Final session: Doing reproducible research

Stats 60 / Psych 10

Ismael Lemhadri

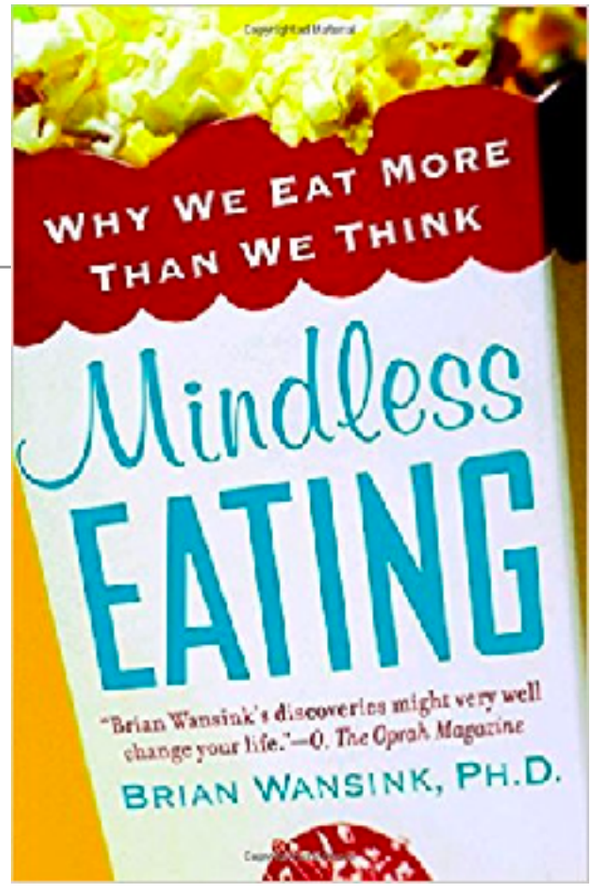
This time

- What are the problems with reproducibility?
  - p-hacking
  - HARKing
  - low power
- Guest lecture: **Rob Tibshirani**

The classical view of how science should work

- You start with a hypothesis
  - Branding with popular characters should cause children to choose “healthy” food more often
- You do an experiment
  - You offer children the choice between a cookie and an apple with either an Elmo-branded sticker or a control sticker
- You do statistics to test the null hypothesis
  - “The preplanned comparison shows Elmo-branded apples were associated with an increase in a child’s selection of an apple over a cookie, from 20.7% to 33.8% (χ^2=5.158; P=.02)“ (Wansink, Just, & Payne, 2012, JAMA Pediatrics)

How science actually works (sometimes)



Brian Wansink

Director, Cornell Food and Brand Lab Author, Mindless Eating

Speaking Topics:

Author, Business, Education, Food

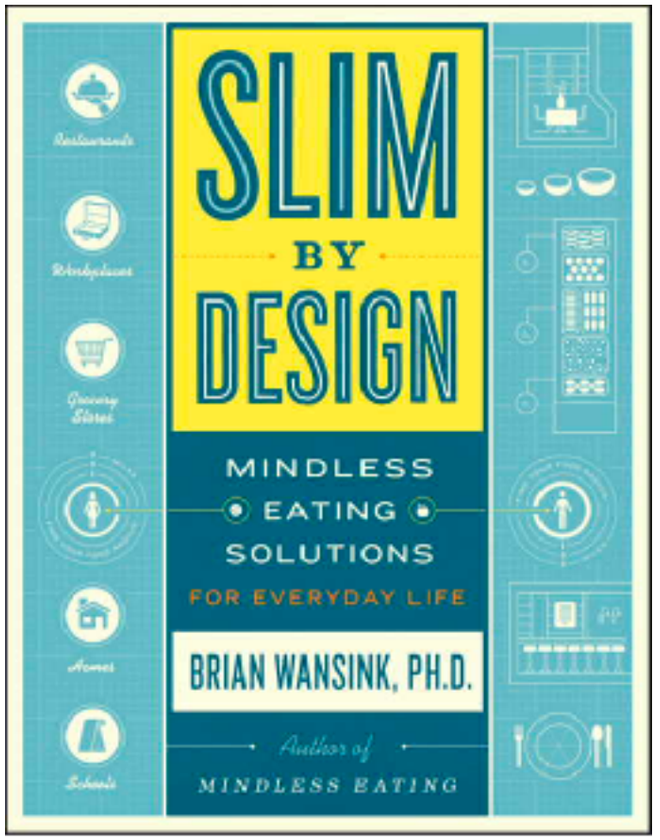
Travels From:

New York, NY, USA

Fee Range:

$30,000-50,000

http://speakerbookingagency.com/talent/brian-wansink/



How science actually works (sometimes)

...back in September 2008, when Payne was looking over the data soon after it had been collected, he found no strong apples-and-Elmo link — at least not yet.

"I have attached some initial results of the kid study to this message for your report," Payne wrote to his collaborators. "Do not despair. It looks like stickers on fruit may work (with a bit more wizardry)."

Wansink also acknowledged the paper was weak as he was preparing to submit it to journals. The p-value was 0.06, just shy of the gold standard cutoff of 0.05. It was a "sticking point," as he put it in a Jan. 7, 2012, email.

"It seems to me it should be lower," he wrote, attaching a draft. "Do you want to take a look at it and see what you think? If you can get the data, and it needs some tweeking, it would be good to get that one value below .05."

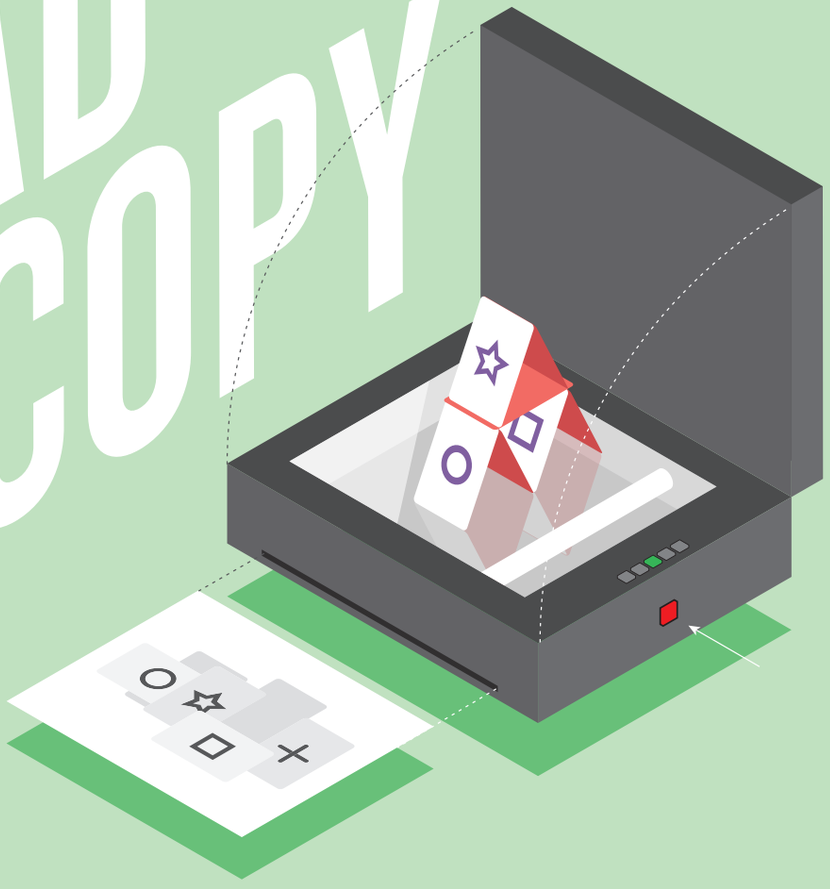
Later in 2012, the study appeared in the prestigious JAMA Pediatrics, the 0.06 p-value intact. But in September 2017, it was retracted and replaced with a version that listed a p-value of 0.02. And a month later, it was retracted yet again for an entirely different reason: Wansink admitted that the experiment had not been done on 8- to 11-year-olds, as he'd originally claimed, but on preschoolers.

https://www.buzzfeed.com/stephaniemlee/brian-wansink-cornell-p-hacking?utm_term=.gtAVwLX2GM#.fep9L6pw78

Science in crisis (?)

298 | NATURE | VOL 485 | 17 MAY 2012

BAD COPY



IN THE WAKE OF HIGH-PROFILE CONTROVERSIES, PSYCHOLOGISTS ARE FACING UP TO PROBLEMS WITH REPLICATION.

BY ED YONG

Rigorous replication effort succeeds for just two of five cancer papers

By Jocelyn Kaiser | Jan. 18, 2017, 1:00 PM

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

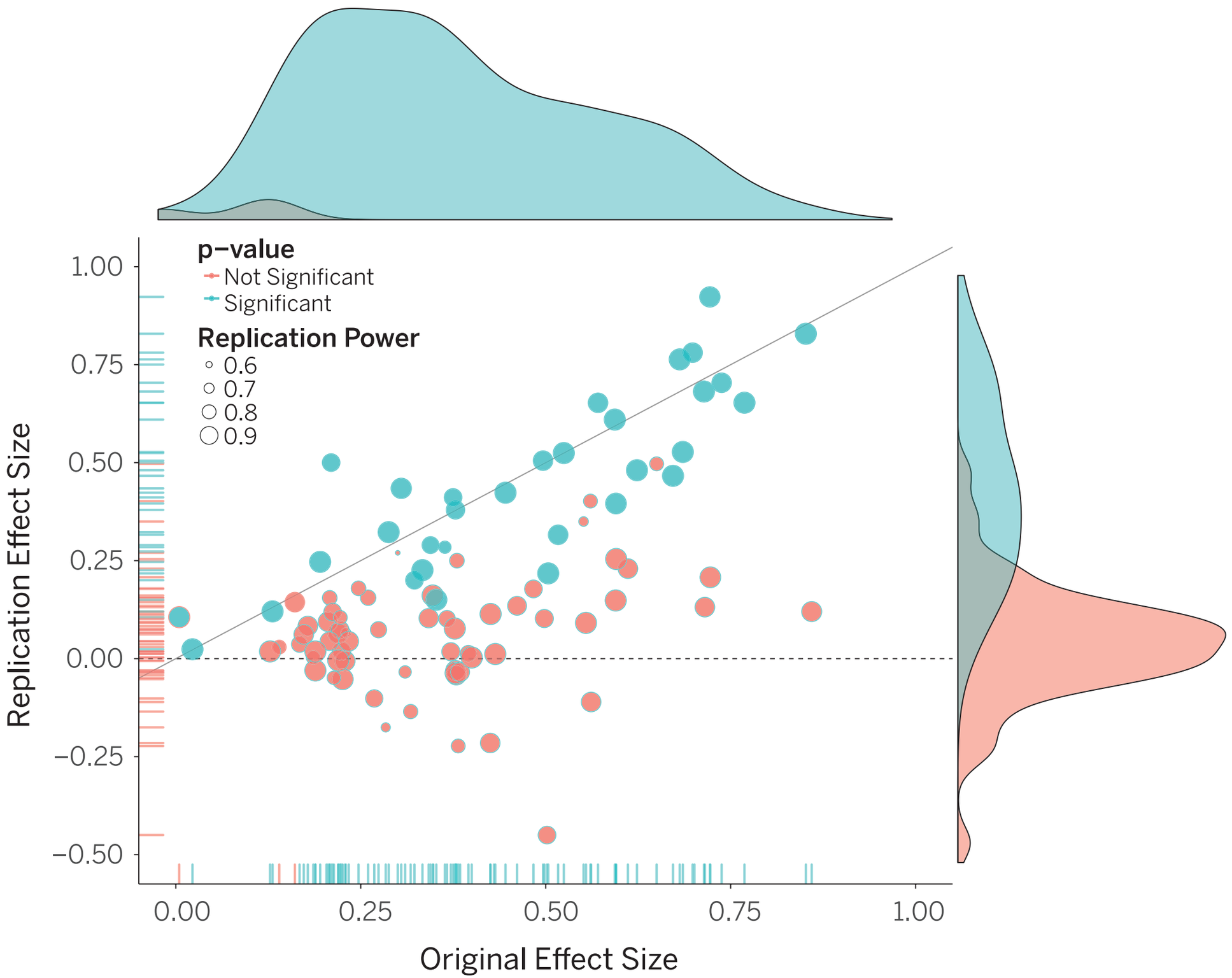
Estimating the reproducibility of psychological science

Open Science Collaboration*

SCIENCE sciencemag.org 28 AUGUST 2015 • VOL 349 ISSUE 6251

We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available.

Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Open Science Collaboration, 2015

FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

Cognitive biases in statistical/scientific reasoning

- "The first principle is that you must not fool yourself and you are the easiest person to fool"
  - R. Feynman
- We pay more attention to information that confirms our hypotheses or biases versus those that disconfirm them
  - We are more likely to overlook errors that confirm our pre-existing ideas
- We fail to consider alternative hypotheses that could explain the data

Is NHST causing an epidemic of false results?

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine August 2005 | Volume 2 | Issue 8 | e124

“There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. … Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. “

John Ioannidis

How likely is a true result?

- Positive predictive value (PPV)

$$PPV = \frac{number\,of\,true\,positives}{number\,of\,true\,positives + number\,of\,false\,positives}$$

$$PPV = \frac{pTrue * (1 - \beta)}{pTrue * (1 - \beta) + (1 - pTrue) * \alpha}$$

$$\alpha = false\,positive\,rate$$

$$\beta = false\,negative\,rate = 1 - power$$

pTrue = prevalence of true relations amongst those tested

$$PPV = \frac{pTrue * (1 - \beta)}{pTrue * (1 - \beta) + (1 - pTrue) * \alpha}$$

Take a field where most of the hypotheses being tested are true (pTrue=0.8), and where the study is well powered ($\boldsymbol{\beta}$=0.2) with the standard alpha of 0.05

$$PPV = \frac{0.8 * (1 - 0.2)}{0.8 * (1 - 0.2) + (1 - 0.8) * 0.05} = 0.98$$

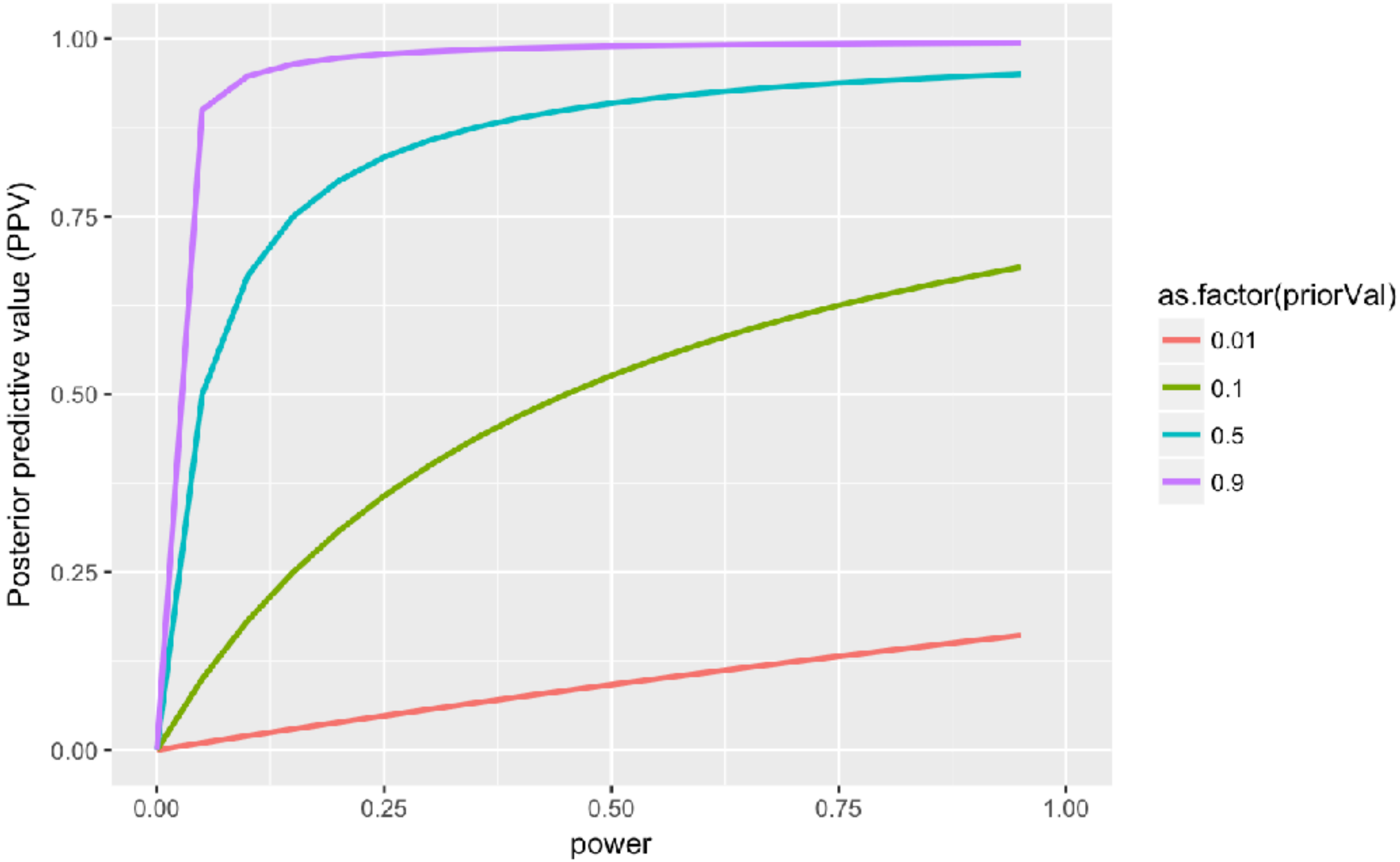
If most hypotheses are true, then is the science interesting?

$$PPV = \frac{pTrue * (1 - \beta)}{pTrue * (1 - \beta) + (1 - pTrue) * \alpha}$$

Now take a field where most of the hypotheses being tested are false (pTrue=0.1), and where the study is poorly powered ($\boldsymbol{\beta}$=0.8) with the standard alpha of 0.05

$$PPV = \frac{0.1 * (1 - 0.8)}{0.1 * (1 - 0.8) + (1 - 0.1) * 0.05} = 0.307$$

In such a field, only 1/3 of statistically significant results would actually be true!



see notebook for simulation

Statistical power remains low in many areas of science

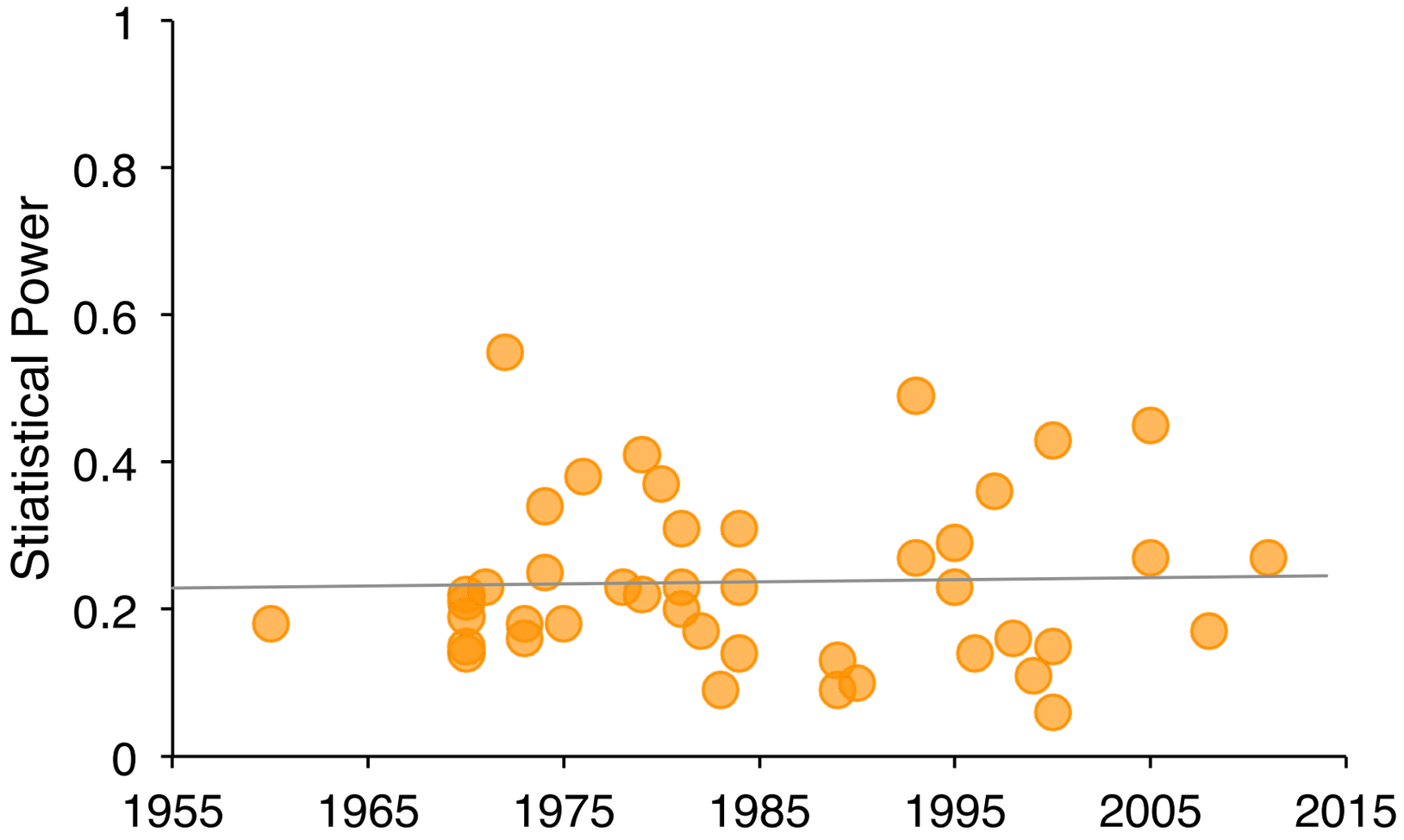


FIGURE 1. Average statistical power from 44 reviews of papers published in journals in the social and behavioral sciences between 1960 and 2011. Data are power to detect small effect sizes ($d = 0.2$), assuming a false positive rate of $\alpha = 0.05$, and indicate both very low power (mean = 0.24) but also no increase over time ($R^2 = 0.00097$).

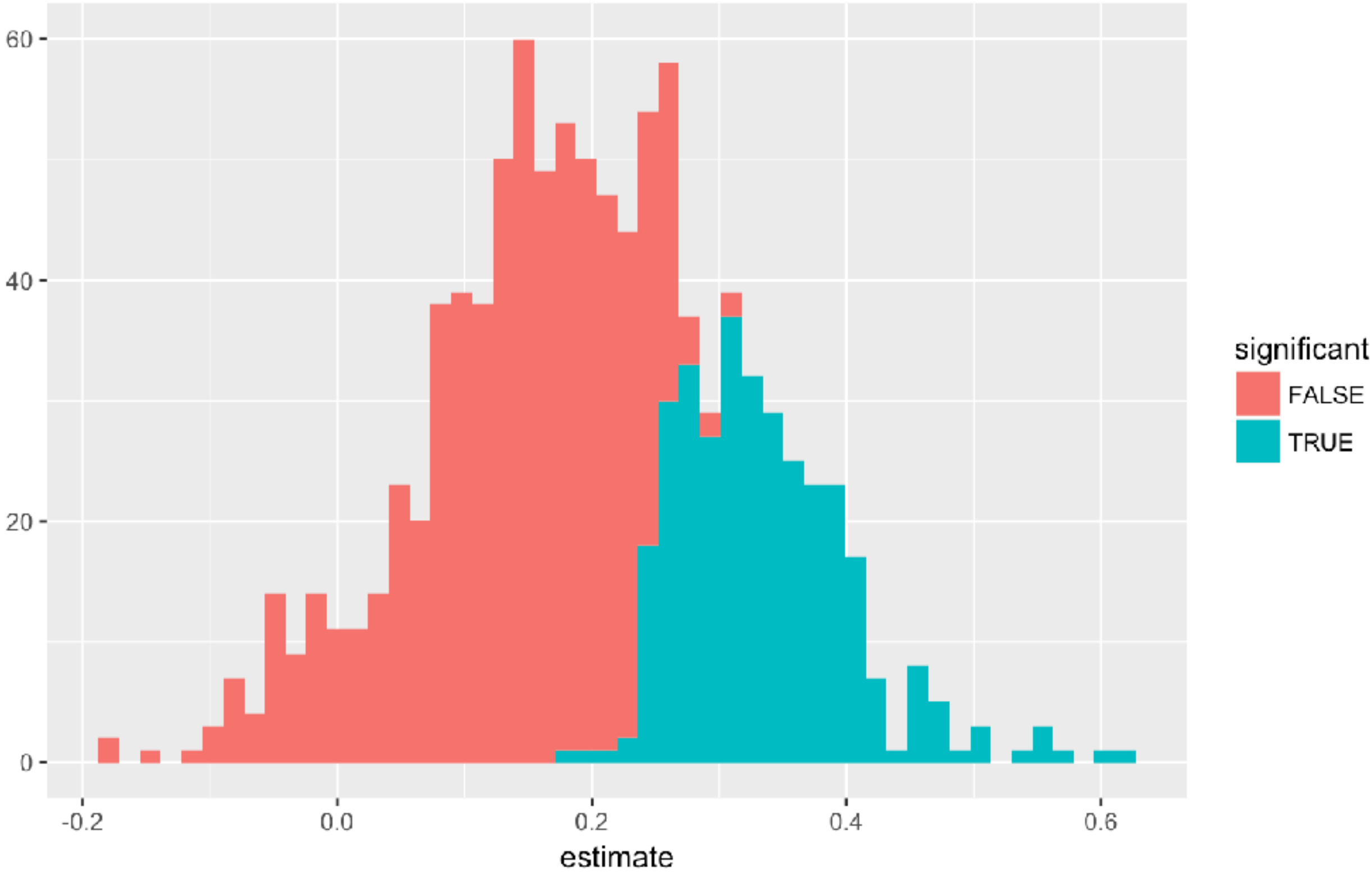
Smaldino & McElreath, 2016, https://arxiv.org/abs/1605.09511

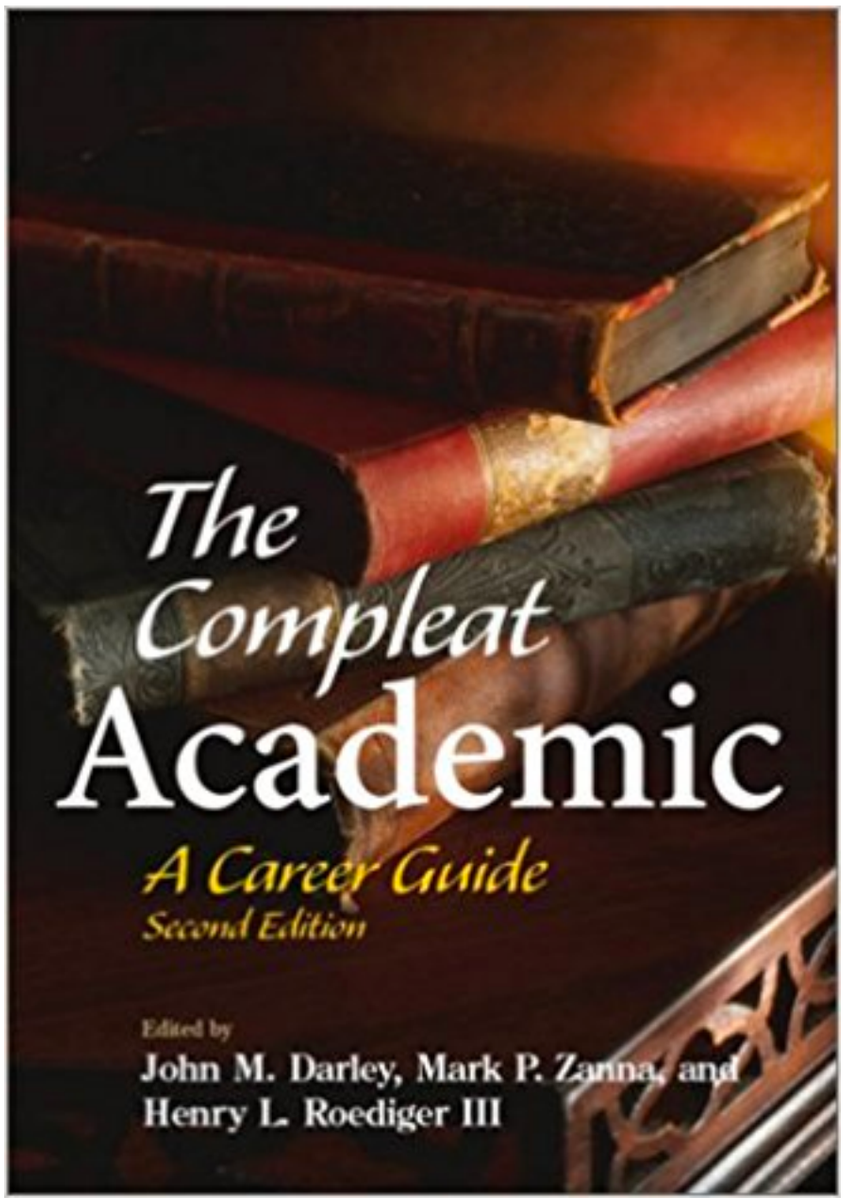
The winner's curse: How the size of estimated effects is inflated by NHST

- In economics:
  - For certain types of auctions (where the value is the same for everyone, like a jar of quarters, and the bids are private), the winner almost always pays more than the good is worth
- In statistics:
  - The effect size estimated from significant results (i.e. the winners) is almost always an overestimate of the true effect size

True effect size: 0.2

Mean effect size of significant effects: 0.33

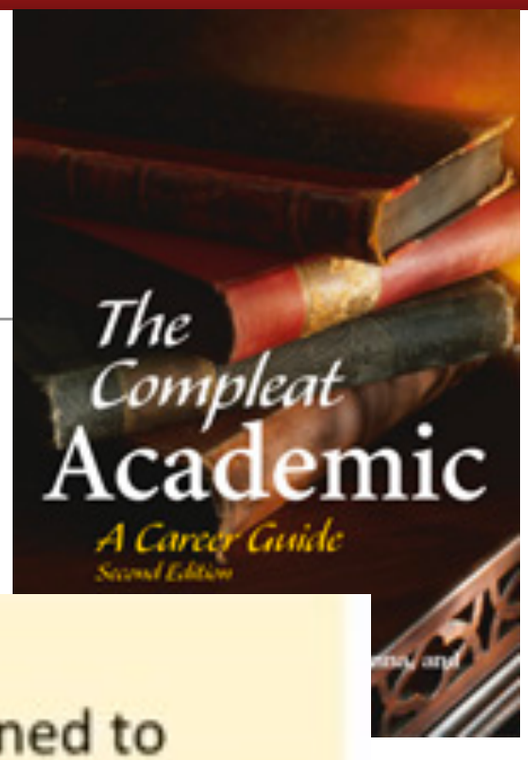




A new career in academia can be a challenge. While academia's formal rules are published in faculty handbooks, its implicit rules are often difficult to discern. Like the first edition, this new and expanded volume of *The Compleat Academic* is filled with practical and valuable advice to help new academics set the best course for a lasting and vibrant career.

https://www.apa.org/pubs/books/4316014.aspx

Career advice from Daryl J. Bem



HARKing

Which Article Should You Write?
There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).

p-hacking

re Data Analysis: Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something— anything —interesting.

http://neuroanatody.com/2017/11/oxford-reproducibility-lectures-dorothy-bishop/

HARKing

- "Hypothesizing after the results are known" (Kerr, 1988)
- Why is this a problem?
  - It can turn Type I errors into theory
    - A post-hoc conclusion gets re-framed as an a priori hypothesis
    - a theory that is re-written to fit the facts is not a very powerful theory!
  - It becomes impossible to disconfirm bad ideas

“P-hacking”

- Doing many analyses and only reporting those that achieve p<.05
- Ways to P-hack
  - Analyze data after every subject, and stop collecting data once p<.05
  - Analyze many different variables, but only report those with p<.05
  - Collect many different experimental conditions, but only report those with p<.05
  - Exclude participants to get p<.05
  - Transform the data to get p<.05

https://www.bitss.org/wp-content/uploads/2015/12/False-Positives-p-Hacking-Statistical-Power-and-Evidential-Value-Leif-Nelson.pdf

Anything can become significant via p-hacking

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

-Simmons et al., 2011, Psychological Science

Exercise

- Go to:
  - https://projects.fivethirtyeight.com/p-hacking/
- If your last name starts with A-L:
  - Find evidence that the U.S. economy is better when Republicans are in office.
- If your last name starts with M-Z:
  - Find evidence that the U.S. economy is better when Democrats are in office.
- Raise your hand once you have a significant effect

Guest Lecture: Rob Tibshirani