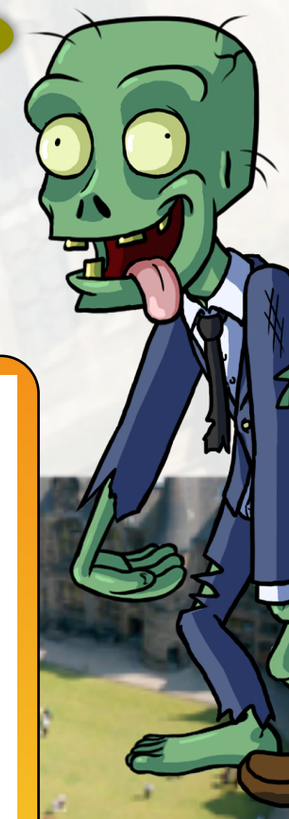
University of Glasgow | Institute of Neuroscience & Psychology

CENTRE FOR COGNITIVE NEUROIMAGING

A simple cure to the p < 0.05 disease

Guillaume Rousselet
@robustgar
https://garstats.wordpress.com

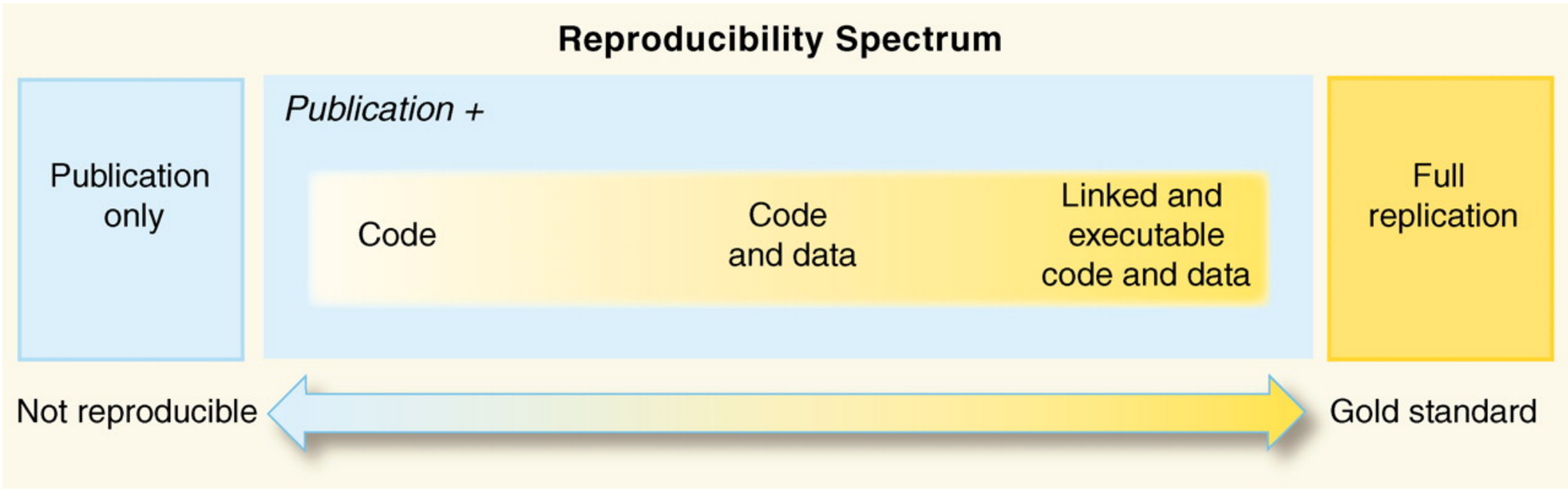


"The strategy of run-a-crappy-study, get p less than .05, come up with a cute story based on evolutionary psychology, and PROFIT . . . well, it does not work anymore. OK, maybe it still can work if your goal is to get published in PPNAS, get tenure, give Ted talks, and make boatloads of money in speaking fees. But it will not work in the real sense, the important sense of learning about the world."

Andrew Gelman, 2018, The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*

Definitions

- "We define **reproducibility** as the ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline."



- "The **replicability** of a study is the chance that an independent experiment targeting the same scientific question will produce a consistent result."

Jeffrey T. Leek and Roger D. Peng (2015) Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *PNAS*
Roger D. Peng (2011) Reproducible Research in Computational Science. *Science*

Errors…

False positives

Forstmeier, W., Wagenmakers, E.J. & Parker, T.H. (2016) **Detecting and avoiding likely false-positive findings – a practical guide.** *Biol Rev Camb Philos Soc.*

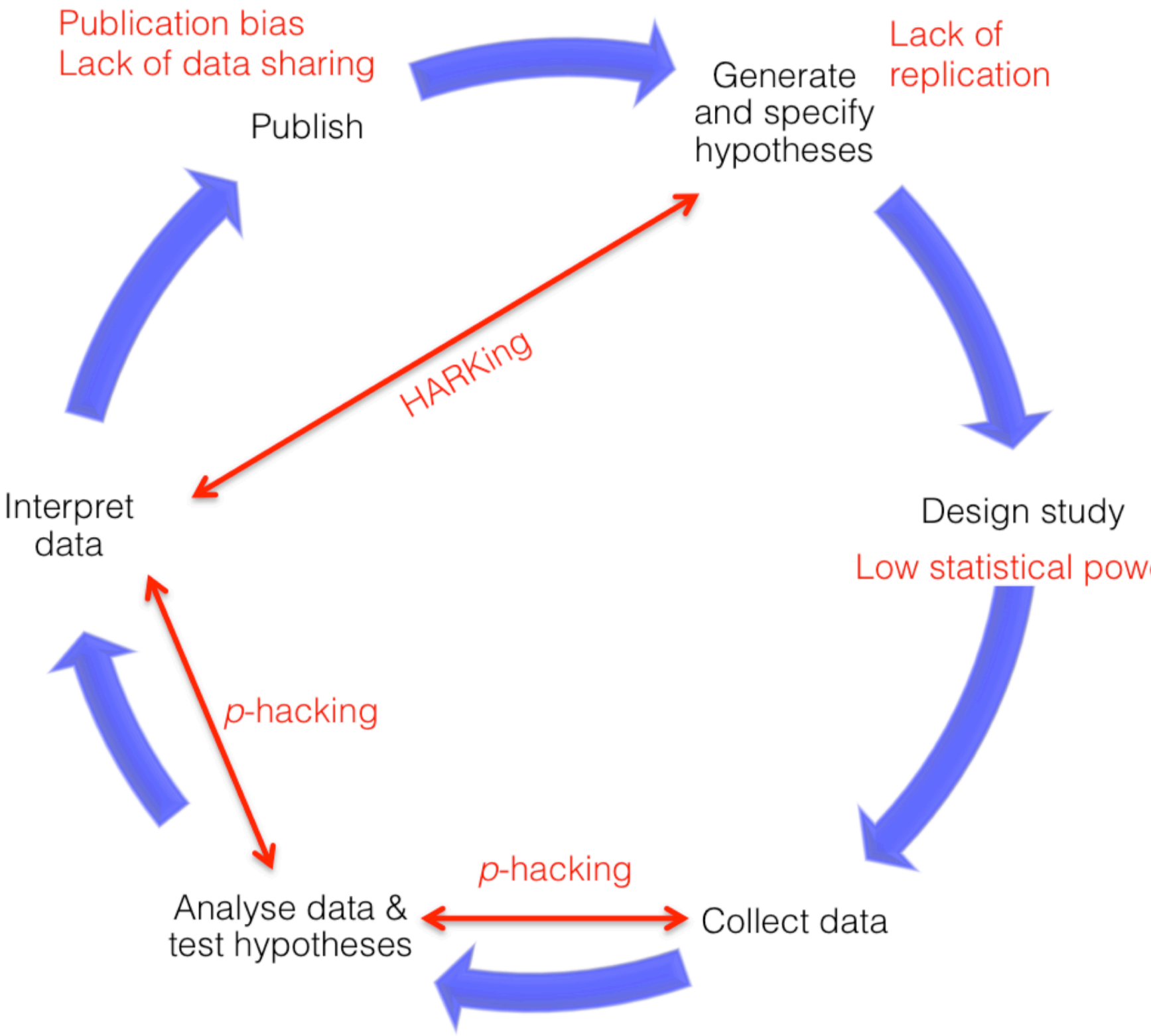
False negatives

statistical power

robust statistics

Precision (type M & S errors)

Gelman, Andrew, and John Carlin (2014).**Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors**. *Perspectives on Psychological Science* 9, no. 6: 641–51

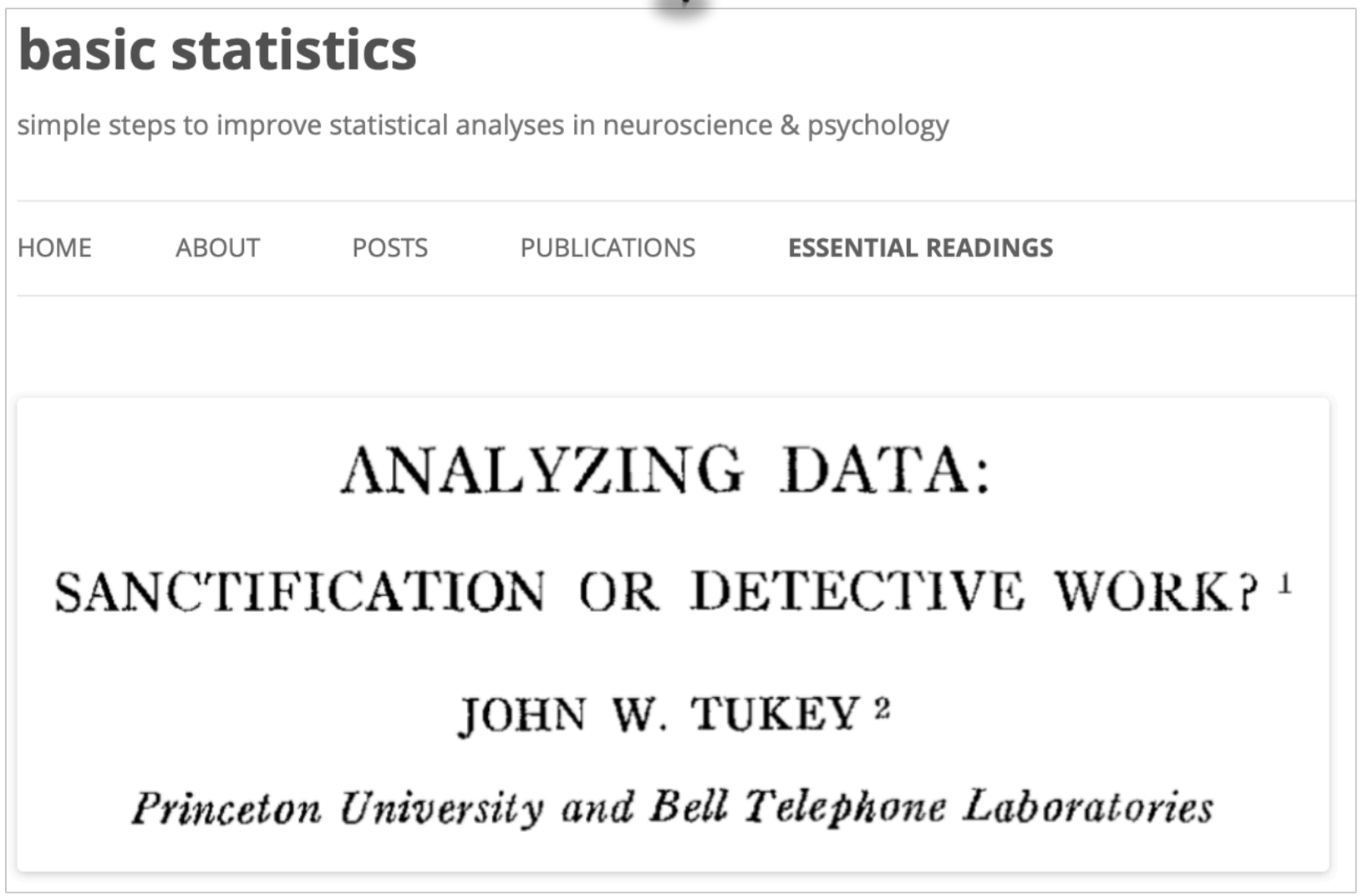


Chambers, Christopher D., Feredoes, Eva, Muthukumaraswamy, Suresh Daniel and Etchells, Peter 2014. **Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond.** *AIMS Neuroscience* 1 (1) , pp. 4-17. 10.3934/Neuroscience2014.1.4

symptoms of diseases

- scientists are not immune to cognitive biases
- training issues (methods, stats, philosophy)
- incentives (stupid metrics, publish or perish)
- publishing system (prestigious journals make careers)
- …

Solutions



basic statistics

simple steps to improve statistical analyses in neuroscience & psychology

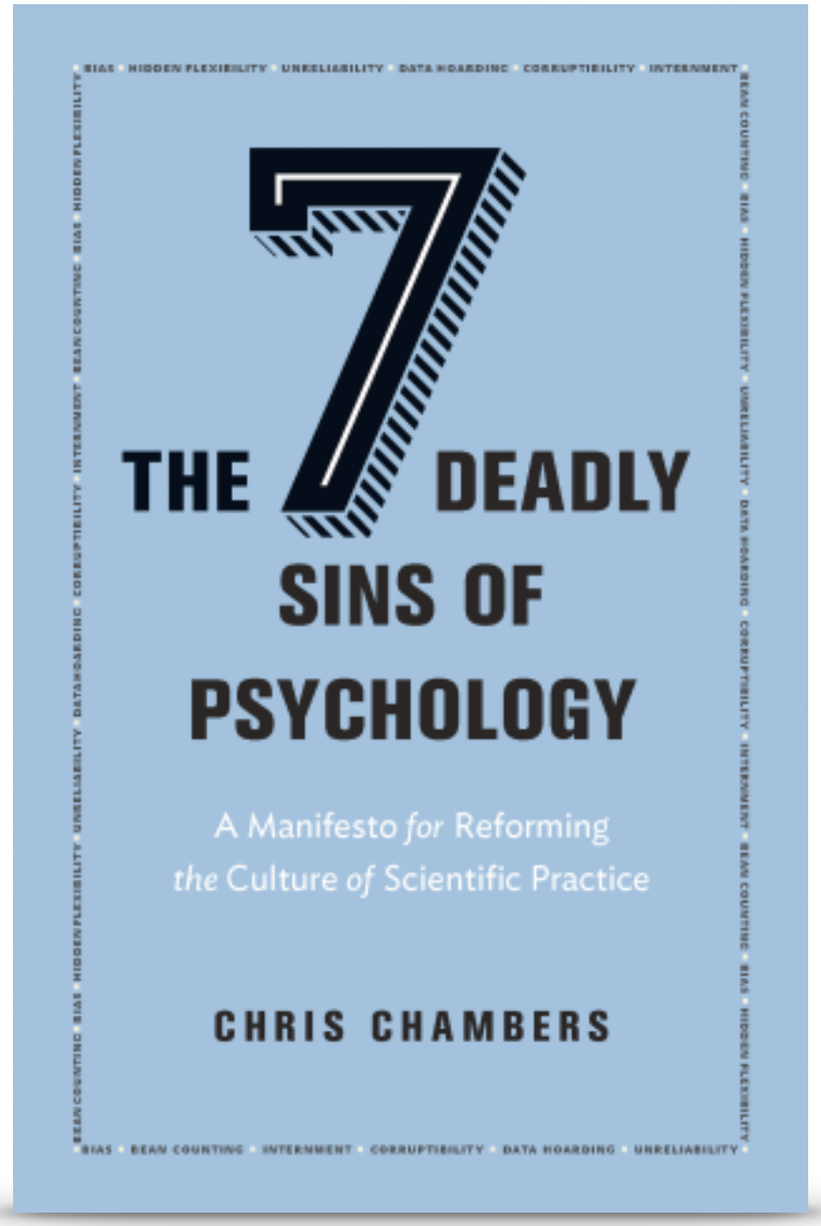
HOME ABOUT POSTS PUBLICATIONS **ESSENTIAL READINGS**

ANALYZING DATA:

SANCTIFICATION OR DETECTIVE WORK? [1]

JOHN W. TUKEY [2]

Princeton University and Bell Telephone Laboratories



Andrew Gelman (2018) **The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It.** *Personality and Social Psychology Bulletin* 44, no. 1: 16–23

typical paper in my experience = NHST + …

- “difference between A and B was significant ($p<0.05$)”
- ($p=0.07$) “borderline significant, approaching significance”…
- “A and B did not differ ($p>0.05$)” (not significant…)
- … discussion of binary outcomes within study and between studies

This obsession with **p<0.05** is a core problem, leading to bad science

> “if the alternative is correct and the actual power of two studies is 80%, the chance that the studies will both show $P \leq 0.05$ will at best be only 0.80(0.80) = 64%; furthermore, the chance that one study shows $P \leq 0.05$ and the other does not (and thus will be misinterpreted as showing conflicting results) is 2(0.80)0.20 = 32% or about 1 chance in 3. Similar calculations taking account of typical problems suggest that one could anticipate a “replication crisis” even if there were no publication or reporting bias, simply because current design and testing conventions treat individual study results as dichotomous outputs of “significant”/“nonsignificant” or “reject”/“accept.” ”

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. & Altman, D.G. (2016) **Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.** *Eur J Epidemiol*, 31, 337-350.

“ if the alternative is correct
studies is 80%, the chanc
≤ 0.05 will **at best**
furthermore, the chance th
the other does not (and th
showing conflicting results
chance in 3. Similar calcul
problems suggest that one
crisis” even if there were n
simply because current de
treat individual study resul
“significant”/“nonsignifican

assuming:

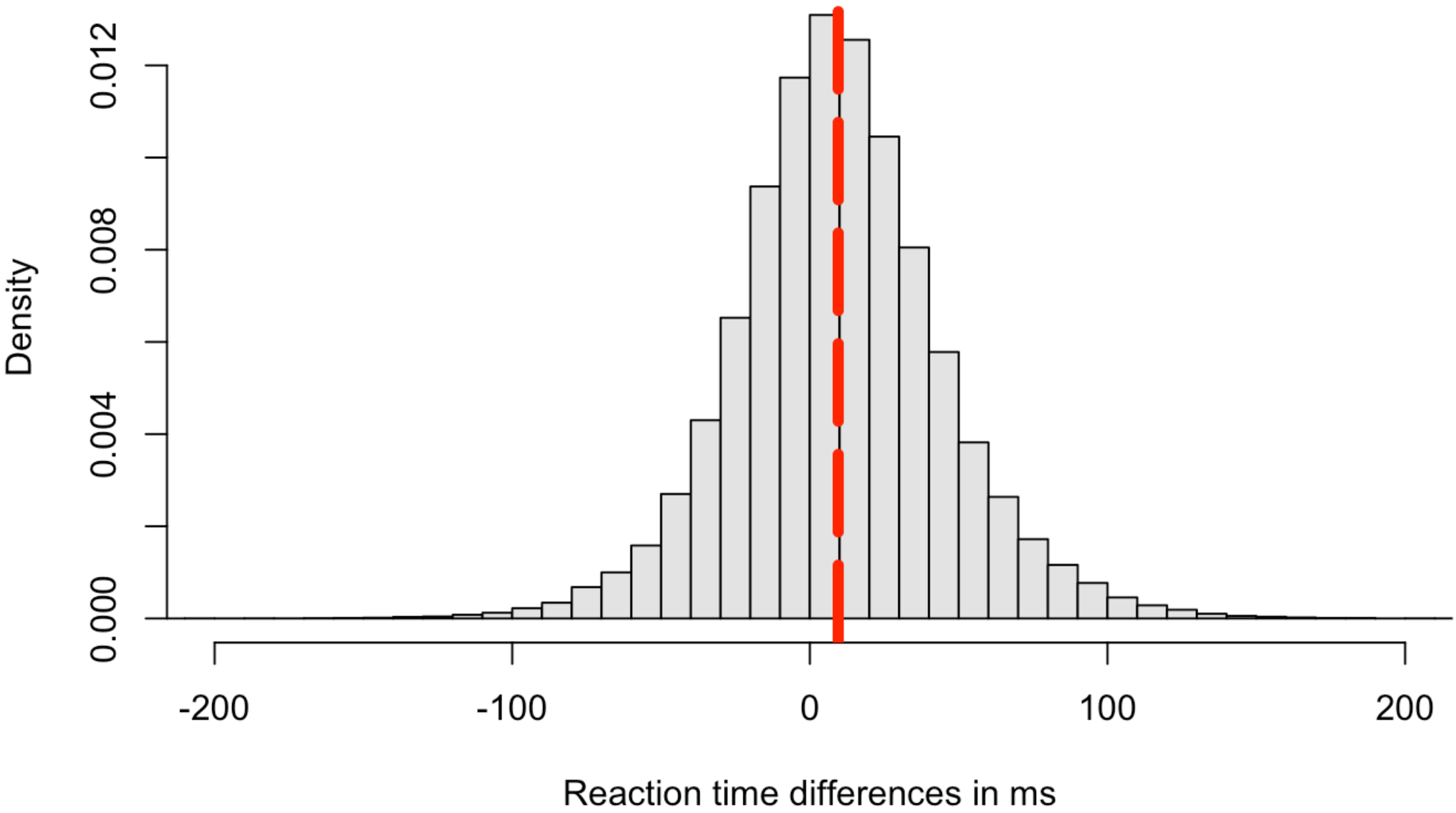
- all goes well
- no measurement noise
- test assumptions are met
- effect size estimation is correct

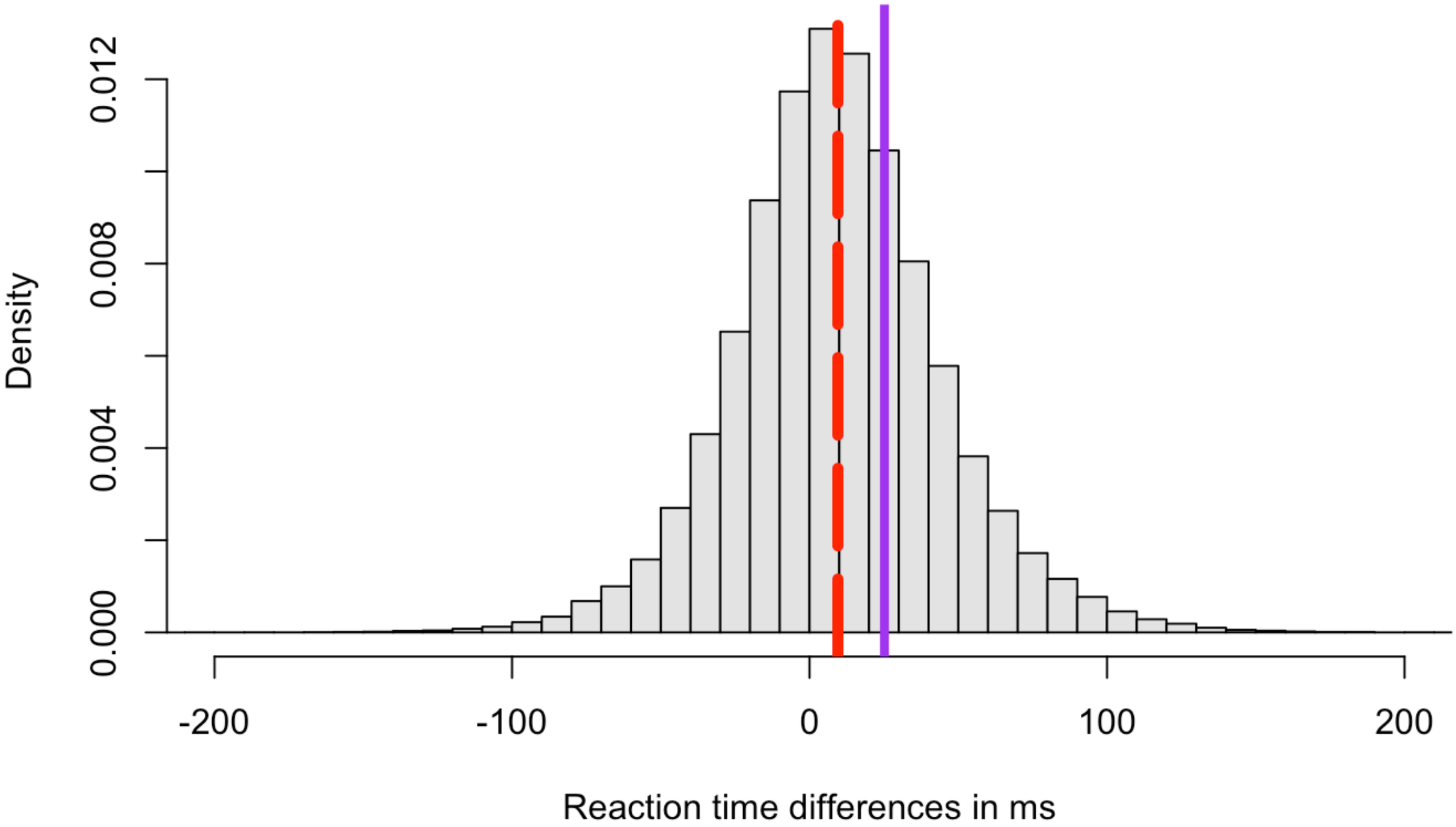
P

d

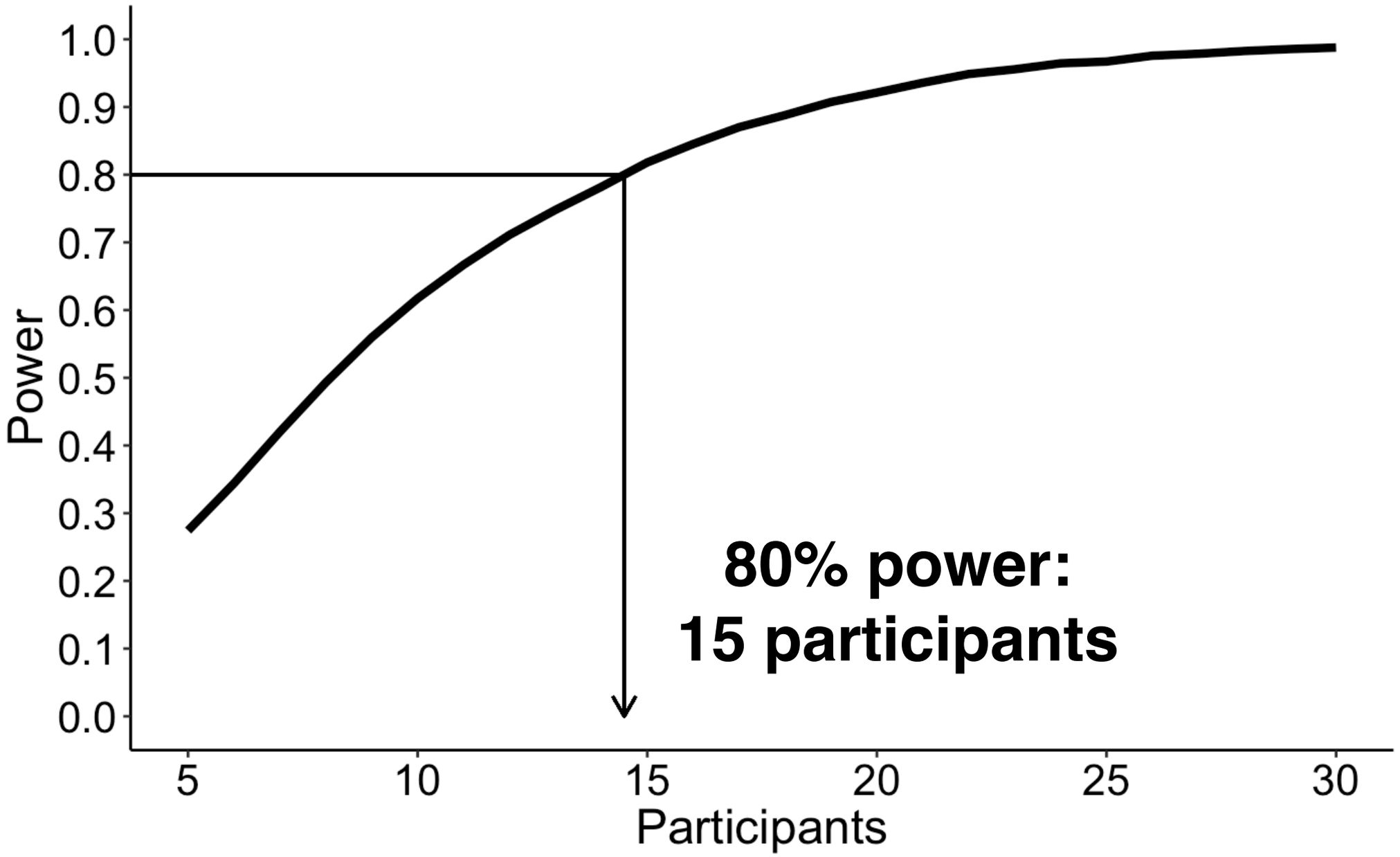
1

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. & Altman, D.G. (2016) **Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.** *Eur J Epidemiol*, 31, 337-350.

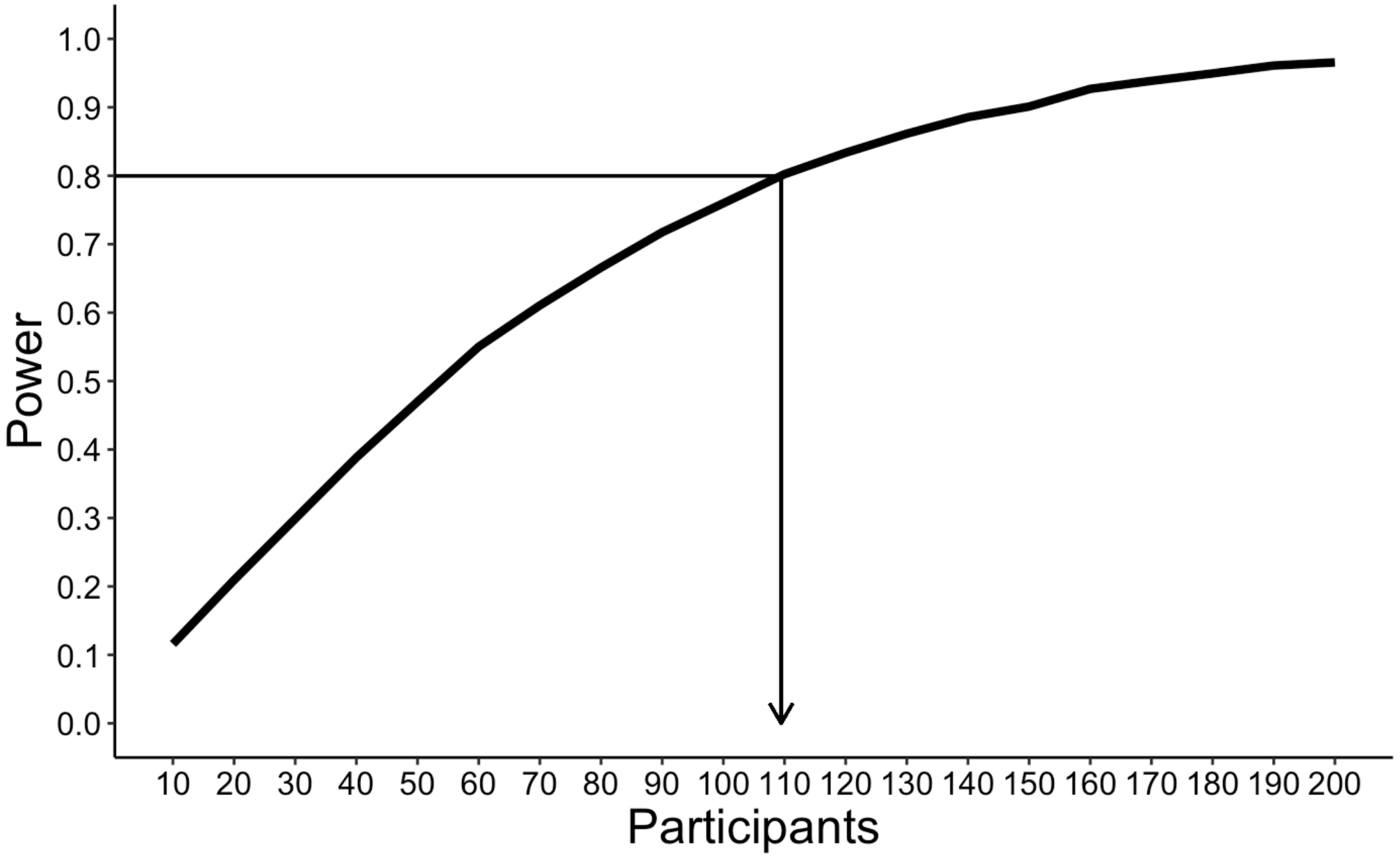

Population values (mean=9.5, sd=35.4, es=0.27)
Density
0.000
0.004
0.008
0.012
-200
-100
0
100
200
Reaction time differences in ms


Population values (mean=9.5, sd=35.4, es=0.27)
Density
0.000
0.004
0.008
0.012
-200
-100
0
100
200
Reaction time differences in ms

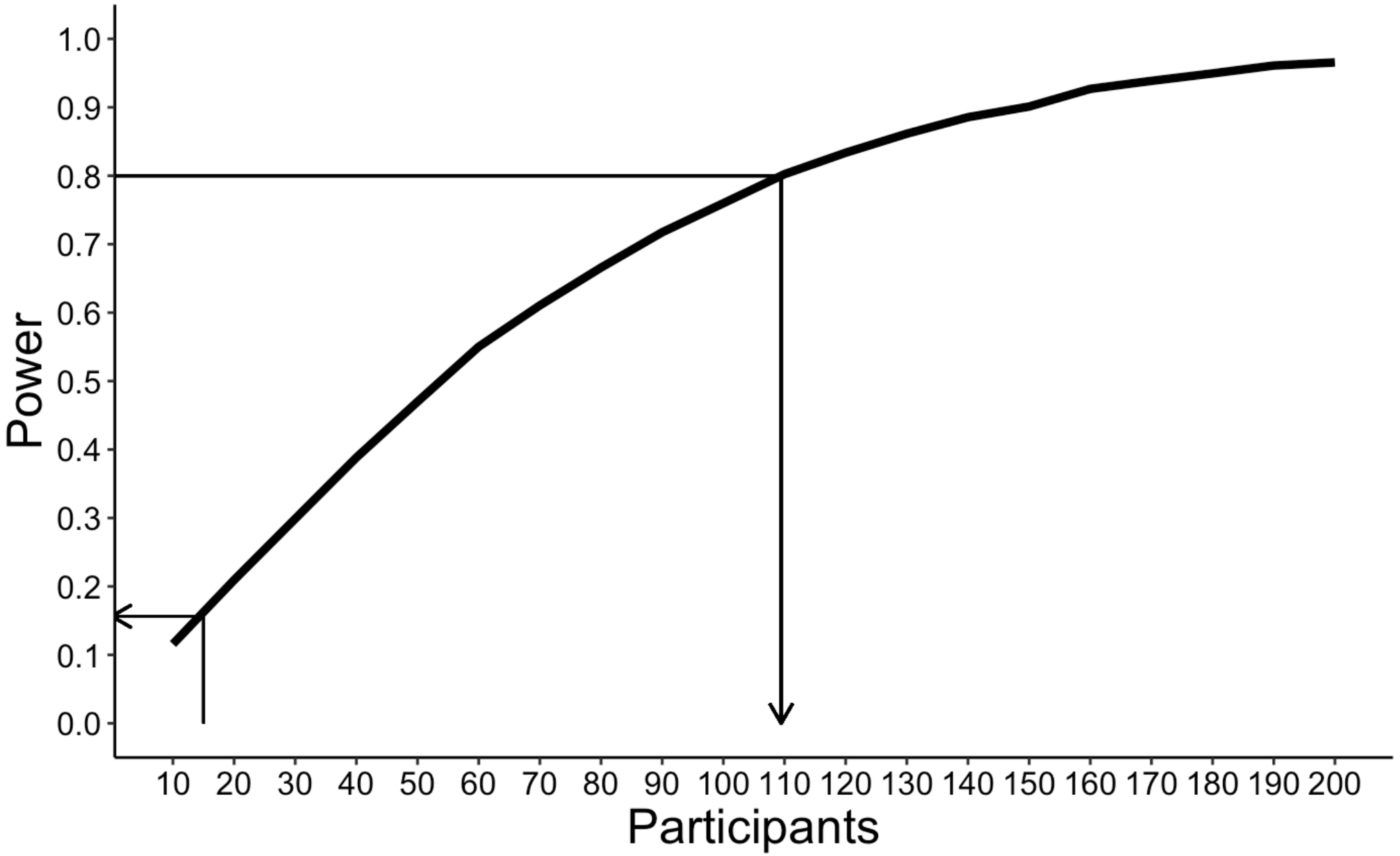
Power curve for expected effect



Power curve for real effect



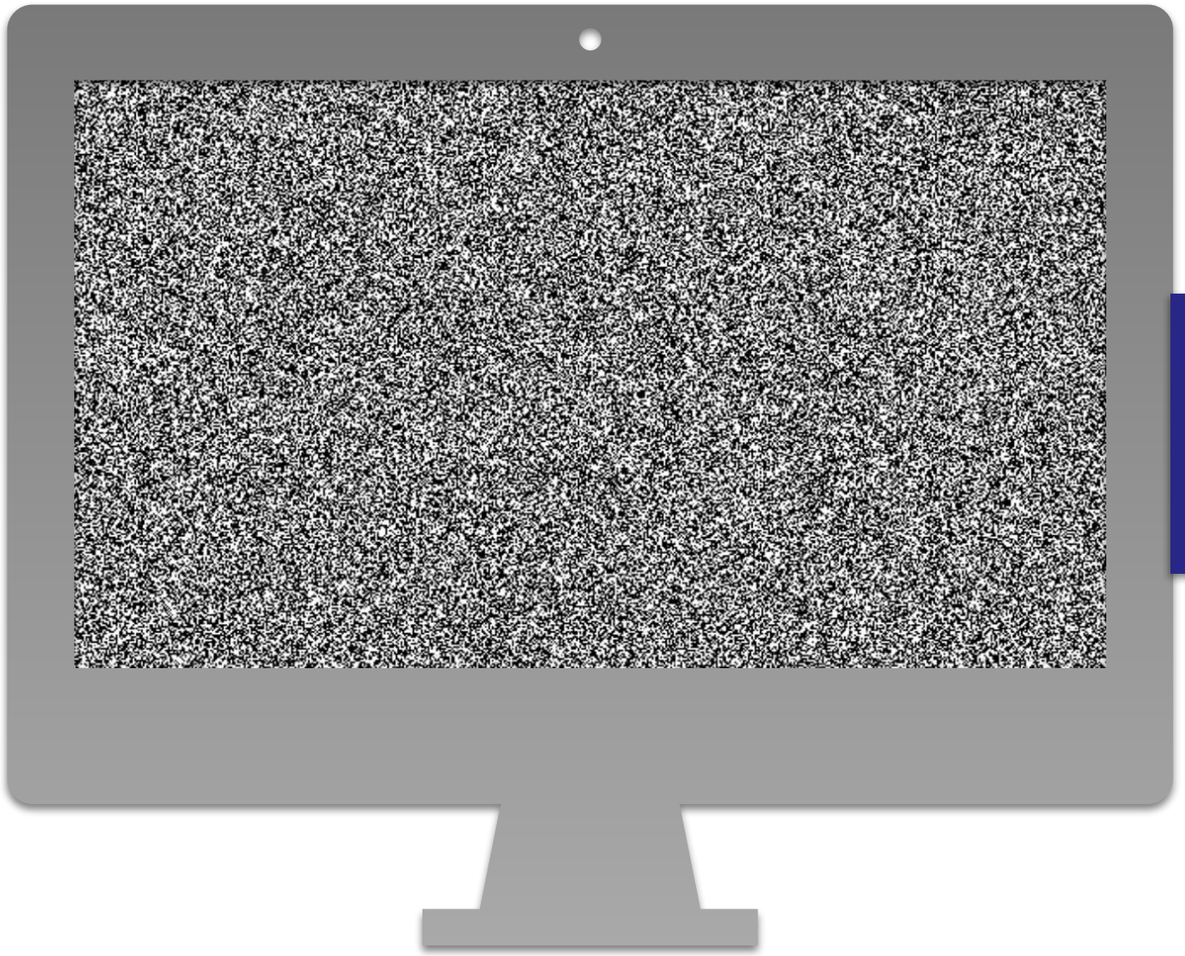
Power actually achieved



Is there a replication crisis?

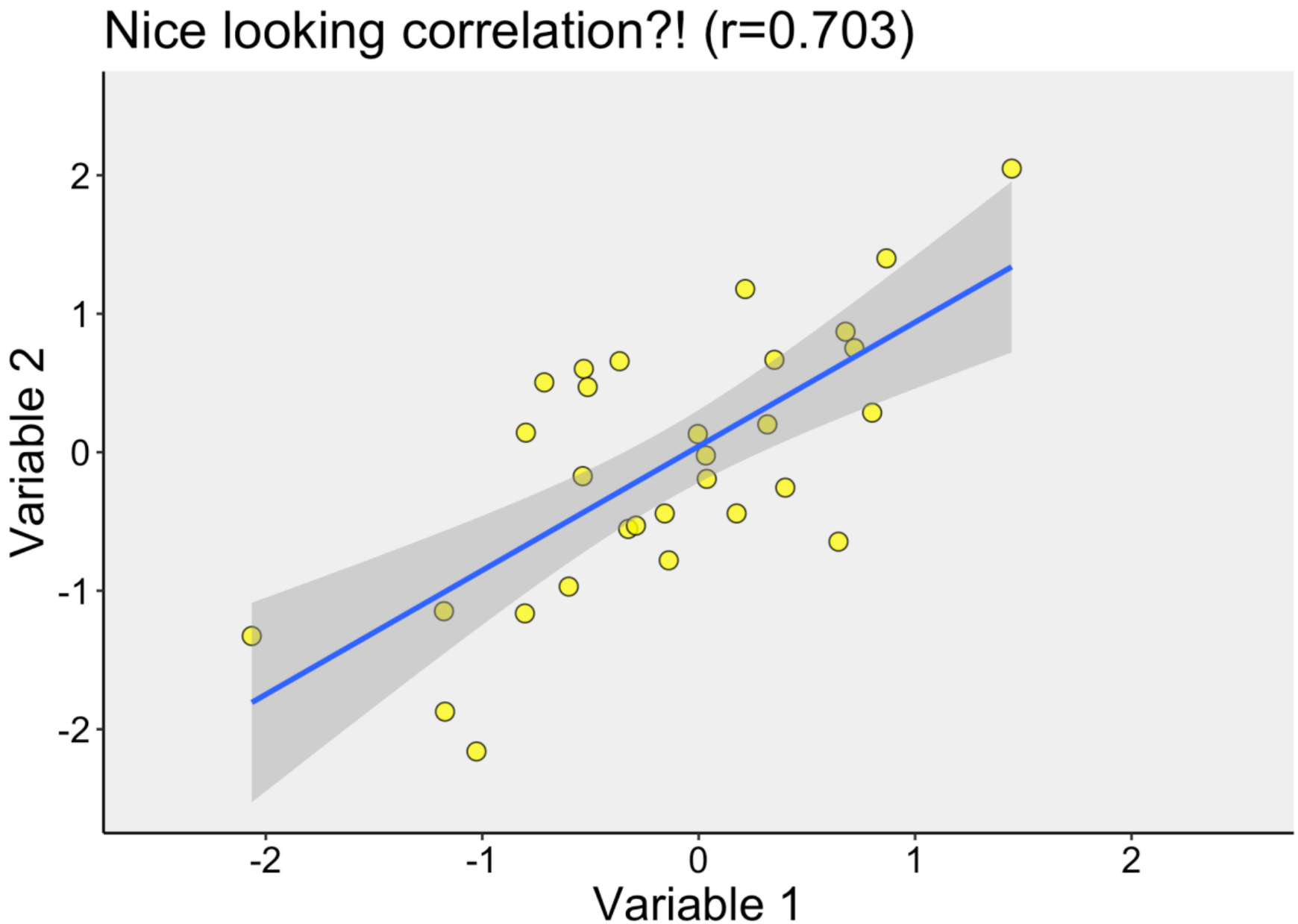
Worship arbitrary thresholds

Magic land of **p<0.05**

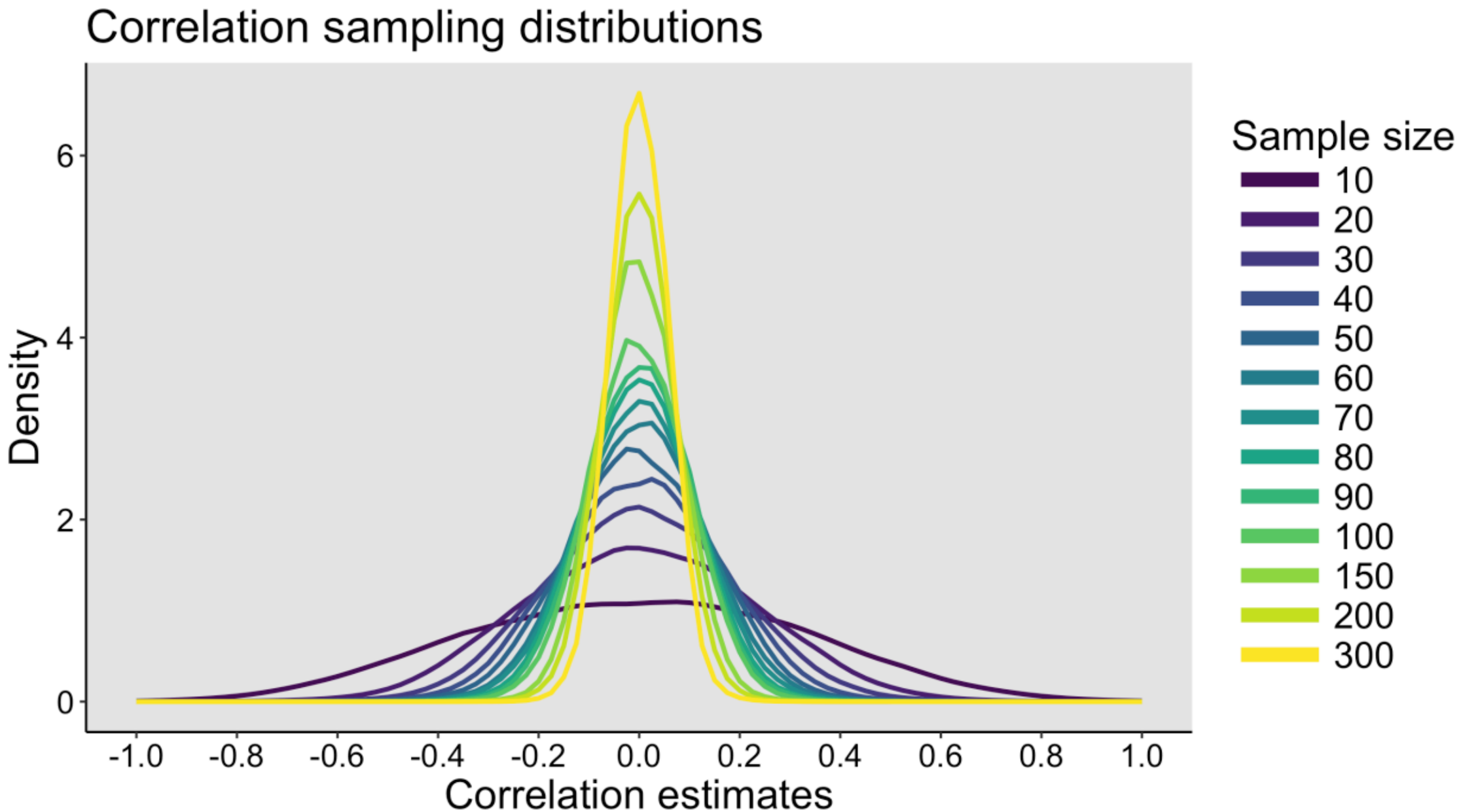


p<0.05

certainty
discoveries
articles
grant applications
press releases
…
patients die

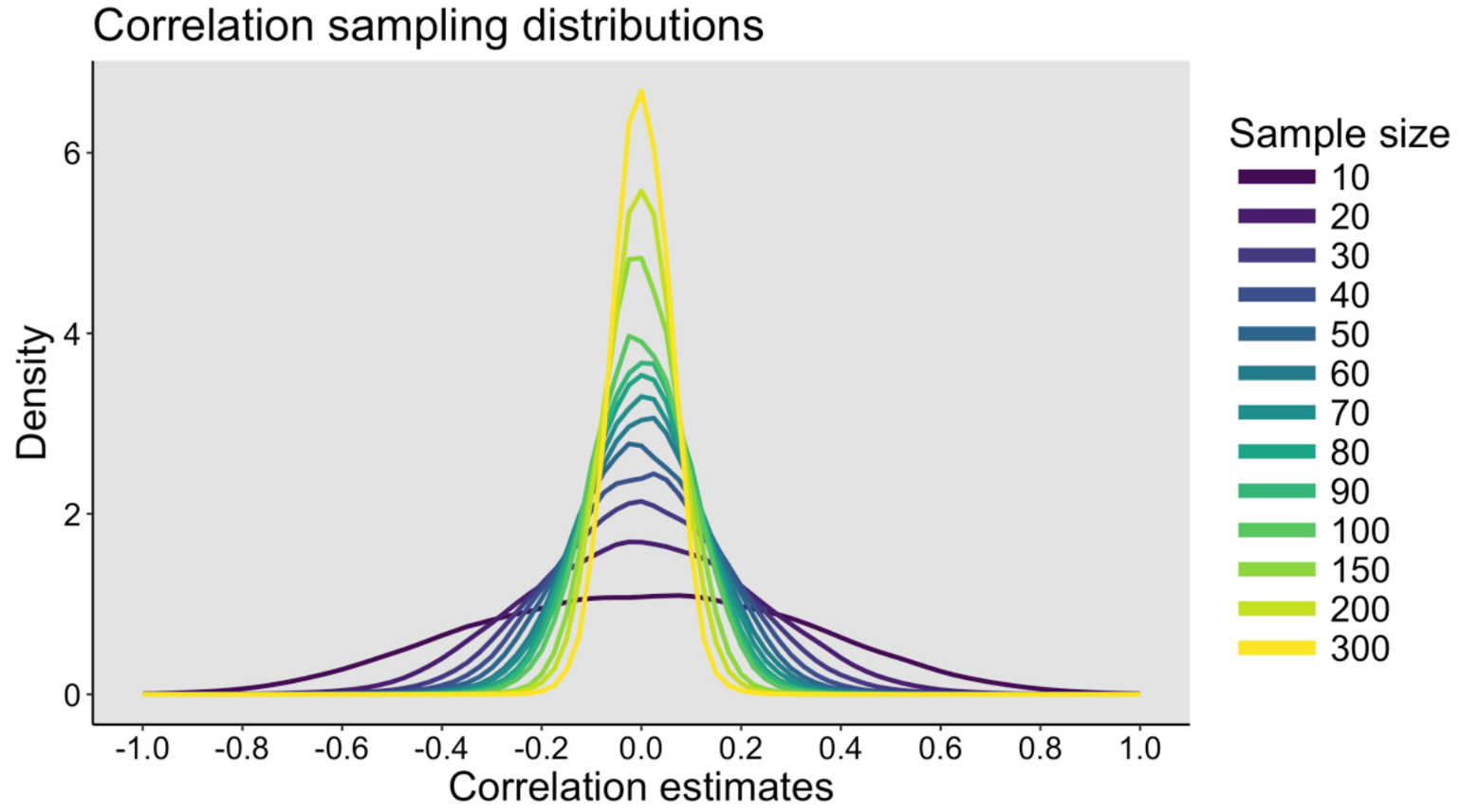


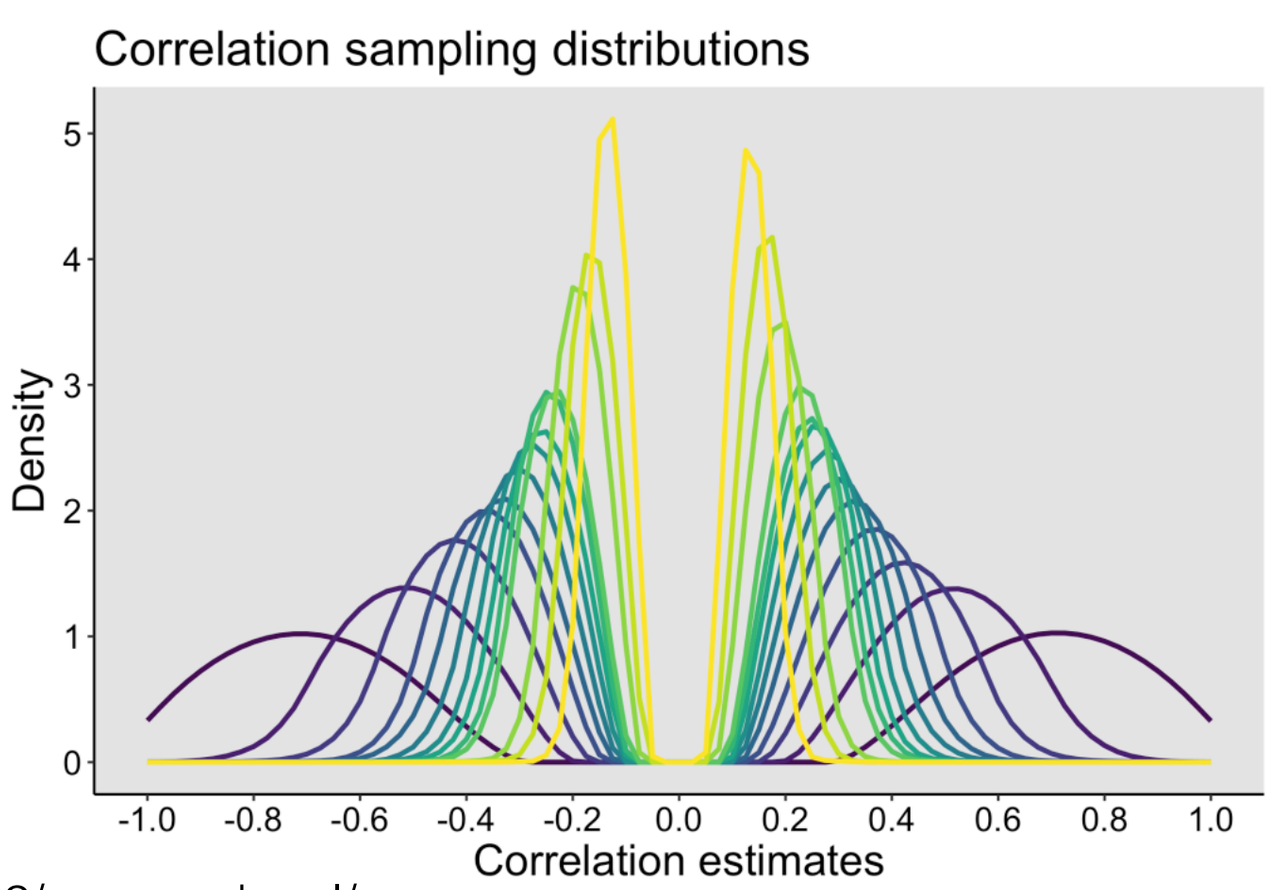
https://garstats.wordpress.com/2018/06/01/smallncorr/



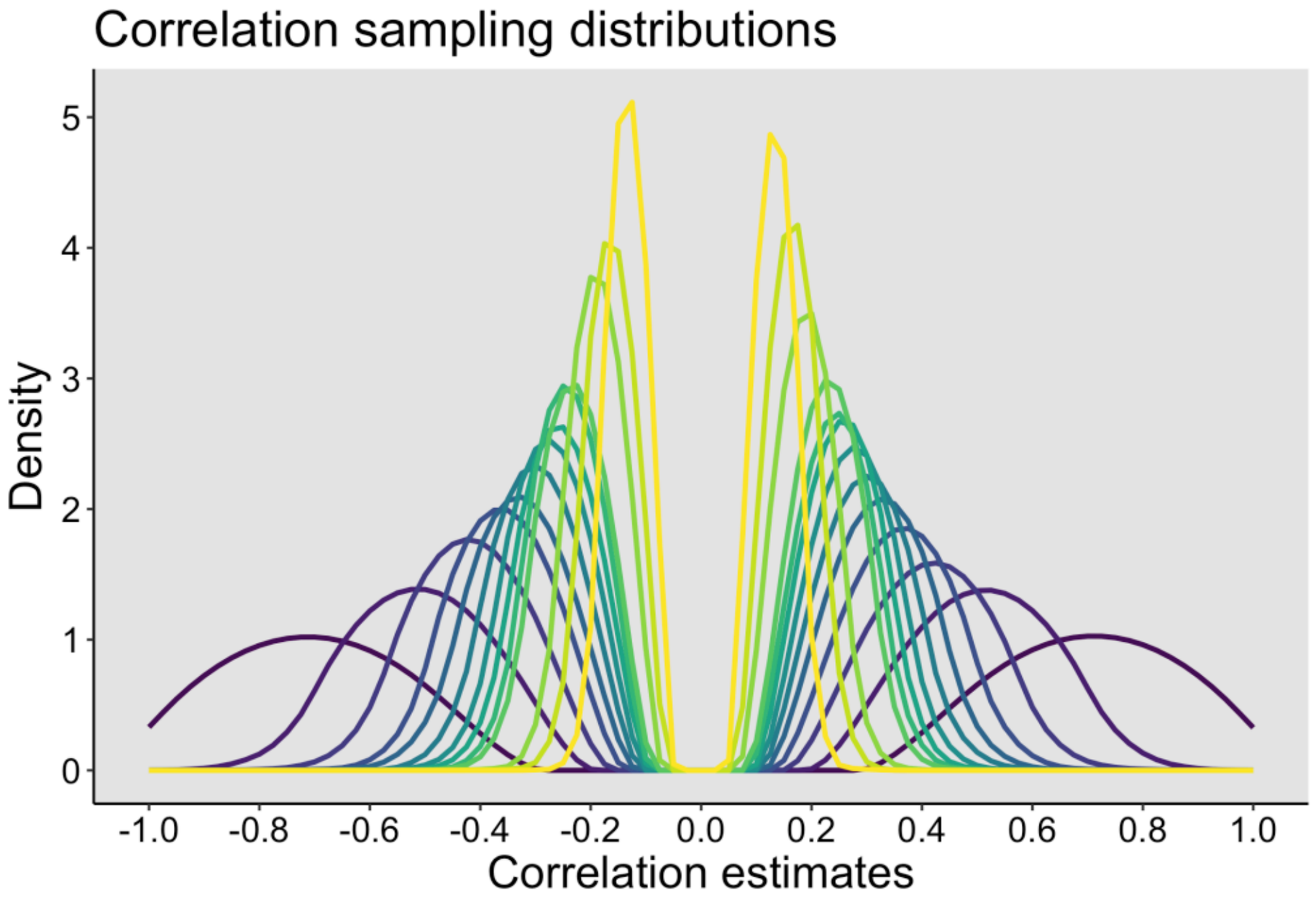
https://garstats.wordpress.com/2018/06/22/corrcondpval/

Sampling distributions conditional on p<0.05



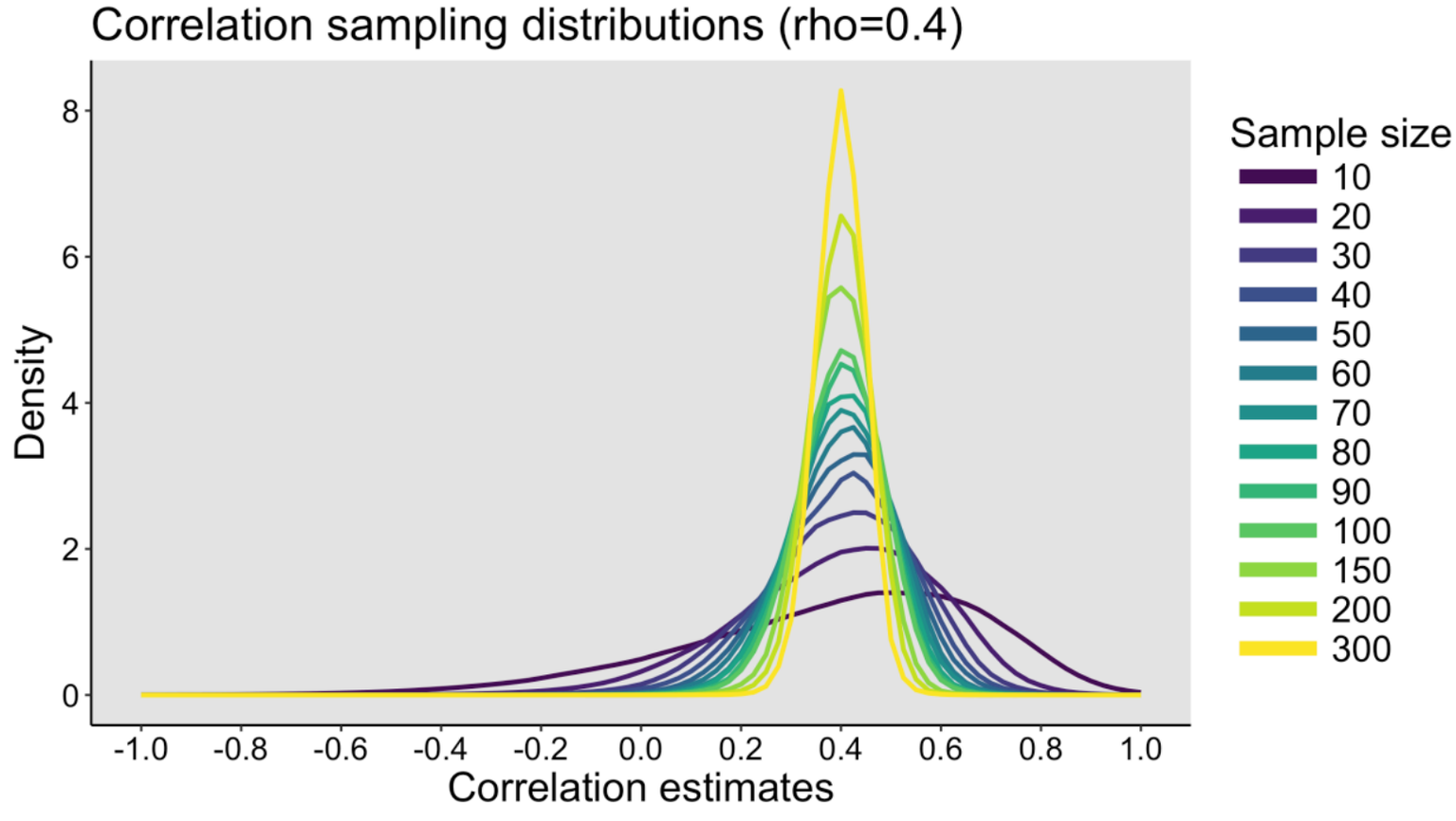


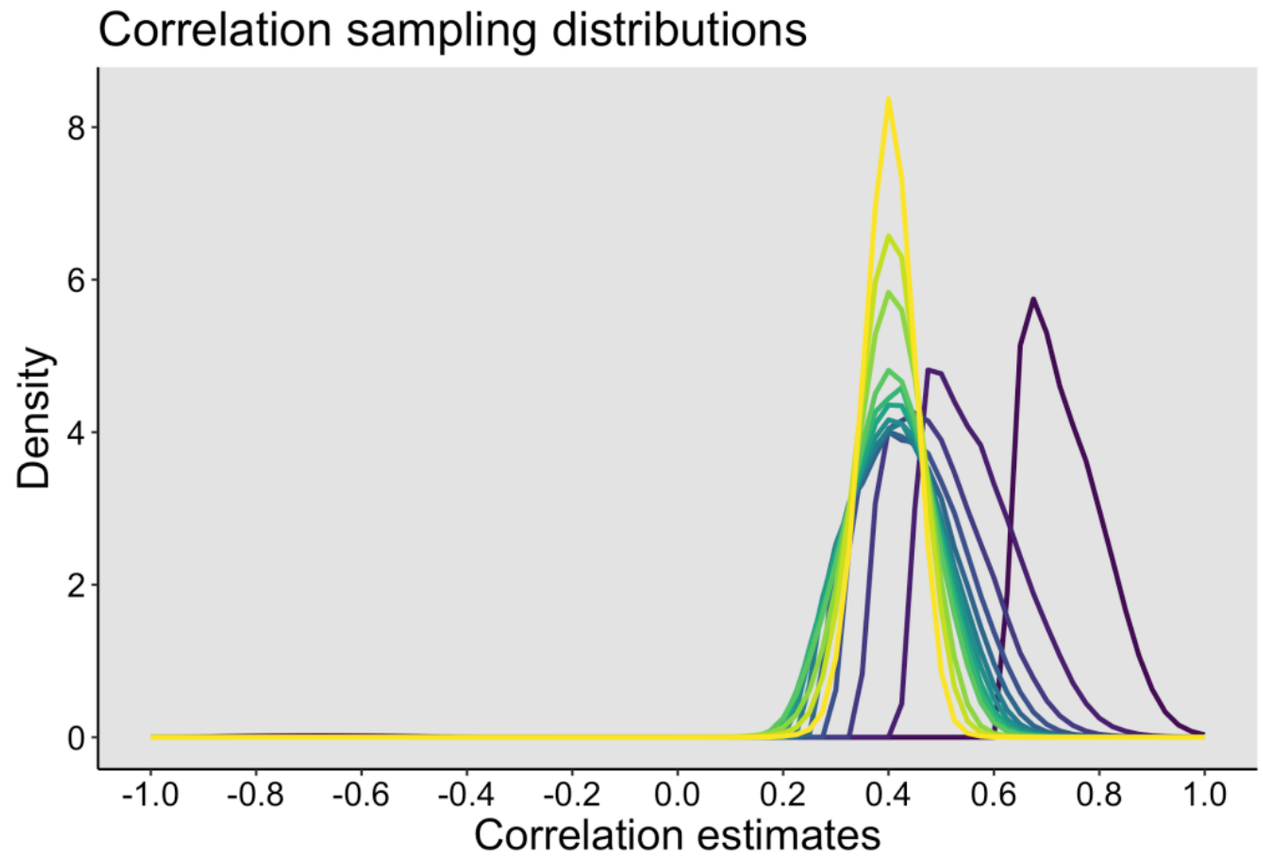
https://garstats.wordpress.com/2018/06/22/corrcondpval/



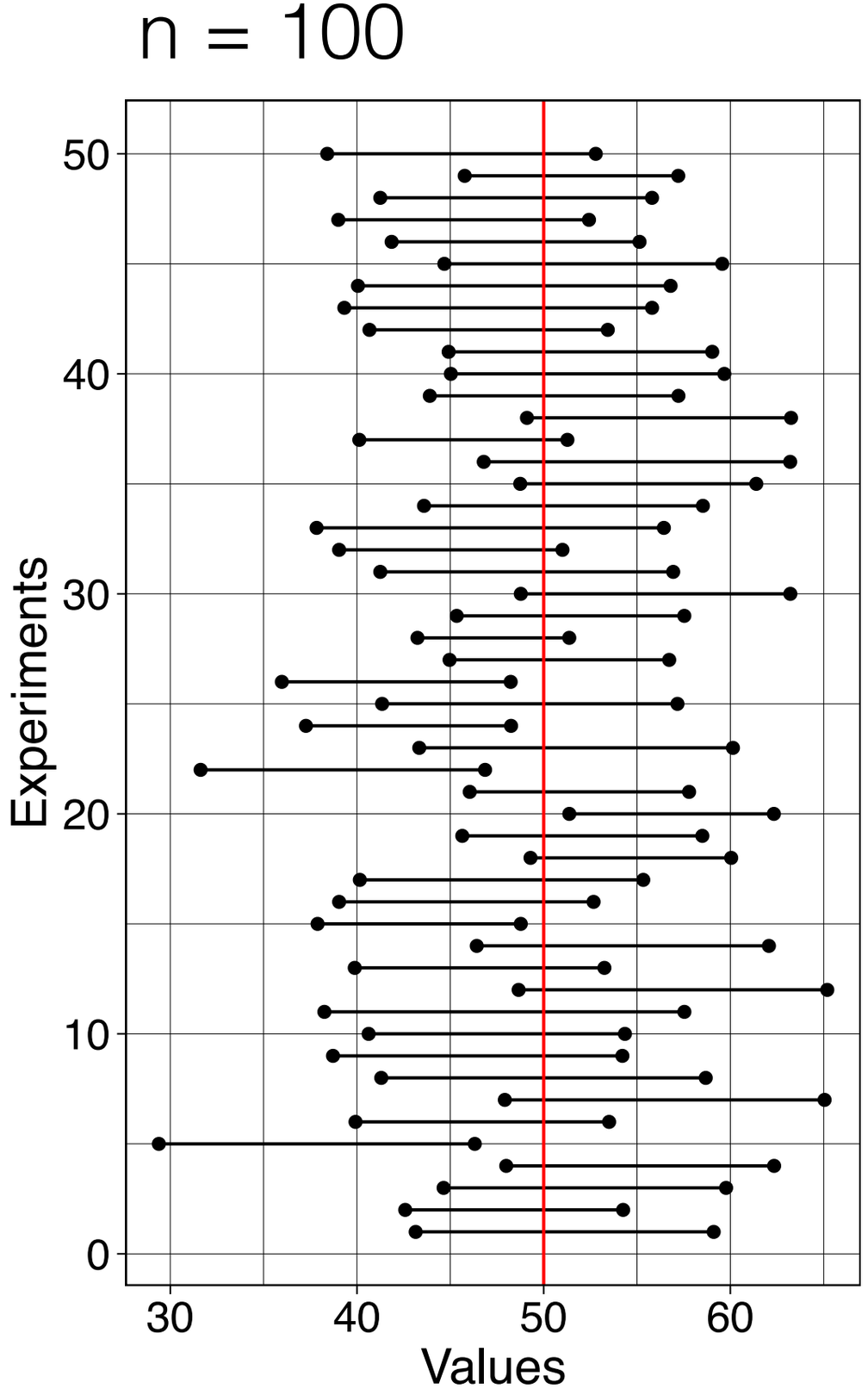
https://garstats.wordpress.com/2018/06/22/corrcondpval/

Sampling distributions conditional on p<0.05





Dance of the confidence intervals



The ASA's statement 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

“the P value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility, and in this sense may be viewed as measuring the fit of the model to the data.”

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. & Altman, D.G. (2016)
Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.
Eur J Epidemiol, 31, 337-350.

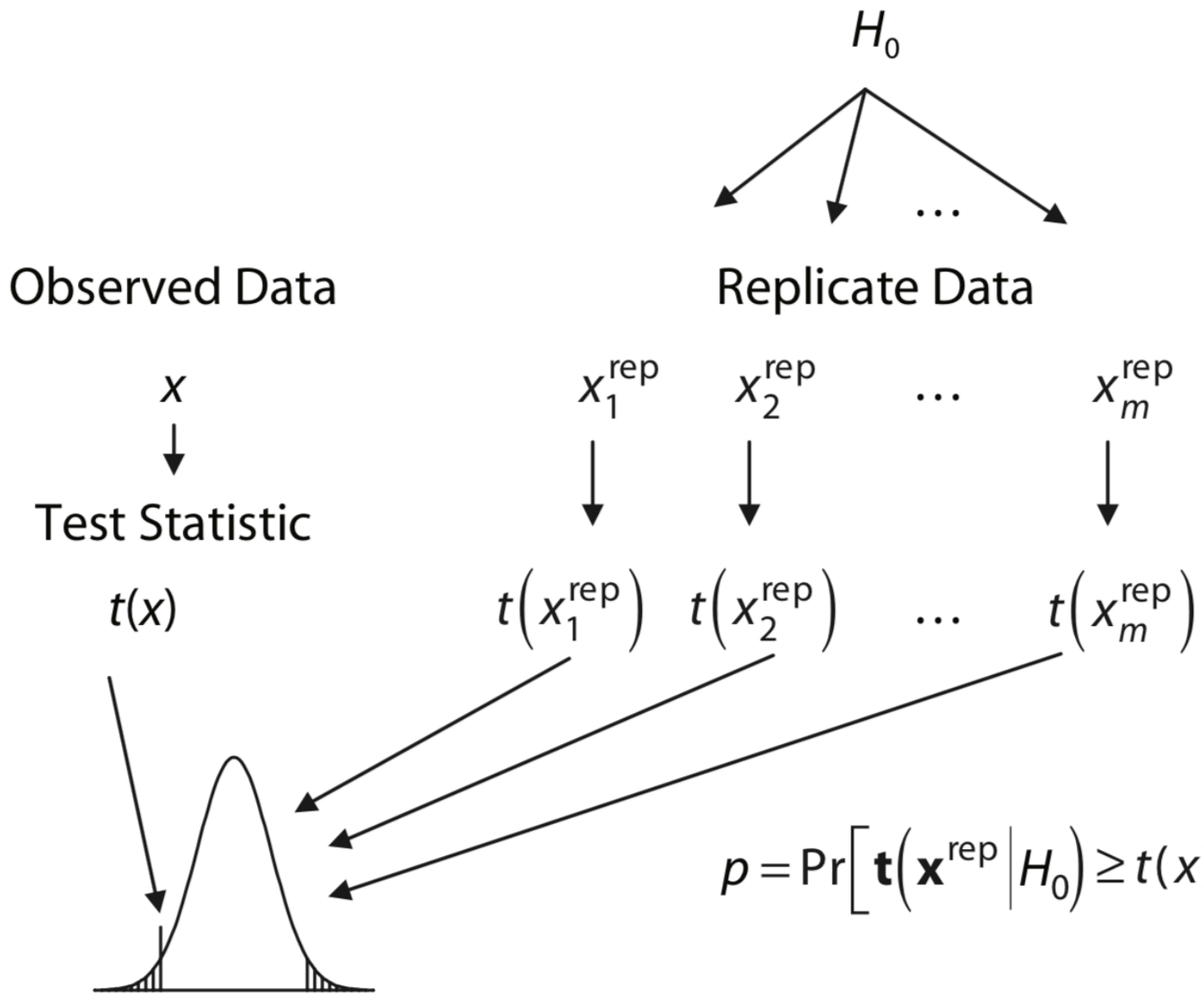


Figure 1. A schematic overview of *p* value statistical null-hypothesis testing. The distribution of a test statistic is constructed from replicated data sets generated under the null hypothesis. The two-sided *p* value is equal to the sum of the shaded areas on either side of the distribution; for these areas, the value of the test statistic for the replicated data sets is at least as extreme as the value of the test statistic for the observed data.

Wagenmakers, E.J. (2007) **A practical solution to the pervasive problems of p values**. *Psychonomic bulletin & review*, **14**, 779-804

sampling

n trials?
n participants?
screen used?
response button used?

pre-processing

coding of variables?
outlier removal?
data transformation?

conditional on:

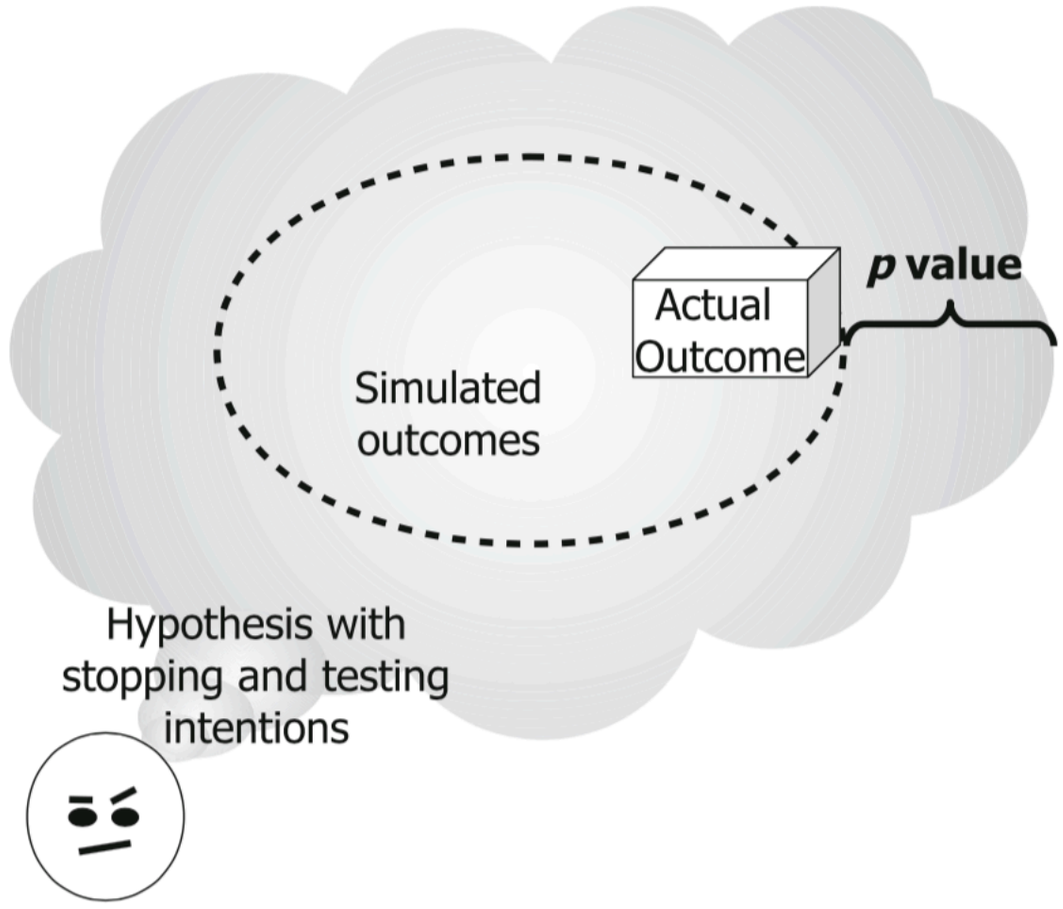
- **mood?**
- **external events?**
- **looking at the data (without clear plan)?**
- **looking at the results?**

analyses

violations of assumptions?
estimator used?
parametric / non-parametric?

final sampling distribution

p value



$$p \text{ value} \equiv p\left(T(D_{simulated}) \succeq T(D_{actual}) \,\middle|\, \mu, I\right)$$

Kruschke, J.K. & Liddell, T.M. (2018)
The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective
Psychon Bull Rev, 25, 178-206.

P values depend on intentions

no pre-registration =

ambiguous *P* values

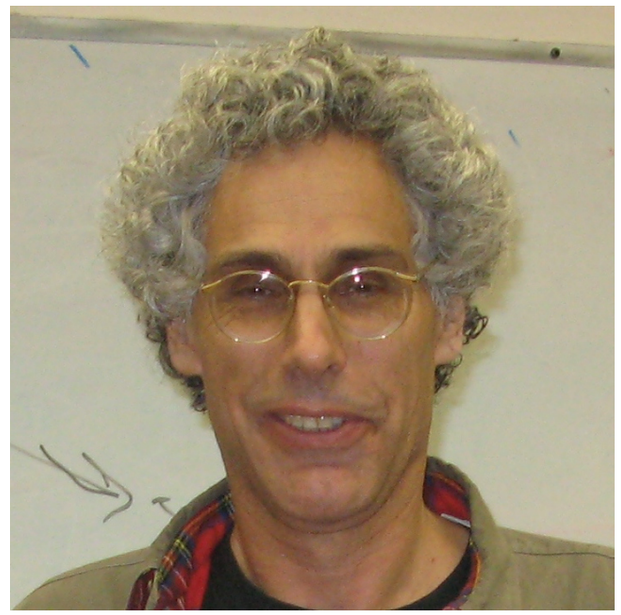
- Wagenmakers, E.J. (2007) **A practical solution to the pervasive problems of p values**. *Psychonomic bulletin & review*, 14, 779-804.

- de Groot, A.D. (1956/2014) **The meaning of "significance" for different types of research**. Acta Psychol (Amst), 148, 188-194.

- Kruschke, J.K. (2015) **Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan**. Academic Press, San Diego, CA.

Description of frequentist statistics

Frank Harrell



Sander Greenland

Language for communicating frequentist results about treatment effects

https://discourse.datamethods.org/t/language-for-communicating-frequentist-results-about-treatment-effects/

Amrhein V, Trafimow D, Greenland S. (2018)
Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication.
PeerJ Preprints 6:e26857v4 https://doi.org/10.7287/peerj.preprints.26857v4

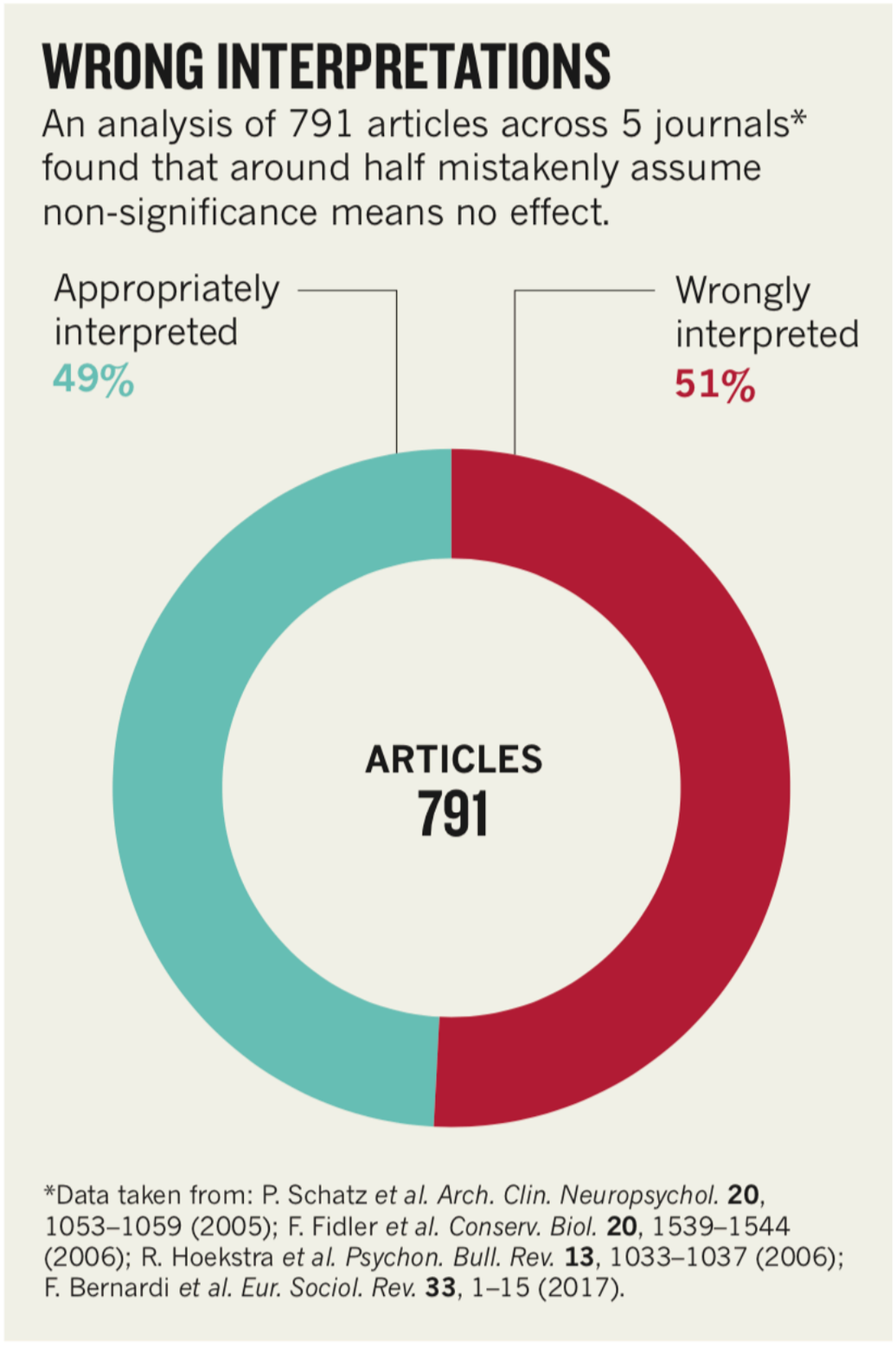
A = 134, B = 130

difference = -4 [-13, 5], p = 0.4

“not significant” **CONFUSING**

“do not differ” “no effect” **INCORRECT**

“the money was spent” **BRUTALLY HONEST**



Amrhein, Greenland, and McShane (2019) **Scientists Rise up against Statistical Significance**. *Nature*

A = 134, B = 130

difference = -4 [-13, 5], p = 0.4

"**Assuming our model**, the probability is 0.4 that another study would yield a test statistic for comparing two means that is more impressive that what we observed in our study, if A and B had exactly the same true mean."

"**Given our model**, mean differences compatible with our data ranged from -13 to 5."

Put values in context, discuss model, illustrate results…

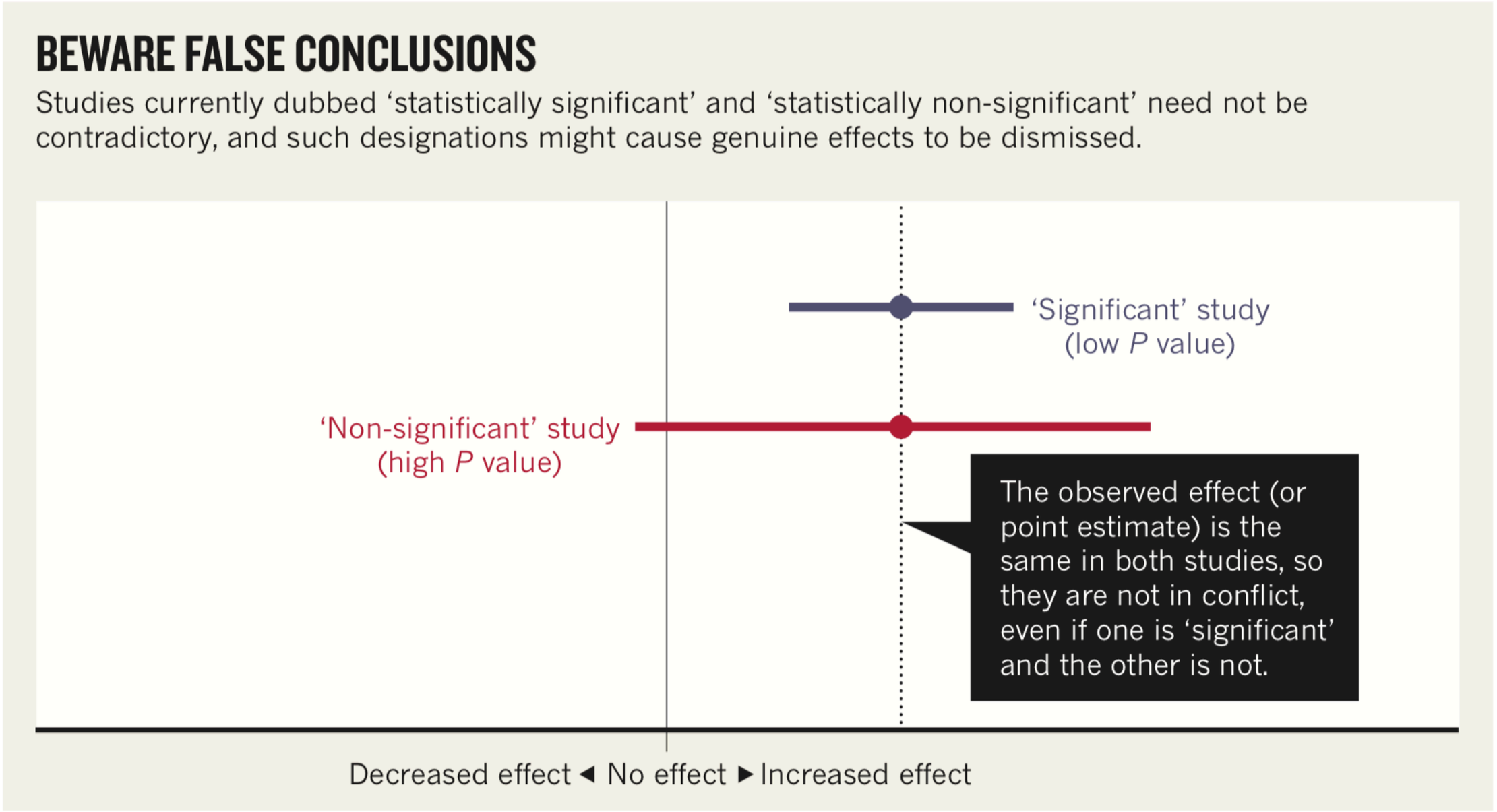
# P<0.05

“significant” **CONFUSING**

“A & B differed”
“there is an effect”
“we proved”
“we demonstrated” **INCORRECT**

“problem in my model?” **HEALTHY**

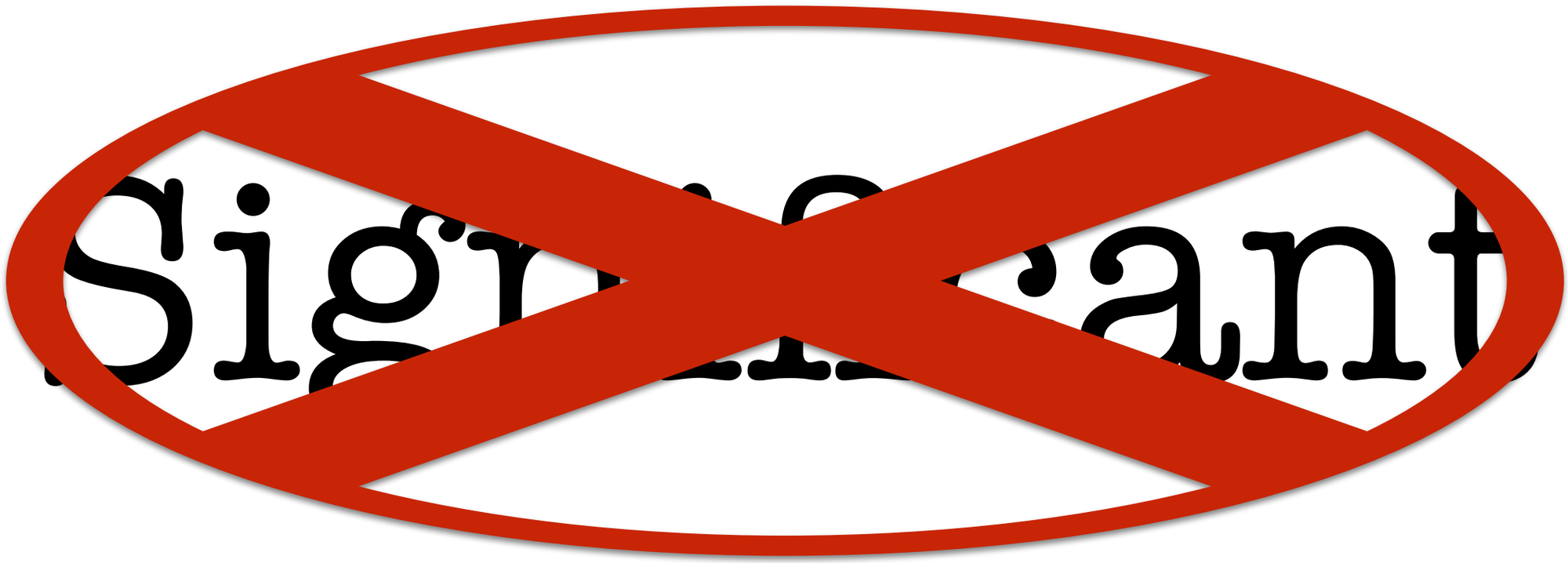
“Our effect is inconsistent with previous results”



Amrhein, Greenland, and McShane (2019) **Scientists Rise up against Statistical Significance**. *Nature*

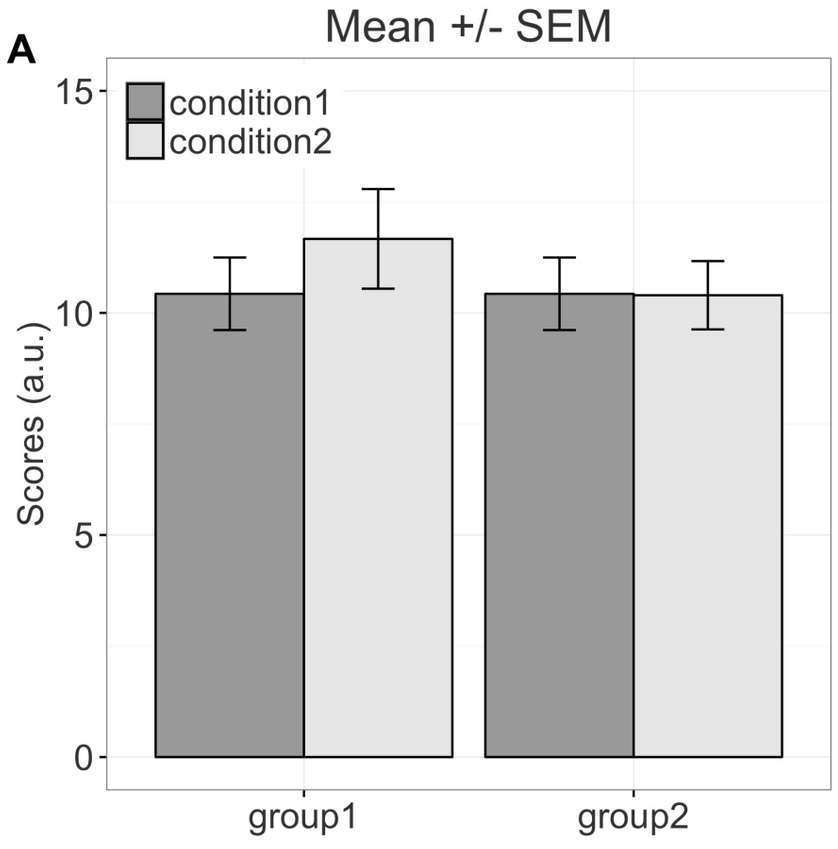
“statistical significance is neither necessary nor sufficient for determining the scientific or practical significance of a set of observations.”

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. & Altman, D.G. (2016)
Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.
Eur J Epidemiol, 31, 337-350.



Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, Jennifer L. Tackett (2018) **Abandon Statistical Significance**. *arXiv*

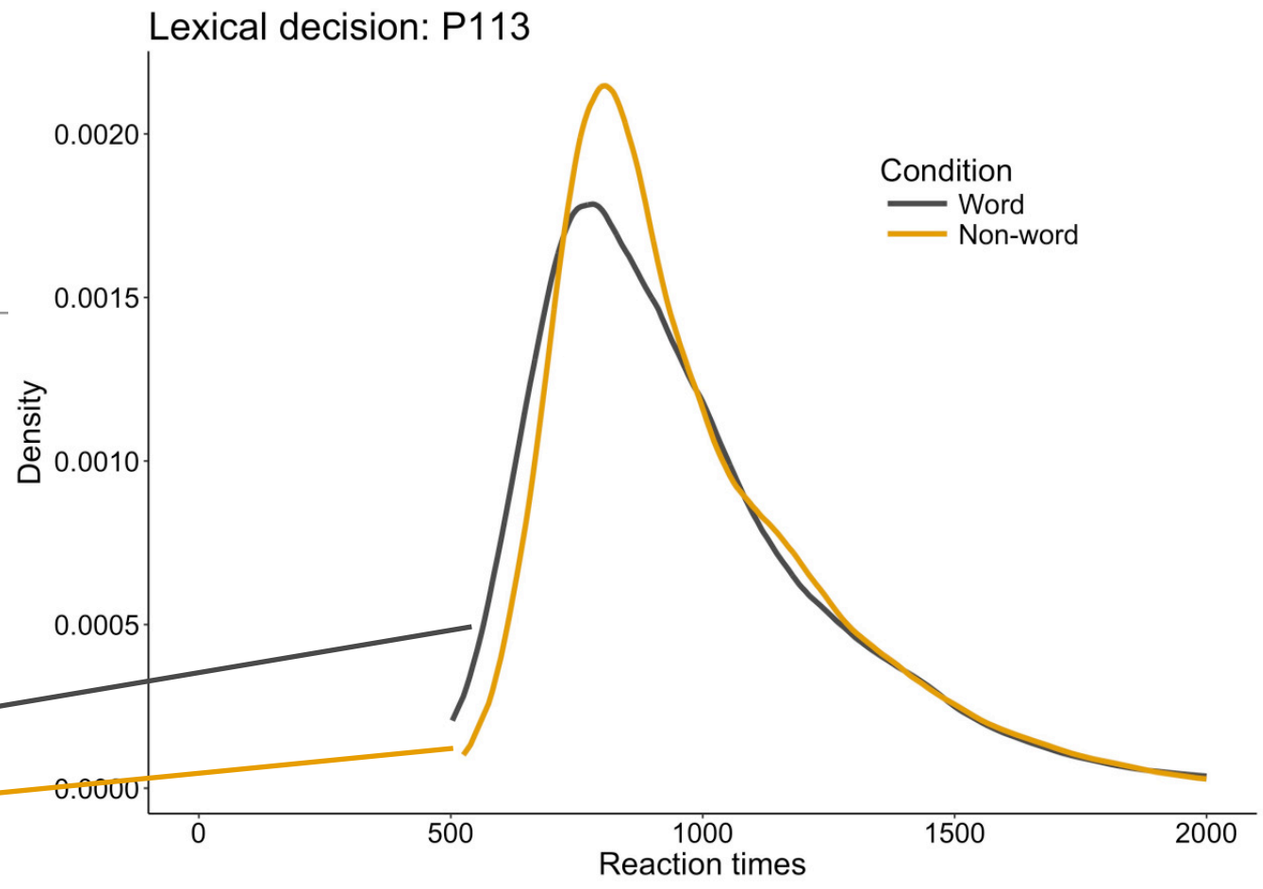
Valentin Amrhein , David Trafimow & Sander Greenland (2018) **Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication**. *PeerJ Preprints*

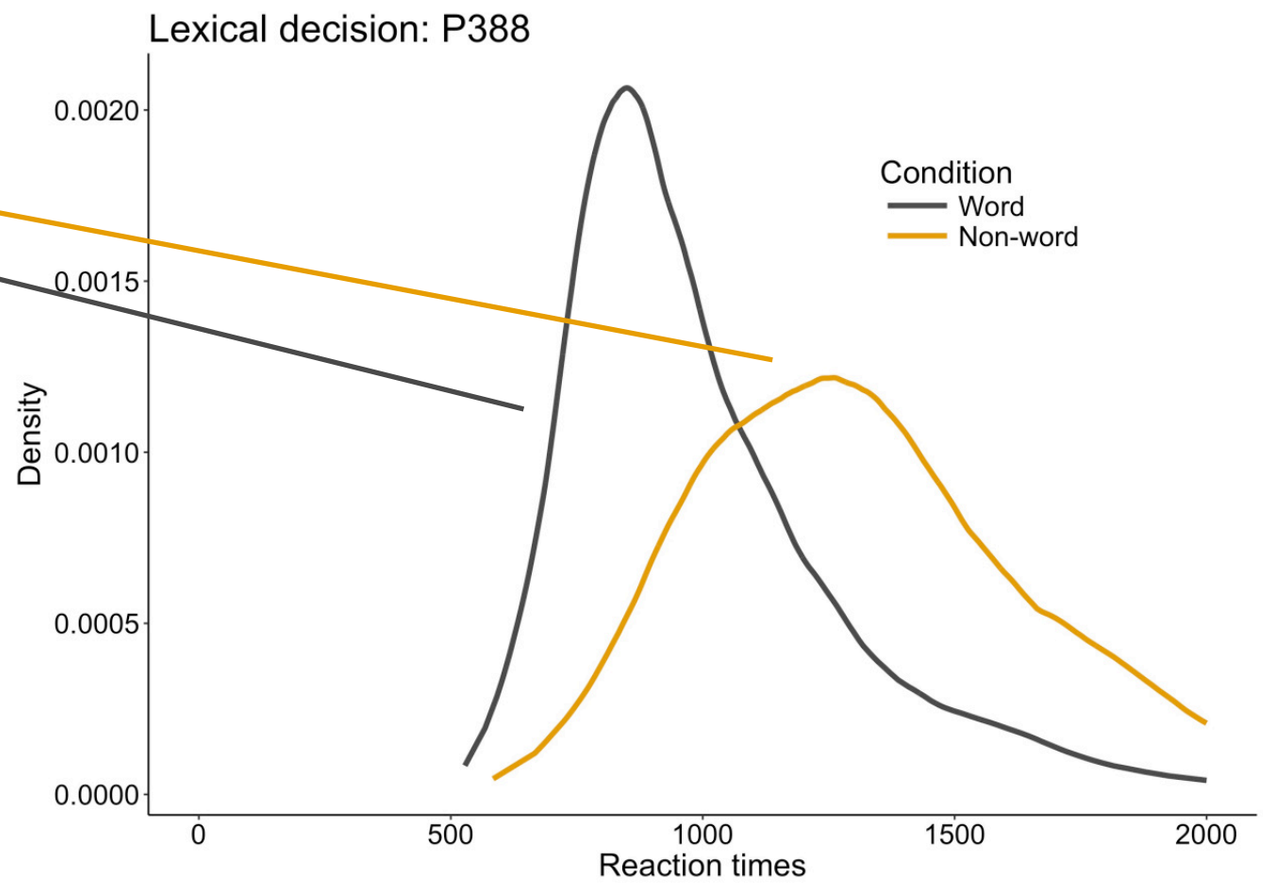

A
Mean +/- SEM
condition1
condition2
Scores (a.u.)
15
10
5
0
group1
group2

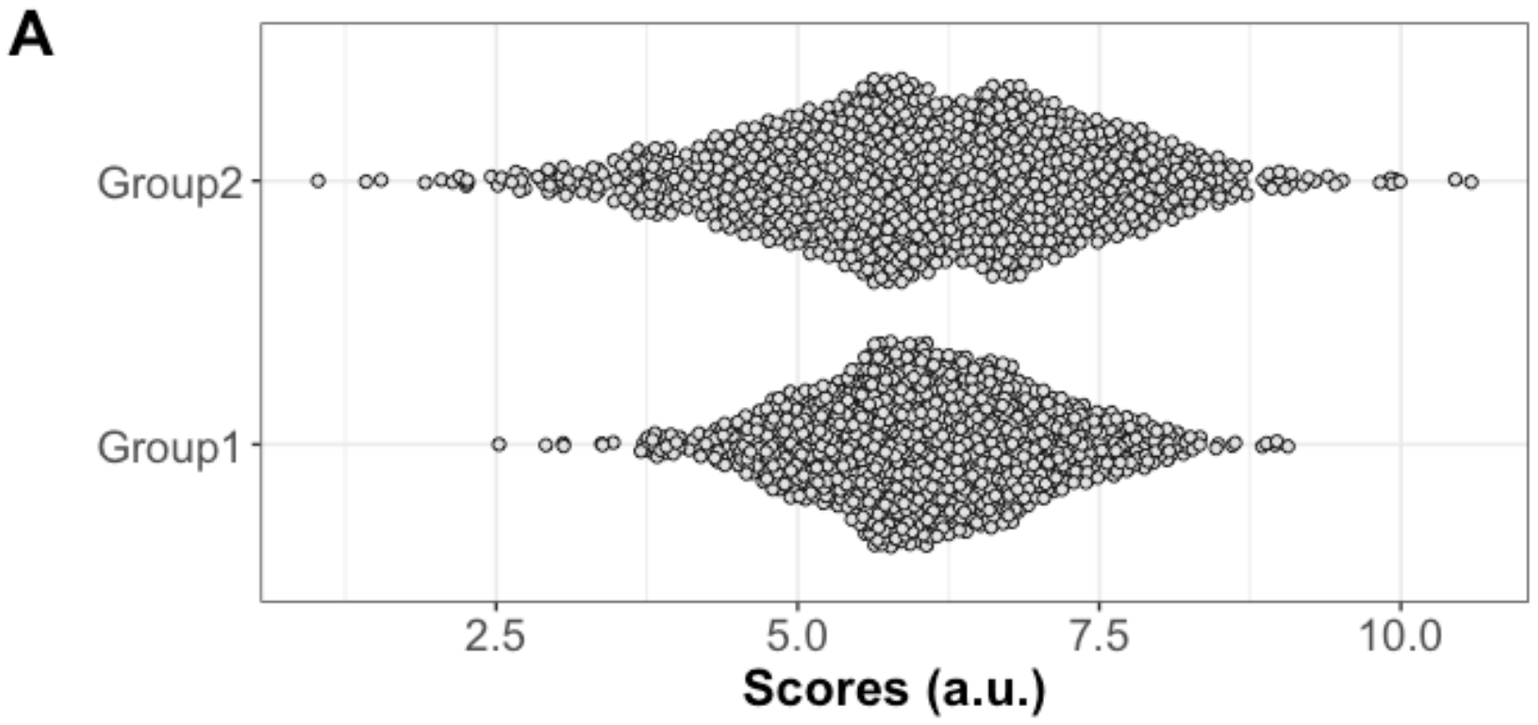
Reaction time data

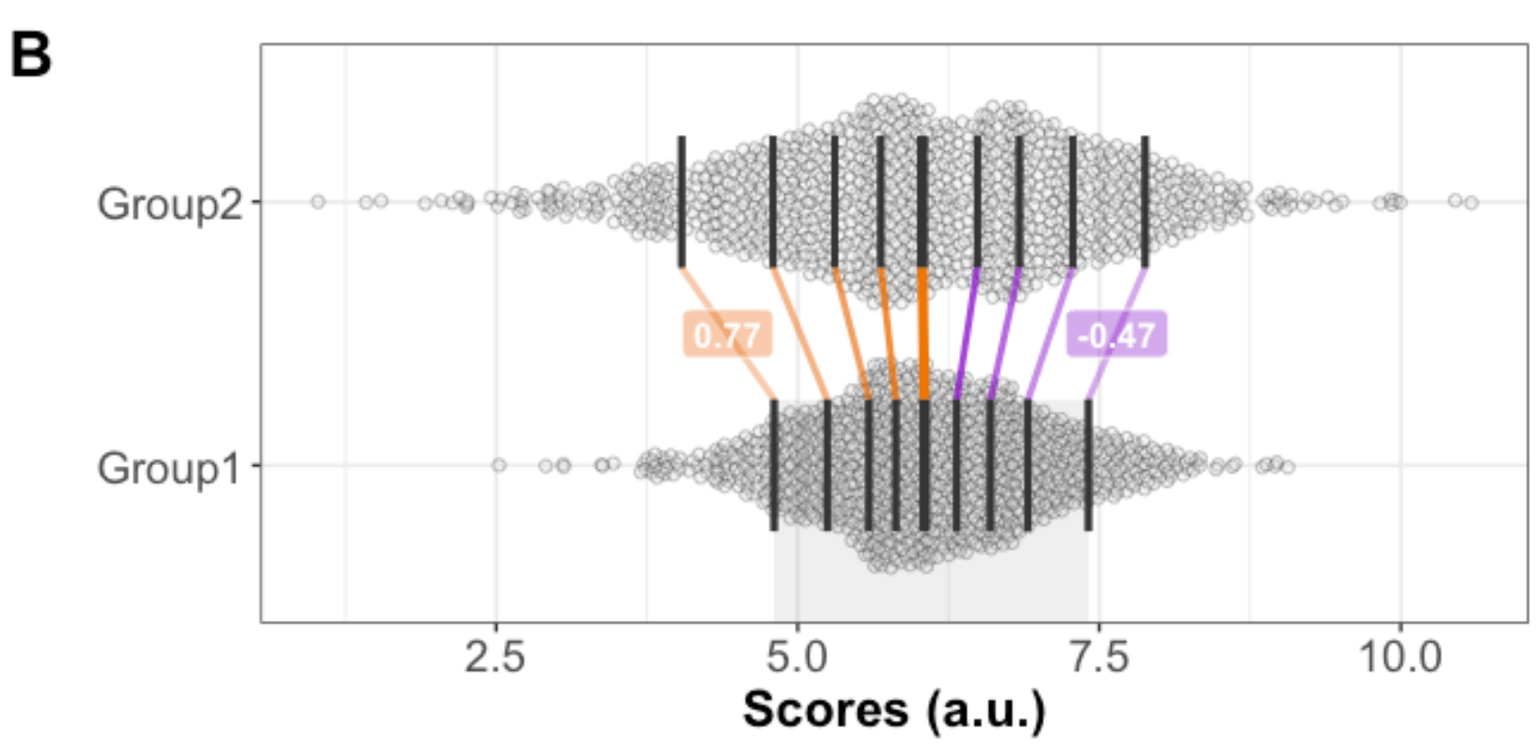
	C1	C2
P1	$\bar{x}_{11}$	$\bar{x}_{12}$
P2	$\bar{x}_{21}$	$\bar{x}_{22}$
P3	$\bar{x}_{31}$	$\bar{x}_{32}$
...		

t.test(C1,C2)

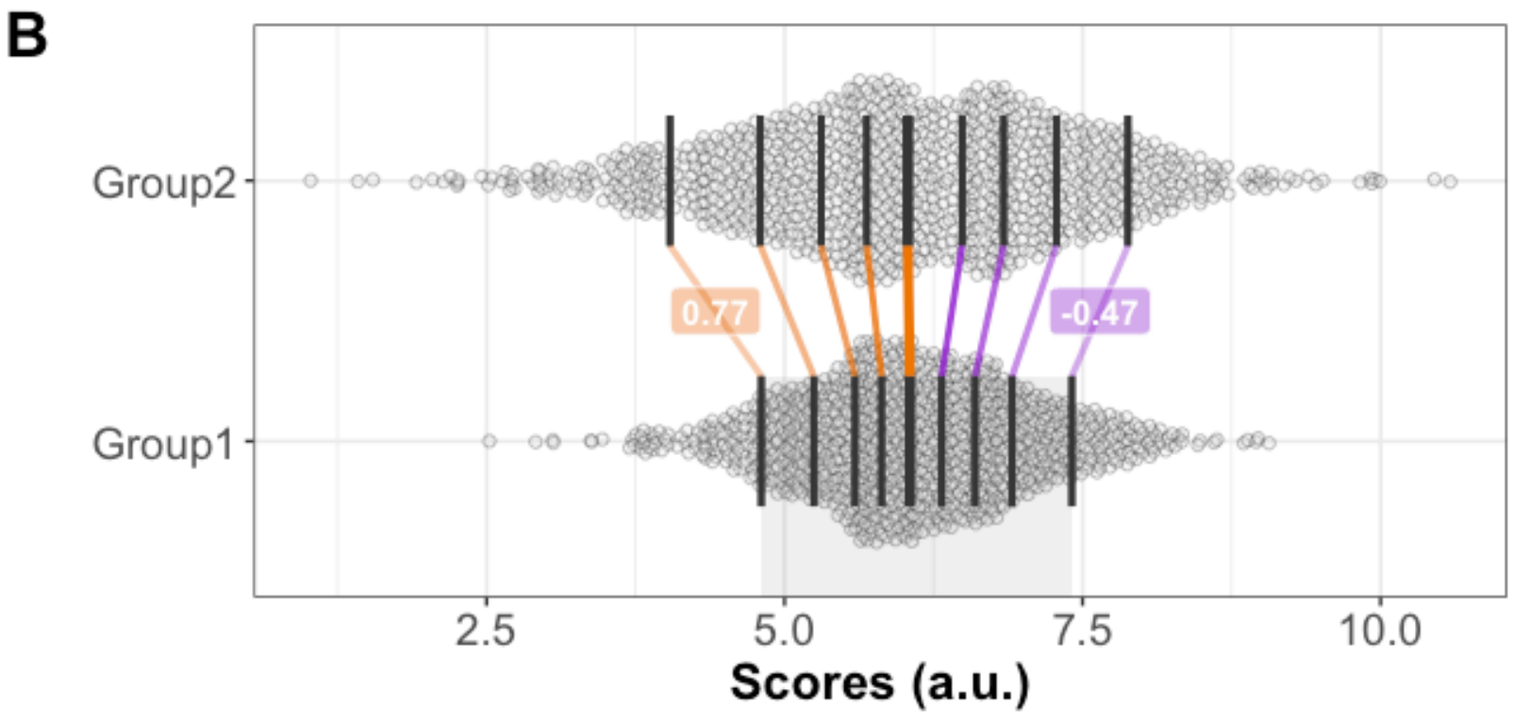


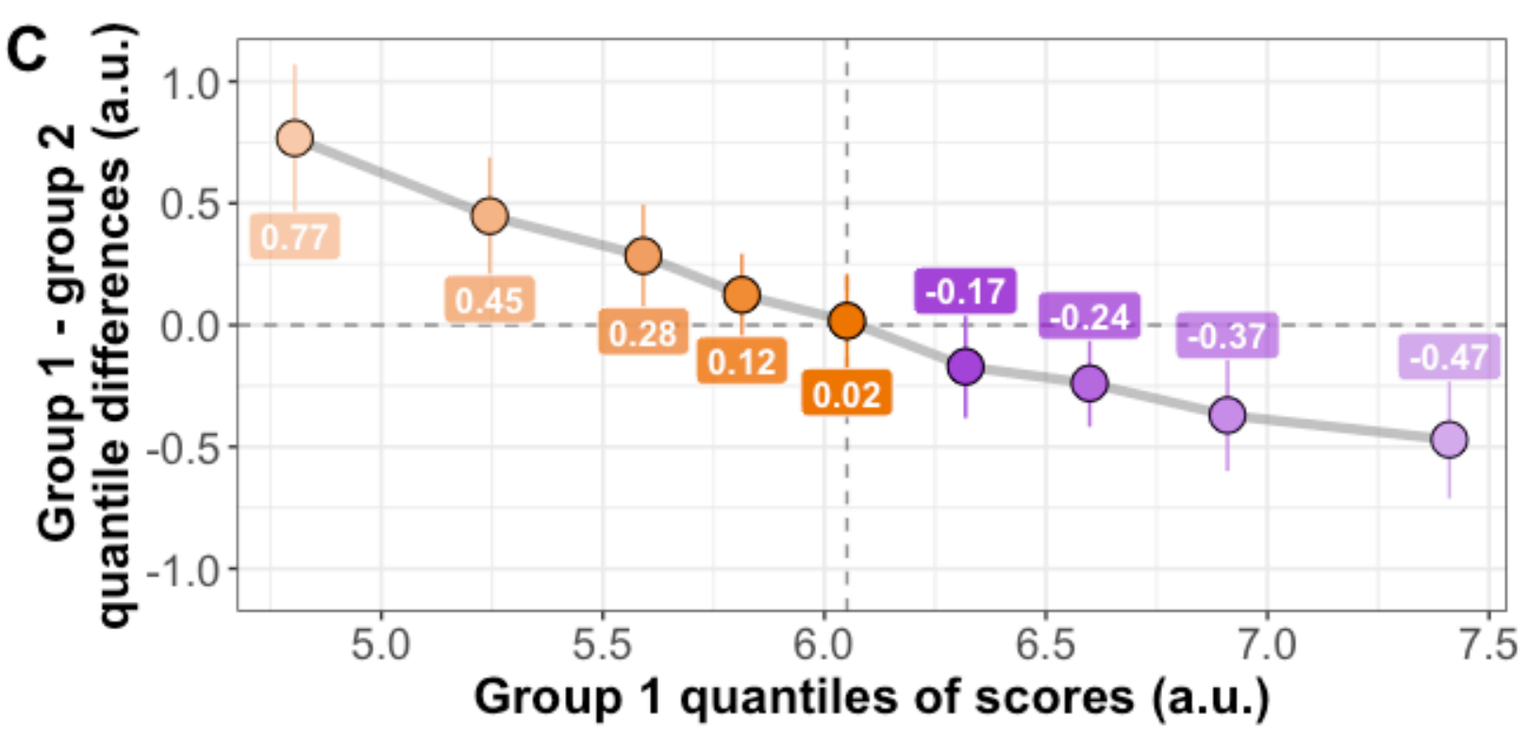




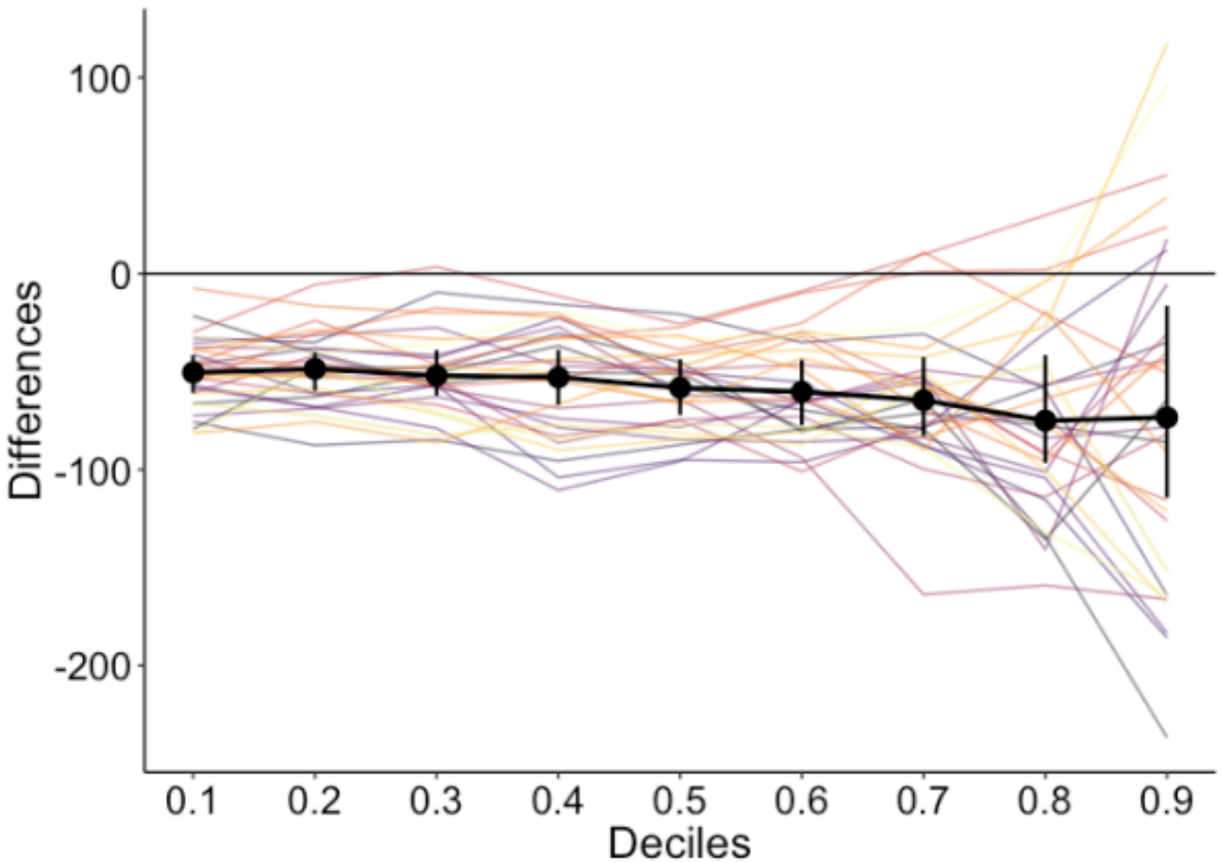


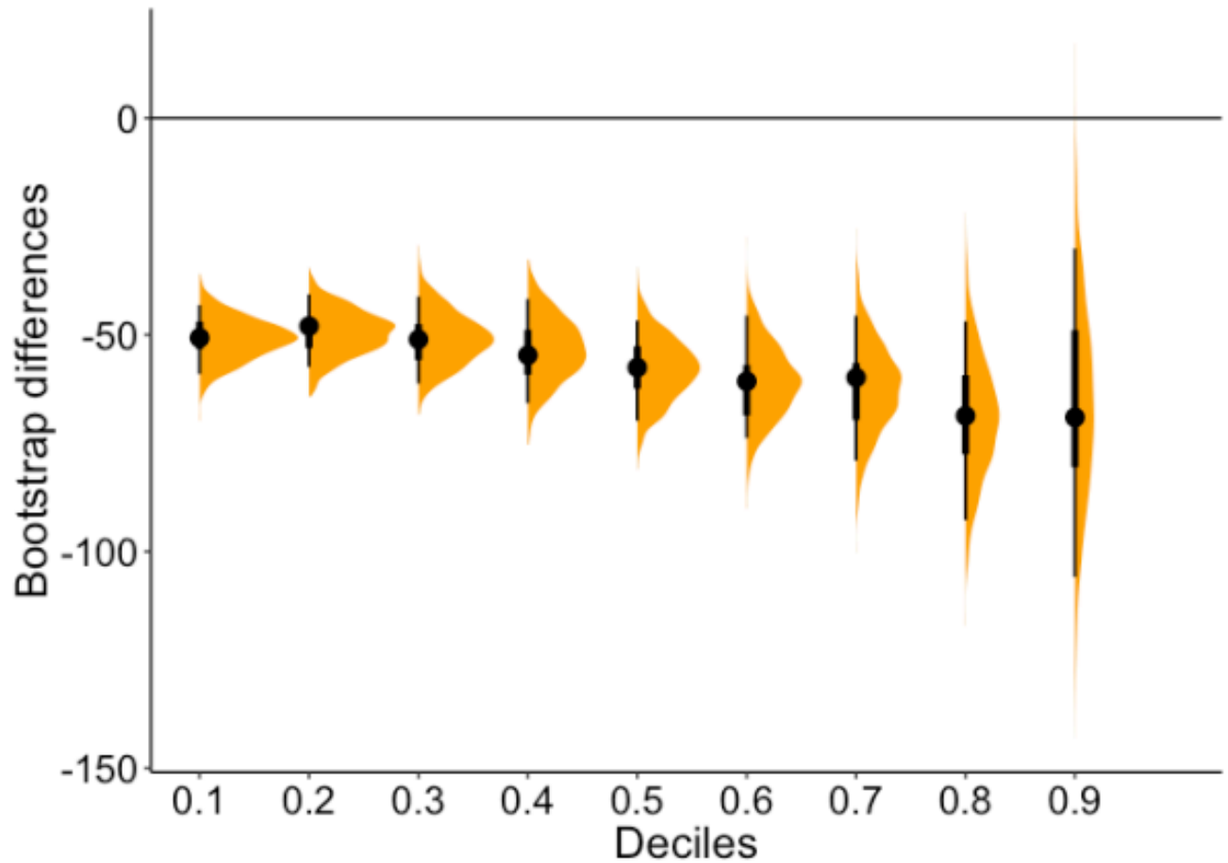
https://github.com/GRousselet/rogme


B
Group2
Group1
0.77
-0.47
2.5
5.0
7.5
10.0
Scores (a.u.)


C
Group 1 - group 2
quantile differences (a.u.)
1.0
0.5
0.0
-0.5
-1.0
0.77
0.45
0.28
0.12
0.02
-0.17
-0.24
-0.37
-0.47
5.0
5.5
6.0
6.5
7.0
7.5
Group 1 quantiles of scores (a.u.)

hierarchical shift function





https://garstats.wordpress.com/2019/02/21/hsf/

Discovery of a new effect is a matter for a *research programme*, not a single experiment.
There is no statistical criterion that can establish a "discovery".

Richard D. Morey

When the statistical tail wags the scientific dog

https://medium.com/@richarddmorey/when-the-statistical-tail-wags-the-scientific-dog-d09a9f1a7c63

“Forget about getting definitive results from a single experiment; instead embrace variation, accept uncertainty, and learn what you can.”

Andrew Gelman 2018

Andrew Gelman (2018) **The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It.** *Personality and Social Psychology Bulletin* 44, no. 1: 16–23

"So when can we be confident that we know something? This is the topic of the vast domains of epistemology, scientific inference, and philosophy of science […]. Nonetheless, a successful theory is one that survives decades of scrutiny. If every study claims to provide decisive results […], there will be ever more replication failures, which in turn will further undermine public confidence in science. We thus believe that decision makers must act based on cumulative knowledge – which means they should preferably not rely solely on single studies or even single lines of research […]."

Amrhein V, Trafimow D, Greenland S. (2018)
Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication.
PeerJ Preprints 6:e26857v4 https://doi.org/10.7287/peerj.preprints.26857v4

The Problem Is Epistemology, Not Statistics

T: Main substantive theory of interest;
A_x: Auxiliary theories relied on in the experiment;
C_p: *Ceteris paribus* clause ("other things being equal");
A_i: Instrumental auxiliaries (devices relied on for control and observation);
C_n: Realized particulars (conditions were as the experimenter reported);
O_1, O_2: Observations or statistical summaries of observations;

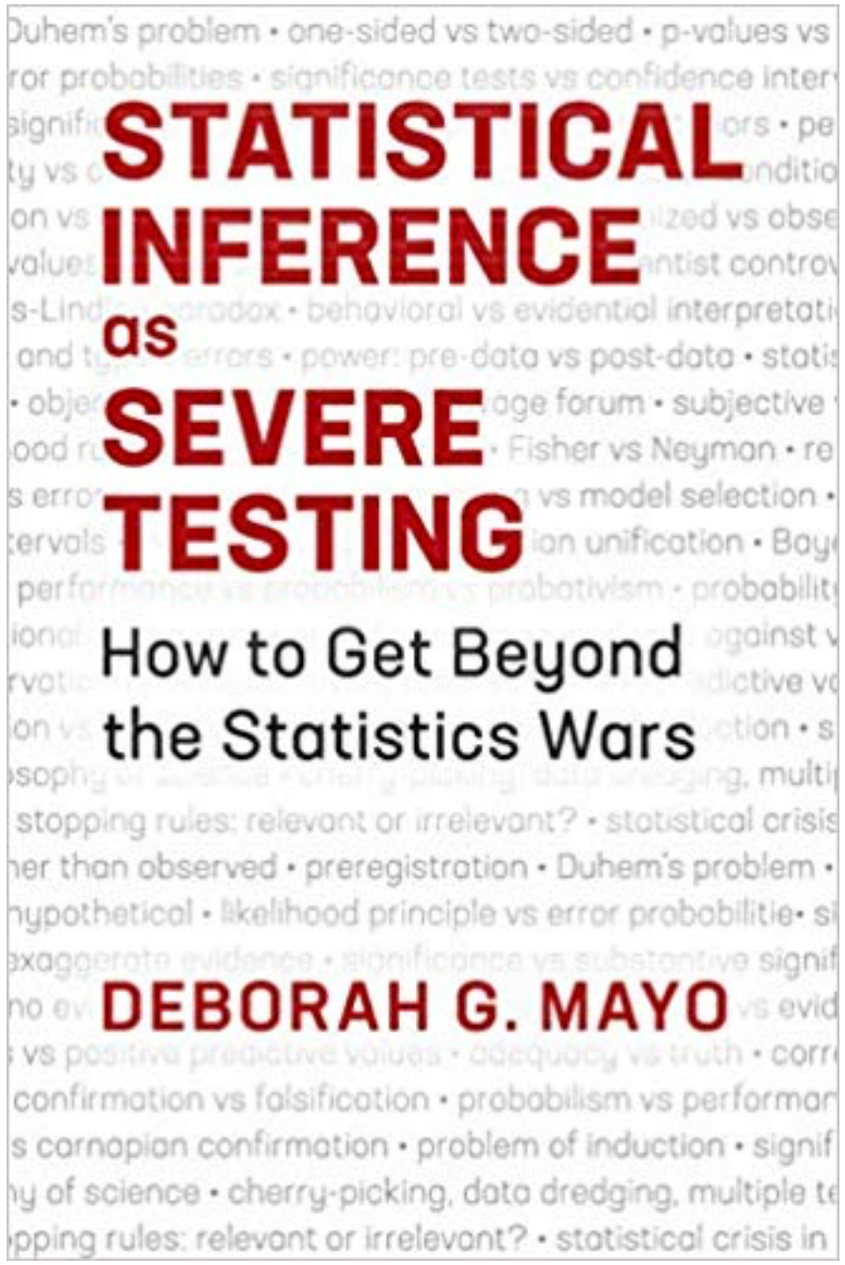
then the logical structure of a test of a theory is the conceptual formula:

$$(T \cdot A_x \cdot C_p \cdot A_i \cdot C_n) \vdash (O_1 \supset O_2)$$

where dots " · " are conjunctions ("and"), turnstile " $\vdash$ " is deductive derivability (entailment, "follows that . . ."), and the horseshoe " $\supset$ " is the material conditional ("If . . . then . . .").

Meehl, P.E. (1997) The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. In: Harlow, L., Mulaik, S.A. and Steiger, J.H., Eds., What If There Were No Significance Tests? Erlbaum, Mahwah, NJ, 393-425.

Severe testing



Statistical Inference in the 21st Century: A World Beyond p < 0.05. The American Statistician, Volume 73, Issue sup1, March 2019

Roadmap: focus on

- **estimation: robust and informative**
- **measurement precision: quality of measurements**
- **description: detailed graphical representations**
- **sharing: data and code**
- **embrace uncertainty: replication is the key**
- **honesty: exploratory / confirmatory research**
- **modesty**

relax: enjoy the fish!

Thank you!