The statistical significance filter leads to overconfident expectations of replicability

Shravan Vasishth and Andrew Gelman

Universität Potsdam, Germany
Columbia University, USA
Contact: vasishth@uni-potsdam.de
http://www.ling.uni-potsdam.de/∼vasishth

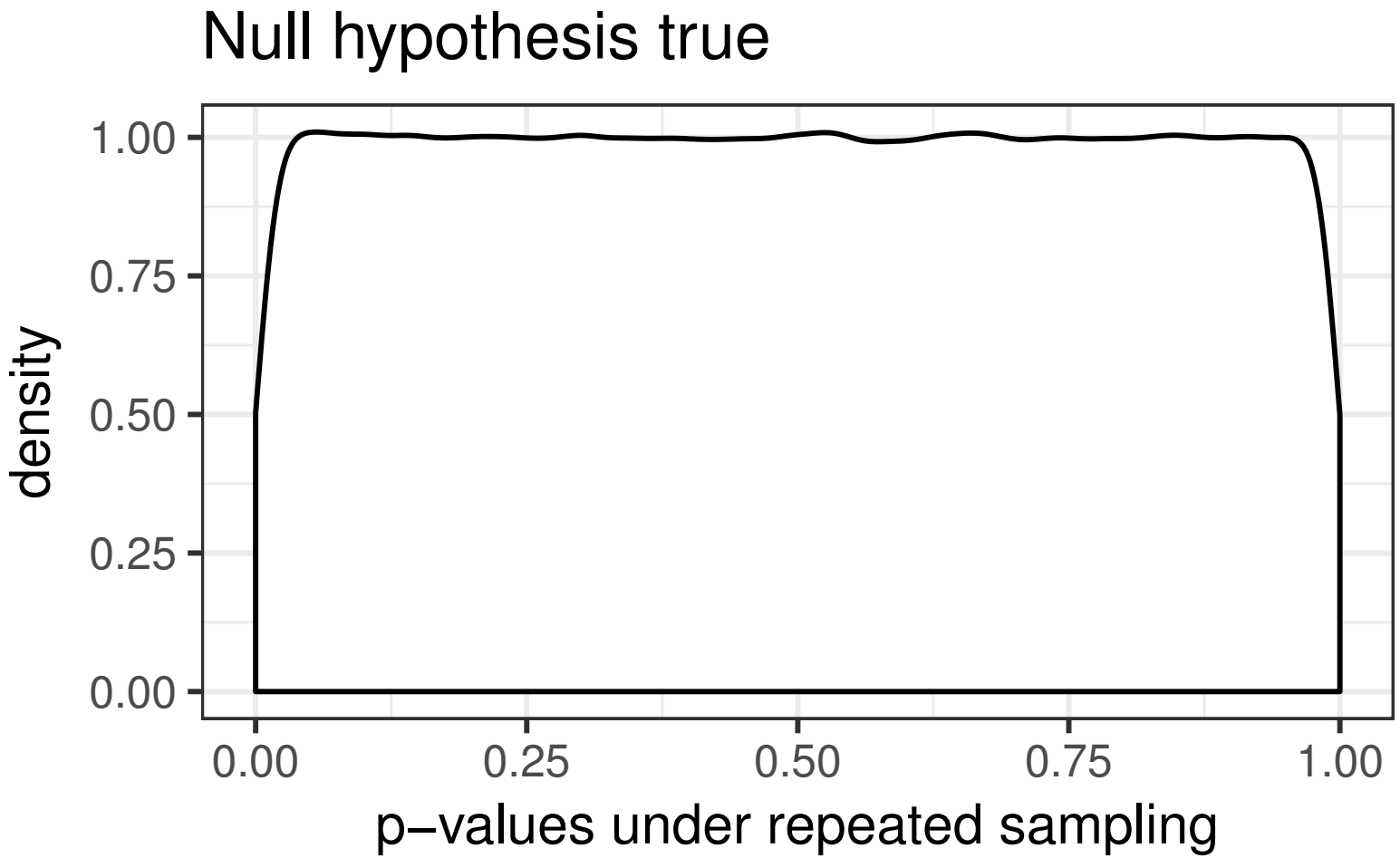
July 15, 2017

Introduction

> *"'. . . in [an]. . . academic environment that only publishes positive findings and rewards publication, an efficient way to succeed is to conduct low power studies. Why? Such studies are cheap and can be farmed for significant results, especially when hypotheses only predict differences from the null, rather than precise quantitative differences and trends."*
> *Smaldino and McElreath*

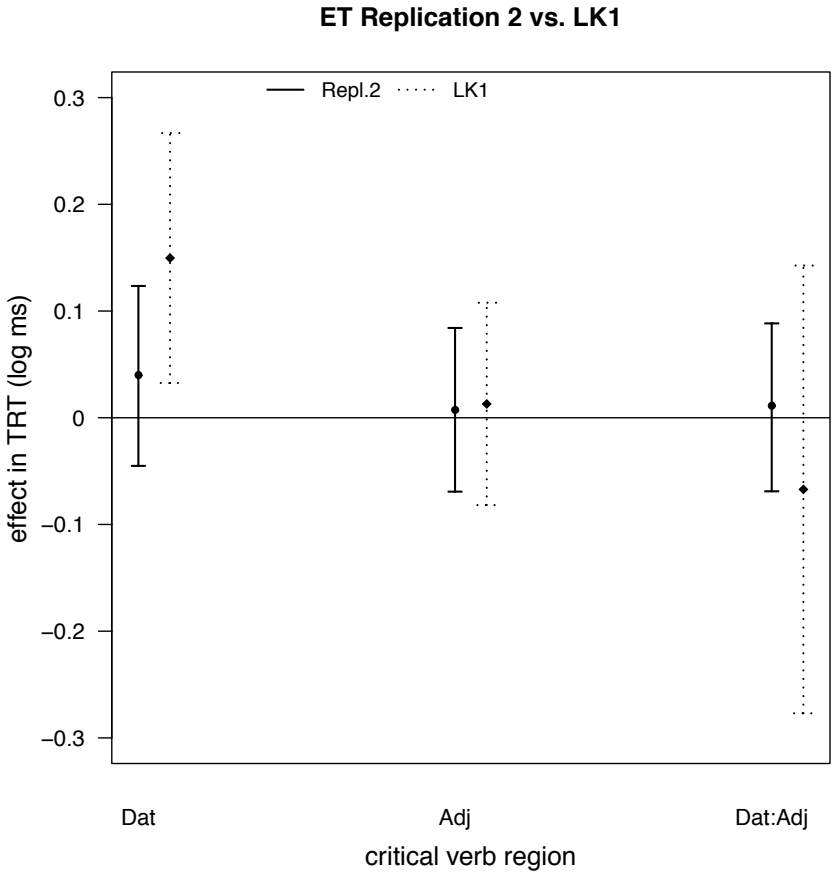
This is the state of affairs in psycholinguistics, linguistics, psychology, and more generally, cognitive science.

Statistical significance conveys no information if null hypothesis is known to be true



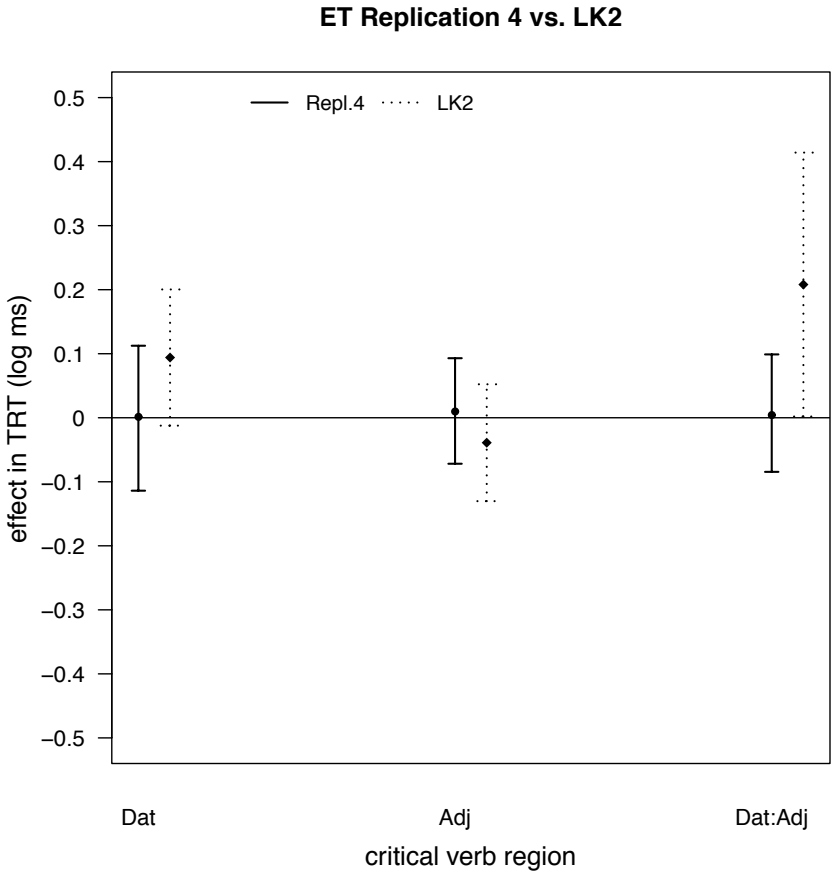
Replication attempts of Levy and Keller 2013, JML, Expts 1 and 2 (Mertzen et al)

Farming for significance also leads to over-enthusiatic expectations of **future** success in replication.



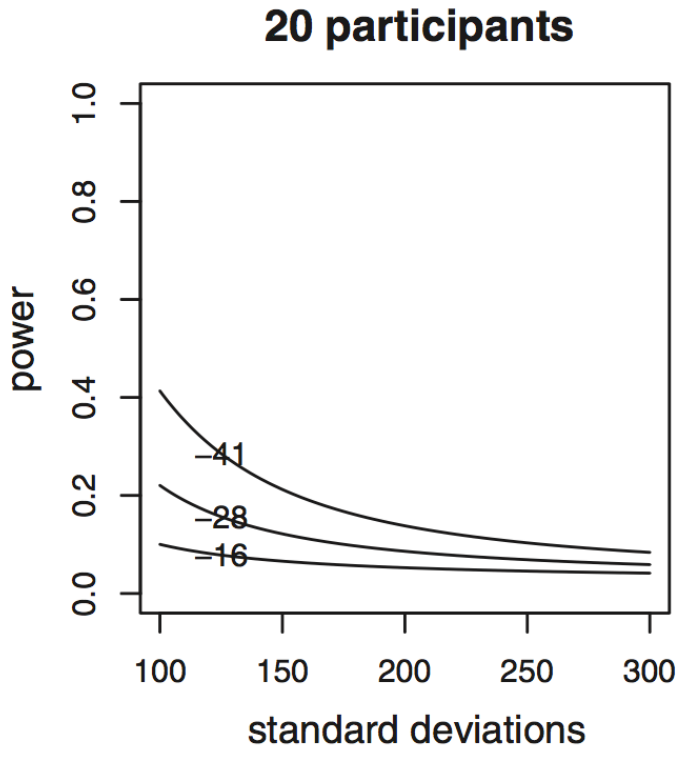
Replication attempts of Levy and Keller 2013, JML, Expts 1 and 2 (Mertzen et al)

Farming for significance also leads to over-enthusiatic expectations of **future** success in replication.



Low power: a staple of psycholinguistics

- This overconfidence is also evident in reading research in psycholinguistics, where it is routine to run experiments with sample sizes ranging from 20 to 40 participants.
- 20-40 participants will yield very low power (Jäger et al., 2017, JML).



Consider a low-power experiment

Consider an experiment with power 0.07: d=0.05, sd=1, two sided, one-sample t-test.

```
power.t.test(d=.05,sd=1,alternative="two.sided",n=100,
             type="one.sample")$power

## [1] 0.07148842
```

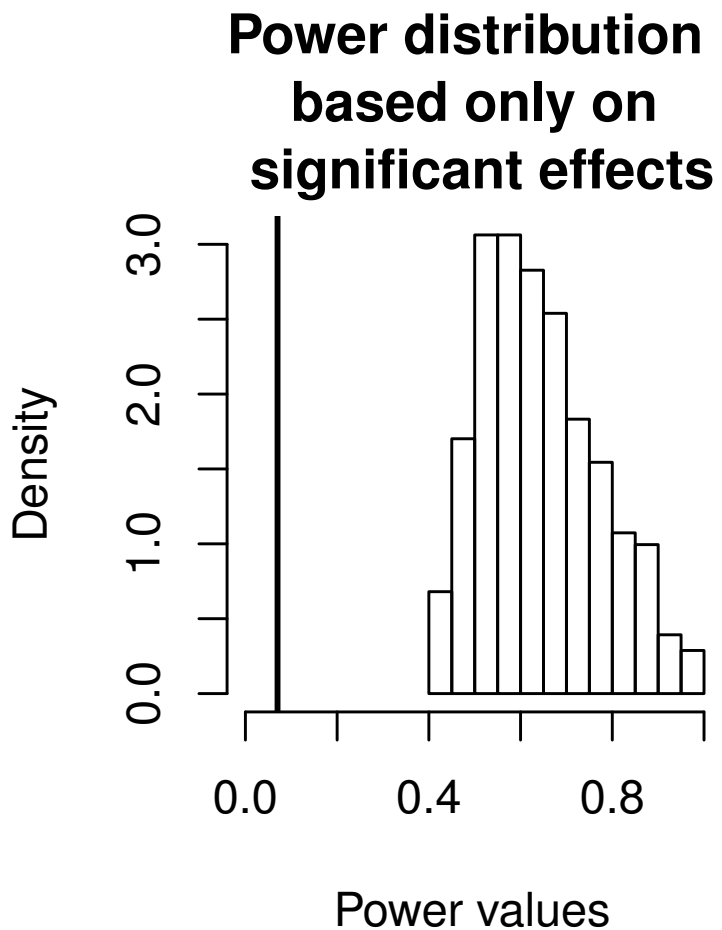
Suppose we conduct 10,000 replications, and publish either

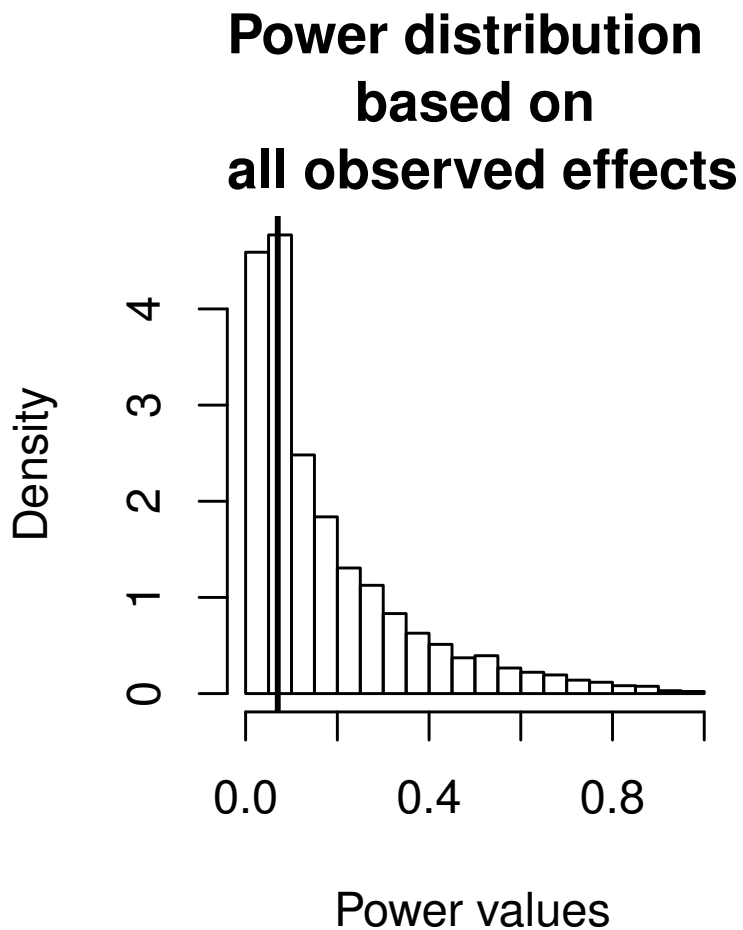
1. only significant results
2. all results

Computing power based on scenario (1) will lead us to an overestimate.

Introduction

How publication bias leads us to be overconfident about our findings





Agreement attraction

(1) a. The key to the cabinets are on the table.

b. The key to the cabinet are on the table.

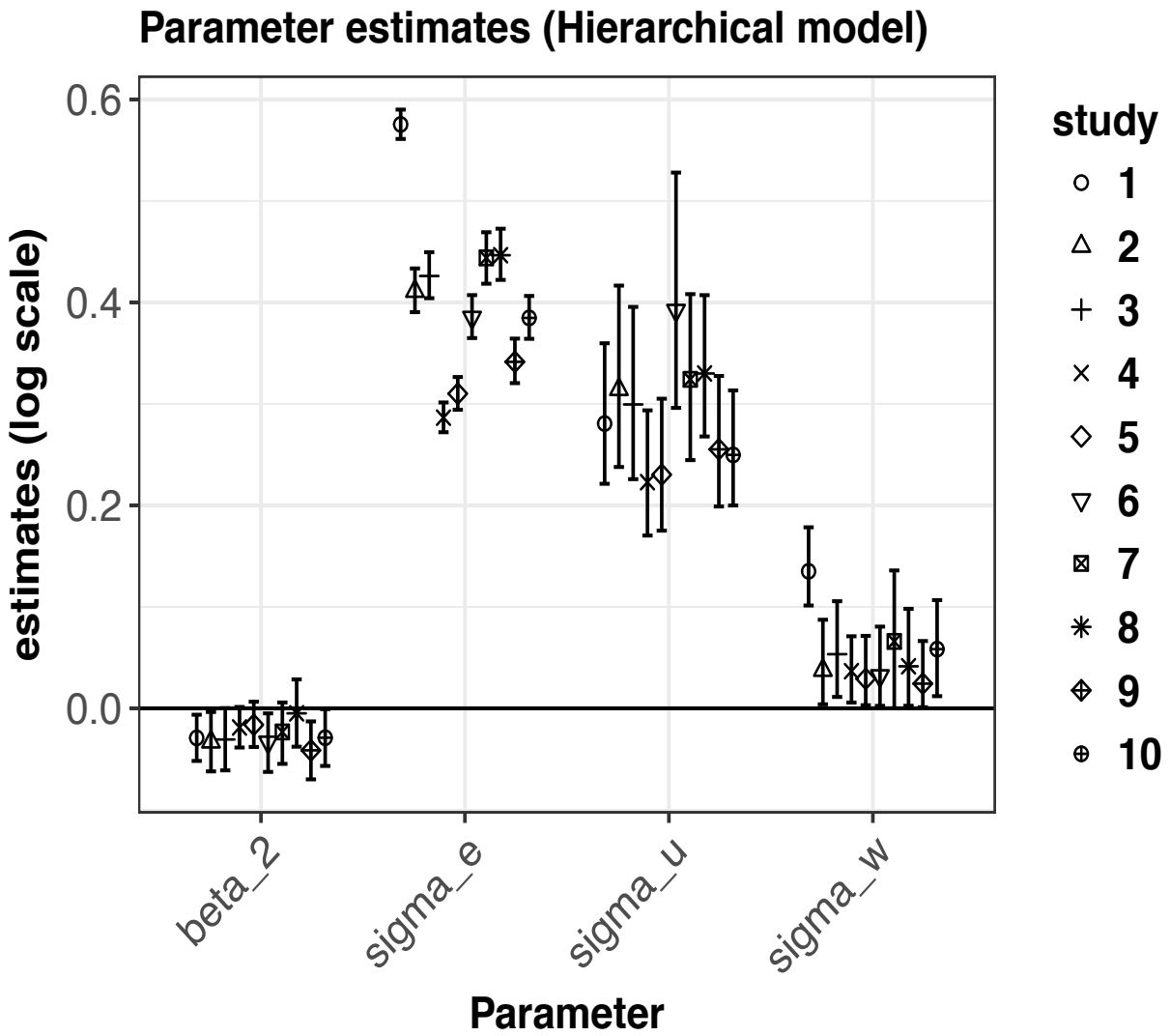
- Both sentences are ungrammatical, but the first one sounds distinctly better.
- Eyetracking and self-paced reading studies consistently show faster reading times at the auxiliary verb *are* in (1a) compared to (1b).

Agreement attraction

Reanalysis of log reading time data at auxiliary from 10 published experiments:

	t	d	n	se	s	pval	power
1	-1.9	-0.1	40	0.0	0.2	0.1	0.3
2	-3.1	-0.1	32	0.0	0.1	0.0	0.6
3	-1.5	-0.0	32	0.0	0.2	0.2	0.2
4	-2.1	-0.0	32	0.0	0.1	0.0	0.3
5	-1.7	-0.0	32	0.0	0.1	0.1	0.2
6	-2.6	-0.1	28	0.0	0.2	0.0	0.4
7	-1.6	-0.0	60	0.0	0.2	0.1	0.2
8	-3.2	-0.1	44	0.0	0.2	0.0	0.6
9	-1.9	-0.1	60	0.0	0.2	0.1	0.3
10	-2.6	-0.0	114	0.0	0.2	0.0	0.5

Summary of linear mixed model fits to the 10 studies



Random-effects meta-analysis

We can obtain a posterior distribution of the agreement attraction effect given these data by pooling information from all the studies (also see Jäger, Engelmann, Vasishth, 2017, JML).

Random-effects meta-analysis

- Let y_i be the effect size in log milliseconds in the i-th study, $i = 1, \ldots, n$.
- Let μ be the true (unknown) effect in log ms.
- Let μ_i be the true (unknown) effect in each study.
- Let σ_i log ms be the true standard error of each study.
- Let τ represent between-study variability.

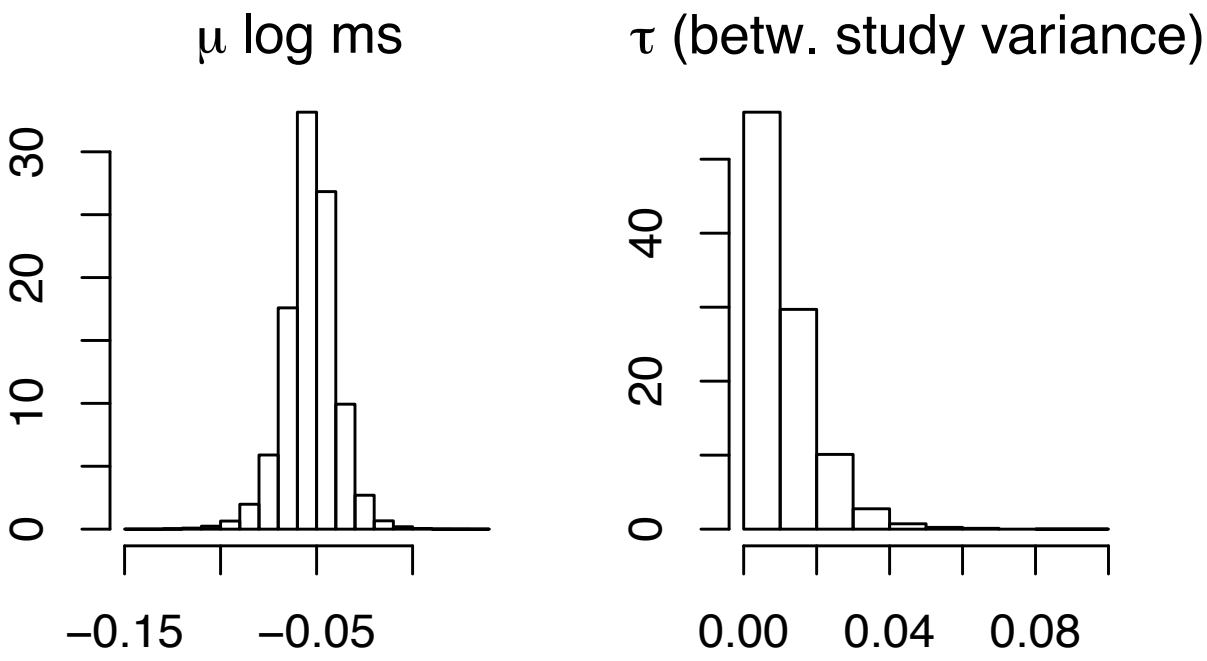
Likelihoods:

$$\begin{aligned} y_i \mid \mu_i, \sigma_i^2 &\sim Normal(\mu_i, \sigma_i^2) \quad i = 1, \ldots, n \\ \mu_i \mid \theta, \tau^2 &\sim Normal(\mu, \tau^2), \end{aligned}$$

Priors: (1)

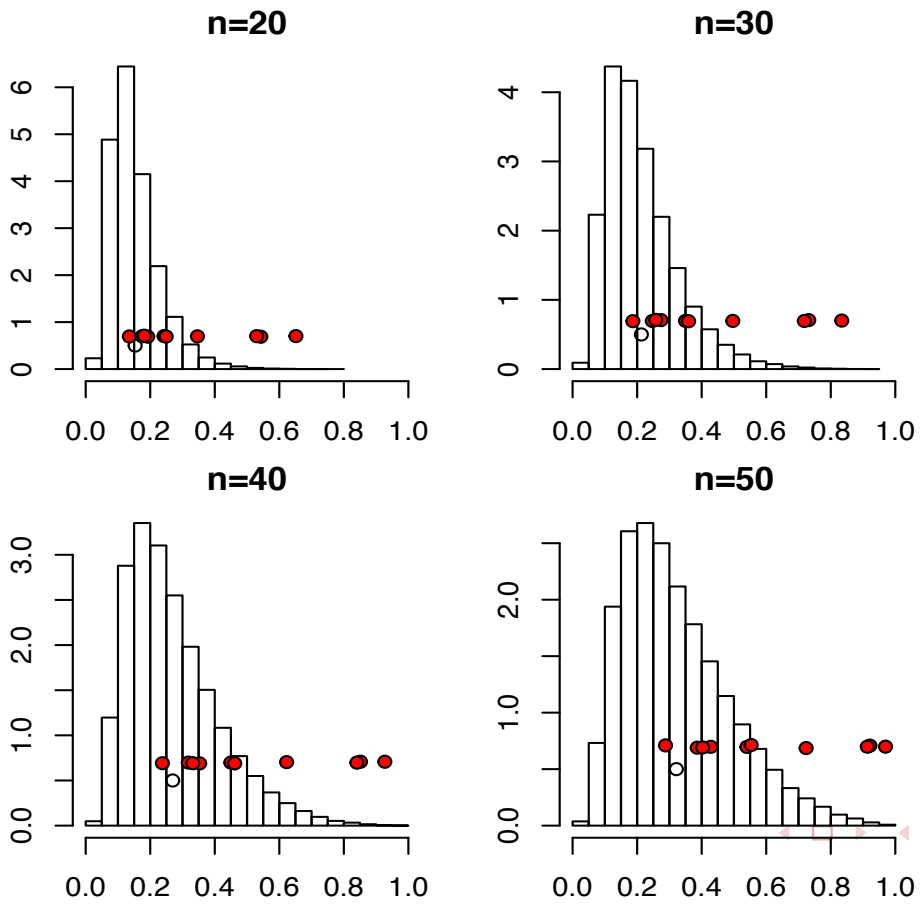
$$\begin{aligned} \mu &\sim Cauchy(0, 2.5), \\ \mu_i &\sim Cauchy(0, 2.5), \\ \tau &\sim Cauchy(0, 2.5), \tau > 0 \end{aligned}$$

Random-effects meta-analysis: Results



Power distributions based on meta-analysis posterior

Red dots show estimates from individual studies



Power inflation index

For a **given** sample size n, compared to the power distribution computed using the meta-analysis ($power_{meta,n}$), how much inflation in power do we get if we compute power using an individual study's effect ($power_{study,n}$)?

We can quantify this by computing the ratio:
Power inflation index$_n$ $= \frac{power_{study,n}}{power_{meta,n}}$

Notice that this index will be a distribution because $power_{meta}$ is a distribution.

Power inflation index

	n=20		n=30		n=40	
Study	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
1	1.37	4.64	0.98	3.95	0.76	3.47
2	3.67	12.45	2.63	10.61	2.04	9.30
3	1.08	3.67	0.78	3.13	0.60	2.74
4	1.95	6.61	1.40	5.64	1.08	4.94
5	1.41	4.76	1.01	4.06	0.78	3.56
6	3.06	10.37	2.19	8.83	1.70	7.75
7	0.76	2.59	0.55	2.20	0.42	1.93
8	2.99	10.12	2.14	8.62	1.65	7.56
9	0.98	3.34	0.71	2.84	0.55	2.49
10	1.02	3.47	0.73	2.96	0.57	2.59

Concluding remarks

- The statistical significance filter leads to over-optimistic expectations of replicability of published research.
- Even if the researcher doesn't conduct any formal power analyses, they can fall prey to this illusion.
  They express this certainty through phrases like "*the effect was reliable*."
- We illustrated the illusion of replicability through a case-study involving 10 published experimental comparisons.

Concluding remarks

- Many psychology journals are beginning to require that power analyses be included in submitted manuscripts.
  JEP:HPP requirement:
  Did you provide explicit information to support that each of your studies has sufficient power and/or precision for confirmatory analyses?
- But such analyses, which invariably are based on previously published work, will tend to provide overestimates of power.

Concluding remarks

To resolve or at least reduce this problem, we offer two suggestions.

- First, abandon the concept of power, which is based on the idea that "$p < .05$" is a win, an attitude that fails miserably when effect sizes are small and measurements are noisy.
- Second, when performing design analysis, consider a range of reasonable effect sizes (Gelman & Carlin, 2014).
- Aim to achieve as high a precision of the parameter estimate as possible.
- Don't rely on the **imagined** frequentist properties of an experiment—these properties are a fantasy. Aim to **actually** replicate the observed pattern of results.

Full paper: https://arxiv.org/abs/1702.00556