LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

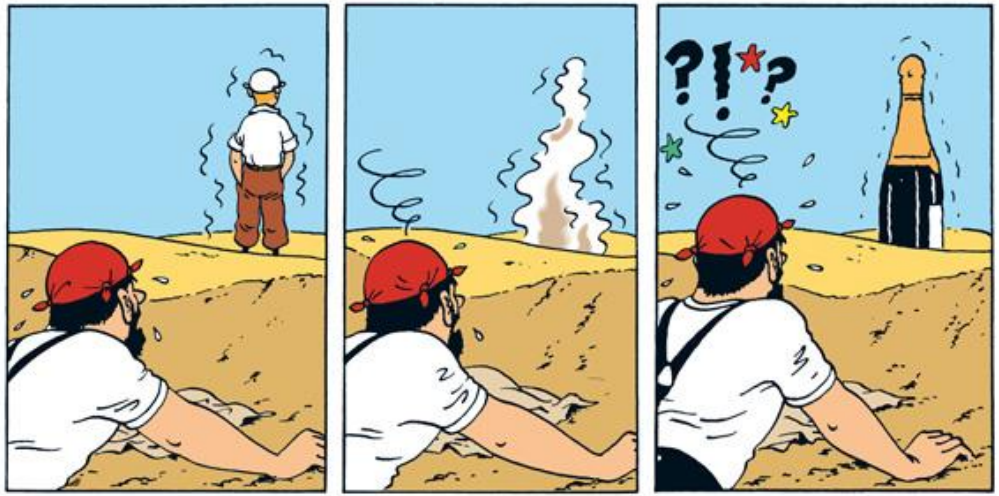
Are most published research findings false?

Trends in statistical power, publication bias and p-hacking as well as the false discovery rate in psychology (1975–2017)

Analytical Sociology: Theory and Empirical Applications, November 16th 2020

Andreas Schneck (LMU Munich)

Problem in a Nutshell



From: Tintin - The crab with the golden claws (1941)

True negative

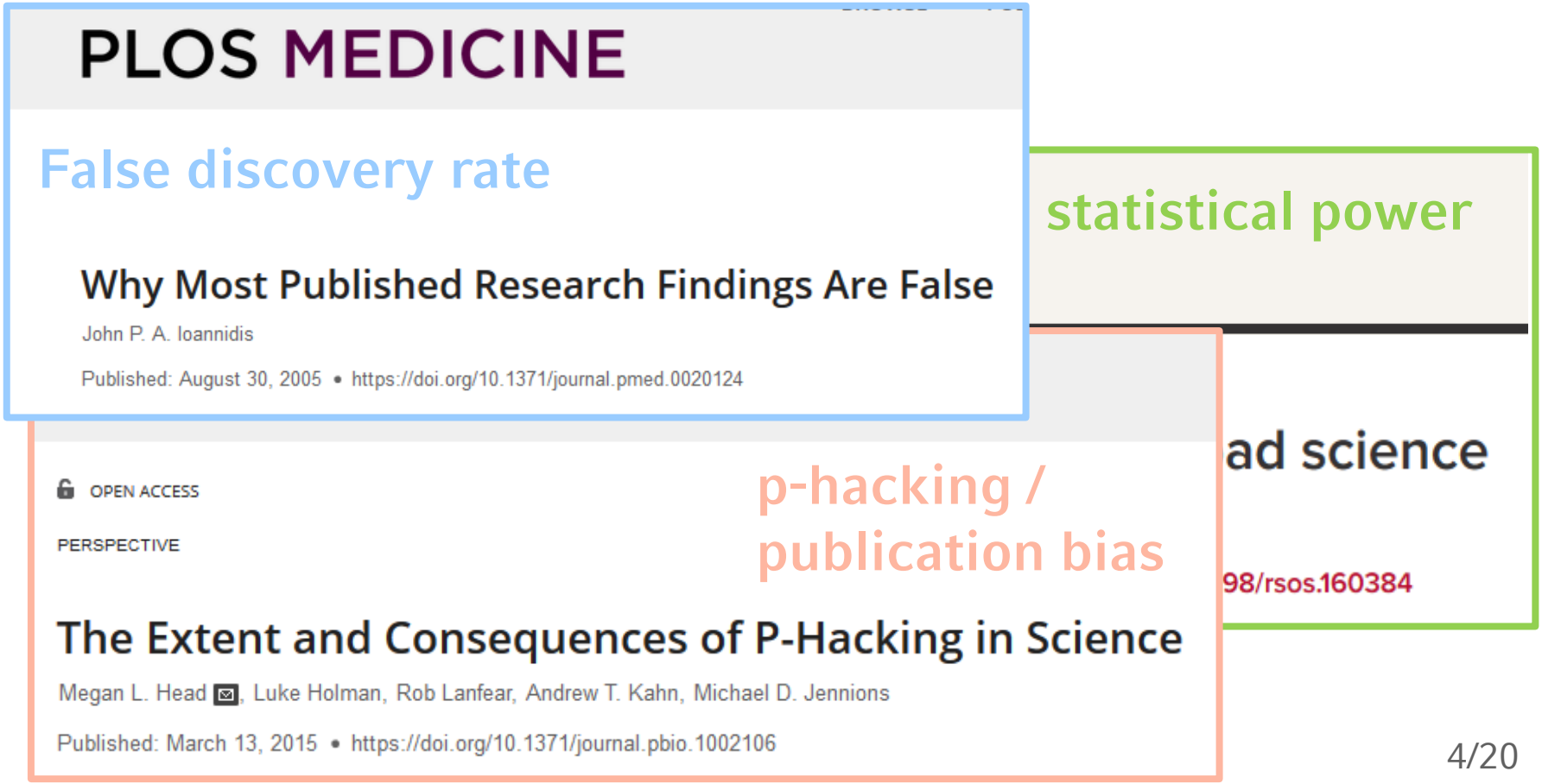
False positive (desired)

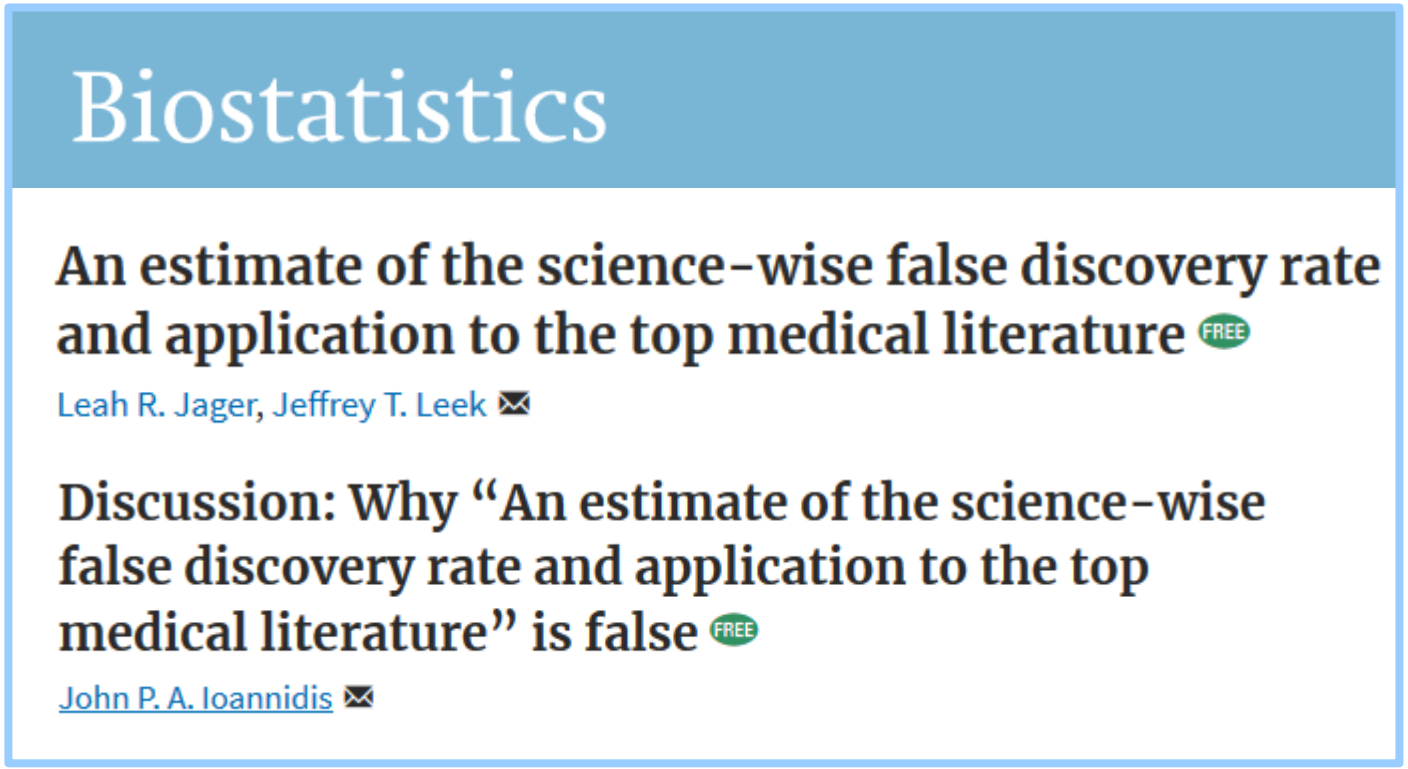
- In Science: How common are true positive effects and ‚mirage'-like false positives?

Quality Criteria

		Truth: Effect	Truth: No Effect	
Estimator	Sign.	True Positive (TP)	False Positive (FP)	False Discovery Rate (**FDR**) = FP / (TP + FP)
	N. sign.	False Negative (FN)	True Negative (TN)	
		Statistical Power (**pow**) = TP / (TP + FN)	False Positive Rate (**FPR**) = FP / (FP + TN)	

- High FDR as direct consequence of publication bias (inflated FPR) and low statistical power


PLOS MEDICINE
False discovery rate
Why Most Published Research Findings Are False
John P. A. Ioannidis
Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124
statistical power
ad science
98/rsos.160384
OPEN ACCESS
PERSPECTIVE
p-hacking /
publication bias
The Extent and Consequences of P-Hacking in Science
Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions
Published: March 13, 2015 • https://doi.org/10.1371/journal.pbio.1002106


Biostatistics
An estimate of the science-wise false discovery rate and application to the top medical literature
FREE
Leah R. Jager, Jeffrey T. Leek
Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false
FREE
John P. A. Ioannidis

Data

Challenges of past research

- Manual coding of articles immense time consuming… (stat. power)
- … or focus on (selective) abstracts (selection bias)

Needs

- Accessible & relevant test-values…
- … that allow for automatic extractions…
- … over a substantial period of time

APA reporting-guideline (1974-)

- All relevant results have to be mentioned along with the test statistic
- In-text reporting very common and standardized (e.g. F(1, 4) = 3.25)
- $p < 0.05$ as first significance threshold

➢ Automatic export of test-values via web-scraping (PsycArticles) and text-mining in Python

➢ **In total 648.578 test-values from 39.218 articles (1975-2017)**

Example:

The statistical comparison of visuospatial working memory performance of patients with ADHD (ADHD1, $n = 48$) with that of the healthy comparison group revealed a significantly lower performance of large size among patients with ADHD, Wilk's $\lambda = 0.825$, $F(2, 93) = 9.847$, $p < .001$, $\eta^2 = .175$ (see Table 6). Univariate comparisons demonstrated medium and significant effects regarding errors, $F(1, 94) = 8.170$, $p = .005$, $d = 0.74$, and total mean response time of correct responses, $F(1, 94) = 12.617$, $p = .001$, $d = 0.72$, indicating a poorer visuospatial working memory performance in patients with ADHD.

Measures

Publication Bias/ p-hacking

General definition:

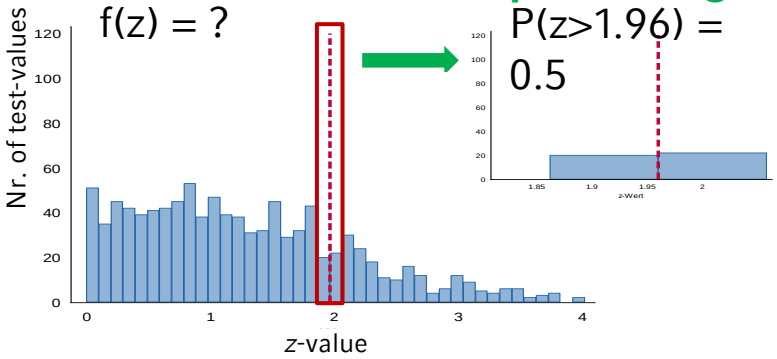
- „a tendency toward preparation, submission and publication of research findings based on the **nature** and **direction** of the research results“ (Dickersin 2005: 13)
- Either sign (direction) or significance (nature) can be the target

- **Publication Bias:** Repeated data collection in case of non-significant results (Rosenthal 1979)
- **p-hacking:** Achieve significant results via changes in the modelling strategy (Simonsohn et al. 2014)

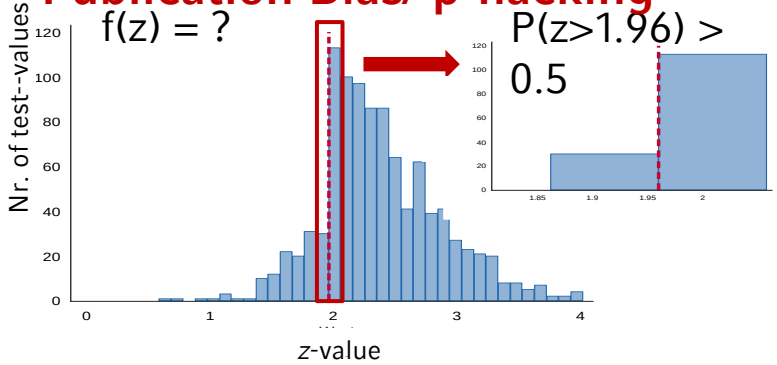
➢ Both increase false positives substantially!

Caliper Test (Gerber & Malhotra 2008a,b)

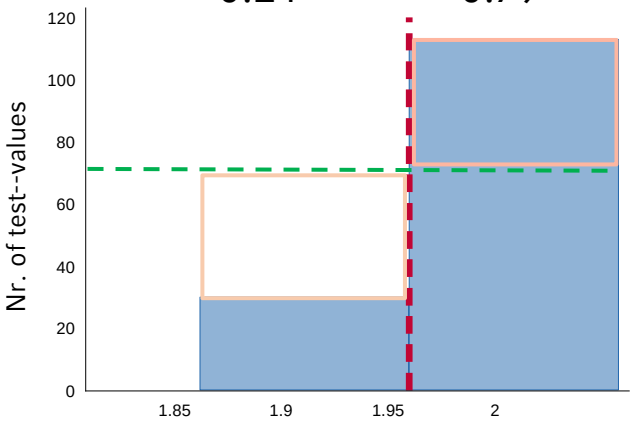
No Publication Bias/ p-hacking



Publication Bias/ p-hacking



$P(x_z = 0) = 0.21$ $P(x_z = 1) = 0.79$



$$\rho = \left(\frac{P(x_z = 1)}{0.5}\right) - 1 = 0.58$$

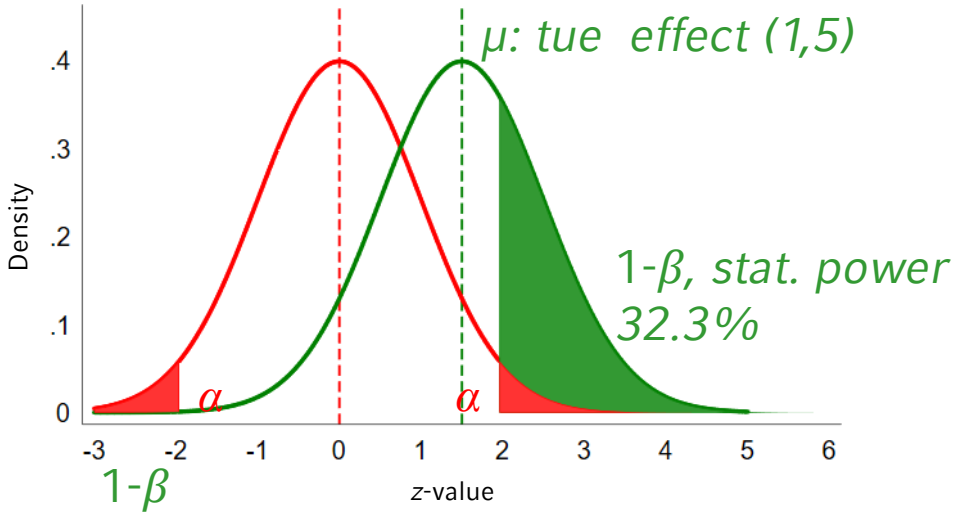
$$FPR_{inf} = FPR + \rho * (1 - FPR)$$

Statistical power (1-β): How many true effects are actually detected?

$$pow = \Phi\left(\Phi^{-1}(0.025) - \left(\frac{\mu}{\sigma_i}\right)\right) + \left(1 - \Phi\left(\Phi^{-1}(0.975) - \left(\frac{\mu}{\sigma_i}\right)\right)\right)$$

- σ: Precision of study (example: 1)

- μ unknown but can be approximated by the mean effect $\hat{\mu}$ (cp. Ioannidis et al. 2017)
- Meta-analyses by subdisciplines (PIC-codes in Psychology)



False discovery rate (FDR): How many significant results are actually false?

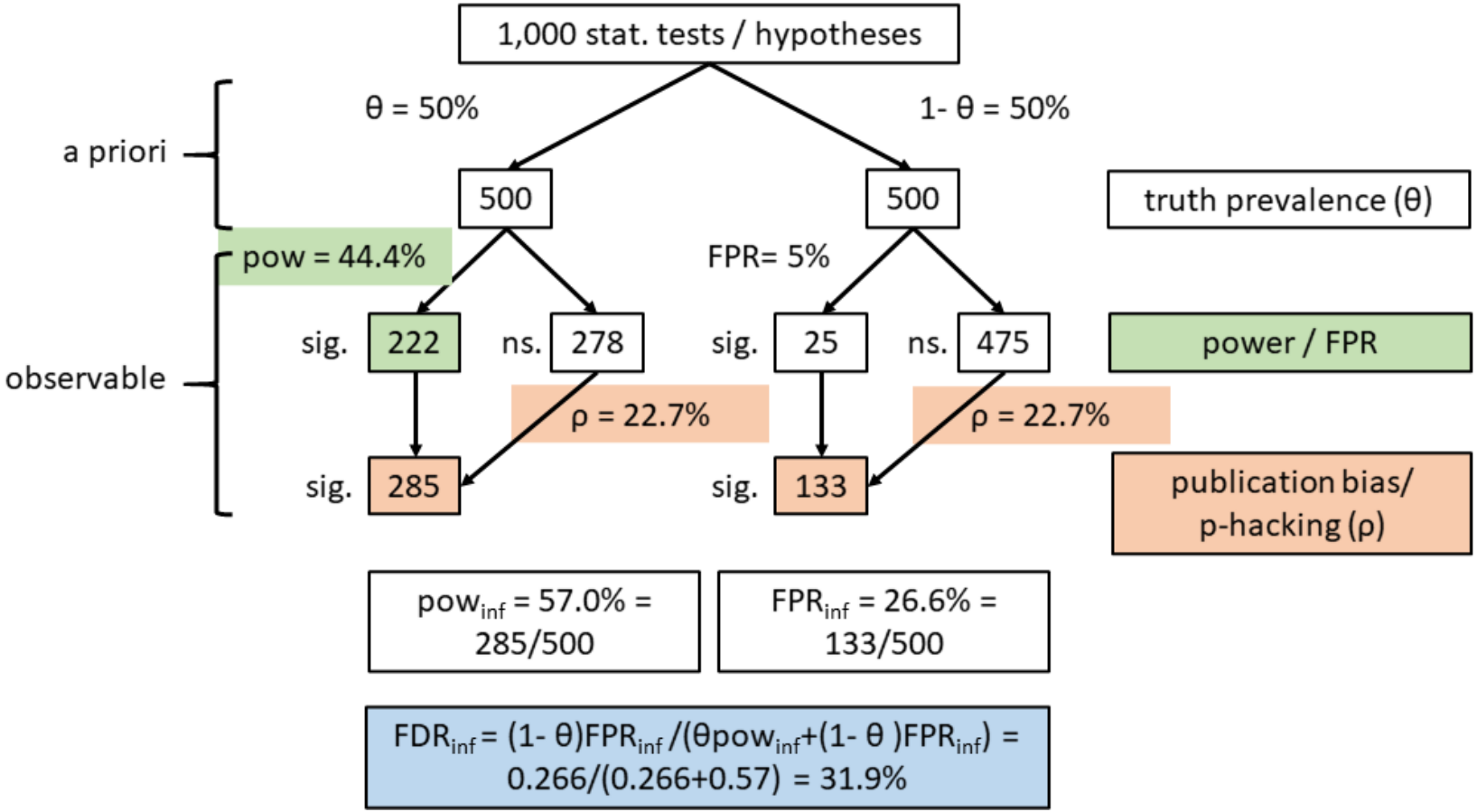
- Dependent on the statistical power (*pow*) and the false-positive rate (FPR)
- Additionally an *a priori* probability (θ), that the research hypothesis is true has to be specified

➢ 50% assumed (but also computed for 10% & 20%) – theoretically sound (Diekmann 2011)

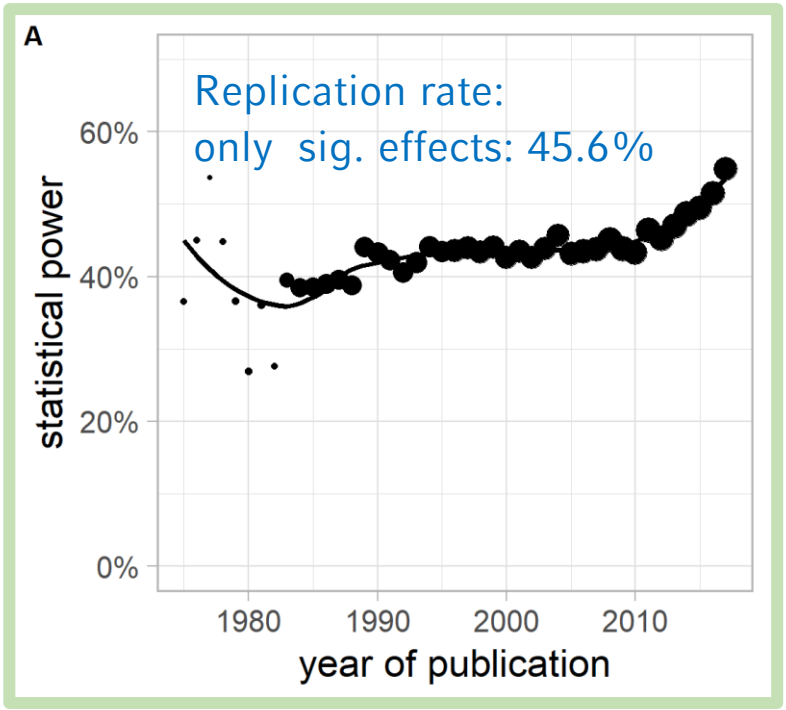
$$FDR = \frac{(1-\theta)FPR}{\theta pow + (1-\theta)FPR}$$

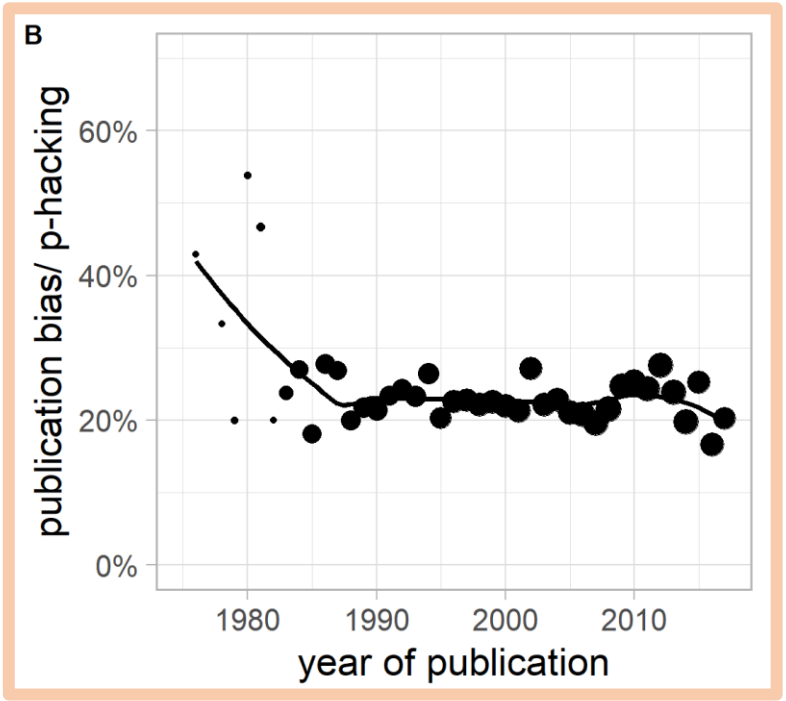
Results

Aggregate Results



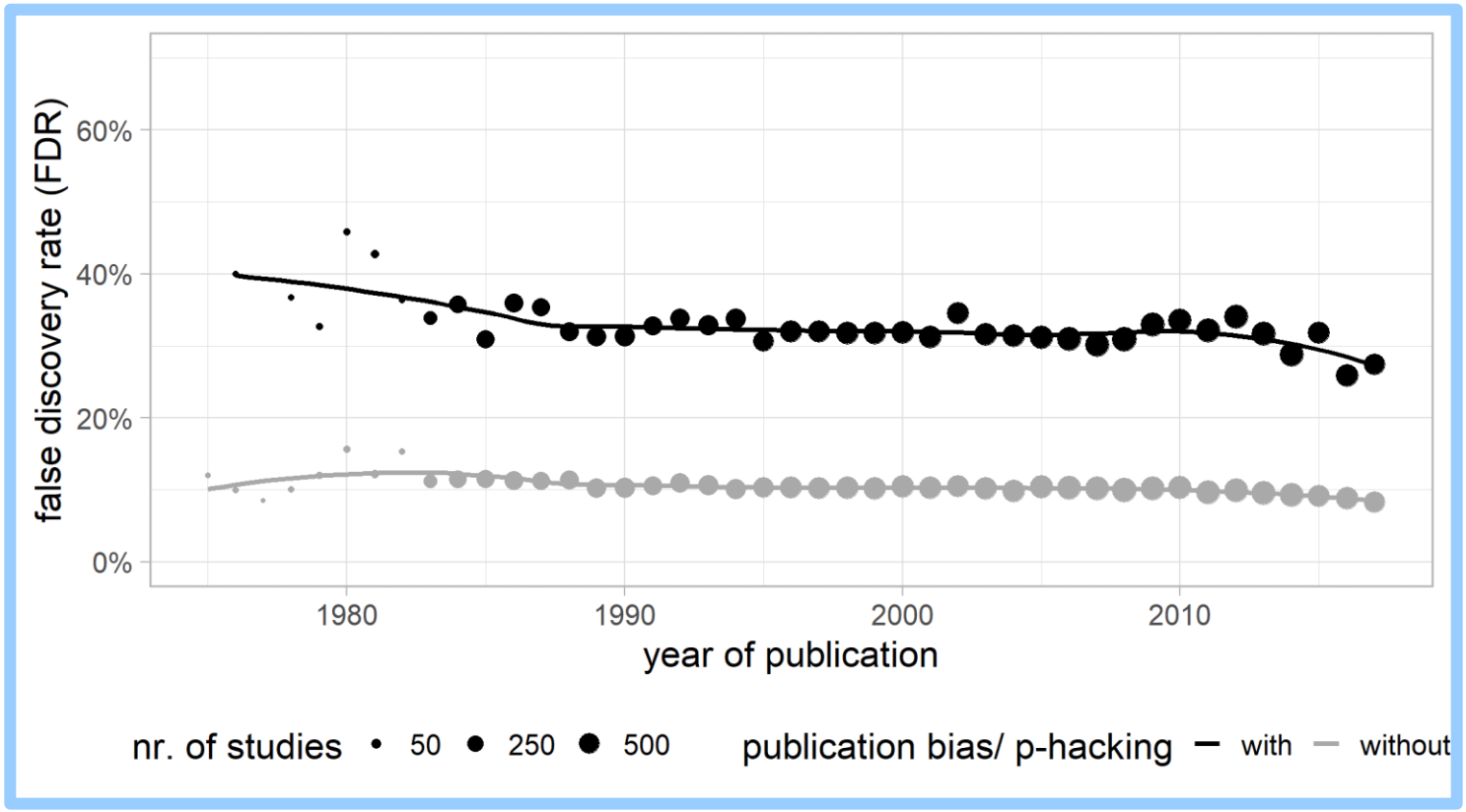
Statistical Power & Publication Bias/ p-hacking by Year





number of studies • 20 • 100 ● 500 ● 750

FDR by Year

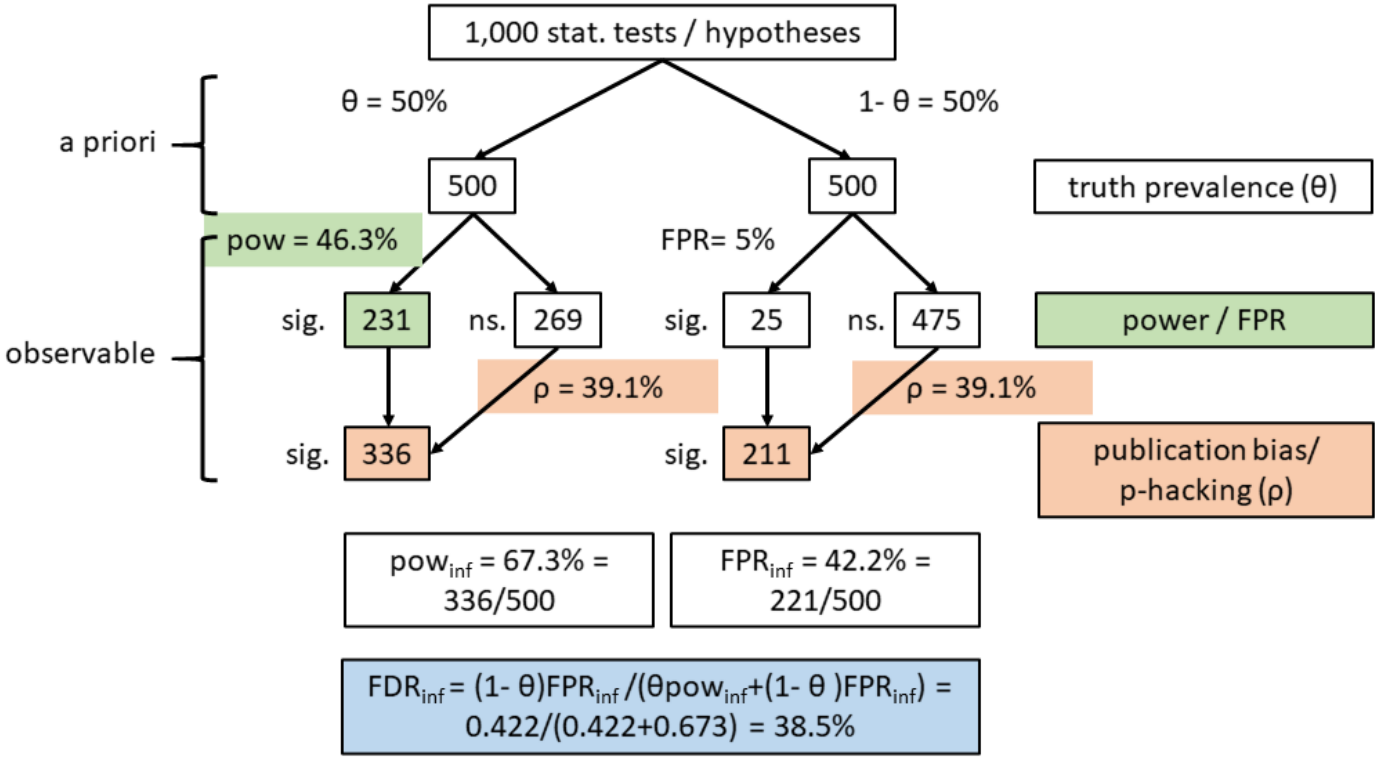

false discovery rate (FDR)
60%
40%
20%
0%
1980
1990
2000
2010
year of publication
nr. of studies
50
250
500
publication bias/ p-hacking
with
without

Disclaimer

- Psychology is examined only because of its strict reporting guidelines that allow for such large scale analyses
- There are no indications that other disciplines are better off!

But what about sociology?

First results preliminary…



Summary

- Publication Bias/ p-hacking is substantial
- Statistical power is way too low

➢ As a consequence, around 32% (Psy)/ 38.5% (Soc) of all statistically significant results are likely to be false

Publication bias/ p-hacking as influential factor:

➢ **Preregistration of the research design along with a complete model specification (lower publication bias)** (Miguel et al. 2014)

Literature

Chatterjee, M., & Chakraborty, A. K. (2016). A Simple Algorithm for Calculating Values for Folded Normal Distribution. *Journal of Statistical Computation and Simulation, 86*, 293-305.

Dickersin, K. (2005). Publication Bias: Recognizing the Problem, Understanding Its Origins and Scope, and Preventing Harm. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (pp. 11-33). Oxford: Blackwell Science.

Diekmann, A. (2011). Are Most Published Research Findings False? Jahrbücher für Nationalökonomie und Statistik, 231, 628-636.

Francis, G. (2012). The Same Old New Look: Publication Bias in a Study of Wishful Seeing. *i-Perception, 3*, 176-178.

Gerber, A. S., & Malhotra, N. (2008). Publication Bias in Empirical Sociological Research. *Sociological Methods & Research, 37*, 3-30.

Gerber, A. S., & Malhotra, N. (2008). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science, 3*, 313-326.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of p-Hacking in Science. *PLOS Biology, 13*, e1002106.

Ingre, M., & Nilsonne, G. (2018). Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate. *Royal Society Open Science, 5*, 181190-181190.

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLOS Med, 2*, e124.

Ioannidis, J. P. (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics, 15*, 28-36; discussion 39-45.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. The Economic Journal, 127, F236-F265.

Jager, L. R., & Leek, J. T. (2014). An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature. *Biostatistics, 15*, 1-12.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). Behavior Research Methods, 48, 1205-1226.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin, 86*, 638-641.

Smaldino, P. E., & McElreath, R. (2016). The Natural Selection of Bad Science. *Royal Society Open Science, 3*.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-Regression Approximations to Reduce Publication Selection Bias. *Research Synthesis Methods, 5*, 60-78.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A Key to the File Drawer. *Journal of Experimental Psychology-General, 143*, 534-547.

Wilson, B. M., & Wixted, J. T. (2018). The Prior Odds of Testing a True Effect in Cognitive and Social Psychology. Advances in Methods and Practices in Psychological Science, 1, 186-197.