOI: 10.1098/rsos.140216



The results of the test suggest 575 people have the condition. But 495 of these are false positives. So 86% of the people who produced a positive result actually **don't** have the condition.

Traditional NHST basics...

- For a design with two experimental groups:
  - Null hypothesis (H0) - there is no statistically significant difference between those experimental groups.
  - Experimental hypothesis (H1) - there is a statistically significant difference between two experimental groups.

- We typically reject H0 that if we find that the result of a statistical test comparing the two experimental groups is $p < 0.05$ (this is the typical alpha (α) level researchers choose).

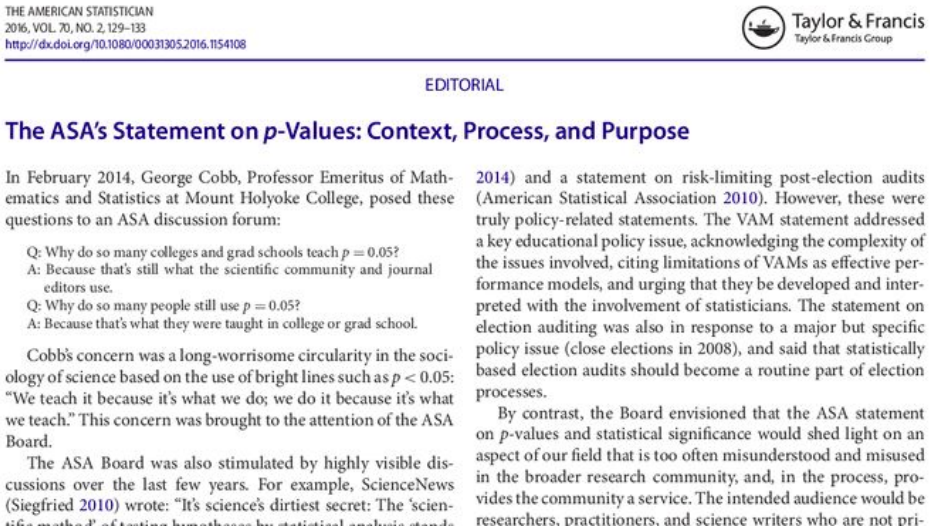
What is statistical significance?

Suppose that a treatment and a placebo are allocated at random to a group of people. We measure the mean response to each treatment, and wish to know whether or not the observed difference between the means is real (not zero), or whether it could plausibly have arisen by chance. If the result of a significance test is $p = 0.05$, we can make the following statement:

If there were actually no effect (if the true difference between means were zero) then the probability of observing a value for the difference equal to, or greater than, that actually observed would be p = 0.05. In other words there is a 5% chance of seeing a difference at least as big as we have done, by chance alone.

Many scientists would not be able to correctly define what is meant by a p-value...worrying!

In 2016 the American Statistical Association had to publish a paper reminding researchers what can be concluded from p-values and what cannot...



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on p-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on p-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not pri-

ASA Principles on *p*-values

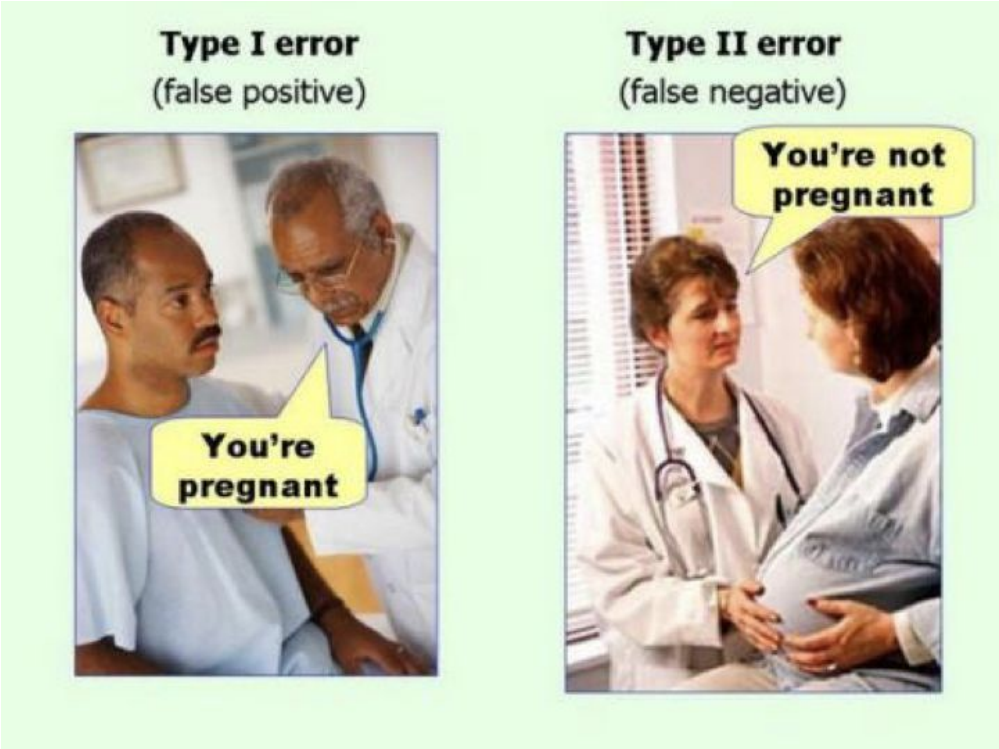
1. *p*-values can indicate how incompatible the data are with a specified statistical model.
2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

Type I and Type II Errors

- With an α-level of 0.05, we have a 5% chance of falsely rejecting the null hypothesis (H0).
- Falsely rejecting H0 is known as a Type I error (i.e., thinking we have found a difference when these isn't one).
- There are also Type II errors which involve failing to find a difference when one is actually present.
- Most of what you have been taught previously will probably have involved trying to avoid Type I errors.

Type I and Type II Errors



Type I and Type II Errors

- Controlling for Type II errors is as important as controlling for Type I errors. The probability of a Type II error is known as Beta (β).

- The probability of arriving at a Type II error (not finding a difference where there is one) is related to the experimental power of your design.

- For any experiment, Power = 1 - β

Is Power That Big a Deal?

- Cohen (1992) describes why power is such a big deal (and what can happen if experiments do not have sufficient power). Low powered studies have a lowered chance of finding a real effect, and along with QRPs also a higher chance of suggesting an effect is present when it is not.
- Reports the results of a review of 1960 volume of Journal of Abnormal and Social Psychology that he conducted at the time and the results of a Sedlmeier and Gigerenzer (1989) review of a 1984 volume of of the same journal.
- In 1960, the average power of the experiments reported in JASP to detect medium effect sizes was 0.48. In 1984, it was 0.25 (in other words only a 25% chance of finding an effect even if it was there!)

Is Power That Big a Deal?

- Button et al. (2013), Nature Reviews Neuroscience, small sample size undermines the reliability of neuroscience. Nord et al., (2017), Journal of Neuroscience, highlight wide heterogeneity in power in neuroscience studies.

Table 2. Median, maximum, and minimum power subdivided by study type

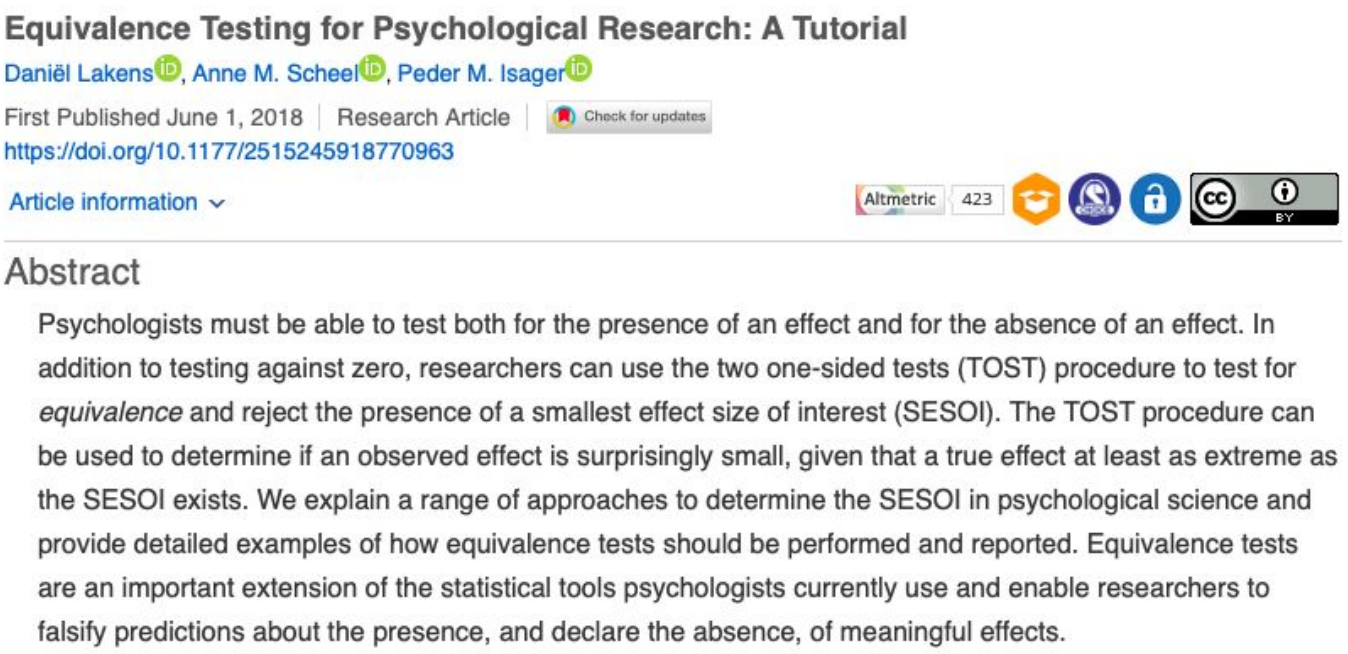
Group	Median power (%)	Minimum power (%)	Maximum power (%)	2.5th and 97.5th percentile (based on raw data)	95% HDI (based on GMMs)	Total N
All studies	23	0.05	1	0.05–1.00	0.00–0.72, 0.80–1.00	730
All studies excluding null	30	0.05	1	0.05–1.00	0.01–0.73, 0.79–1.00	638
Genetic	11	0.05	1	0.05–0.94	0.00–0.44, 0.63–0.93	234
Treatment	20	0.05	1	0.05–1.00	0.00–0.65, 0.91–1.00	145
Psychology	50	0.07	1	0.07–1.00	0.02–0.24, 0.28–1.00	198
Imaging	32	0.11	1	0.11–1.00	0.03–0.54, 0.71–1.00	65
Neurochemistry	47	0.07	1	0.07–1.00	0.02–0.79, 0.92–1.00	50
Miscellaneous	57	0.11	1	0.11–1.00	0.09–1.00	38

Cohen's d

- Power (1-β) is related to:
  - sample size (i.e., N)
  - effect size
  - α
- Cohen (1992) proposes that a reasonable level of Power to aim for should be around 0.8
- Power of 0.8 (with a β of 0.20), alongside an α of 0.05 results in a β:α ratio of 4:1 in terms of the risk associated with respective errors.

	Small Effect	Medium Effect	Large Effect
Cohen's d	0.2	0.5	0.8
Pearson's r	0.1	0.3	0.5

Equivalence Testing



Equivalence Testing for Psychological Research: A Tutorial

Daniël Lakens, Anne M. Scheel, Peder M. Isager

First Published June 1, 2018 | Research Article | Check for updates
https://doi.org/10.1177/2515245918770963

Article information

Altmetric 423

Abstract

Psychologists must be able to test both for the presence of an effect and for the absence of an effect. In addition to testing against zero, researchers can use the two one-sided tests (TOST) procedure to test for *equivalence* and reject the presence of a smallest effect size of interest (SESOI). The TOST procedure can be used to determine if an observed effect is surprisingly small, given that a true effect at least as extreme as the SESOI exists. We explain a range of approaches to determine the SESOI in psychological science and provide detailed examples of how equivalence tests should be performed and reported. Equivalence tests are an important extension of the statistical tools psychologists currently use and enable researchers to falsify predictions about the presence, and declare the absence, of meaningful effects.

https://journals.sagepub.com/doi/full/10.1177/2515245918770963

Data Simulation

faux 0.0.1.0 | Reference | Articles | Changelog

faux

It is useful to be able to simulate data with a specified structure. The `faux` package provides some functions to make this process easier. See the vignettes for more details.

Installation

You can install the development version of faux from GitHub with:

```
devtools::install_github("debruine/faux")
```

Data Simulation

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Volume 7, Issue 4
April 2016
Pages 493-498

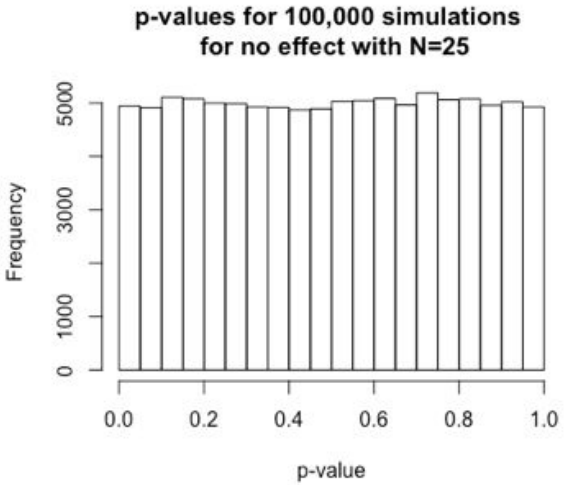
Application | Free Access

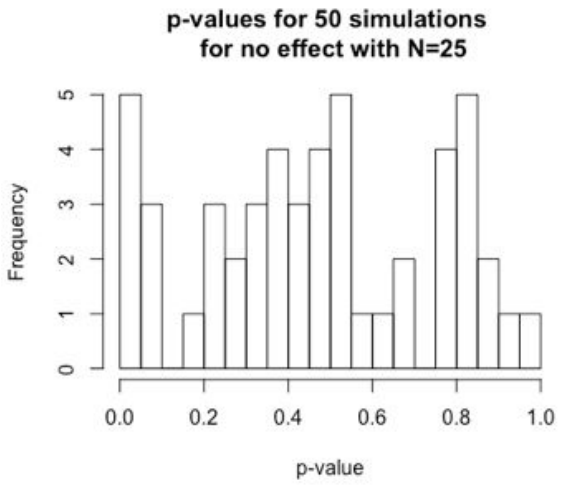
SIMR: an R package for power analysis of generalized linear mixed models by simulation

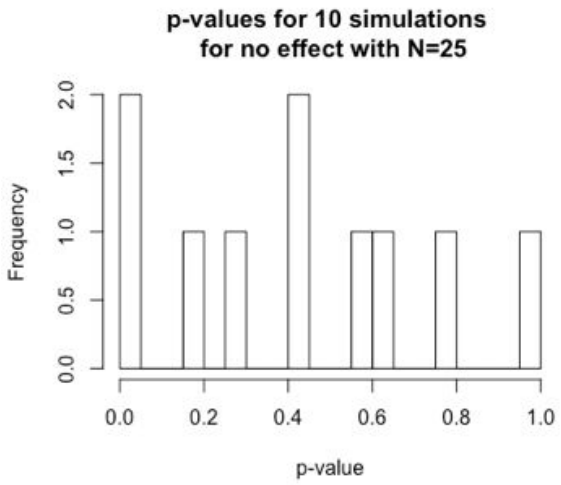
Peter Green, Catriona J. MacLeod

First published:17 November 2015 | https://doi.org/10.1111/2041-210X.12504 | Citations: 210

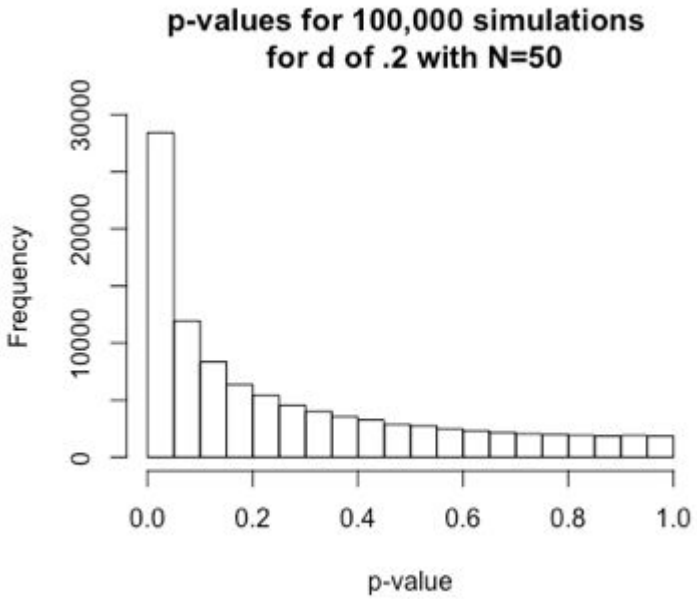
Data Simulation - When No Real Effect Exists





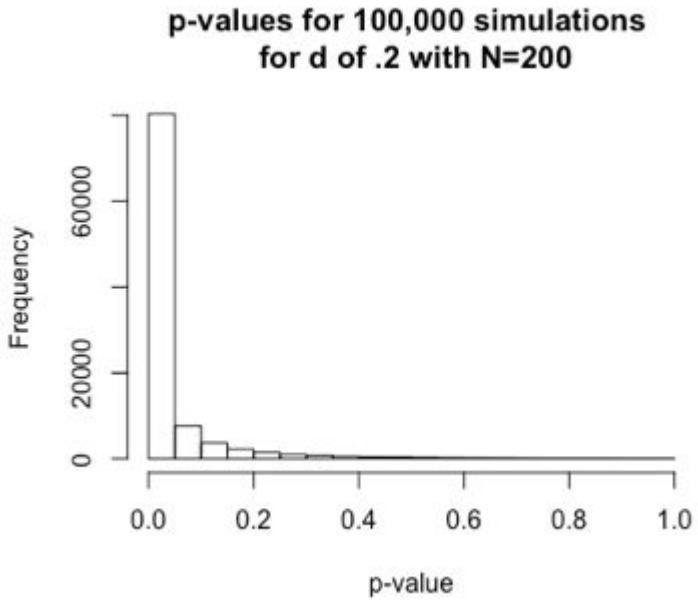


Real Effects Will NOT Always Replicate



Assuming $p < .05$ alpha, N=50 gives us around 30% power, which means that 70% of the time we'll miss the effect (even though it is present).

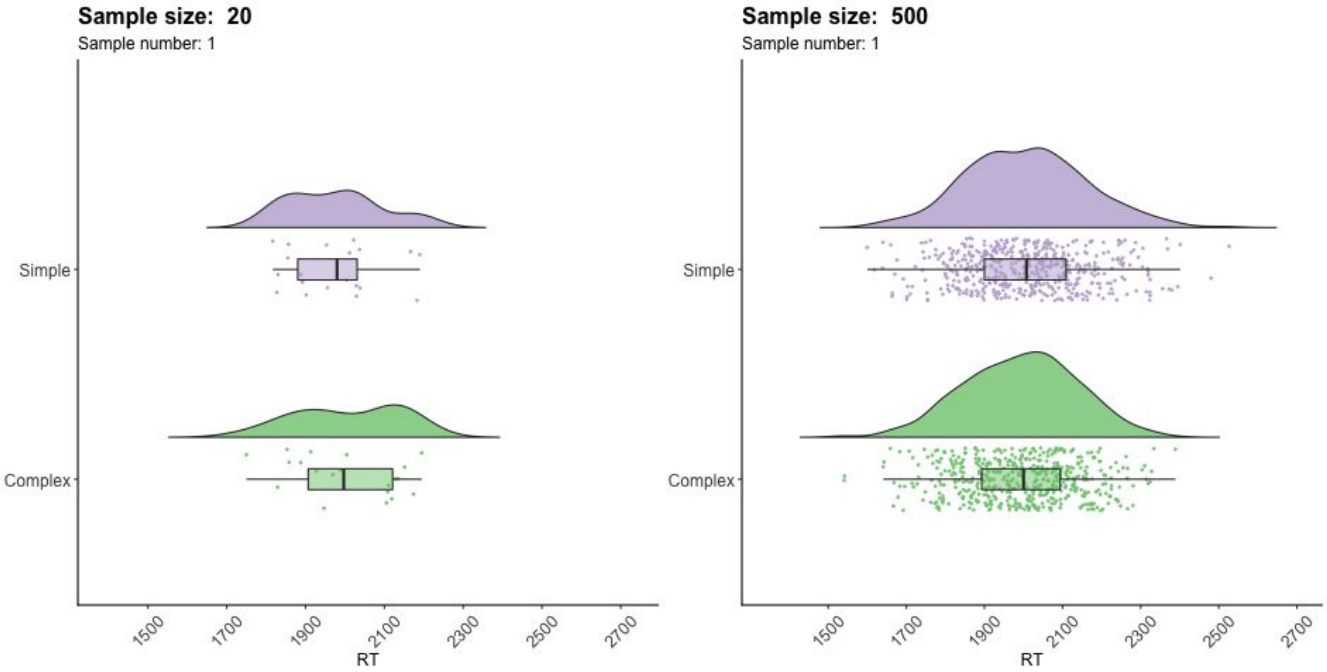
Real Effects Will NOT Always Replicate



N=200 gives us around 80% power, which means that 20% of the time we'll miss finding the effect (even though it is present).

The Problem of Sampling Bias

Samples for conditions “Simple” and “Complex” are drawn from the same population. Due to sampling error, with small samples (e.g., N=20) we might conclude there could be a difference between A and B where there isn’t one (as you can see with the N=500 samples). Enter QRPs...



Summary

- Power is important - underpowered experiments are a waste of time (often yours!), money, and resources such as lab space etc.
- Underpowered experiments combined with questionable research practices (QRPs) and publication bias results in a literature that is full of research articles that are wrong.
- The scientific theories/models you're testing need to allow you to determine what the minimal effect size of interest is - and it is this minimal effect size that you need to power your experiment to find.
- Even in a high powered study (e.g., 80%) sometimes you will fail to find an effect even though it is present - and with NHST just because you might have an absence of evidence for an effect, this is not the same as having evidence of the effect not being there. When our test is non-significant, we cannot conclude an effect is not there - just that we don't have the support to conclude that it **is** there.