

Statistical and Design Concepts in Cognitive Neuroscience

Luis Morís Fernández

12.11.2021

Who am I?

Who am I?

- I work with:
 - Dr. Salvador Soto (Cognitive Neuroscience)
 - Dr. Miguel Ángel Vadillo (Meta Science)
 - Freelance Data Scientist for the Pharma Industry
- Engineer / Statistics / Neuroscience

Responsible Science Consumers

Consuming science

What's the main way of consuming science?

- Journal Articles

When you read a paper, how do you know if there is evidence for their claims?

- It is *complex*
- Let start with **Null Hypothesis Testing**

Null Hypothesis Significance Testing

The example

I have invented a way of measuring the IQ, that ranges from -50 to 50.

I want to test if the average IQ of people in the UB is higher than 0.

I consider you a representative sample of the UB population and I run a test on each of you, and get the individual IQs.





I test if the average of those IQs is higher than 0.

Null Hypothesis Significance Testing





After this test two things can happen:

- **Not significant.** Accept the Null Hypothesis. The average IQ is not higher than 0.
- **Significant.** Reject the Null Hypothesis. Accept the alternative the average IQ is higher than 0.

Outcomes

HYPOTHESIS TESTING OUTCOMES		R e a l i t y	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
R e s e a r c h	The Null Hypothesis Is True	Accurate 	Type II Error β 
	The Alternative Hypothesis is True	Type I Error α 	Accurate 

Null Hypothesis Significance Testing

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
R e s e a r c h	The Null Hypothesis Is True	Accurate 	Type II Error β 
	The Alternative Hypothesis is True	Type I Error α 	Accurate 

- α : Probability of False Positive
 - Given that the IQ is **not higher** than 0 and α 5%. If I run the IQ test many times 5 out of times my test will indicate the IQ is **higher** than 0.
- β : Probability of False Negative
 - Given that the IQ is **higher** than 0 and β 5%. If I run the IQ test many times 5 out of times my test will indicate the IQ is **not higher** than 0

How are α and β established?

Oversimplified!

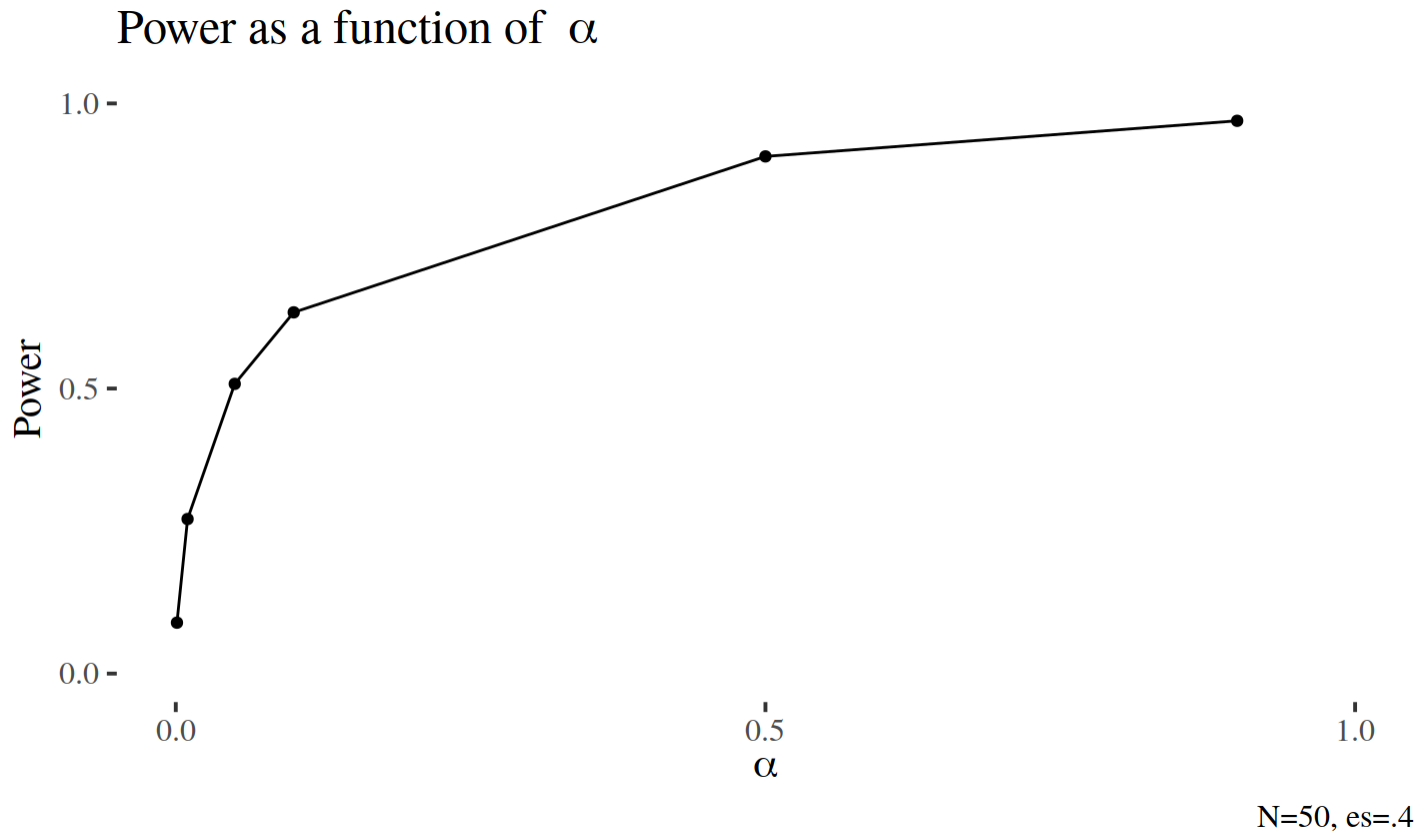
They depend on three things:

- Sample size
- Effect size
- α or β must be fixed
 - Usually α is the one fixed (in many fields is .05)

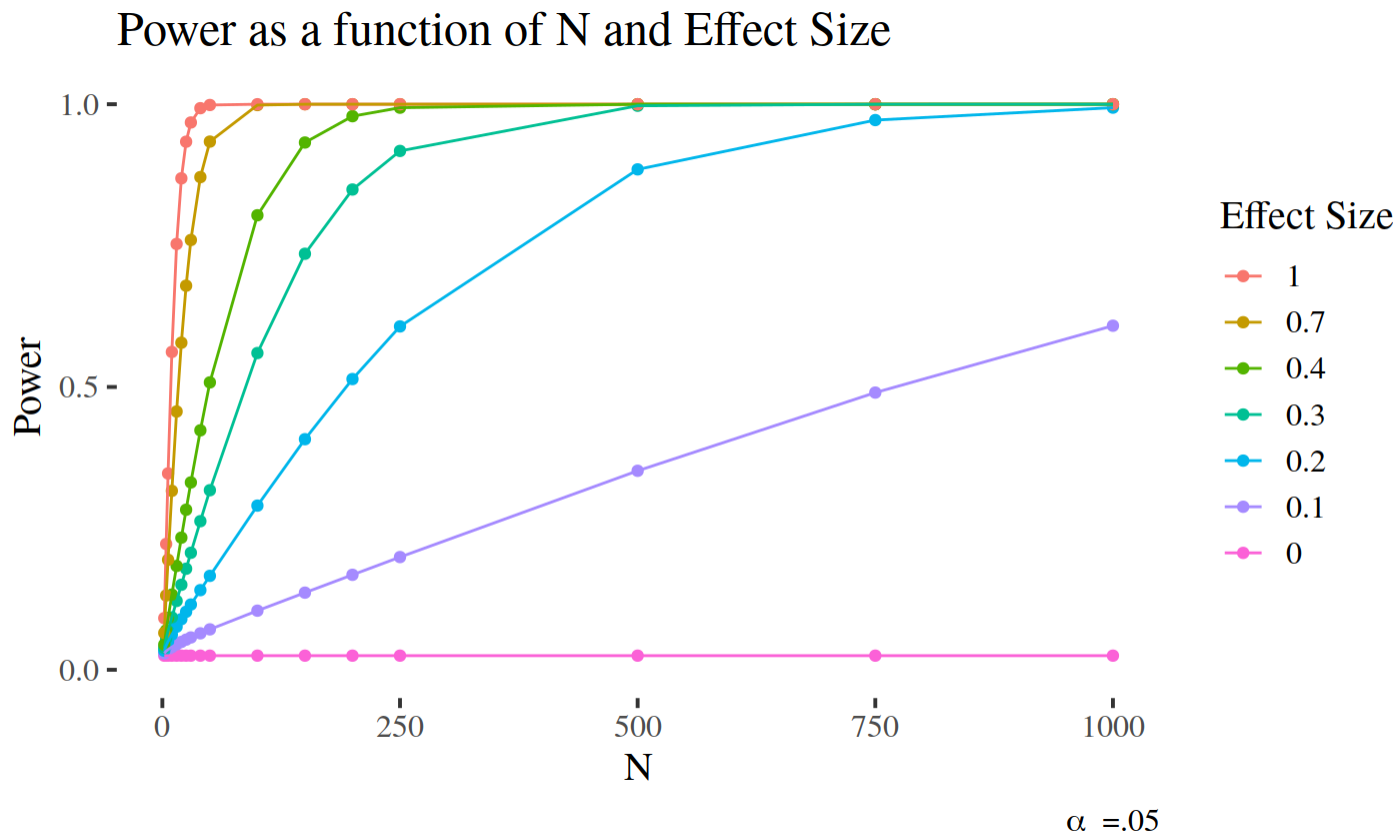
Probability of rejecting H_0 if H_1 is true Power = $1 - \beta$

More in the application

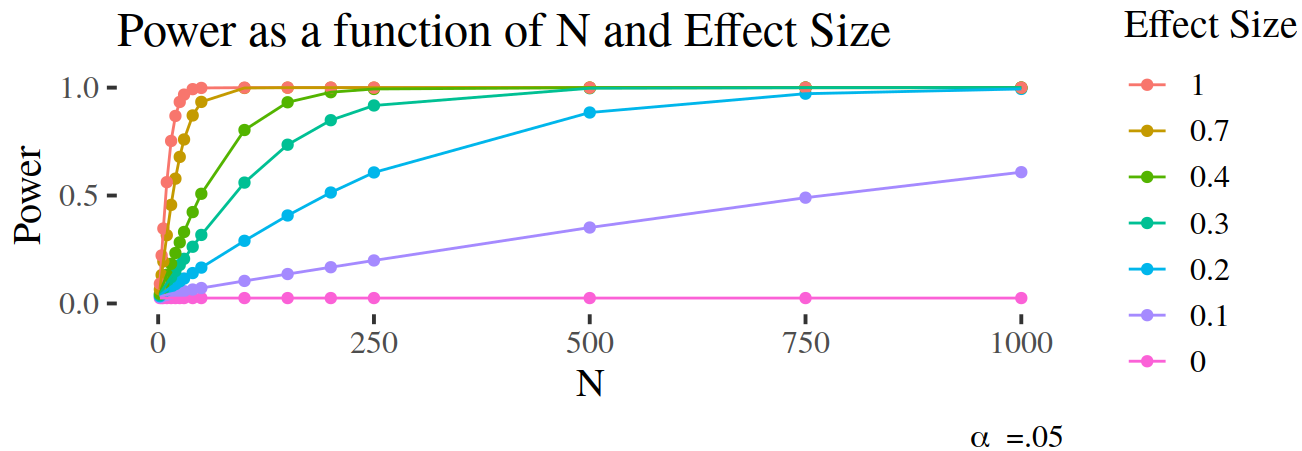
How are α and β established?



How are α and β established?



Some questions



- For an Effect Size of .2, how many samples should I gather to have approximately a Type II error rate of .25?
- If I look after an effect that it is bigger should my Power increase or decrease?
- If I now accept a Type I error rate of .005, what should I do to maintain my power?
- If the effect size of my treatment is .3, and I gather 250 samples, how often will I find a significant effect?
- If the Null Hypothesis is true, how often will I reject it if my α is .25

Some questions

- For an Effect Size of .2, how many samples should I gather to have approximately a Type II error rate of .25?
- If I look after an effect that it is bigger should my Power increase or decrease?
- If I now accept a Type I error rate of .005, what should I do to maintain my power?
- If the effect size of my treatment is .3, and I gather 250 samples, how often will I find a significant effect?
- If the Null Hypothesis is true, how often will I reject it if my α is .25

Some more questions

- Would you run an experiment looking for an effect size of .2 with a sample size of 40 ($\alpha=.05$)? Why? Why not?
- Well that is what many researchers do!

- Why not gather a googplex (10^{100}) samples and then we are done!
 - Money
 - Time

By now:

- Understand NHST
 - Accept an alternative hypothesis, or retaining the null
- Understand the relation α | Power | Effect size | Sample Size
 - Tests are not infallible and errors **must** occur
 - We have control (*more or less*) over those error rates
- Infinite samples are not possible.

Before advancing

Should I believe an article if they tested their theory using NHST?

- Should I discard a theory because they did not find a significant effect and therefore retained H_0 ?
- Should I believe a theory because they found a significant effect and therefore discarded H_0 ?

Problems we should be aware of
when consuming Science



🔍 New study shows|



🕒 [new study shows](#)

Eliminar

🔍 new study shows **vaping causes cancer**

🔍 new study shows **breathing air is linked to staying alive**

🔍 new study shows **how autism can be measured through a non-verbal marker**

🔍 new study shows **psychology**

🔍 new study shows **health**

🔍 new study shows **cats**

🔍 new study shows **spike in violent incidents in ontario's elementary schools**

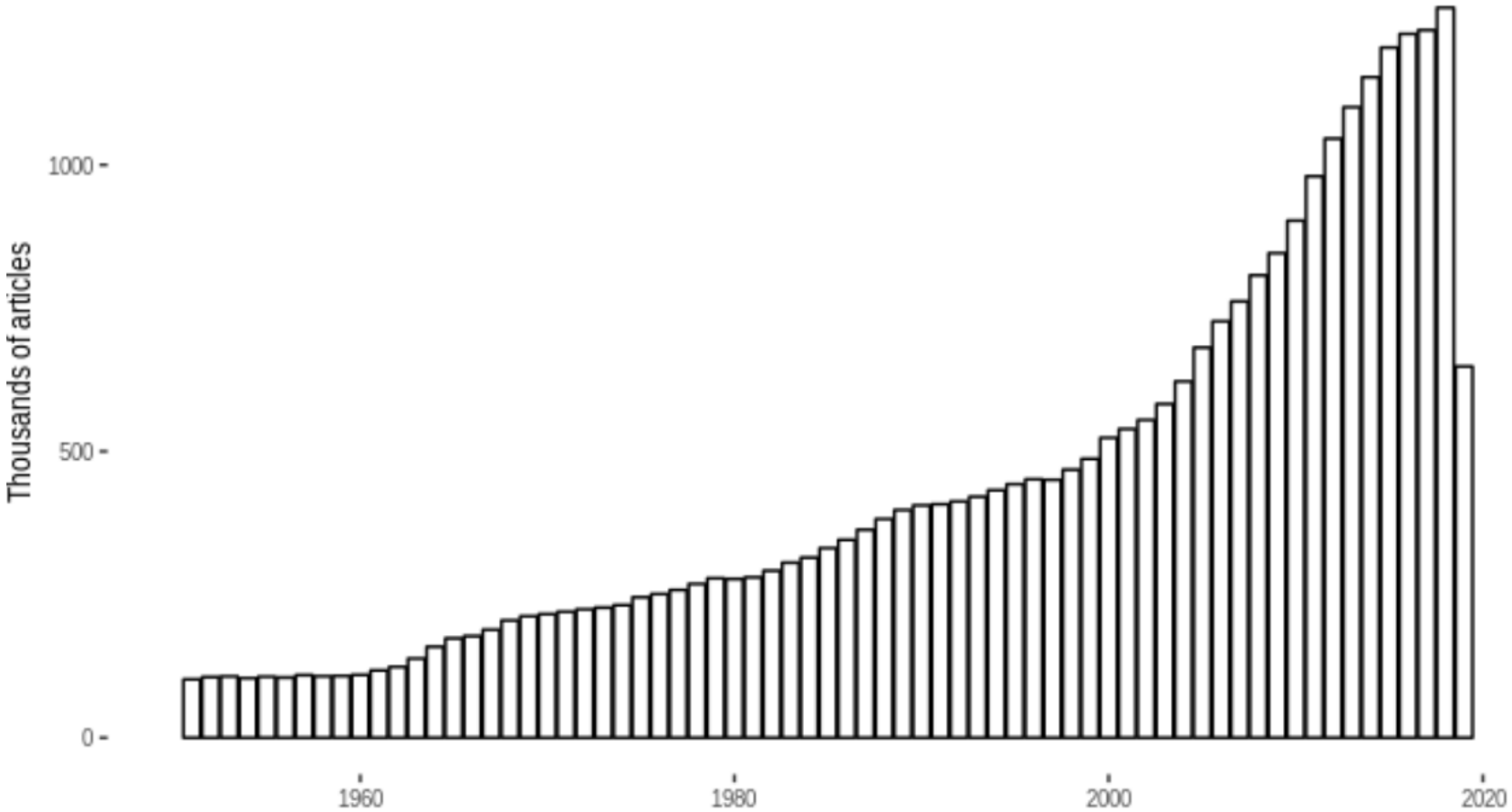
🔍 new study shows **possible link between environmental mercury and autism**

🔍 new study shows **chipotle**

[Denunciar predicciones inadecuadas](#)
[Más información](#)

Herramientas

Number of publications per year since 1950



Problems

- Why are there so many studies demonstrating so many strange things...?
- Do researchers lie?
- Are these studies false?
- ...

Let's take our time machine

Back to 1962!

Things I Have Learned (So Far)

Jacob Cohen *New York University*

Playing with this new toy (and with a small grant from the National Institute of Mental Health) I did what came to be called a meta-analysis of the articles in the 1960 volume of the *Journal of Abnormal and Social Psychology* (Cohen, 1962). I found, among other things, that using the nondirectional .05 criterion, the median power to detect a medium effect was .46—a rather abysmal result. Of course, investigators could not have known how

We shouldn't be finding a lot of significant results... Right?

Back to 1989!

Psychological Bulletin
1989, Vol. 105, No. 2, 309-316

Copyright 1989 by the American Psychological Association, Inc.
0033-2909/89/\$00.75

Do Studies of Statistical Power Have an Effect on the Power of Studies?

Peter Sedlmeier and Gerd Gigerenzer
University of Konstanz, Federal Republic of Germany

The long-term impact of studies of statistical power is investigated using J. Cohen's (1962) pioneering work as an example. We argue that the impact is nil; the power of studies in the same journal that Cohen reviewed (now the *Journal of Abnormal Psychology*) has not increased over the past 24 years. In 1960 the median power (i.e., the probability that a significant result will be obtained if there is a true effect) was .46 for a medium size effect, whereas in 1984 it was only .37. The decline of power is a result of alpha-adjusted procedures. Low power seems to go unnoticed: only 2 out of 64 experiments mentioned power, and it was never estimated. Nonsignificance was generally interpreted as confirmation of the null hypothesis (if this was the research hypothesis), although the median power was as low as .25 in these cases. We discuss reasons for the ongoing neglect of power.

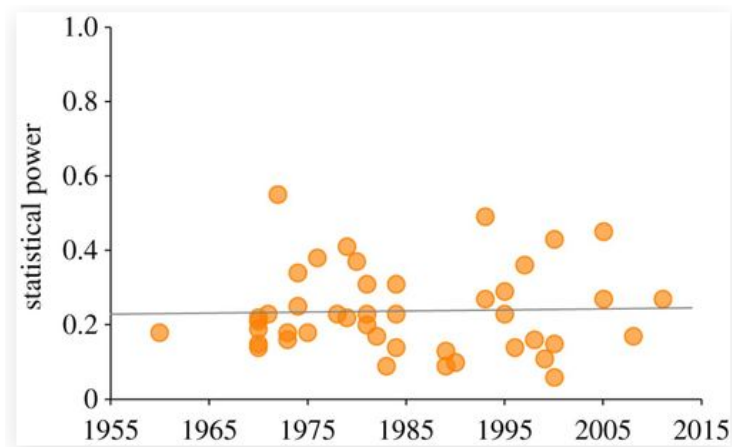
Ok! Message received!

Why is people not aware of this! Journals should be flooded with negative results!

So not many significant results...

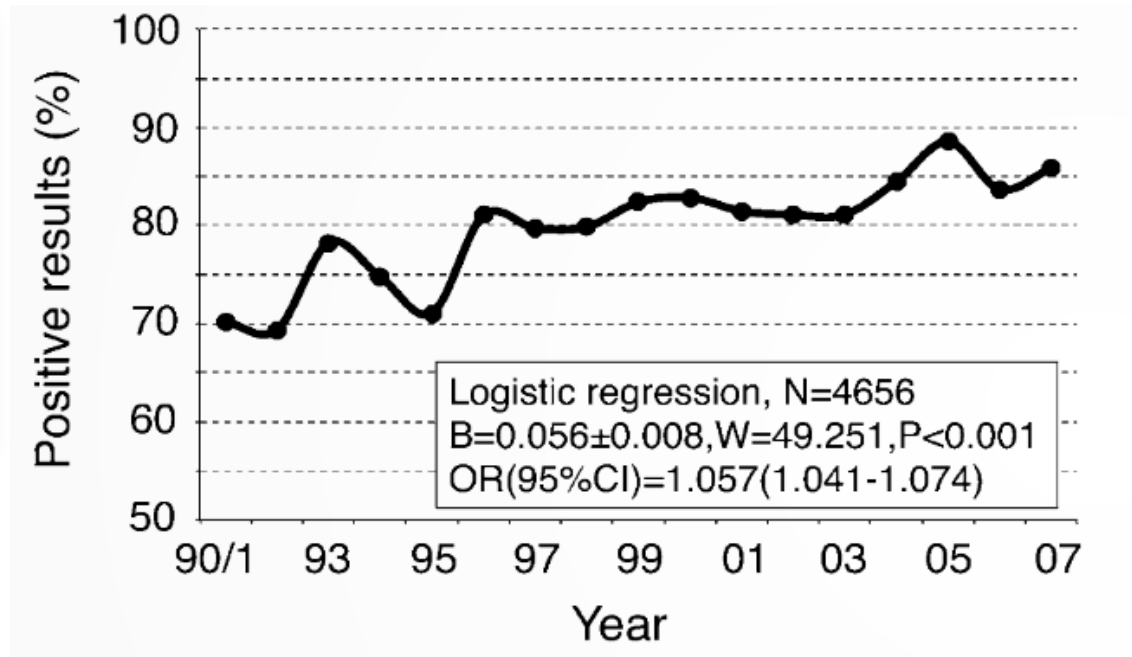
Right?

Because they are not flooded.



Wait wait!...

There should be less positive results! What's happening there!



Percentage of Significant Results per year

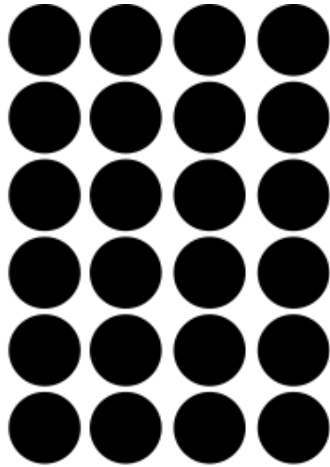
Publication Bias

- For a long time only positive results were published in journals
- Negative results had been deemed irrelevant
 - Poorly conducted studies
 - Low statistical power
 - *No evidence in them*
- New study finds evidence that people have the ability of reading the future! **COOL!**
- New study finds **no** evidence that people have the ability of reading the future! **BORING!**

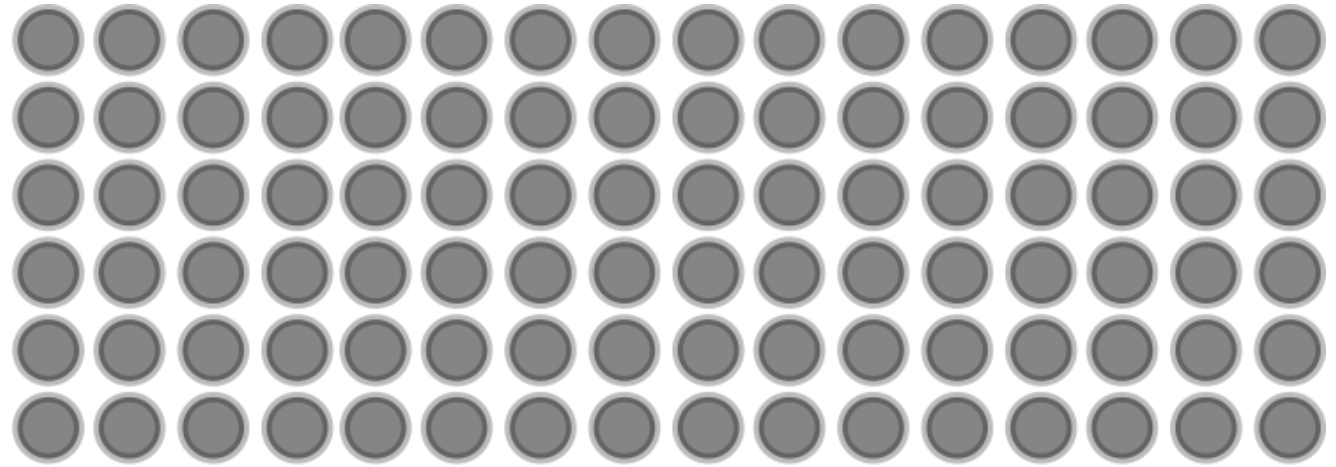
But this is logical I don't want to read the second paper, do I?

Publication Bias

Could it be possible we are only seeing false positive papers?



Published
(Significant)



Unpublished
(Non Significant)

- Imagine all this studies refer to a single effect.
- I could be sure of it existing.
- But in the fact it is all a false-positives!

Publication Bias

It does occur

SOCIAL SCIENCE

Publication bias in the social sciences: Unlocking the file drawer

Annie Franco,¹ Neil Malhotra,^{2*} Gabor Simonovits¹

We studied publication bias in the social sciences by analyzing a known population of conducted studies—221 in total—in which there is a full accounting of what is published and unpublished. We leveraged Time-sharing Experiments in the Social Sciences (TESS), a National Science Foundation–sponsored program in which researchers propose survey-based experiments to be run on representative samples of American adults. Because TESS proposals undergo rigorous peer review, the studies in the sample all exceed a substantial quality threshold. Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up. We provide direct evidence of publication bias and identify the stage of research production at which publication bias occurs: Authors do not write up and submit null findings.

Publication bias occurs when “publication of study results is based on the direction or significance of the findings” (1). One pernicious form of publication bias is the greater likelihood of statistically signif-

icant results being published than of insignificant results being published. Selective reporting is referred to as publication bias because it represents a selection process that results in published research that is not representative of the true population of research results. Publication bias can be thought of as a selection process that results in published research that is not representative of the true population of research results. Publication bias can be thought of as a selection process that results in published research that is not representative of the true population of research results.

the state of knowledge in a field or on a particular topic because null results are largely unobservable to the scholarly community. Publication bias has been documented in various disciplines within the biomedical (3–9) and

thresholds (19, 20). However, these visualization-based approaches are sensitive to using different measures of precision (21, 22) and also assume that outcome variables and effect sizes are comparable across studies (23). Last, methods that compare published studies to “gray” literatures (such as dissertations, working papers, conference papers, or human subjects registries) may confound strength of results with research quality (7). These techniques are also unable to determine whether publication bias occurs at the editorial stage or during the writing stage. Editors and reviewers may prefer statistically significant results and reject sound studies that fail to reject the null hypothesis. Anticipating this, authors may not write up and submit papers that have null findings. Or, authors may have their own preferences to not pursue the publication of null results.

A different approach involves examining the publication outcomes of a cohort of studies, either prospectively or retrospectively (24, 25). Analyses of clinical registries and abstracts submitted to medical conferences consistently find little to no editorial bias against studies with null findings (26–31). Instead, failure to publish appears

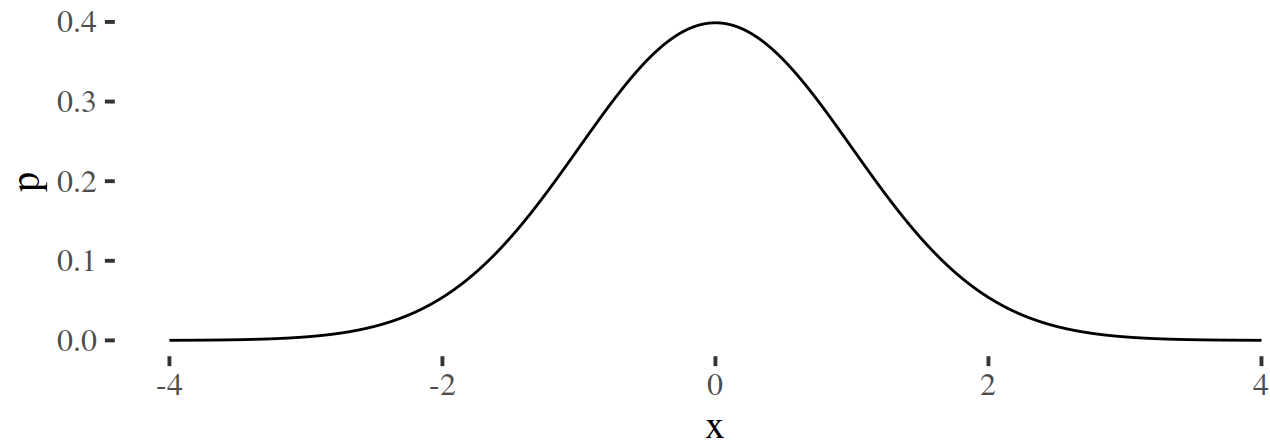
icant results being published than of insignificant results being published. Selective reporting is referred to as publication bias because it represents a selection process that results in published research that is not representative of the true population of research results. Publication bias can be thought of as a selection process that results in published research that is not representative of the true population of research results.

¹Department of Psychology, Stanford University, Stanford, CA, USA
²Department of Psychology, Stanford University, Stanford, CA, USA
*Corresponding author

	Unpublished, not written	Unpublished, written	Published	Book chapter	Missing	Total
Null results	31	7	10	1	0	49
Mixed results	10	32	40	3	1	86
Strong results	4	31	56	1	1	93
Missing	6	1	0	2	12	21
Total	51	71	106	7	14	249

Publication Bias

Study Precision

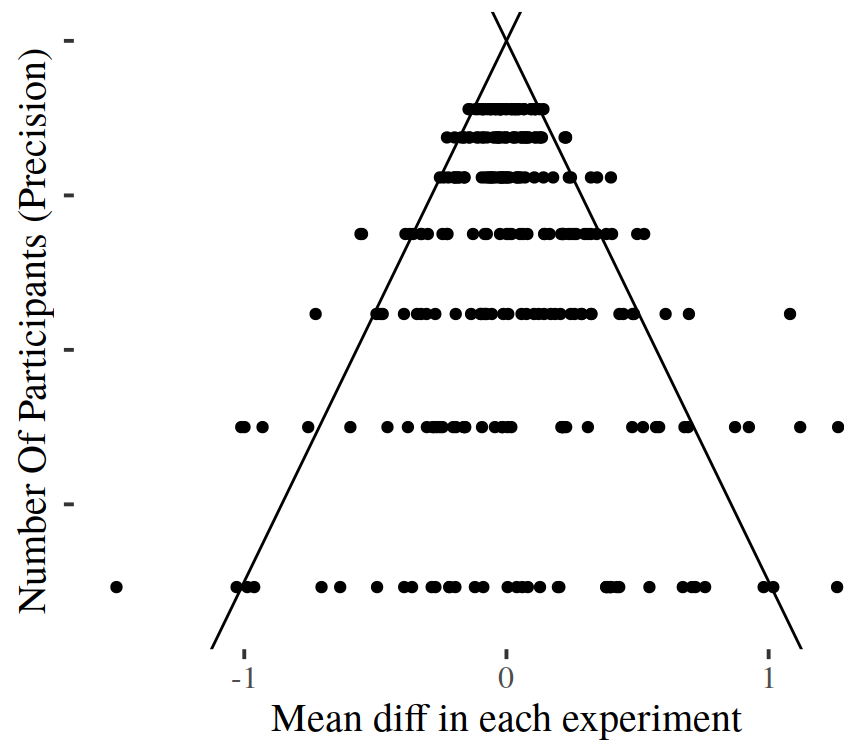


Expected mean?

Publication Bias

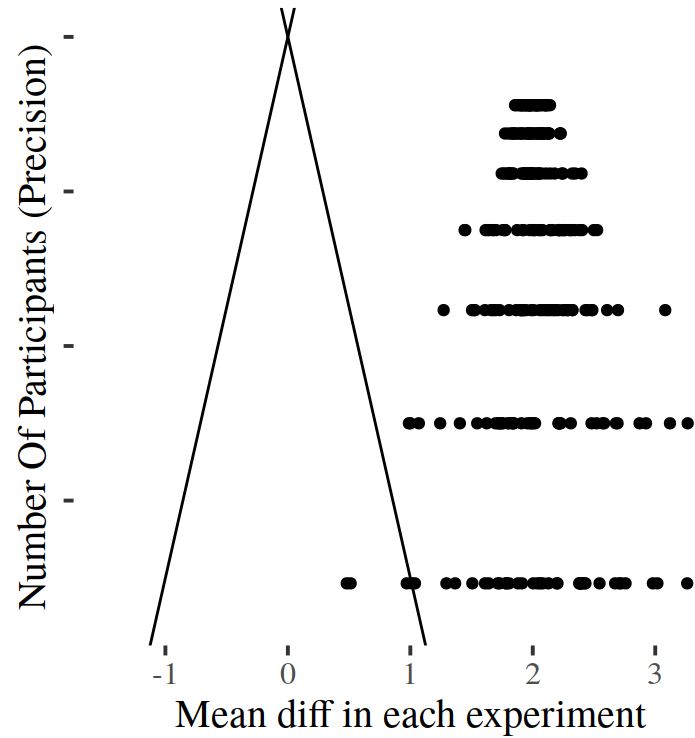
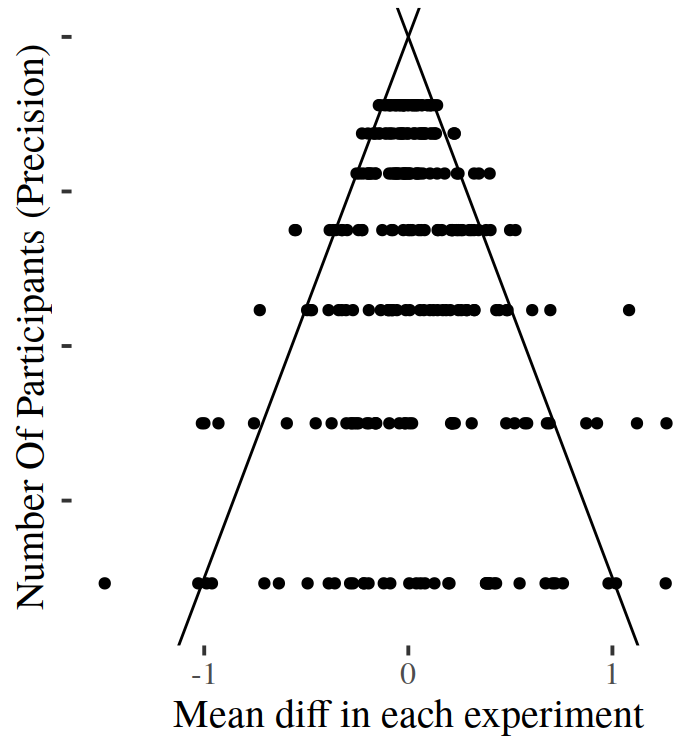
Study Precision

In the long run 0, but for each sample it depends on sample size



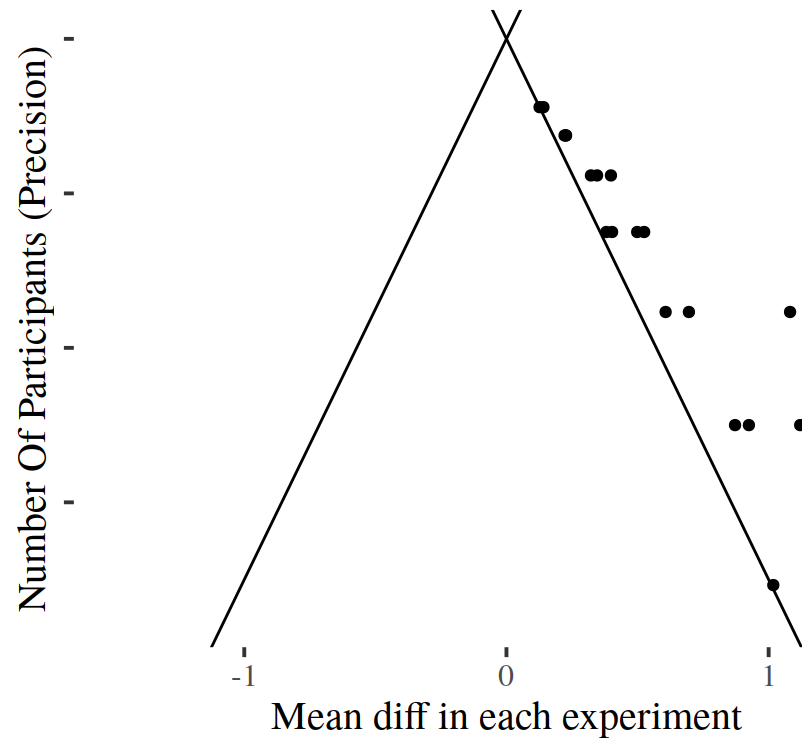
Publication Bias

Study Precision



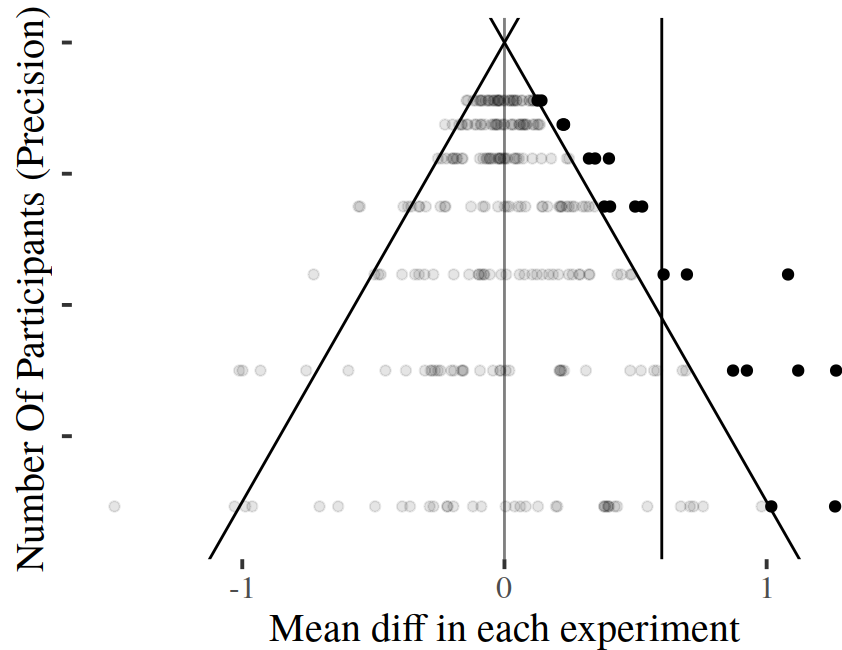
Publication Bias

We gather experiments and see this...



Publication Bias

We gather experiments and see this...



We are not seeing the unpublished studies!

Bias in the mean Effect Size!

We cannot establish if an effect is true or not!

Funnel Plots for Publication Bias

Romance, Risk, and Replication: Can Consumer Choices and Risk-Taking Be Primed by Mating Motives?

David R. Shanks
University College London

Miguel A. Vadillo
King's College London

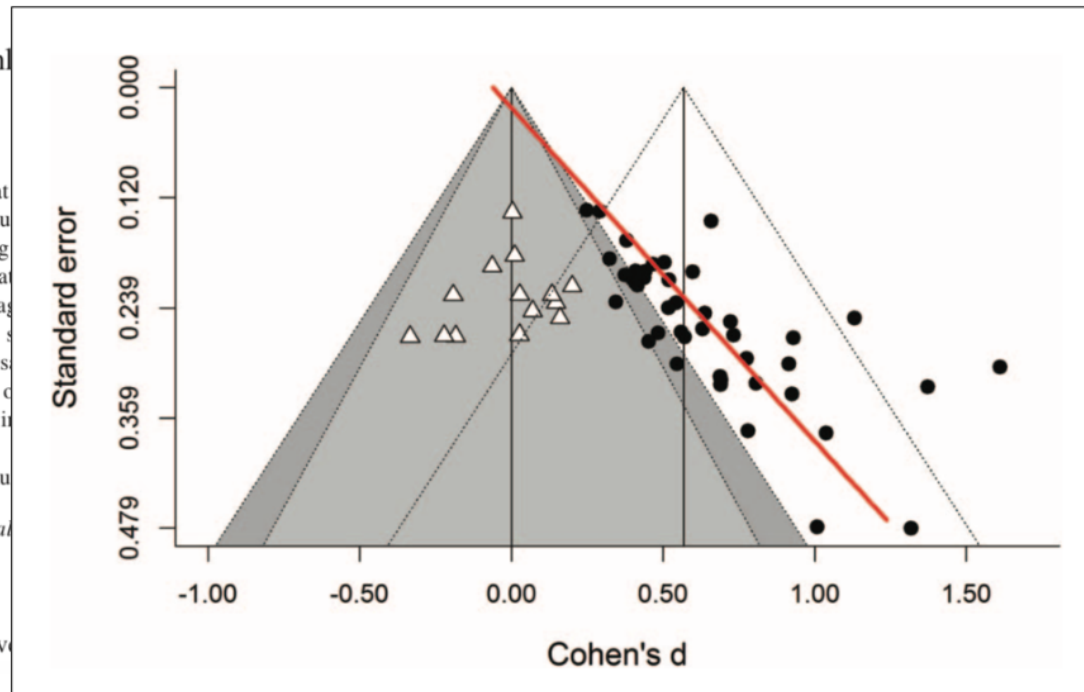
Benjamin Riedel, Ashli

Interventions aimed at
tance. A number of stu
reported that priming
designed to trigger ma
tion items and to enga
this literature reveals s
8 studies with a total s
the studies, including c
romantic primes can i

Keywords: risk, consu

Supplemental material

Extensive efforts have been made in sev



Funnel Plots for Publication Bias

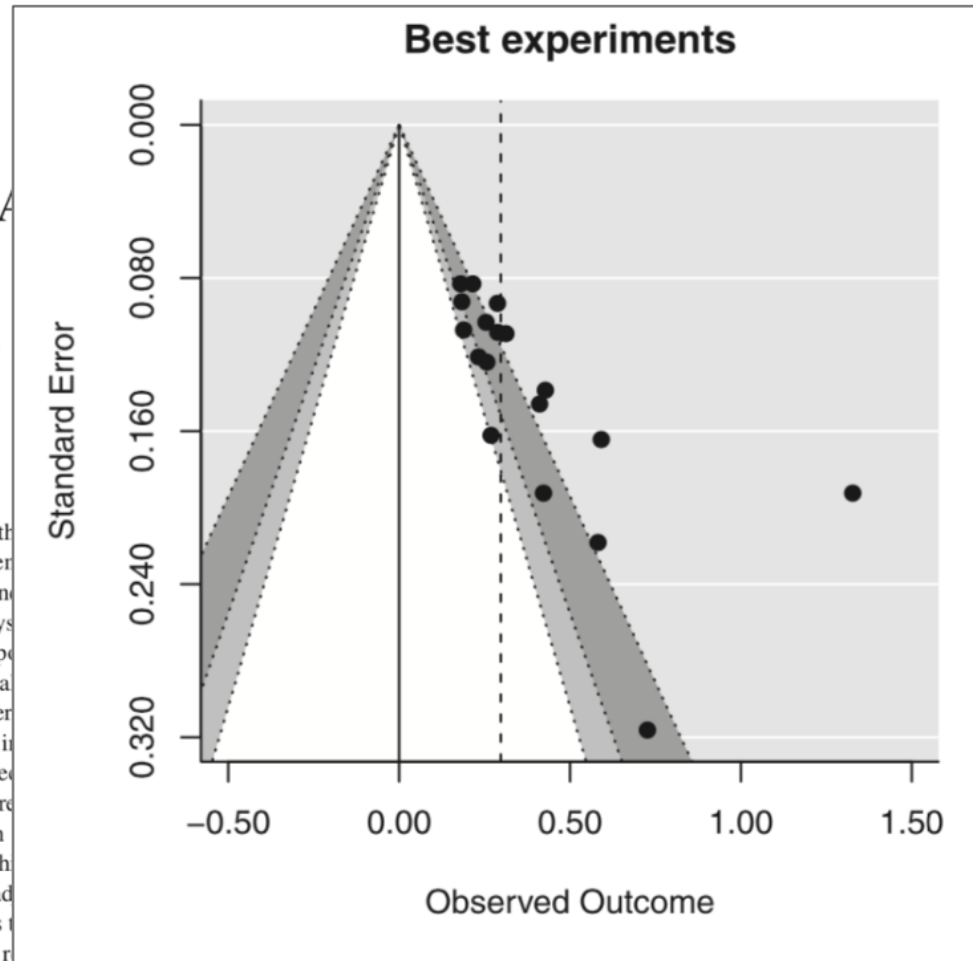
Psychological Bulletin
2017, Vol. 143, No. 7, 757–774

© 2017 American Psychological Association
0033-2909/17/\$12.00 http://dx.doi.org/10.1037/bul0000074

Overstated Evidence for and Behavior: A

Joseph Hilgard
University of Pennsylvania

Violent video games are the
behaviors. Important evidence
colleagues (2010), who found
research. In that meta-analysis
literature, an argument supported
we reexamine their meta-analysis
sizes. Our conclusions differ
substantial publication bias in
and aggressive behavior. See
behavior in experimental research
aggressive affect are much
appear largely unbiased. The
quality do not yield larger ad
bias, indicating that perhaps the
experimental research. The r



Short Interim

Not seeing all results is problematic because:

- Effect sizes are biased/inflated
- Difficult to establish if an effect is true or not

Check on possible solutions:

- Funnel Plots
- Metanalysis

Metaviz package:

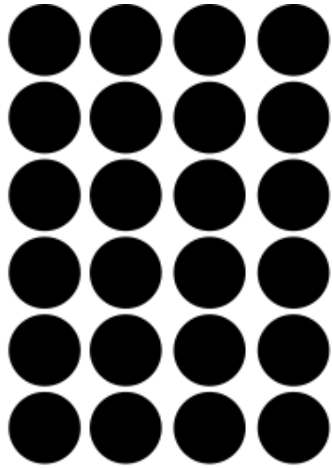
<https://cran.r-project.org/web/packages/metaviz/vignettes/funnelinf.html>

Meta package:

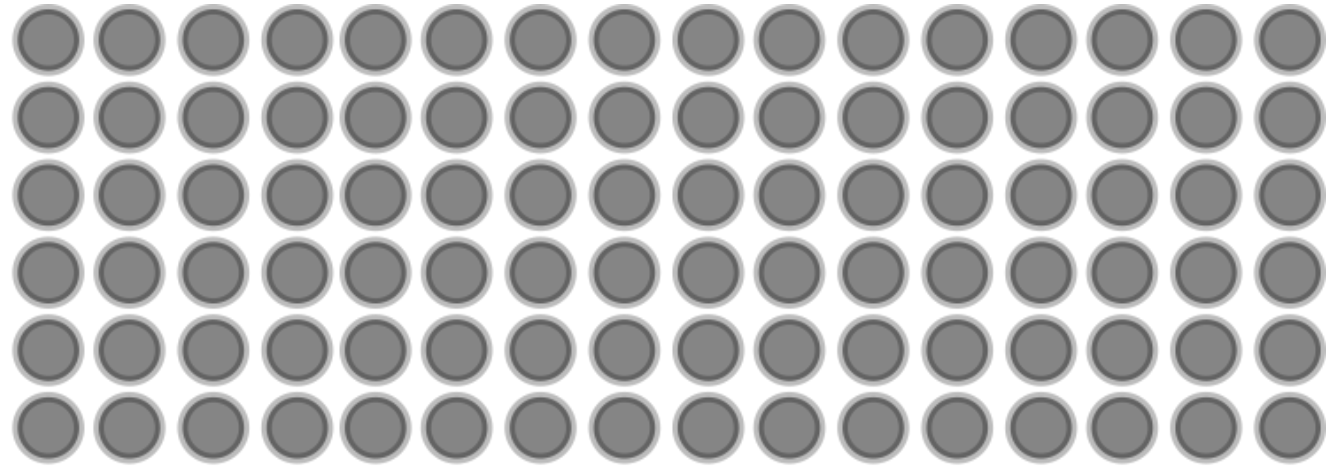
<https://github.com/guido-s/meta>

P-hacking

What is the problem with this figure?



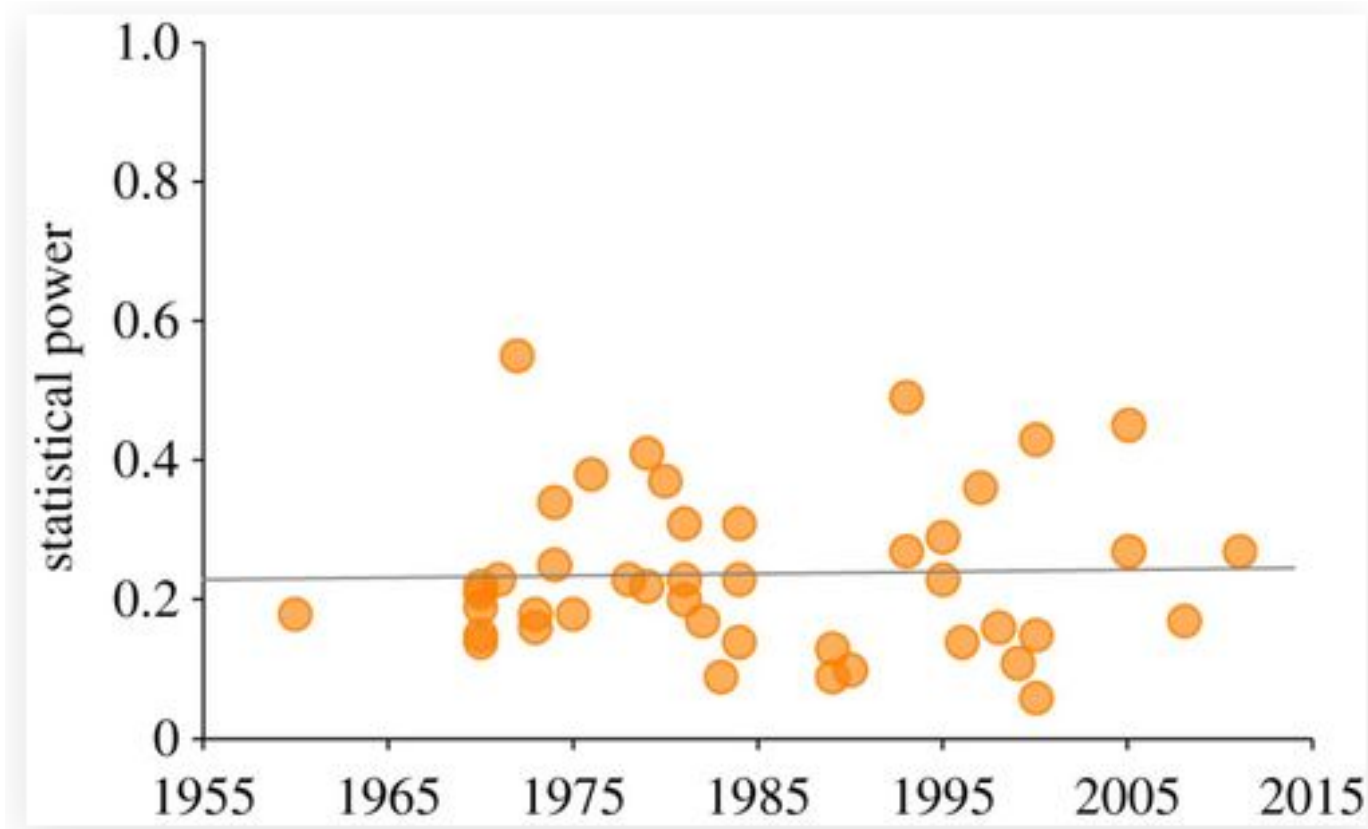
Published
(Significant)



Unpublished
(Non Significant)

- Do researchers really keep all those studies in the file-drawer?
- Do people really run 20 experiments and publish one of them?
- Do I put more than half of my job in a drawer?

Even if not all of them are false-positives



Would people throw away between 20% and 50% of the studies due to non-significant results?

If researchers would have realized that they always obtained non-significant results then they would have done things differently:

"It is one thing for a very young child to believe that 12 peas are enough for dinner and quite another for a chronically starving adult to do so." Nelson & Colleagues (2018)

- Why then?
- Because they were publishing them anyway...

The false positive rate was not 5%!

It was much higher!

Researchers engaged a practice called *p-hacking*

Frequentist tests in the NHST output a parameter called the p-value, if this p-value is below our α (.05) then the test is significant and we accept the alternative hypothesis.

This implies that we should have a 5% false positive rate.

In a typical experiment:

- More than one dependent variable: Accuracy, Reaction Time, Fixation time,...
- More than two conditions: (High Medium Low) Congruent, Incongruent, Neutral,...
- Other variables: Gender, Age, Level of Studies, Language Proficiency,...

I want to assess if people on white shirts are better at distinguishing English backwards from French Backwards .

So I make them listen to many words backwards in English and French. I throw nouns, verbs, adjectives...

And I gather demographic data on them, just in case, because more data cannot hurt me right?

The method that can "prove" almost anything - James A. Smith



The what if game

- White shirts are better than red shirts should be more accurate when distinguishing.
- I do a NHST($\alpha=.05$) on the level of accuracy comparing white and red, **non significant**
- What if the difference is not in accuracy but in how fast distinguishing that, lets check the reaction time. **non significant.**
- What if the effect is really small, so lets look at a subset of the data for more salient words, maybe insults... Or maybe sad words... **non significant.**
- Or maybe the problem was not at all white vs red shirts but happy vs. sad words! **non significant.**
- Maybe this only happen for women? **non significant**

The what if game

- Maybe it is relevant for the subset of participants if their father are fluent in English and in French?
- **AHA! Significant!**
- "A new study demonstrates that women are faster when discriminating sad words from happy words backwards.
- But we kindly forgot to mention, that we tried 30 different things before we found one that was significant.

The what if game

- False Positive Rate?

- Up to
61%

in some cases probably even more.

Harking

Hypothesizing after the results are known

- An additional step here, in many cases the hypothesis is rewritten in a way that the final result seems to have been hypothesized from the beginning.

So ok scientists lie? Everything is a farce?

- We didn't know these effects until recently
- Researchers were just trying to make sense of their data
- Humans suffer from extreme apophenia!
- We love being right
- **We are promoted based on the number of publications**
- **Papers are/were valued according to novelty and significance, not methods**
- All of the above occurs due to a combination of wrong incentives, *being human* and ignorance

Ok so,... How do we fix all of this?

Not a single solution

Increase statistical power by using correct sample sizes, usually bigger.

Better estimations of the effect size, the more participants I have the more precision in the estimation.

Preregistration

- Write your hypothesis, methods, sample size, expected Type I error and Type II error, ...
- **Before starting the experiment**
- That way you cannot change things afterwards inadvertently
- Avoids p-hacking

Registered Report in Journals

- Preregistration plus:
- Evaluate papers based on the interest of the question and the methodology.
- Accept or reject them based on the above, regardless of significance or result
- Avoids publication bias

Not a single solution

Replication

- Truthfulness of an effect is difficult to establish with a single study
- We need to repeat them, by independent groups
- It is not only interesting to know if something is true
- We also want to know how big is the effect (Particularly in clinical applications)
- “A new studies says there is a statistically significant difference of 1 mm between the right and left side of the class”

What should I look for when I read a study

- Is it replicated? By different people?
 - Conceptually or exactly replicated?
 - Don't they change anything? Analysis, measures, paradigm,...
- Have they registered their study before the analysis?
 - No degrees of freedom during the analysis
 - Fixed hypothesis, Fixed methods
- Was their sample size adequate?
- If there are several studies:
 - Is there a metaanalysis?
 - Can I draw a funnel plot?

What should I look for when I read a study

- **Be critic!**
- Do not just, believe, peer review is not enough, being published is not enough
- Avoid confirmation biases (I look for evidence that support my belief)
- If it is too good to be true, then it is probably not true
- Huge claims asks for huge evidence

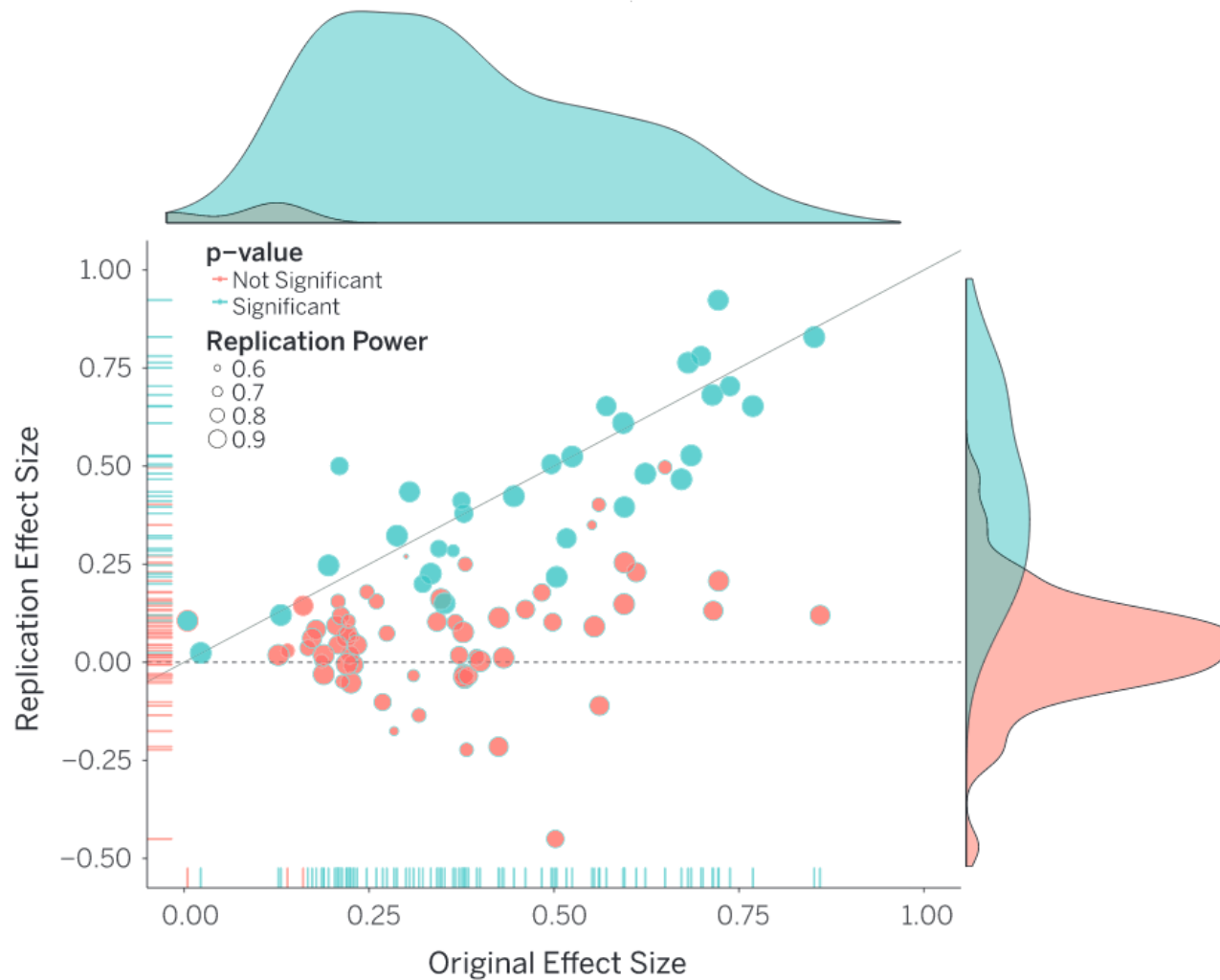
Do you believe me?

Do not believe me!

Were you not listening?

**Show me the evidence of the
problem!**

Open Science collaboration



Approximately 35% to 50% of 100 studies replicated

Literally pre cognition

People can see or change the future

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

© 2011 American Psychological Association
0022-3514/11/\$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

- People can see/change the future and researchers should be more open-minded
- I have looked at my data until I found a subset of participants, and a subset of stimuli, after extensive testing and repeating pilots and studies, that showed **FINALLY** a significant effect. Sadly, I have not been able to replicate it (properly) ever again
- He was not evil, all of the above was written in the paper, we just didn't know well yet
- People were not critic enough

Power Pose

Amy Cuddy: 30 Seconds on Power Poses



TED talks, still out there, lots of criticism and papers againts...

"Good people can do bad science. Indeed, if you have bad data you'll do bad science (or, at best, report null findings), no matter how good a person you are.

[...]being a scientist, and being a good person, does not necessarily mean that you're doing good science. I don't know Cuddy personally, but given everything I've read, I imagine that she's a kind, thoughtful, and charming person.[...] In any case, it's not my job to judge these people nor is it their job to judge me.[...]

Conversely, if Eva Ranehill, or Uri Simonsohn, or me, or anyone else, performs a replication [...] and finds that your data are too noisy for you to learn anything useful, then they may be saying you're doing bad science, but they're not saying you're a bad person."

Andrew Gelman

<https://statmodeling.stat.columbia.edu/2017/10/18/beyond-power-pose-using-replication-failures-better-understanding-data-collection-analysis-better-science/>

- Stapel
- Bargh
- Hans Eysenck
- ...

Ok but it's only in psychology!

Nope...

Ok but it's only in psychology!

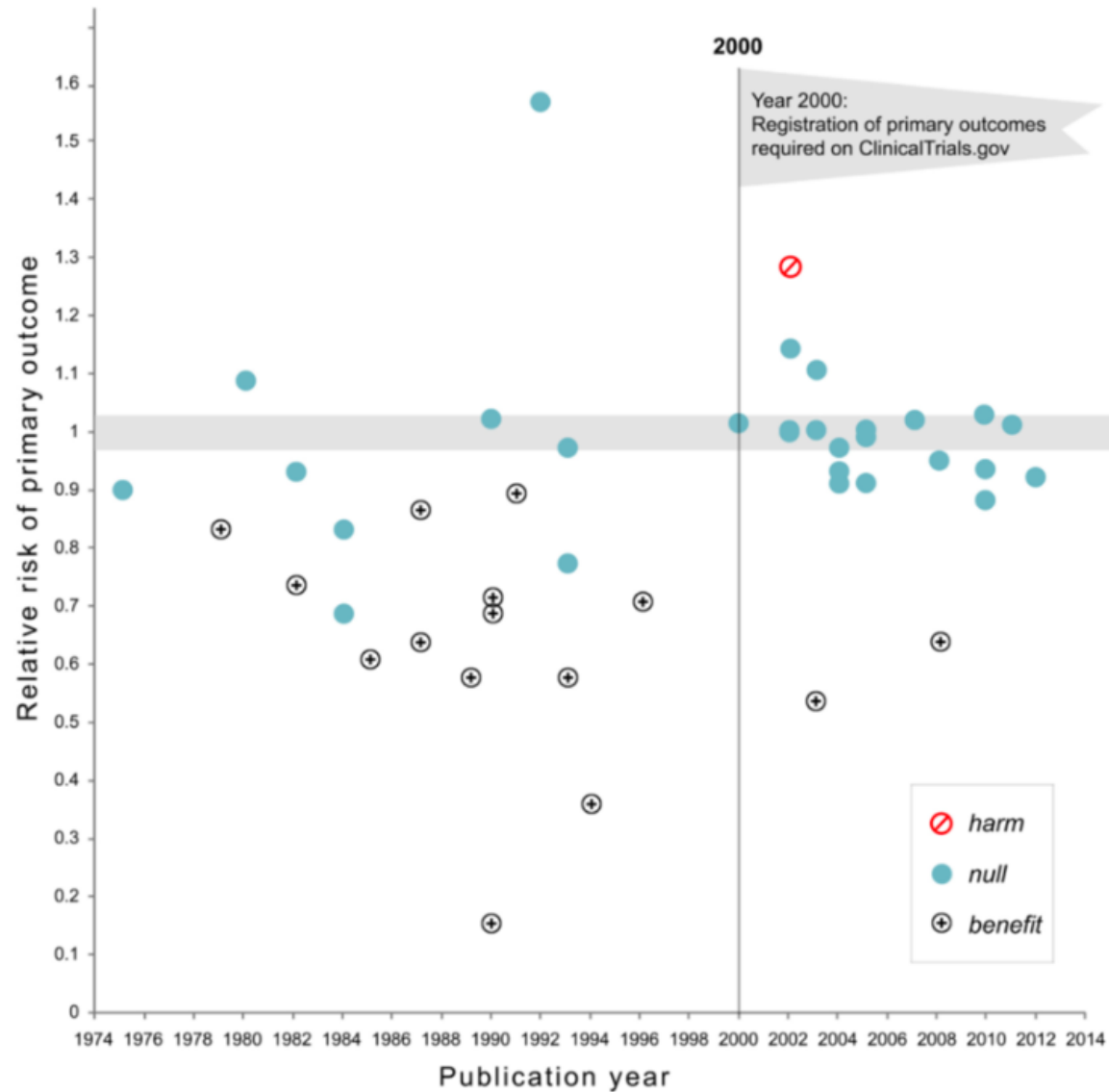
Nope...

Cancer reproducibility project yields first results

In 2011, Bayer researchers made a splash with news that they could only replicate 25% of the preclinical academic projects that they took on ([Nat. Rev. Drug Discov. 10, 712; 2011](#)). Amgen fared even worse when trying to recreate the findings from cancer papers, with just an 11% success rate ([Nature 483, 531–533; 2012](#)).

Ok but it's only in psychology!

Nope...



Open science

Open science

- Until recently data, methods, stimuli was kept in the labs
- People was not very open to share them
- Open Science Foundation and others
 - Public data
 - Public analysis
 - Public Stimuli
 - ...
- Easier to replicate and check results
- Easier to run metaanalyses
- ...

Some questions

- Adding observations, using different measures, dropping conditions,... based on the significance of the statistical test is called:
- p-hacking
- The problem that arises when only significant results are published is called:
- Publication Bias
- The fact that a paper is published in a journal is enough guarantee of its credibility:
- No
- Problems such as p-hacking or publication bias arise from conscious malpractice of the researchers
- No, it arises from ignorance, wrong incentives and making sense of data
- **Open science refers to:**
- Making accessible your data, stimuli, analyses,...

Some questions

- One of the advantages of registered reports is:
 - Evaluate based on methodology and interest of research question, avoid publication bias,...
- One of the advantages of preregistration:
 - It helps to avoid p-hacking
 - Publication bias is helps researchers because it filters only good research:
 - No
 - Adding observation to a sample, because our statistical is not yet significant, doesn't raise the false positive rate
 - FALSE

Questions

The end

References

- Statistical Inference as Severe Testing (Deborah G. Mayo, Book)
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Leif D. Nelson, Joseph Simmons, Uri Simonsohn Psychology's Renaissance *Annual Review of Psychology* 2018 69:1, 511-534
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), 160384. <https://doi.org/10.1098/rsos.160384>
- A year of horrors: <http://ejwagenmakers.com/2012/Wagenmakers2012Horrors.pdf>
- Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, 7(6), 543–554. <https://doi.org/10.1177/1745691612459060>

References

- Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3), 666–678. <https://doi.org/10.3758/s13428-011-0089-5>
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Understanding Psychology as a Science (Zoltan Dienes, book)
- Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, 7(6), 645–654. <https://doi.org/10.1177/1745691612464056>
- Preregistration revolution Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, David T. Mellor *Proceedings of the National Academy of Sciences* Mar 2018, 115 (11) 2600-2606; DOI: 10.1073/pnas.1708274114
- The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Andrew Gelman and Eric Loken (2013)
http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Nuijten, M. B. (2018). Practical tools and strategies for researchers to increase replicability. *Developmental Medicine & Child Neurology*, 61(5), 535–539. <https://doi.org/10.1111/dmcn.14054>

References

- Lakens, D., Adolphi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., ... Zwaan, R. A. (2017, September 18). Justify Your Alpha. <https://doi.org/10.31234/osf.io/9s3y6>
- Better methods can't make up for mediocre theory. Smaldino (2019) Nature editorial