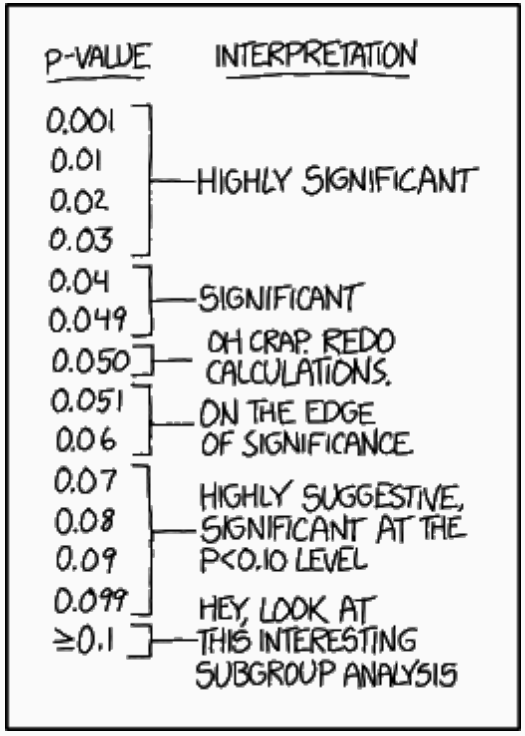
Flexible Data Collection

How do I lie with flexible data collection?

Lennart Stipulkowski



Source: https://xkcd.com/1478/

Seminar: „How do I lie with statistics?“
Supervisor: Prof. Dr. Ulrich Köthe

November 14, 2019
Heidelberg University

Table of contents

Quote

A reader quick, keen, and leery
Did wonder, ponder, and query
When results clean and tight
Fit predictions just right
If the data preceded the theory

- Anonymous

The Problem

Problem

- Main Problem: False-Postive rate
  - Author finds evidence for an effect that does not exist
  - Incorrect rejection of the null hypothesis
- Few revocations of false-positive findings → persist in literature
- Field/Scientists/Journal loses credibility if exposed
- It is unusual to publish null findings
  - Incentive to publish findings with high level of „significance“

Problem

- Despite stated significance of $p < .05$ → higher false-positive rates are likely
    - **Reason:** Influence of data collection and analysis

Problem - Researchers Degrees of Freedom

Researchers Degrees of Freedom

- Amount of data to be collected
- Exclusion of observations
- Selection of combined conditions and which one to compare
- Which control variables?
- Combining measures
- Transforming measures

Problem - Question

Question is: Should/Could one do the decisions before data acquisition/analysis?

- Accepted and common practice to not decide beforehand
- Different alternatives are tested and optimised for the highest „statistically significance“
- Its likely one alternative leads to false positive findings $\geq .05$

Problem - Reasons

Ambiguity of these decisions

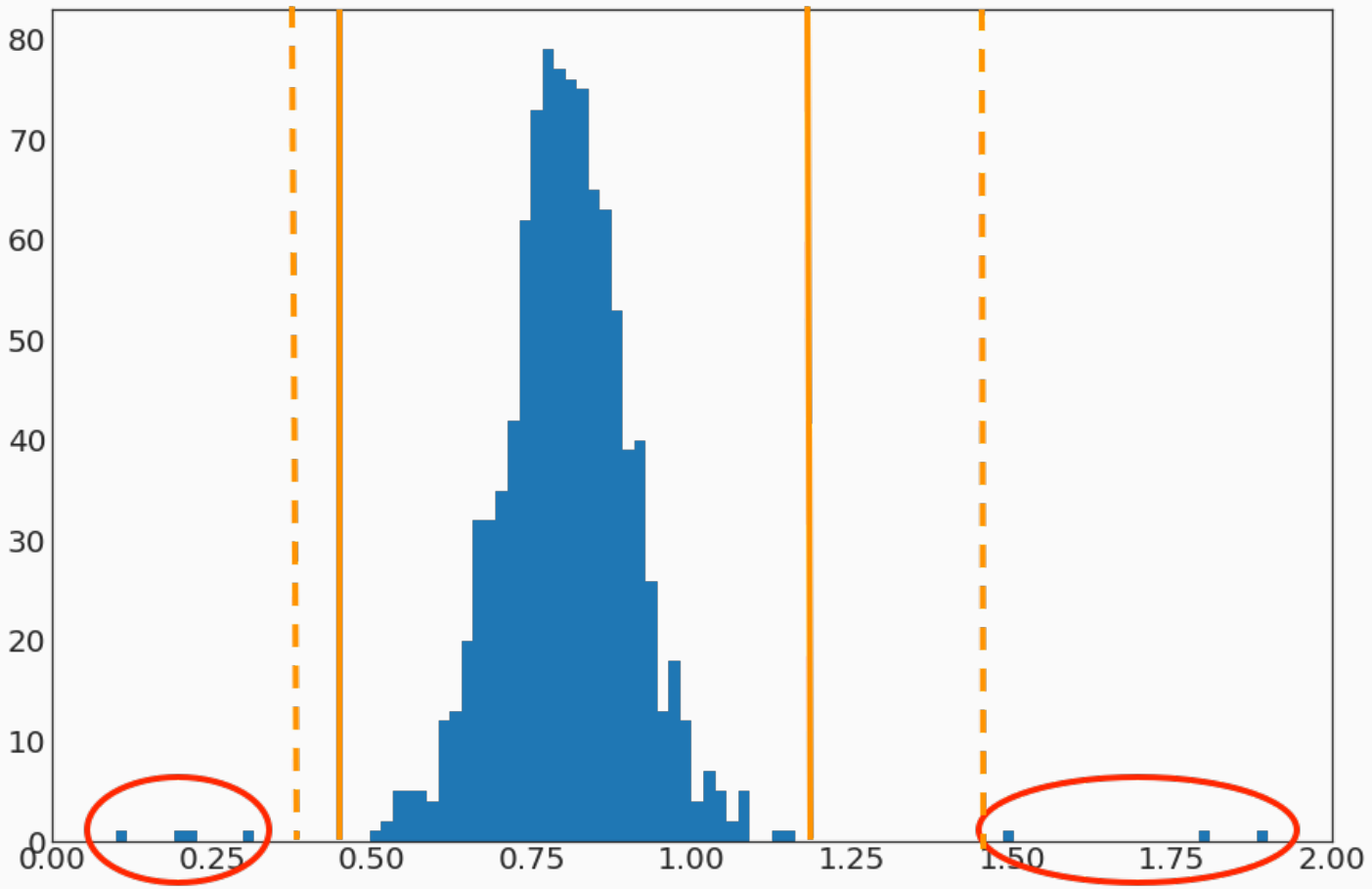
- Intention of the researcher to have the maximum statistical significance
- Ambiguous analytic questions → appropriate decisions are those with statistical significance (convincing self-justification)

Example: How to treat outliers?

Given a study measuring the reaction times of students.

- Researchers have to make a decision: How to treat the outliers (fast/slow reaction times)
- They often tend to decide in favor of high significance
- No common standard to comparable studies → problem of reproducibility

Example: How to treat outliers?



HARKing

HARKing

Hypothesizing After the Results are Known (HARK) vs. Hypothetico-deductive (HD)

Hypothetico-deductive (HD)

- Deductive reasoning based on hypotheses prior the research

Hypothesizing After the Results are Known (HARK)

- Presenting post hoc hypothesis after the results are known
- Presenting like a priori hypothesis

Categorizing Hypotheses

	After Results Are Known	
Before the Study	**Plausible**	**Implausible**
Anticipated & Plausible	a	b
Anticipated & Implausible	c	d
Unanticipated	e	f

Table 1: Cross-Classification of Hypotheses by A Priori and Post Hoc Status [Kerr et al., 1998]

- The HD approach is classified as a or b

„How Bad Can It Be?“ Simulations

„How Bad Can It Be?“ Simulations [Simmons et al., 2011]

- Simulations of the common researcher degrees of freedom
- Four common degrees:
  - (a) choosing among dependent variables
  - (b) choosing sample size
  - (c) using covariates
  - (d) reporting subsets of experimental conditions
- According to a survey: 70% of asked behavioural scientists admitted a flexible sample size
  - Belief of a trivial influence on false-positive rate

„How Bad Can It Be?“ Simulations - Degrees

A: Two dependent variables (r=0.5)

- Variable 1
- Variable 2
- Average. Variable 1 + 2

→ one of three tests below significance level (T-Tests)

B: Addition of observations

- 20 Observations

→ test for significance

- 10 Observations

→ test for significance

„How Bad Can It Be?“ Simulations - Degrees

C: Controlling for gender or interaction of gender with treatment

- Each observation a gender is assigned

→ test for significance

- ANCOVA (analysis of covariance) to „reduce“ the effect of the gender on analyzed effect

→ test for significance

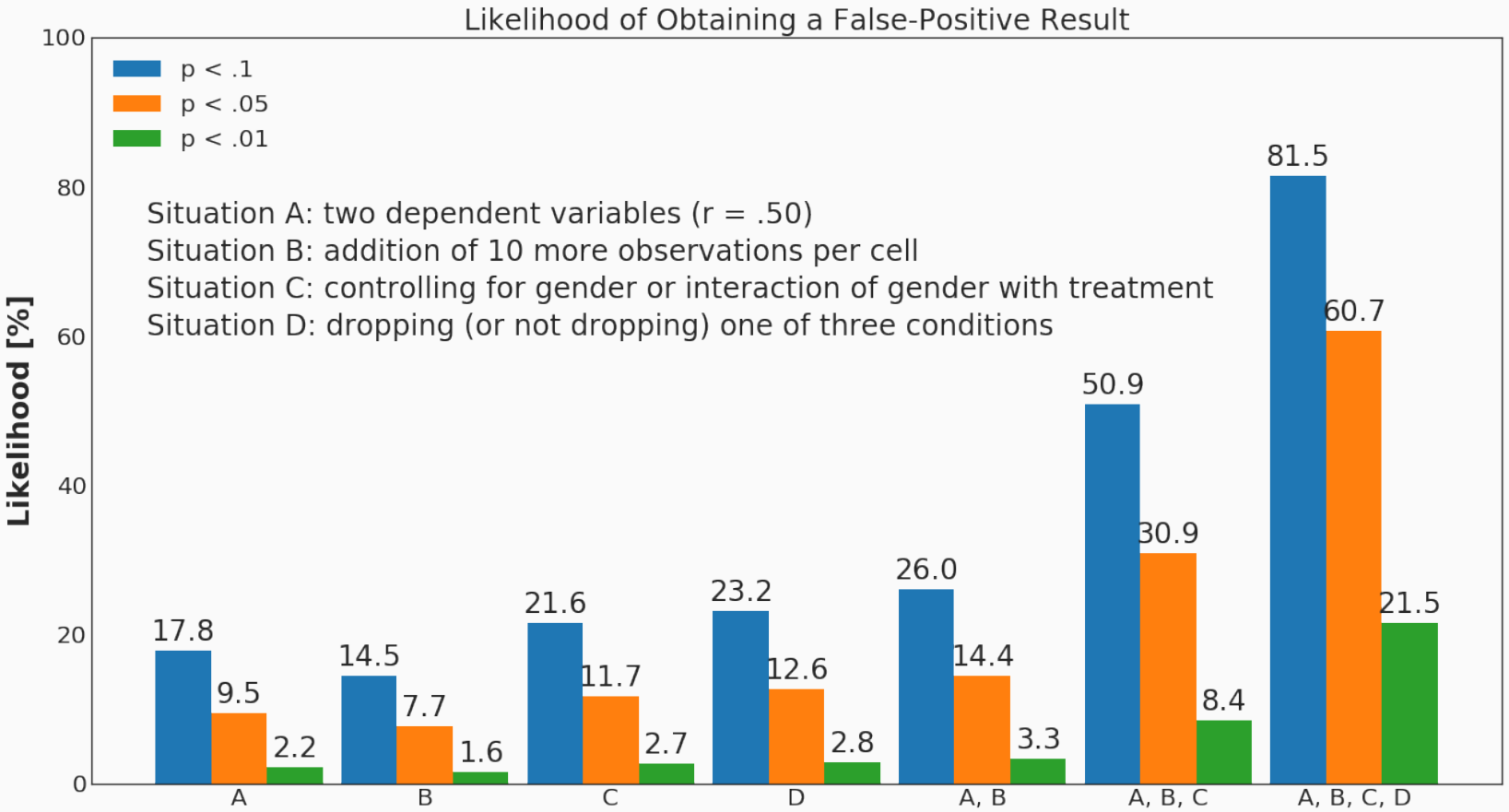
D: Dropping (or not dropping) one of three conditions

→ test for significance

- dropping one of the three conditions

→ test for significance (repeat for each condition dropped)

„How Bad Can It Be?“ Simulations - Results



Simulation: Continuously adding observations

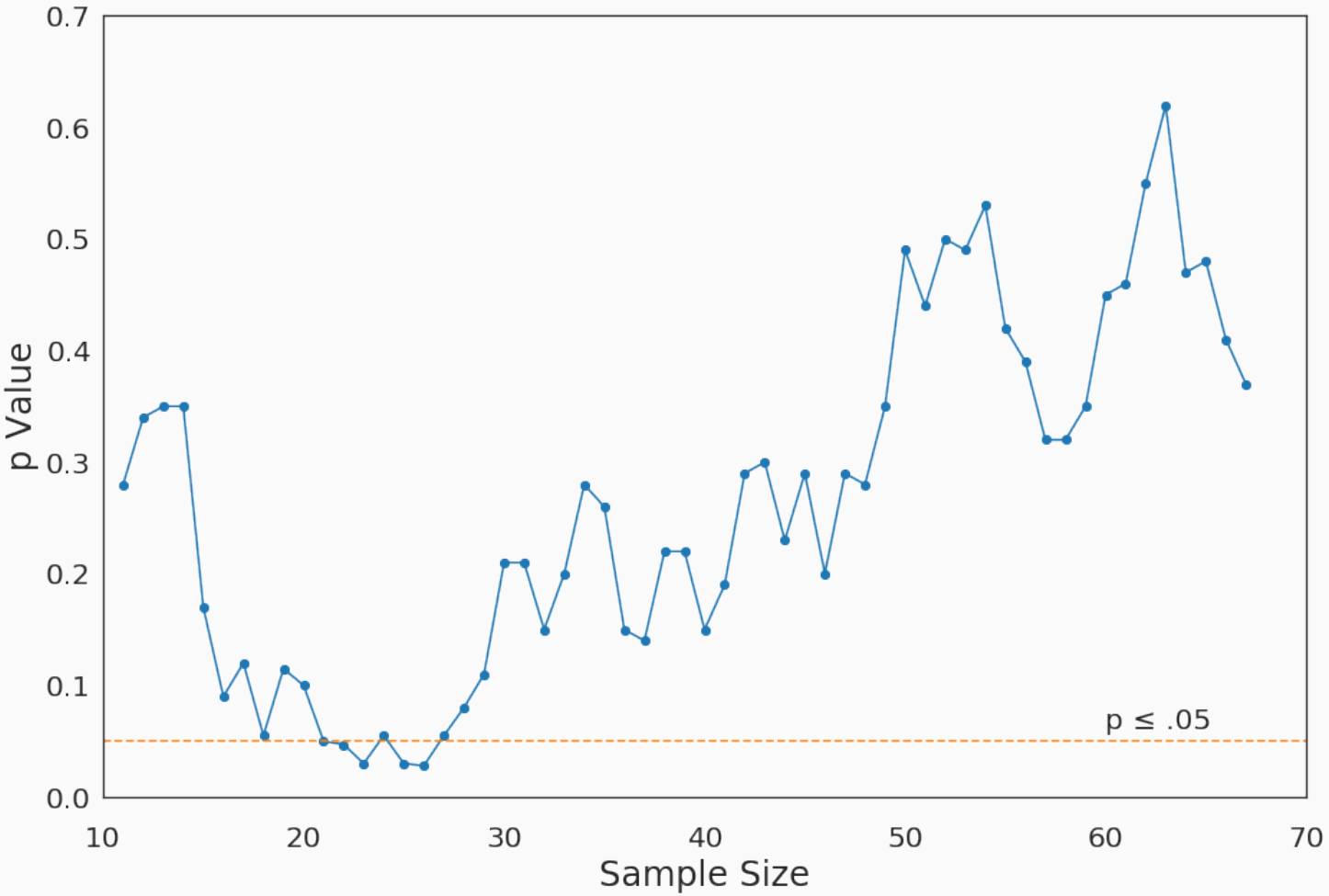
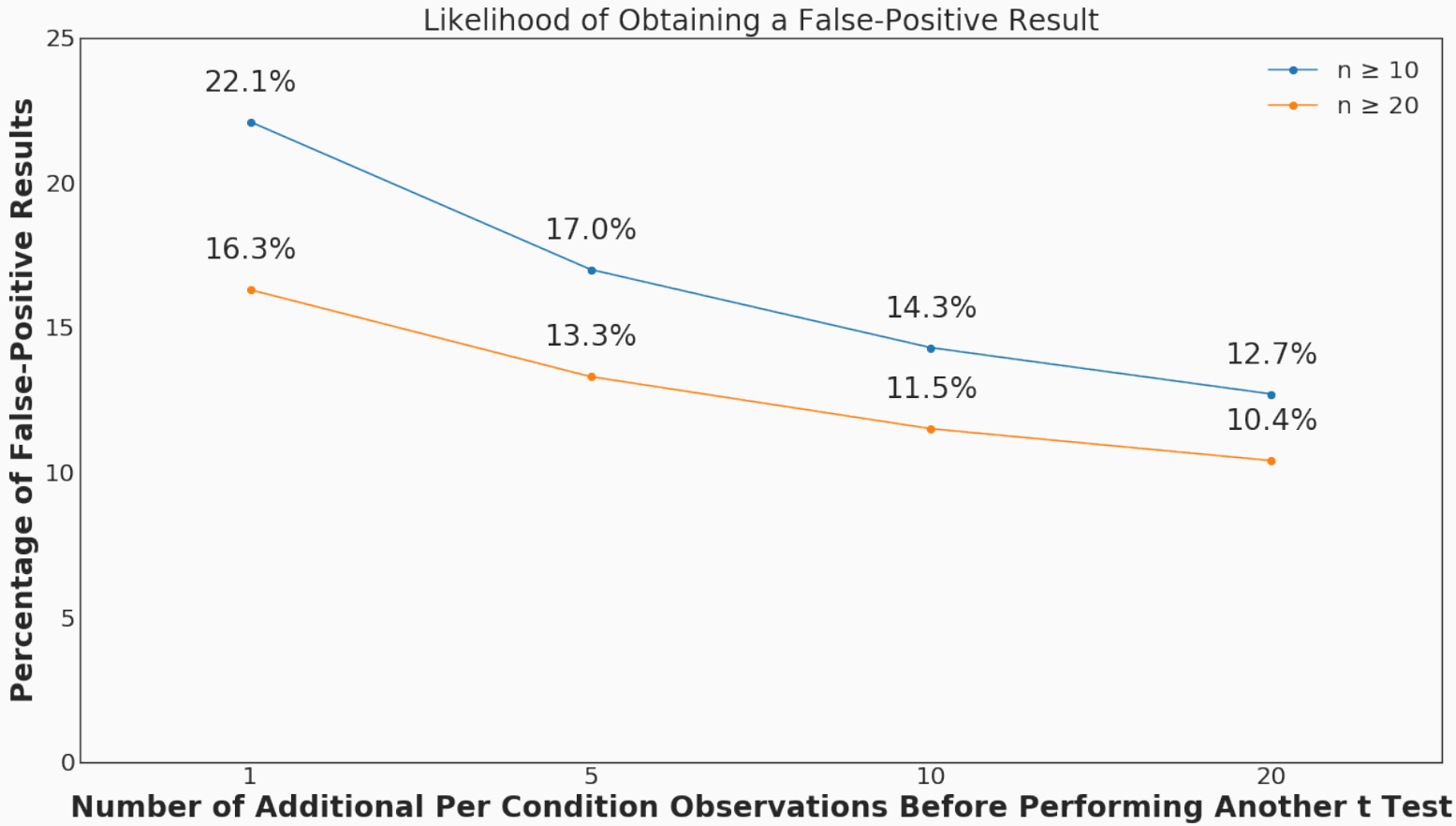


Figure 1: Continuously adding observation

„How Bad Can It Be?“ Simulations - Results



Analyzing P-Hacking (Approach 1)

P-Curve

P-Curve

Distribution of p-values of a given set of studies

- It can be used to determine the effects of p-hacking
- Mainly the effects of:
  - Selection bias / „file drawer effect“
  - Inflation bias / „p-hacking“

Example: XKCD Jelly Beans

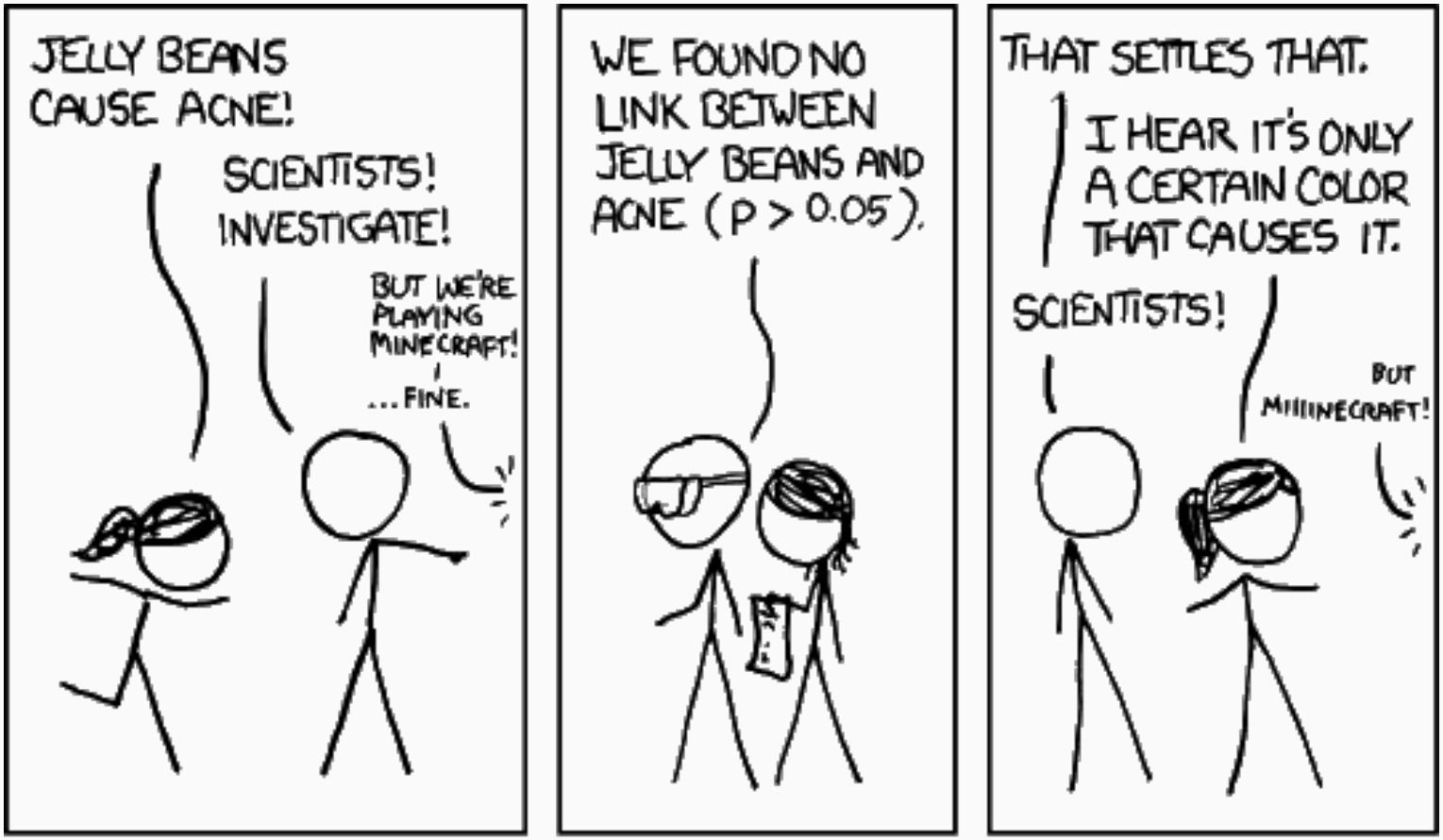


Figure 2: XKCD: https://xkcd.com/882/

Example: XKCD Jelly Beans

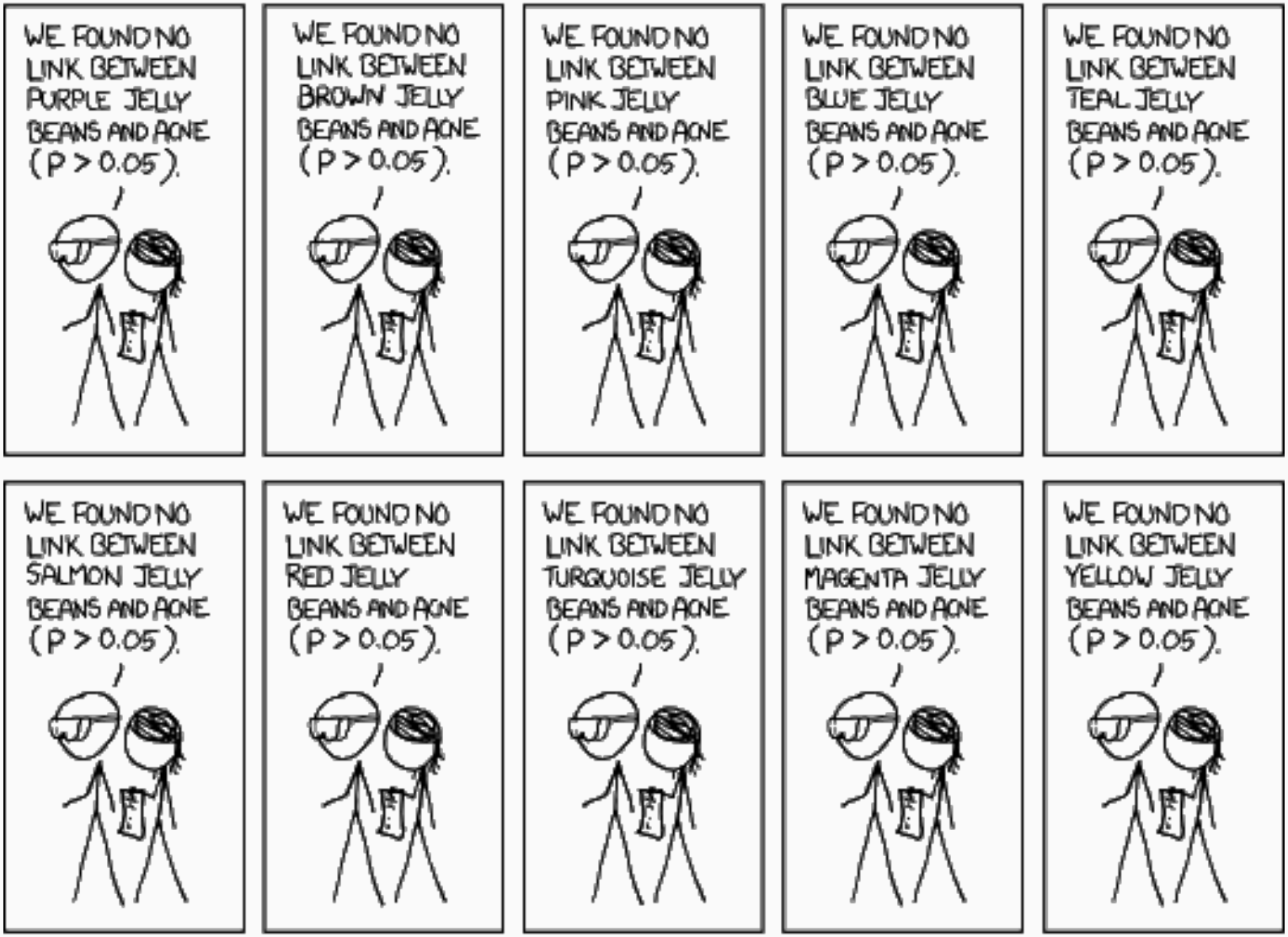


Figure 2: XKCD: https://xkcd.com/882/

Example: XKCD Jelly Beans

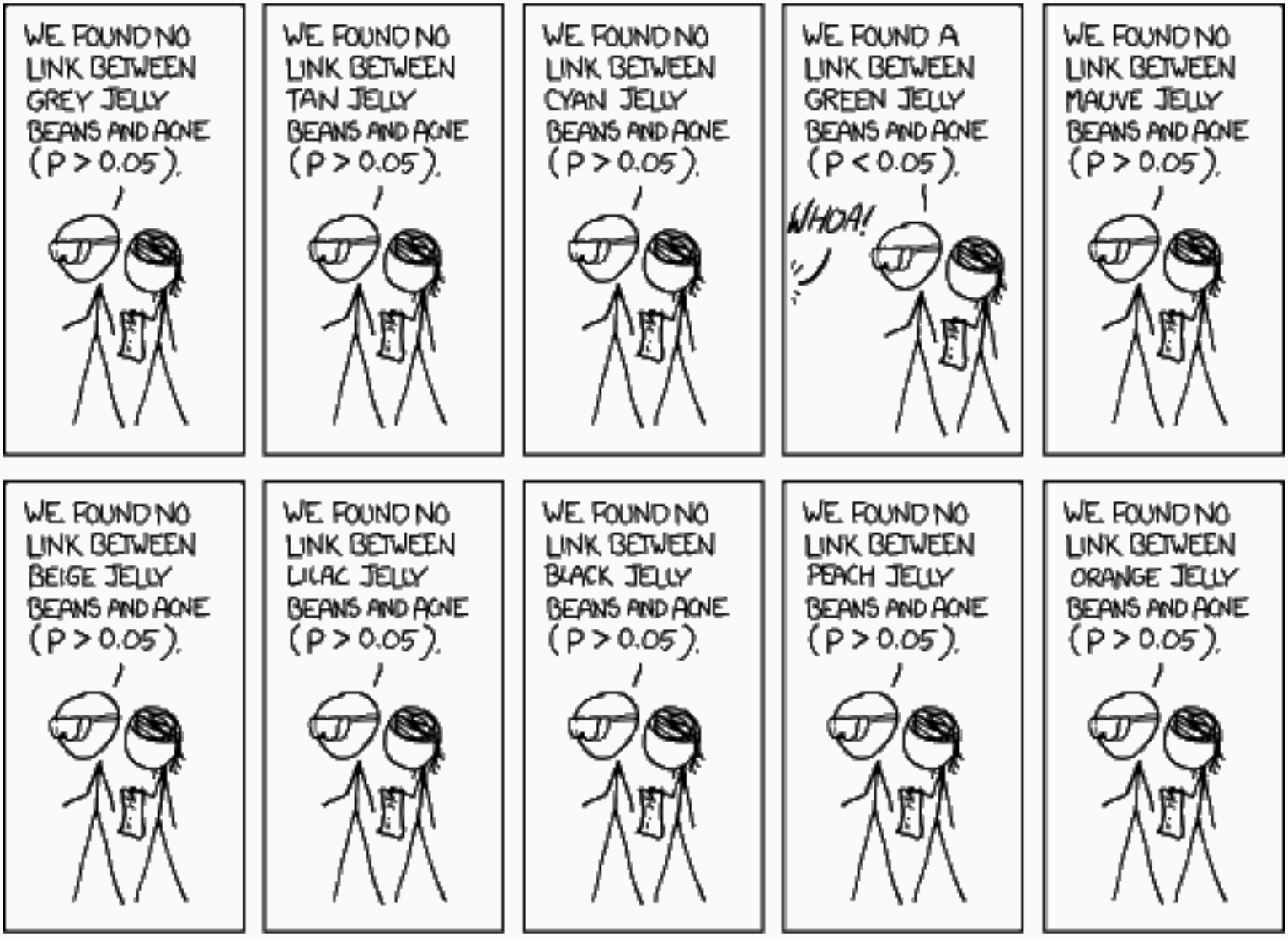


Figure 2: XKCD: https://xkcd.com/882/

Example: XKCD Jelly Beans

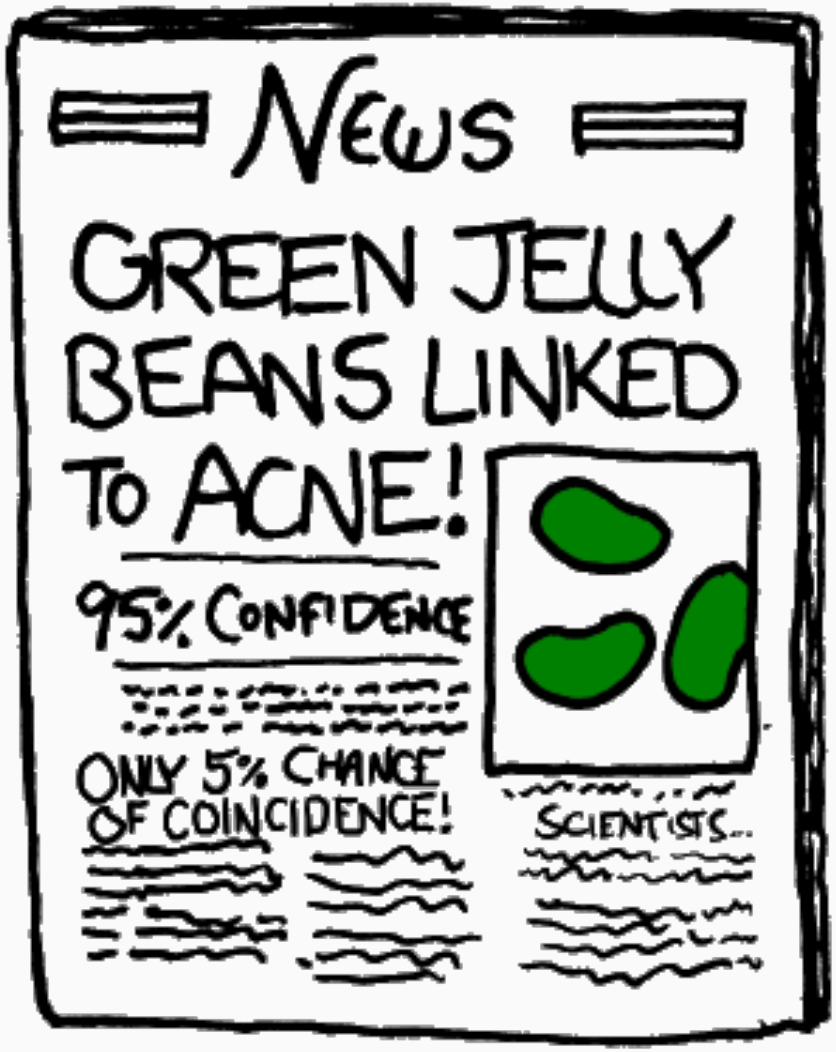


Figure 2: XKCD: https://xkcd.com/882/

P-Curve: Publication bias

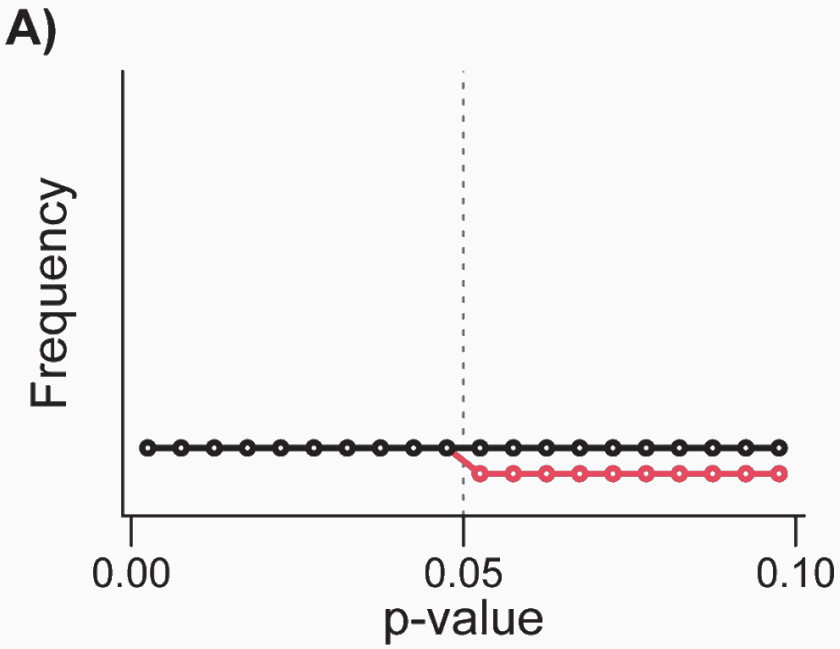


Figure 3: Publication bias / No evidental value [Head et al., 2015]

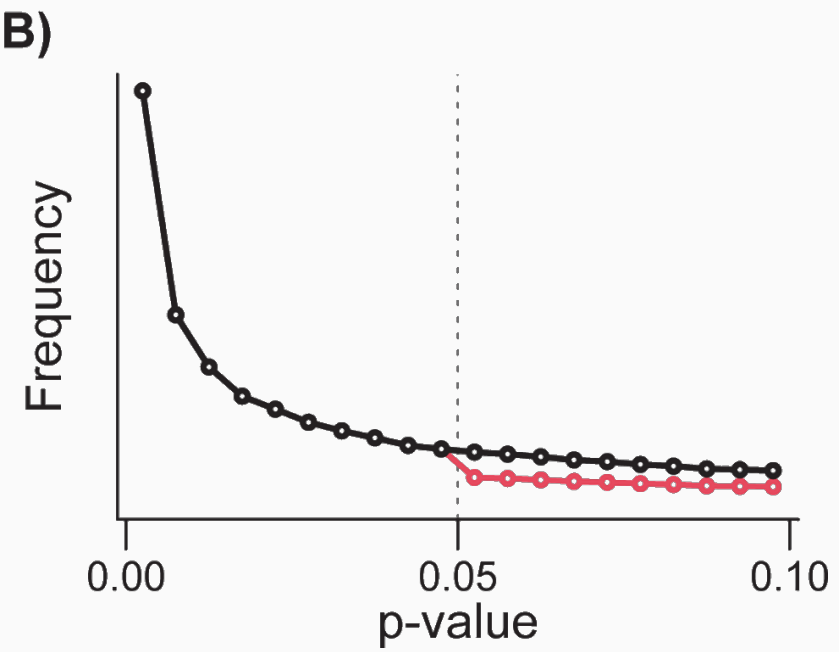


Figure 4: Publication bias / Evidental value > 0 [Head et al., 2015]

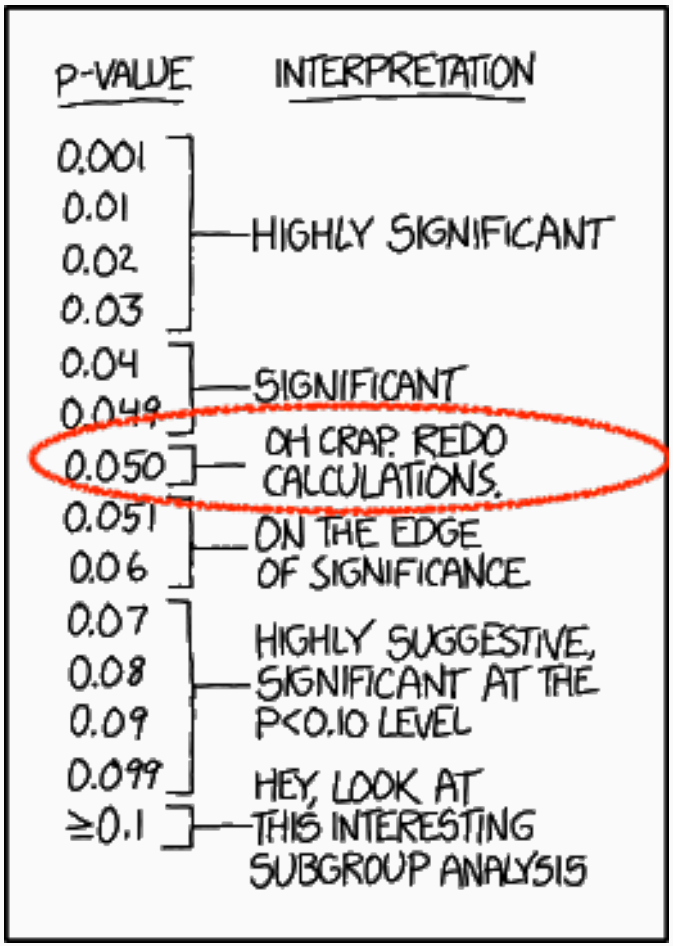


Figure 5: XKCD: https://xkcd.com/1478/

P-Curve: P-Hacking

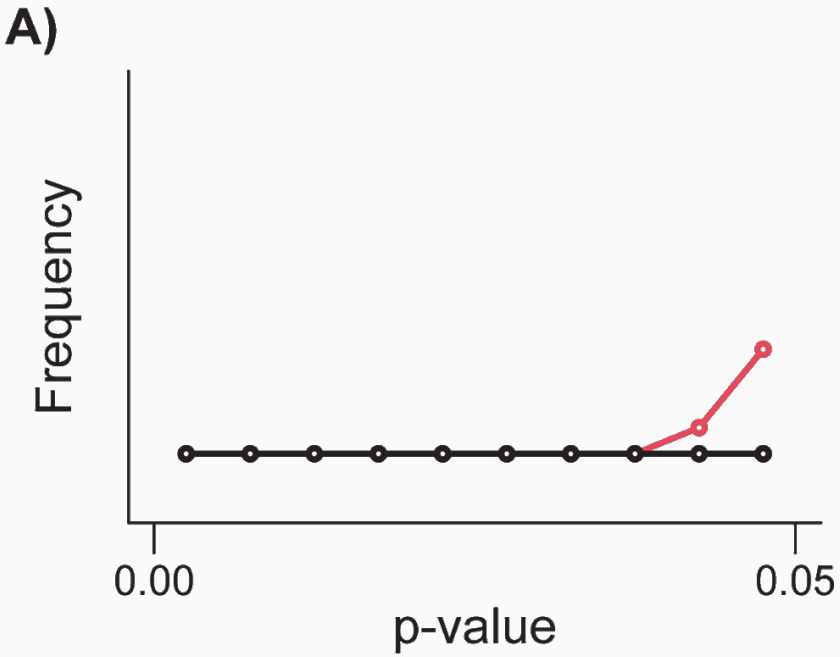


Figure 6: P-hacking / No evidental value [Head et al., 2015]

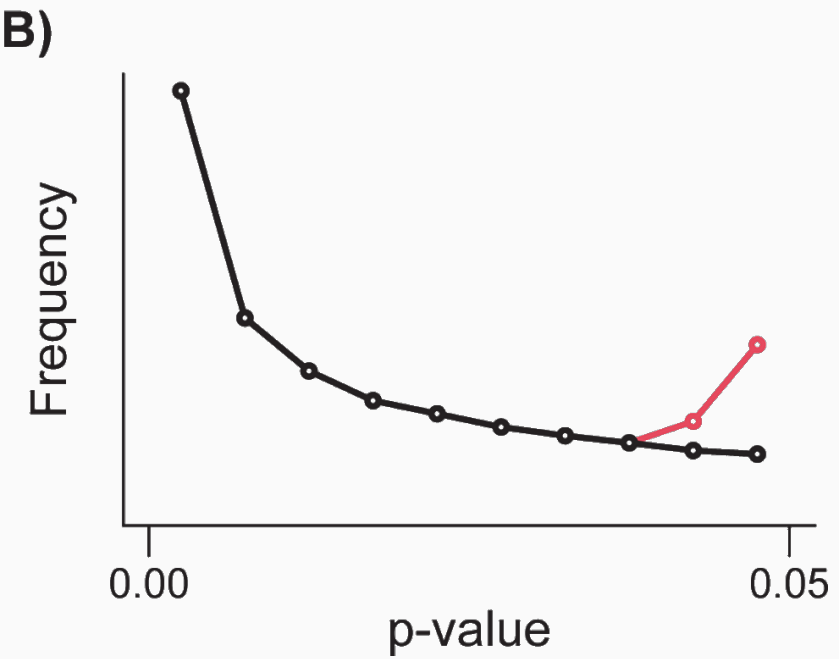


Figure 7: P-hacking / Evidental value > 0 [Head et al., 2015]

Analyzing P-Hacking (Approach 2)

Example Scenario [Shun-Shin and Francis, 2013]

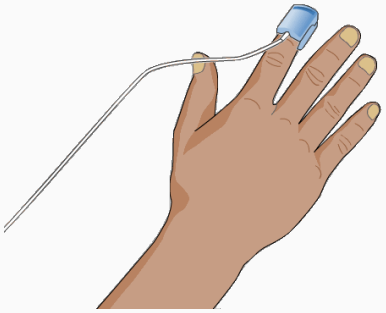


Figure 8: Pulse oxymetry[1]

- Student nurse is about to document a oxygen saturation by pulse oximetry of 85%
- Patient is ambulant, looking pink and feeling well
- All previous values $\geq 97\%$´

Do you:

a) Immeadiately confine to bed, initiate 100% oxygen
b) Document 85% and request tests for possible pulmonary emobolism
c) Remeasure the oxygen saturation yourself, document the new value

[1] Source: Royal College of Nursing, URL: http://rcnhca.org.uk/clinical-skills/observation/oxygen-levels/, November 9, 2019

Effects of Remeasurement, Removal, Reclassification

D'Agostino z-score

$$z \leftarrow \begin{cases} \frac{g(y)-g(x)}{\sqrt{2}} & \text{if } \bar{x} \leq \bar{y} \\ \frac{g(x)-g(y)}{\sqrt{2}} & \text{if } \bar{x} \geq \bar{y} \end{cases}$$

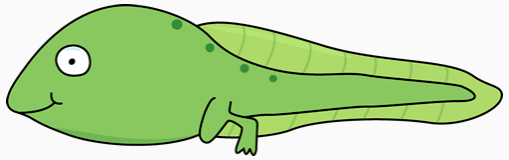


Figure 9: A tadpole

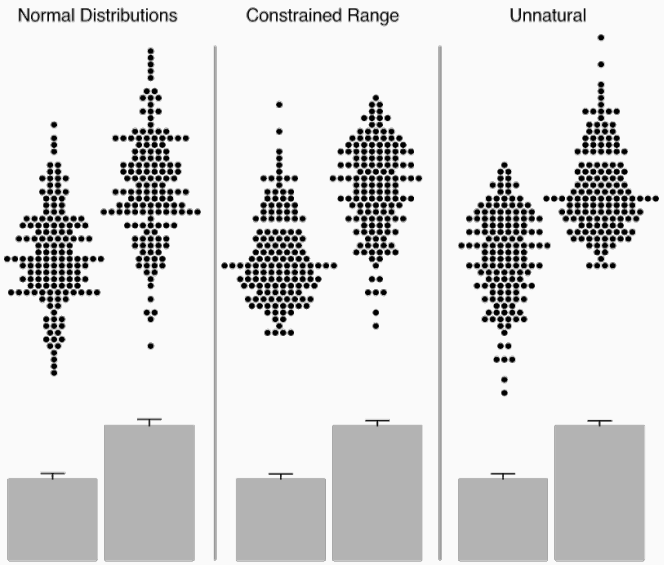


Figure 10: Natural/Unnatural distributions
[Shun-Shin and Francis, 2013]

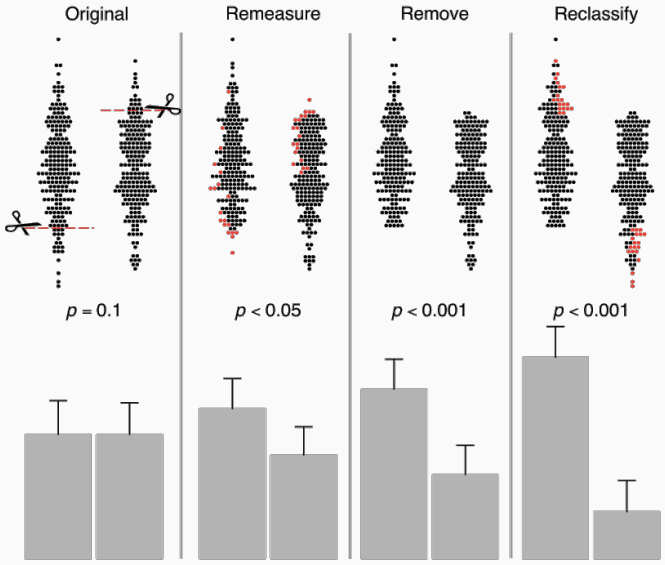


Figure 11: Manipulation of the distribution
[Shun-Shin and Francis, 2013]

Reaching Significance by Remeasurement, Removal, Reclassification

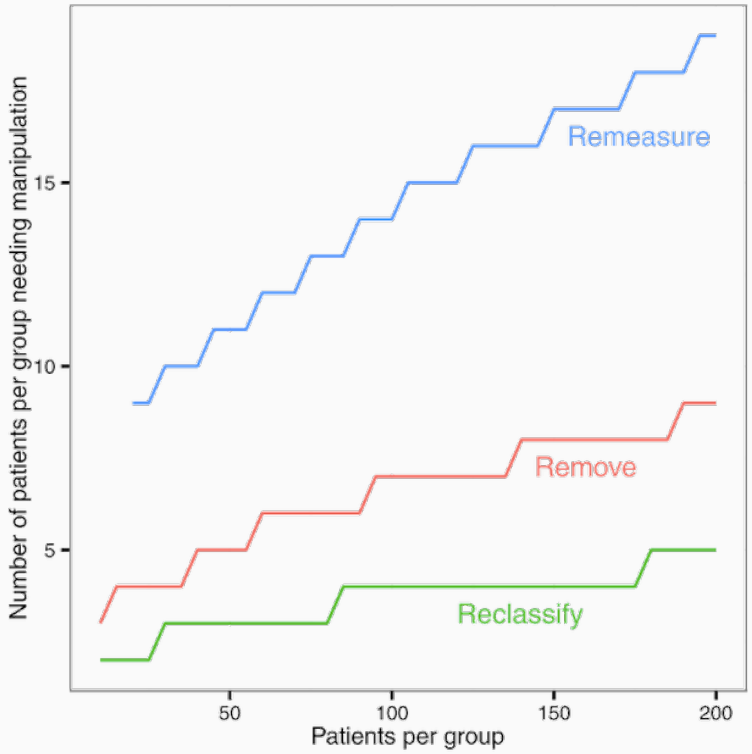


Figure 12: Effects of remeasurement, removal, reclassification on significance
[Shun-Shin and Francis, 2013]

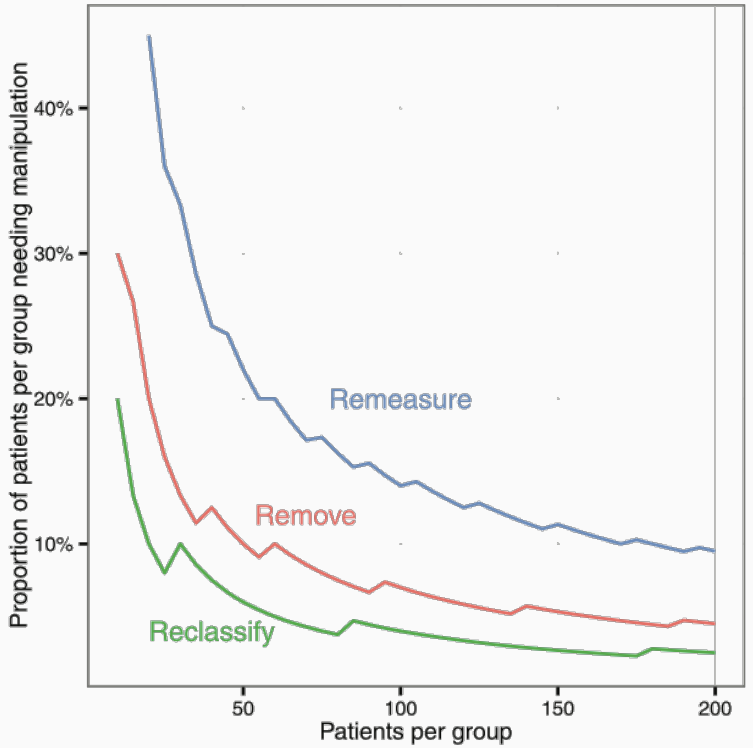


Figure 13: Effects of remeasurement, removal, reclassification on significance
[Shun-Shin and Francis, 2013]

Solutions

Solutions - Rules for the Authors

According to Simmons et. al. [Simmons et al., 2011]

1. Rule for terminating data collection prior collecting
2. Enough observations per cell
3. List all variables collected
4. Report all experimental conditions (e.g. failed manipulations)
5. Report the statistical results if no observations would be excluded
6. If analysis includes covariate → report of the results without covariate

Solutions - Rules for the Reviewers

According to Simmons et. al. [Simmons et al., 2011]

1. Author should follow the authors requirements
2. Tolerance of imperfections in results
3. Require authors to report their analytic decisions
4. If justification of data-collection or analysis are not compelling
   → require authors to conduct exact replication

Registered Reports (Center for Open Science - cos.io)

- Currently used by 210 journals (2019)
- → Peer-review before results are known

Figure 14: Registered Reports process[1]

[1]Source: Center for Open Science, URL: https://cos.io/rr/, November 10, 2019

Registered Reports (Center for Open Science - cos.io)

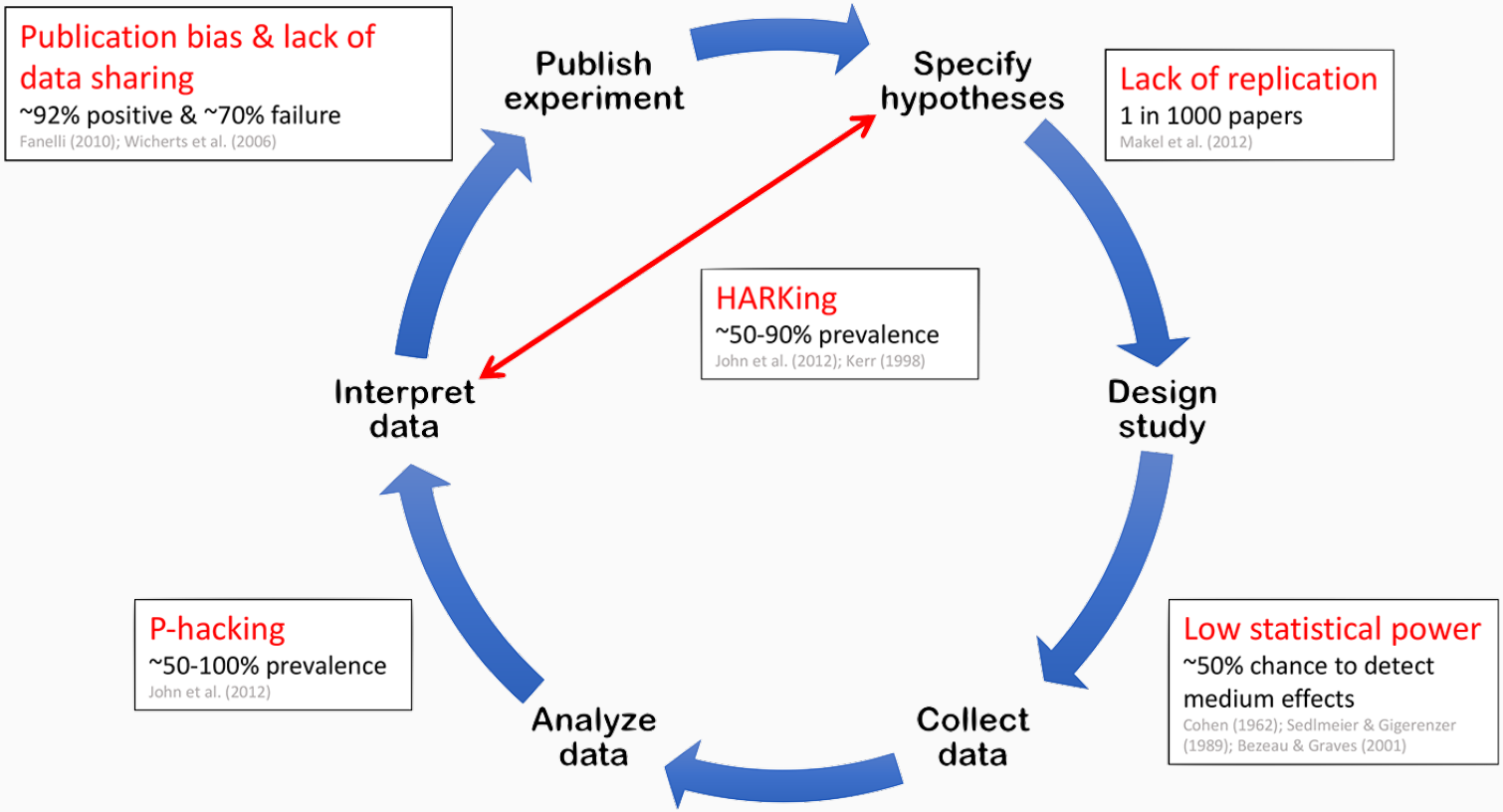


Figure 15: Registered Reports [2]

[2] Source: Center for Open Science, URL: https://cos.io/rr/, November 10, 2019

References i


Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015).
The Extent and Consequences of P-Hacking in Science.
PLoS Biology, 13(3).


Kerr, N. L., Adamopoulos, S., Fuller, T., Greenwald, S., Kiesler, P., Laughlin, D., and Mcglynn, A. (1998).
HARKing: Hypothesizing After the Results are Known.
Technical Report 3.


Shun-Shin, M. J. and Francis, D. P. (2013).
Why Even More Clinical Research Studies May Be False: Effect of Asymmetrical Handling of Clinically Unexpected Values.
PLoS ONE, 8(6).

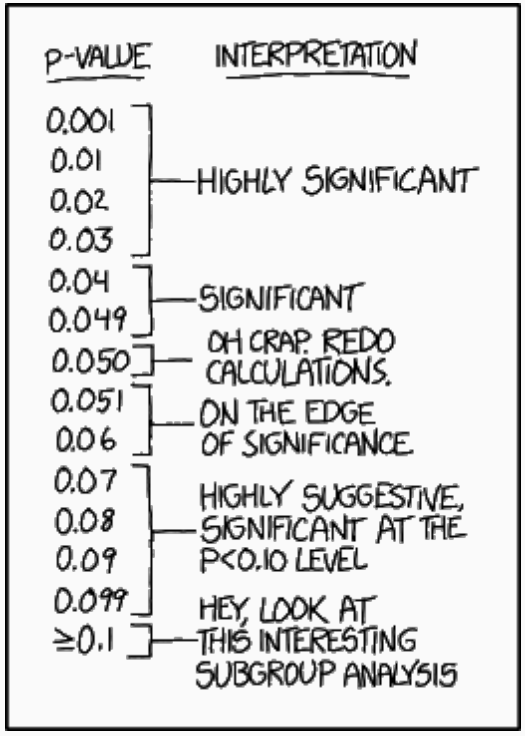
References ii

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011).
False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.
Psychological Science, 22(11):1359–1366.

Flexible Data Collection

How do I lie with flexible data collection?

Lennart Stipulkowski



Source: https://xkcd.com/1478/

Seminar: „How do I lie with statistics?“
Supervisor: Prof. Dr. Ulrich Köthe

November 14, 2019
Heidelberg University