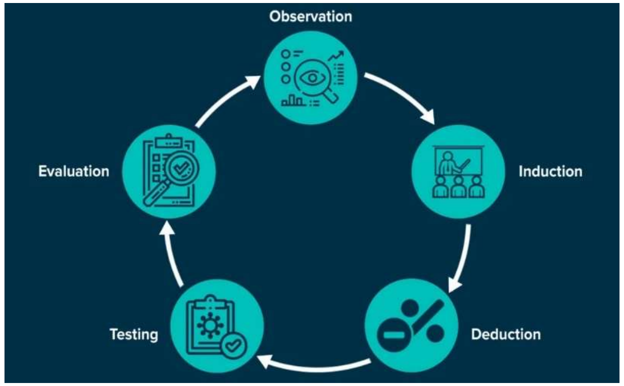
Empirical research in management and economics

The replication crisis and open science

Thorsten Pachur

Technical University of Munich
School of Management
Chair of Behavioral Research Methods



Evaluation is now open

Please take part in the evaluation of the module!

Your feedback is highly appreciated.

Evaluation period: 10–23 January

If you have questions on topics that we covered and that you would like to be clarified in the final lecture …

Submit your question(s) via the following link by 22 January

https://forms.gle/jpcwqhXa6cQPPTwb8

Recap from last week

- What is the goal of a conjoint analysis?
- What are important criteria when designing attribute profiles for a conjoint analysis?
- How does a fractional orthogonal design differ from a full factorial design? What does it mean if a fractional design is orthogonal?
- Give different approaches to collect preference data
- What is the structure of a conjoint analysis model?
- How does effect coding differ from dummy coding?
- Describe the calculation of partworth utilities
- Describe the calculation of attribute importances

Agenda for the semester

Session	Date	Topic
1	13 October	Introduction
2	20 October	Descriptive data analysis
3	27 October	Hypothesis development and measurement
4	3 November	Inferential data analysis I
5	10 November	Inferential data analysis II
6	17 November	Simple regression
7	24 November	Multiple regression
8	1 December	Logistic regression
9	8 December	Factor analysis
10	15 December	Cluster analysis
11	12 January	Conjoint analysis
12	**19 January**	**The replication crisis and open science**
13	26 January	Summary and questions
	11 February	Exam

Goals for this week

- You are familiar with the replication crisis in the behavioral sciences
- You know what *questionable research practices* are (e.g., p-hacking, HARKing)
- You can describe why questionable research practices are problematic and can lead to nonreplicable findings
- You know the principles of open science, including the steps of conducting a preregistration
- You are familiar with principles underlying the Bayesian approach to data analysis

How replicable is research in the behavioral sciences?

→ Psychology

- Sampling scheme
  - 100 articles from three leading psychology journals (published in 2008)
  - Always last experiment reported in the article
  - 97 of the 100 sampled experiments had significant results
- Replication teams from different labs: 270 authors in total

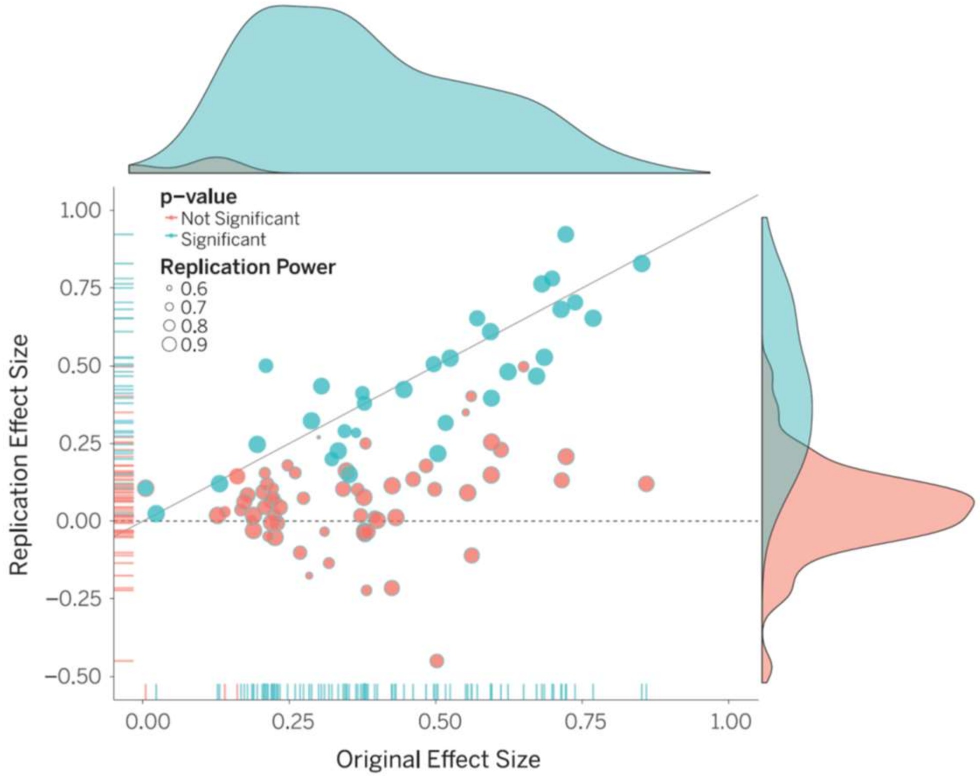
Brian Nosek

Open Science Collaboration (2015)

Replicability of research in psychology

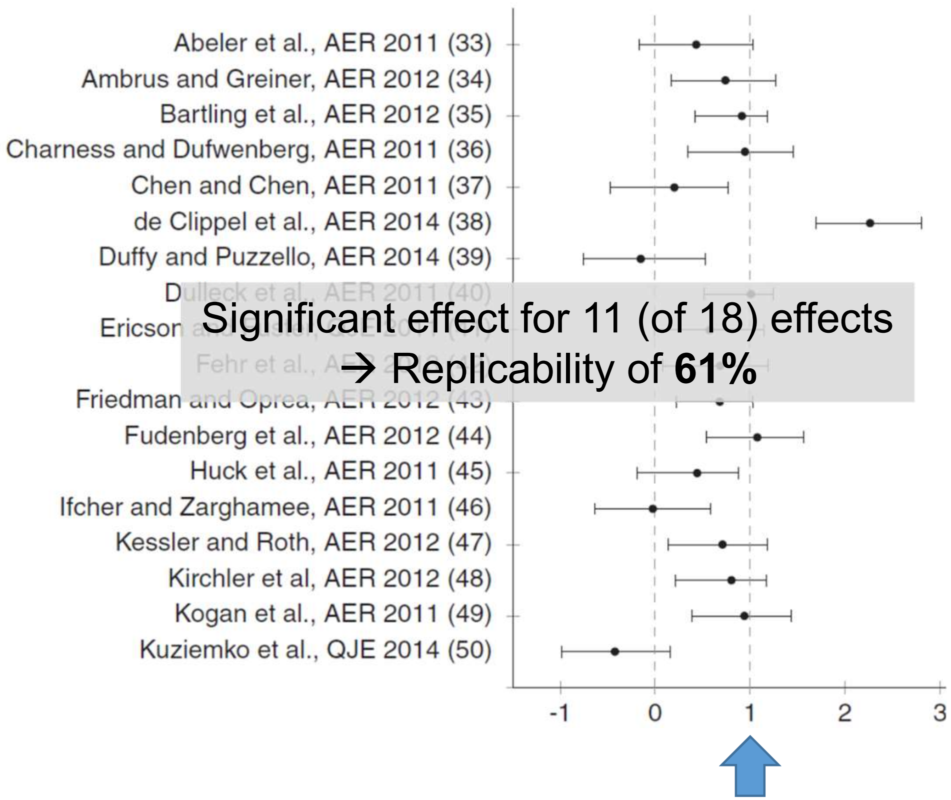
	Replications $P < 0.05$ in original direction	Percent	Mean (SD) original effect size	Mean (SD) replication effect size
Overall	35/97	36	0.403 (0.188)	0.197 (0.257)
JPSP, social	7/31	23	0.29 (0.10)	0.07 (0.11)
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	0.27 (0.24)
PSCI, social	7/24	29	0.39 (0.20)	0.21 (0.30)
PSCI, cognitive	8/15	53	0.53 (0.2)	0.29 (0.35)

Significant effect for 35 (of 97) effects
→ Replicability of **36%**



Open Science Collaboration (2015)

Replicability of research in experimental economics



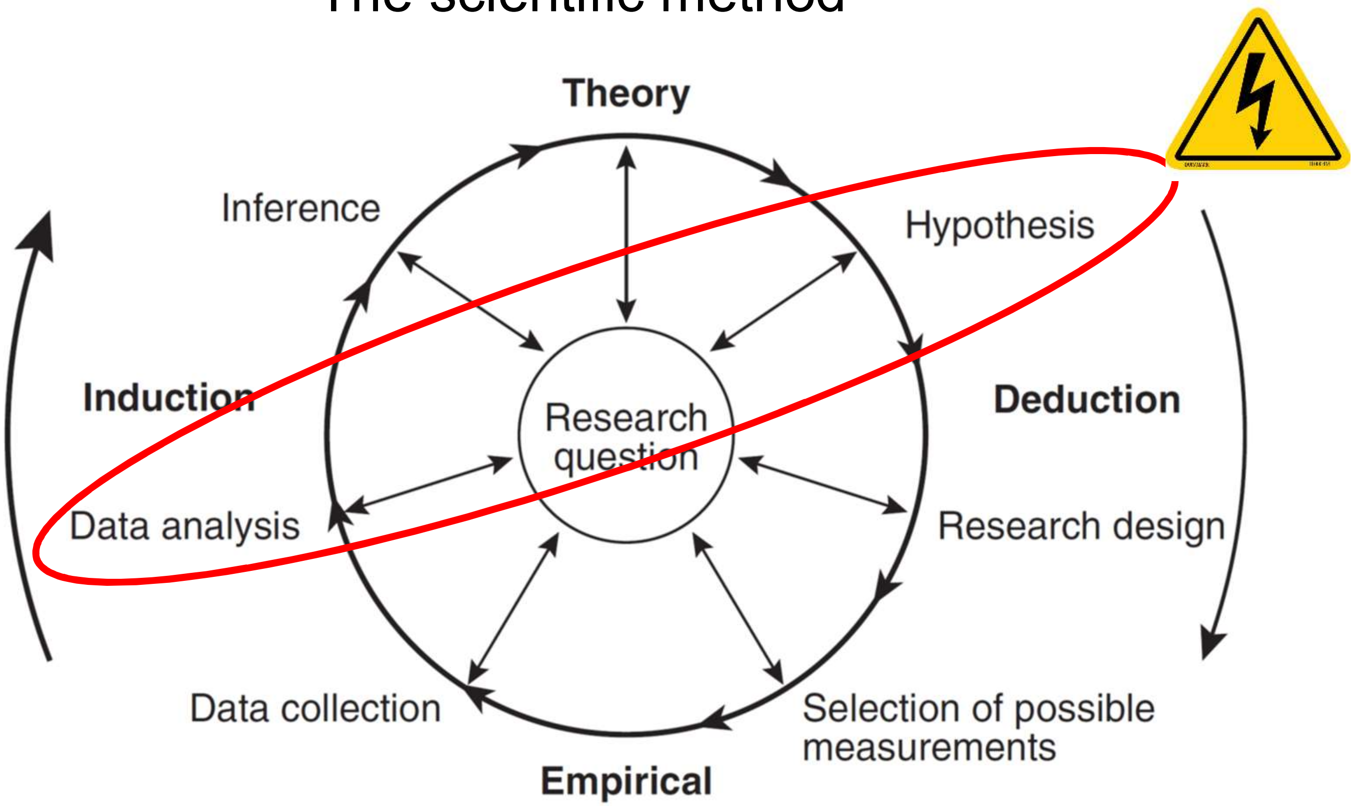
Significant effect for 11 (of 18) effects
→ Replicability of **61%**

1 = original effect size

Camerer et al. (2010)

Hypothetico-deductive model

"The scientific method"

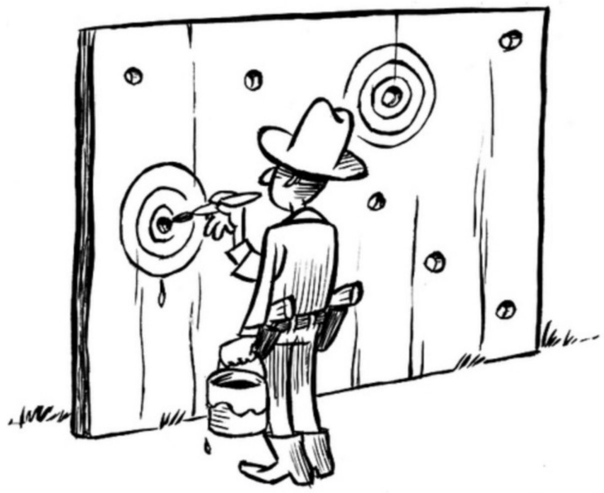


HARKing

Hypothesizing **A**fter the **R**esults are **K**nown (Kerr, 1998)

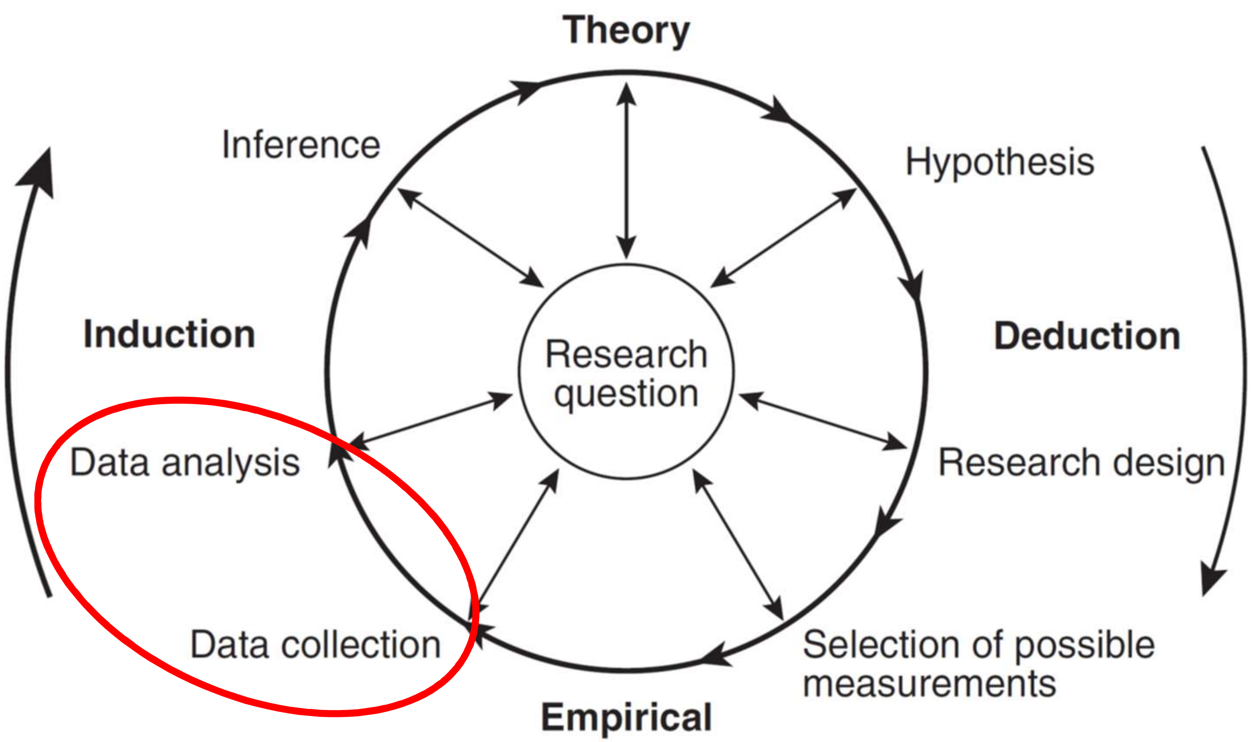
Presenting a post-hoc hypothesis that is informed by a result of the data analysis as if the hypothesis was, in fact, developed prior to data analysis.

Example: Performing unplanned subgroup analyses, finding an effect in one subgroup and writing the introduction with a "hypothesis" that matches these results.



Hypothetico-deductive model

"The scientific method"



Considerable degrees of freedom

Degrees of freedom during data collection and data analysis

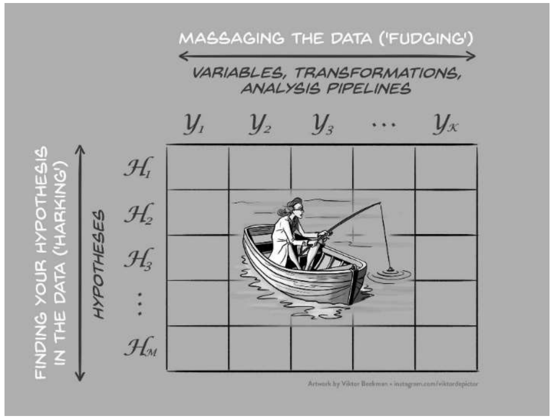
Decisions that have to be made

- When to stop recruiting sampling participants?
- How to define outliers in the data (that are then excluded)?
- Which covariates or moderators should be used?
- On which measures (operationalizations) are the key conclusions based? Which measures are reported?
- (How) Are data transformed (e.g., log-transformation)?
- Which statistical test(s) to run (e.g., regression vs. ANOVA; Pearson vs. Spearman correlation)?
- …

p-hacking

Making relevant decisions contingent on the outcome of the data analysis (i.e., whether it yields a significant result) rather than making them beforehand.

Example: Performing multiple types of analyses and reporting only those that are significant (i.e., $p < 0.05$), selectively removing data until $p < 0.05$ or selecting variables for use in analyses based on whether they yield significant results.

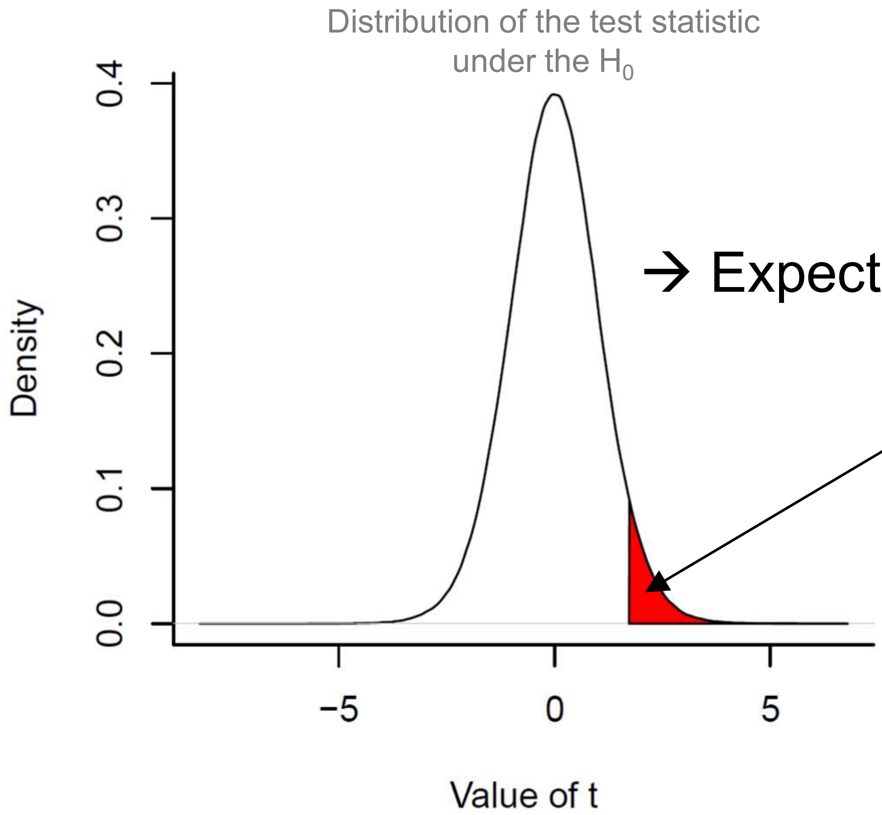


Questionable research practices

- HARKing: Reporting an unexpected finding as having been predicted from the start
- p-hacking
  - Selective reporting of (dependent) variables, experimental conditions, and statistical tests
  - Deciding whether to collect more data after looking to see whether the results are significant
  - Reporting selectively studies that worked
  - Rounding off a p-value just above .05 (e.g., .054) and claim that the results are significant

Why are questionable research practice problematic and can lead to low replicability?

Expected false-positivity rate in standard null hypothesis statistical testing



$\alpha = 5\%$

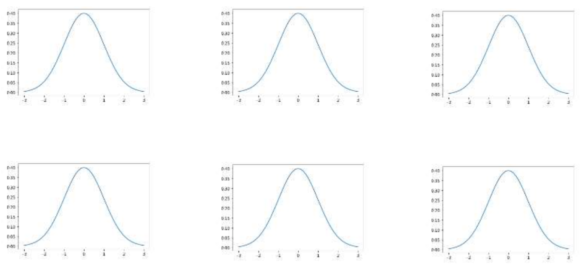
→ Expected false positive rate is 5%

Result-contingent decisions can inflate the false-positive rate

Condition 1	Condition 2	Condition 3
N=20	*N*=20	*N*=20

DV1

DV2

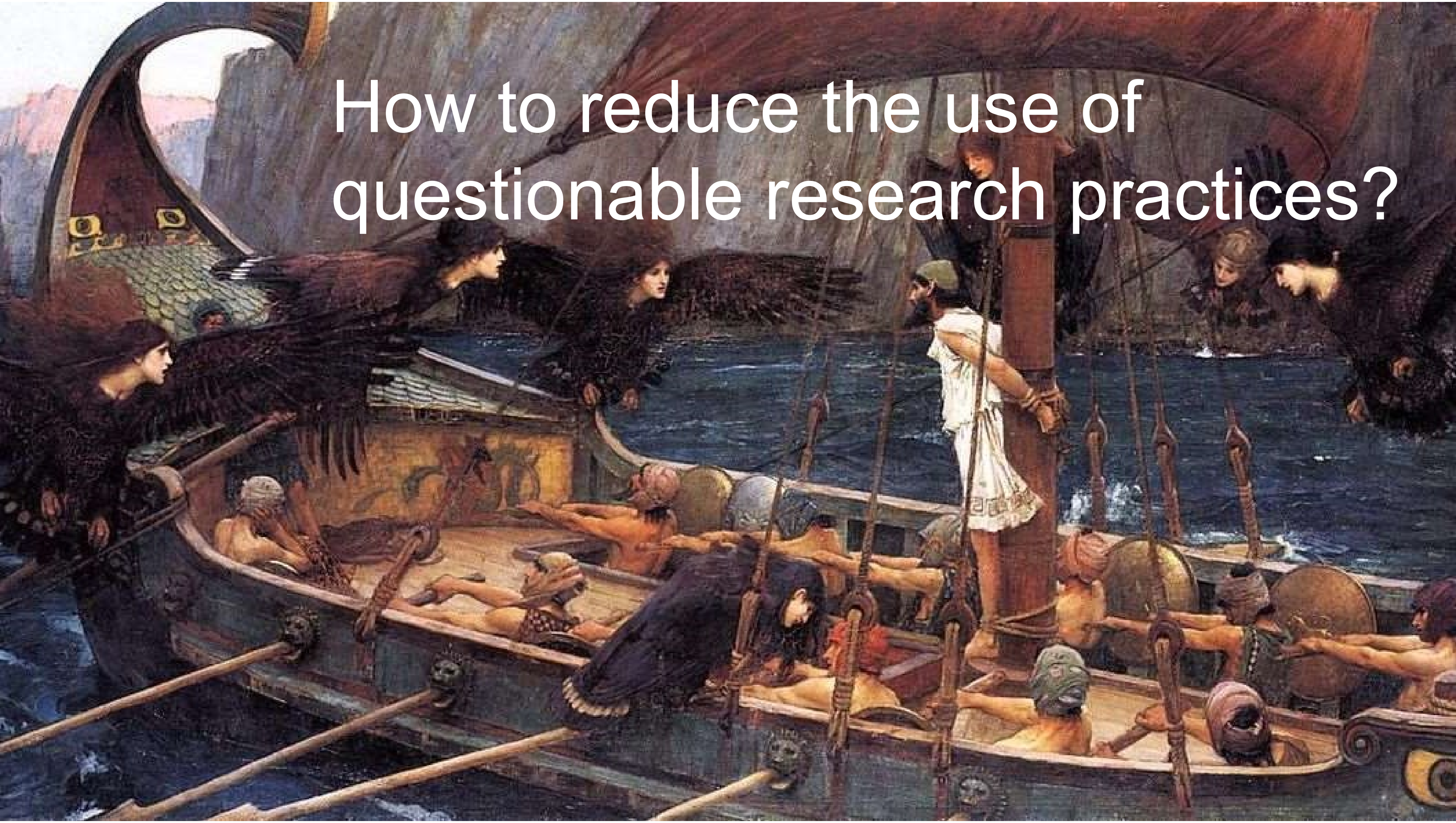


Probability of obtaining a false-positive result (p < .05)

Researcher degree of freedom	
Situation A: Having two dependent variables	9.5%
Situation B: Adding 10 more observations per condition (if with original sample size difference is non-significant)	7.7%
Situation C: Controlling for a third (binary) variable (e.g., gender)	11.7%
Situation D: Having an additional condition and dropping one if it yields nonsignificant results	12.6%
Combine A and B	14.4%
Combine A, B, and C	30.9%
Combine A, B, C, and D	60.7%

Simmons, Nelson, & Simonsohn (2011)

How to reduce the use of questionable research practices?



Preregistration

Specification and public documentation of key methodological decisions prior to data collection

- Research question(s)
- Hypotheses
- Variables (dependent, independent) and how they are connected to the research question(s) or hypotheses
- Study design
- Data collection plan (how participants are recruitment)
- Sample size (e.g., based on a power analysis)
- Statistical models for addressing research question(s)/hypotheses (incl. data transformations)
- Exclusion and inclusion criteria

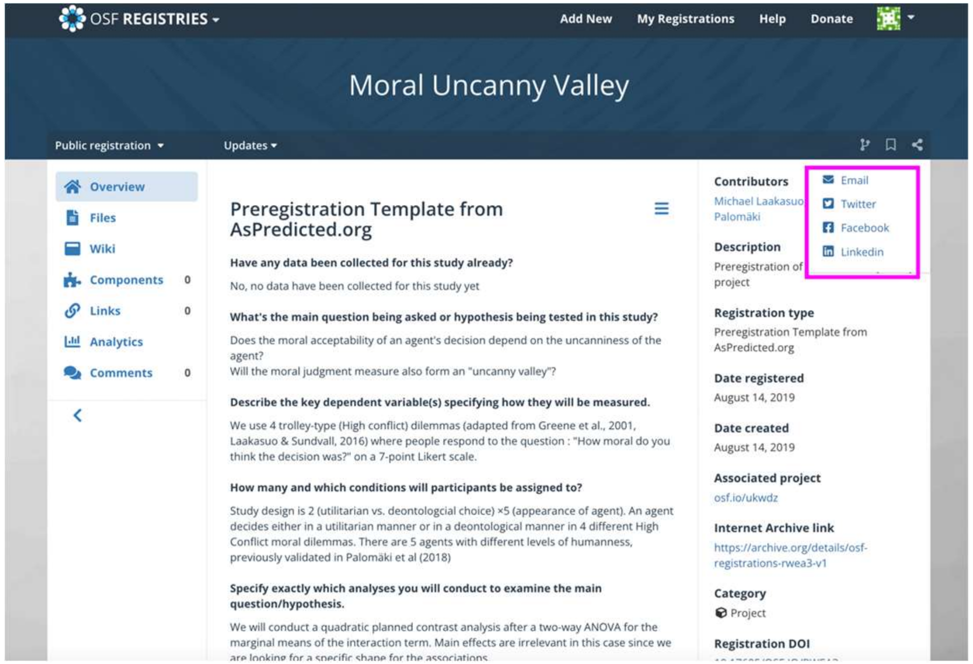
→ Once submitted, the research plan is time-stamped and can no longer be modified

Preregistration

Repositories for submitting a preregistration

ASPREDICTED

OSF template



Reporting a preregistered study

- Report the results of all preregistered analyses, regardless of the outcome
- Deviations from the preregistration are possible
  - Running different analyses
    → Report the change transparently and provide a justification
  - Running additional (i.e., unpreregistered) analyses, conducting unplanned data transformation (if data turn out to show a skewed distribution)
    → Transparently label them as “unregistered” or “exploratory”

Homework for Exercise this week

Watch this brief tutorial (< 9 min) on preparing a preregistration on ASPREDICTED

https://www.youtube.com/watch?v=7ic_BgpmakI

Open data

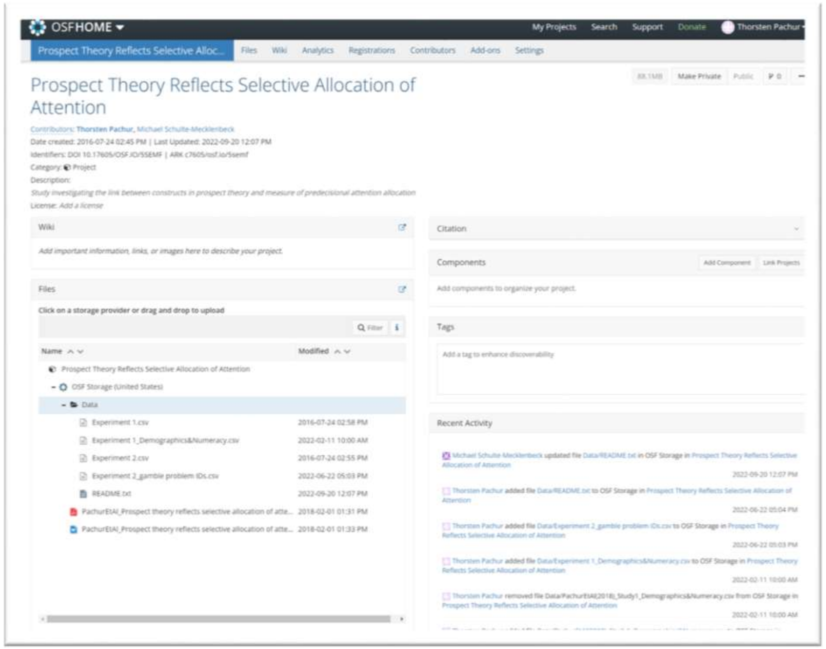
Open Science Framework (OSF)

https://osf.io/

→ Center for Open Science

CENTER FOR OPEN SCIENCE

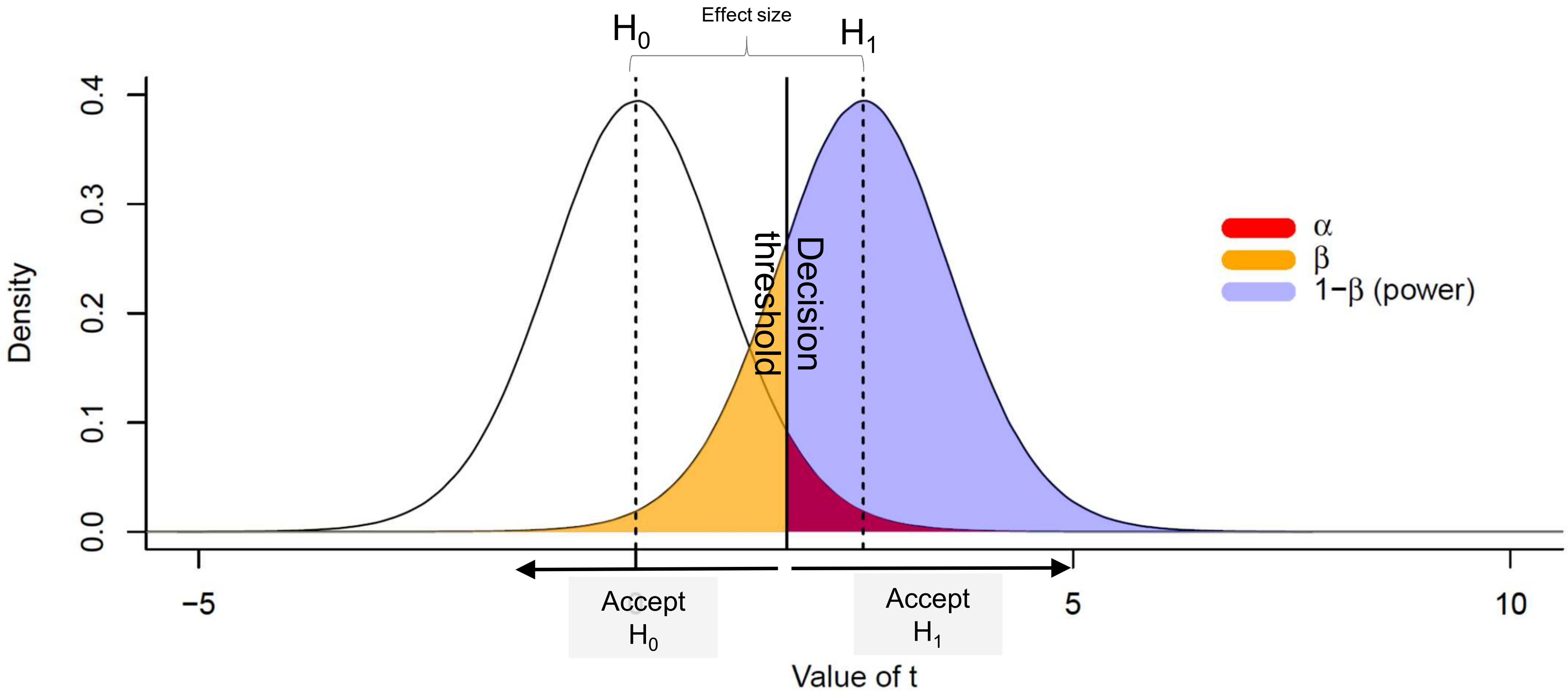
- Storage of
  - Materials
  - Preregistration file
  - Data, analysis code



Brian Nosek

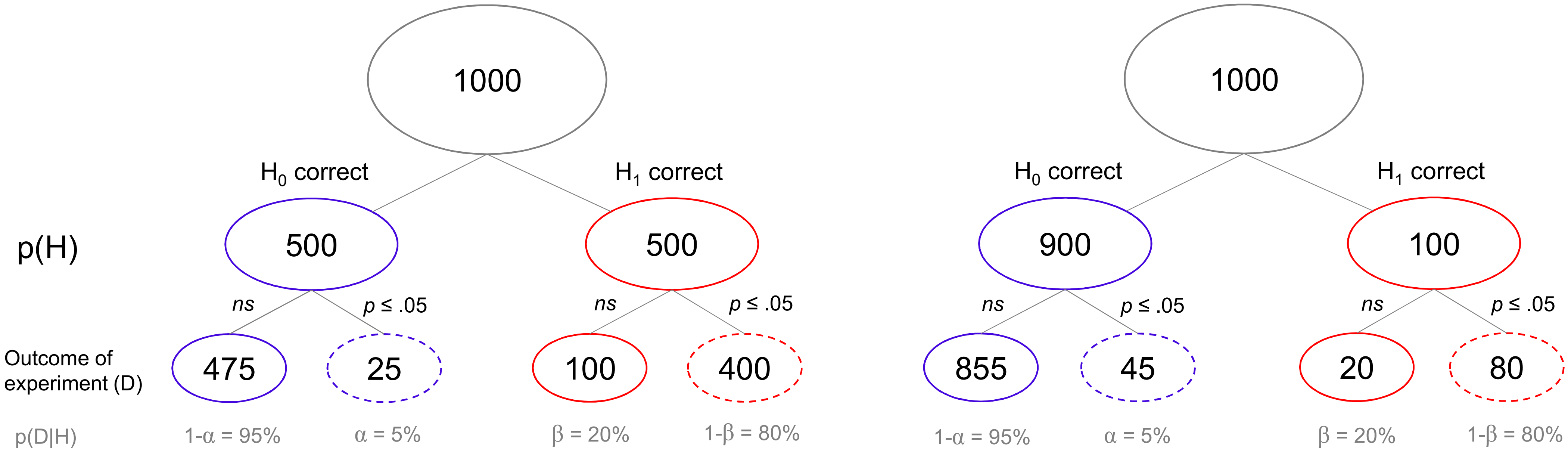
Toward better data analysis

Statistical inference



H_0 and H_1 are ***equally likely*** a priori

H_0 is ***less likely*** than H_1 a priori



How often ist H_1 correct when $p \leq .05$?
→ $p(H_1|D)$

$$\frac{400}{25 + 400} = .94$$

$$\frac{80}{45 + 80} = .64$$

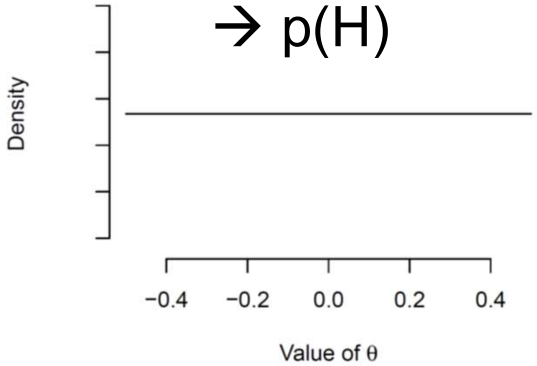
→ How strongly we can trust our results depends on the priors of the hypotheses!

Bayesian data analysis

→ How probable is a given **hypothesis**?

Prior distribution

→ p(H)



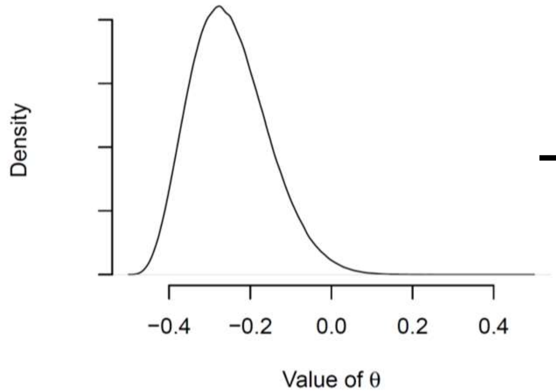
updating

Data

$$p(\theta|D) = \frac{p(D|\theta) \times p(\theta)}{p(D)}$$

Bayes' rule

Posterior distribution



→ p(H|D)

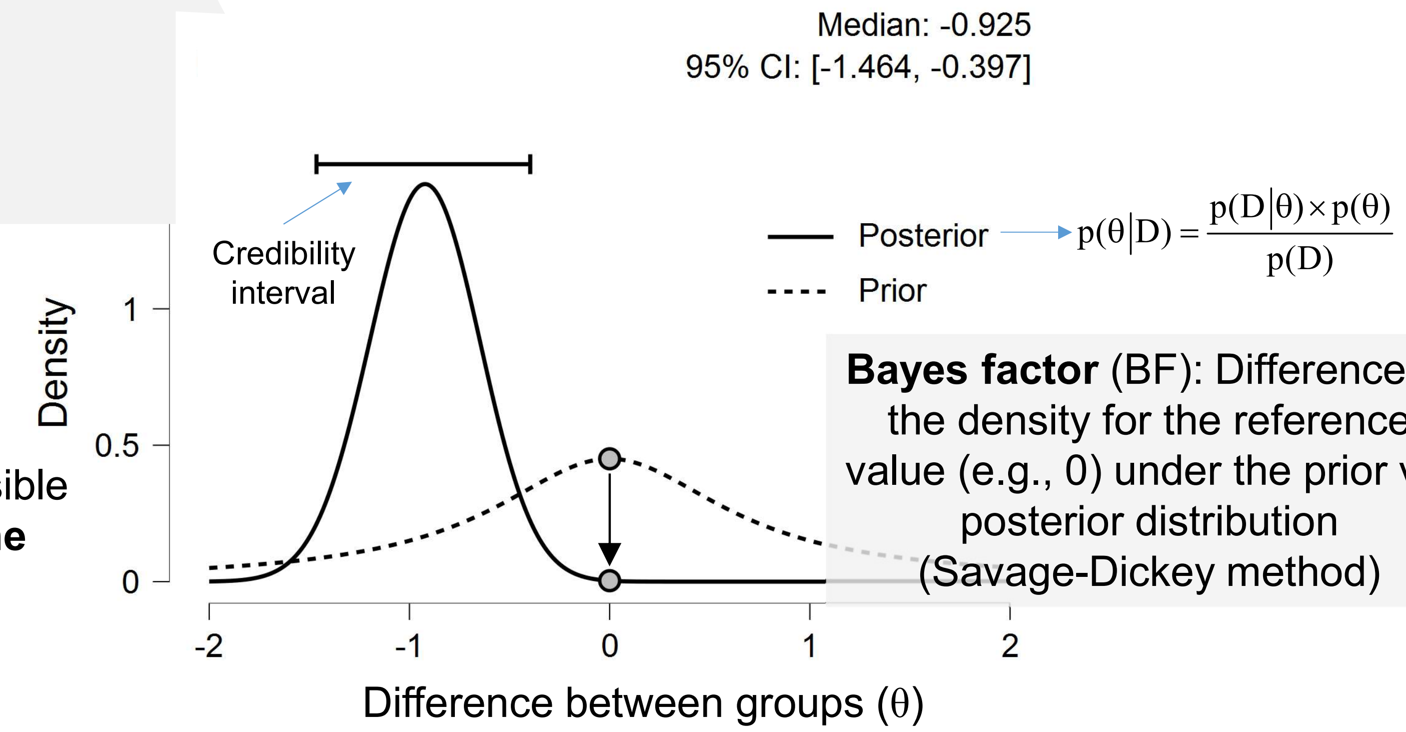
Hypothesis = value of θ

Comparing two means: Bayesian t-test

BF_{10}: Evidence for H_1 relative to evidence for H_0
BF_{01}: Evidence for H_0 relative to evidence for H_1

$$BF_{01} = \frac{1}{BF_{10}}$$

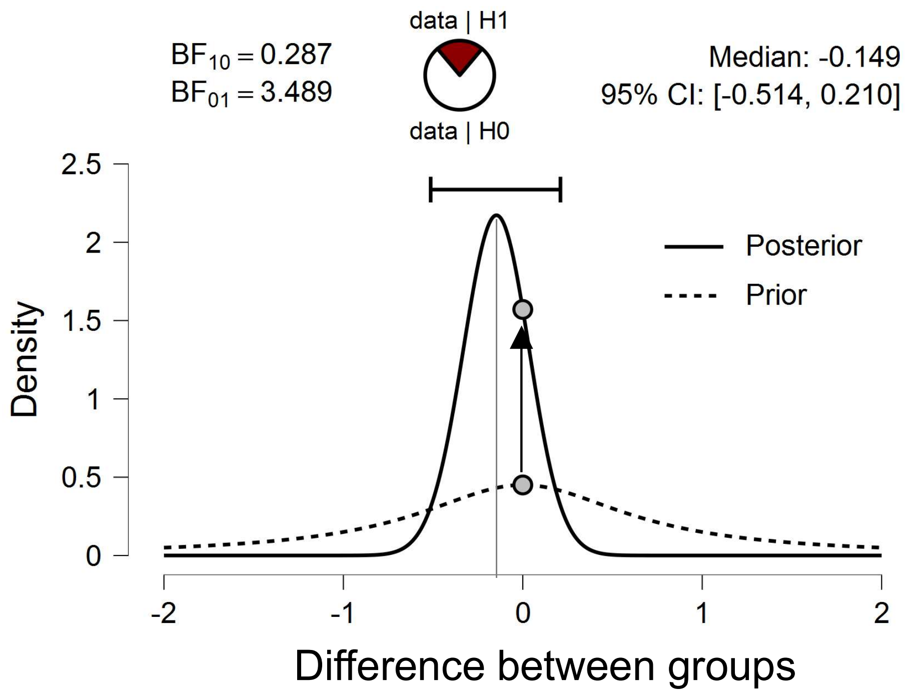
$$\frac{p(H_1|D)}{p(H_0|D)} = \underbrace{\frac{p(D|H_1)}{p(D|H_0)}}_{BF_{10}} \times \frac{p(H_1)}{p(H_0)}$$



Bayes factor (BF): Difference of the density for the reference value (e.g., 0) under the prior vs. posterior distribution (Savage-Dickey method)

With the Bayes factor it is possible to quantify the **evidence for the null hypothesis**!

Comparing two means: Bayesian t-test



How to interpret a Bayes factor?

Bayes factor (BF_{10})	Interpretation
>100	Decisive evidence for H_1
30–100	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Substantial evidence for H_1
1–3	Anecdotal evidence for H_1
1	No evidence
1/3–1	Anecdotal evidence for H_0
1/10–1/3	Substantial evidence for H_0
1/30–1/10	Strong evidence for H_0
1/100–1/30	Very strong evidence for H_0
<1/100	Decisive evidence for H_0

Data in favor of alternative hypothesis

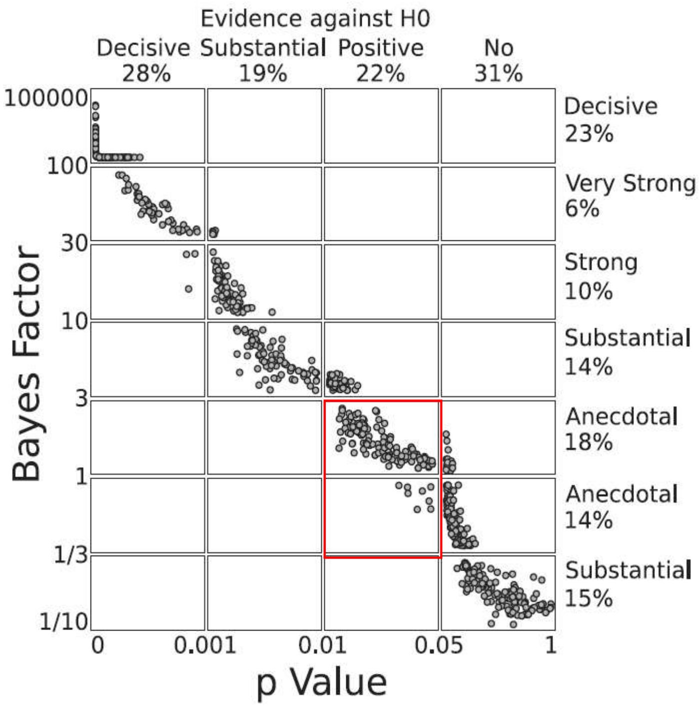
No inference on either hypothesis possible (→ collect more data!)

Data in favor of null hypothesis

Jeffreys (1961)

Do not trust p-values

Example: 855 *t* tests in experimental psychology



70% chance of only anecdotal evidence

Wetzels et al. (2011)

Self-quiz questions

- What is meant by the replication crisis?
- What is meant by HARKing—and why is it problematic?
- Describe research practices that represent p-hacking
- Why might p-hacking have contributed to the replication crisis?
- Which aspects of a research project are specified in a preregistration?
- What is the logic of Bayesian data analysis?
- What does a Bayes factor represent? Give an advantage of Bayes factors over p-values in statistical inference.

If you wanto to learn more about the topics…

Coming up in SS 2027

- Module "*Open Science: Best Practices in Empirical Research*" (Thorsten Pachur)
- Module "*Introduction to Bayesian Data Analysis*" (Linus Hof)