Statistics 2

Good statistics, bad statistics

Casper Albers & Jorge Tendeiro

Lecture 13, 2019 – 2020

Overview

Introduction

Questionable research practices

Example

Replication

Solutions

Literature for this lecture

Read:

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2001). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*, 1359-1366. doi:10.1177/0956797611417632

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*, 524-532. doi:10.1177/0956797611430953

Goals of this course

Main goal:

$\boxed{\text{To learn how to do statistics properly}}$

Secondary goal:

$\boxed{\text{Recognizing flawed reasoning and statistical cheating}}$

Why?

- Statistics is hard. Flawed reasoning is omnipresent.
- Scientists are humans. Cheating is omnipresent.
- Obviously, such mistakes bias the results of studies.
- Thus, important to be able to separate the good, the bad, and the ugly.

Flawed reasoning

Recall from Lecture 3 some forms of flawed statistical reasoning:

- Confusing correlation with causality.
  Example: Films with Nicholas Cage and pool drownings
- Forgetting to include ‘hidden moderators’.
  Example: Strong correlations between ice cream sales and sea drownings.
  Hidden moderator: Whether it’s a summer’s day or not.
- Coincidence.
  Your result could just be a lucky hit → replication is important.

But there’s more. . .

Questionable Research Practices (QRP)

- John, Loewenstein, Prelec (2012). Psychological Science, 23(5).
- A Questionable Research Practice is following a wrong statistical/methodological procedure which leads to biased results.
- This is not necessarily intentional; could be due to ignorance.
- It's not just good or bad, it's a grey area.
- QRPs are prevalent in empirical research.

(There are also non-statistical QRPs, such as not crediting coworkers and plagiarism.)

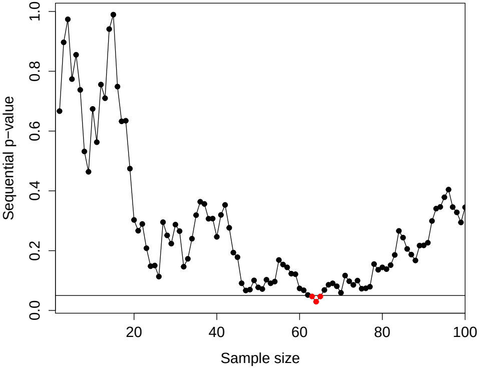
Questionable Research Practices (QRP)

Anonymous survey by John, Loewenstein, Prelec (2012). Psychological Science, 23(5).

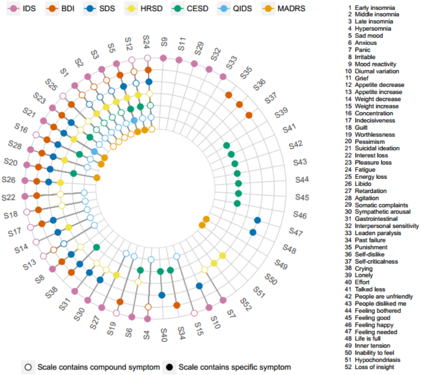
- Failing to report all dependent measures: 64%
- Deciding whether to increase *n* after seeing preliminary results: 56%
- Not reporting all conditions: 28%
- Selectively reporting studies that ‘worked’: 46%
- Excluding outliers because it made *p* smaller: 38%
- Presenting exploratory results as confirmatory: 27%
- Falsifying data: 1%

QRP: Sequential sampling

- Regular *p*-value is constructed for a single test
- With sequential sampling: multiple testing → chance capitalization → adjustments required



QRP: Selectively choosing variables



Source: Fried, E. (2017)[1]. The 7 most common depression scales contain no less than 52 symptoms. Each scale will give a different measure, yet all are interpreted as 'level of depression'.

[1] Fried, E. (2017). The 52 symptoms of major depression. Journal of Affective Disorders. 10.1016/j.jad.2016.10.019

QRP: Report findings on a misleading scale

Relative risk:

New drug reduced cancer incidence by 50%!!!

Absolute risk:

New drug reduced cancer incidence from 2 in 1000 to 1 in 1000

People easily confuse percentages and percentage points. Be clear to send the right message.

QRP: HARKing

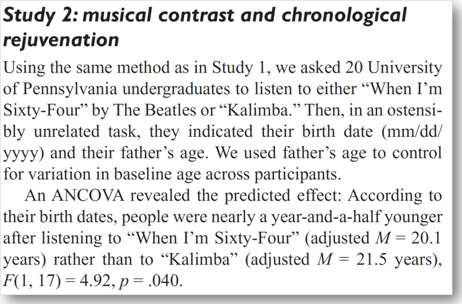
- HARKing: Hypothesizing after the results are known
- Term coined by Kerr (1988), Personality and Social Psychology Review

How to HARK:

1. Study a handful of variables: A lot of p-values: Main effects, two-way interactions, three-way interactions, etc.
2. Ignore chance capitalization.
3. Pick a relation based on a significant *p*-value.
4. Write your paper on that relation as if you intended to do that from the start.

QRPs are hard to detect

Simmons, Nelson & Simonsohn (2001)[2] demonstrate how QRPs are done and ‘hidden’.



Study 2: musical contrast and chronological rejuvenation

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either “When I’m Sixty-Four” by The Beatles or “Kalimba.” Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father’s age. We used father’s age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), $F(1, 17) = 4.92$, $p = .040$.

Seemingly sound text.

[2] False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22(11), 1359-1366. doi:10.1177/0956797611417632

QRPs are hard to detect

The full story:

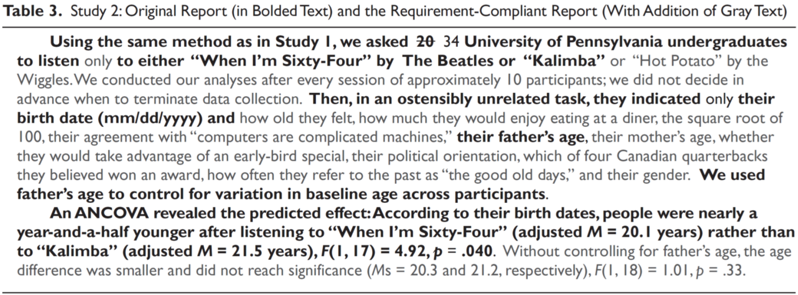


Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked ~~**20**~~ 34 **University of Pennsylvania undergraduates to listen** only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age**, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), *F*(1, 18) = 1.01, *p* = .33.

QRPs:

- Removal of participants
- Extra condition
- Didn't determine number of participants beforehand
- Additional dependent variables
- Entire results depend on covariate

Consequences of QRP

Huge increase in number of false positive findings:

You claim that something is significant when it actually isn't.

Failure to replicate

Many famous psychological experiments failed to replicate, or have been found out to be based on QRPs.

Some examples:

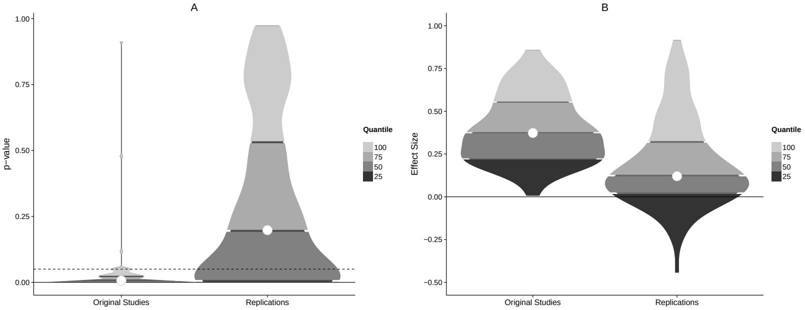
- The Blocking Effect
- Diffusion of responsibility
- Ego-Depletion
- Facial-Feedback Hypothesis
- Learning Styles
- The Marshmallow Test
- The Mozart Effect
- Power Posing
- The Robber's Cave
- The Stanford Prison Experiment

(Source: `https://bit.ly/2y7qeer`)

The Reproducibility Project

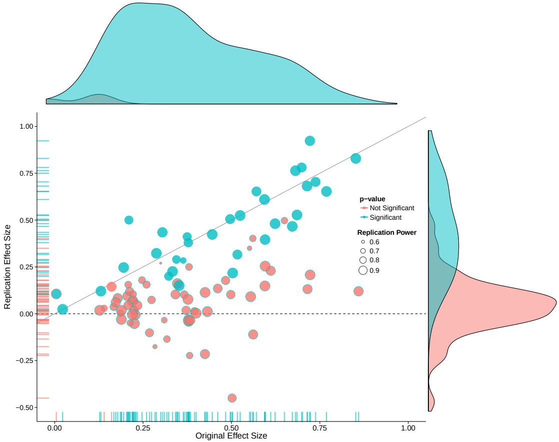
- Open Science Collaboration (2015), Estimating the reproducibility of psychological science, Science.
- 100 empirical studies from publications in three psychological top journals in 2008.
- Studies were replicated (done again) under as similar conditions as possible (nr. of participants, experimental setings, analysis method, etc.).
- Goal: Compare the effect sizes and p-values of the original and the replicated studies.
- If everything is fair, you expect that roughly half of the replication experiments have lower p-value or effect size than the original studies, half have higher.

The Reproducibility Project



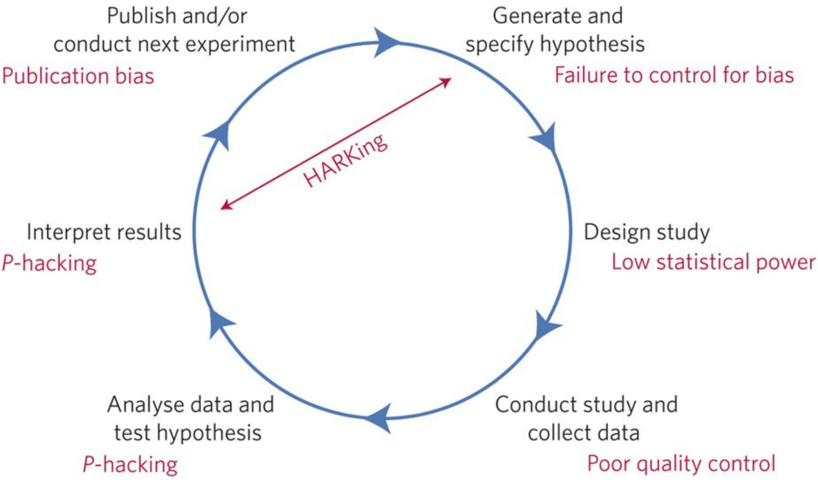
Evidence for 'file drawering' effect: Only positive results get published.

The Reproducibility Project



Evidence for 'file drawering' effect: Only positive results get published.

QRPs and the circle of science



Publish and/or conduct next experiment
Publication bias
Generate and specify hypothesis
Failure to control for bias
HARKing
Interpret results
P-hacking
Design study
Low statistical power
Analyse data and test hypothesis
P-hacking
Conduct study and collect data
Poor quality control

(Mufanò et al., Nature Human Behaviour, 2017)

Solutions

How to reduce the amount of QRPs and non-replicable findings in scientific literature?

Trust but verify

Better (and more) **methodological training**. Both of scientists and the public.

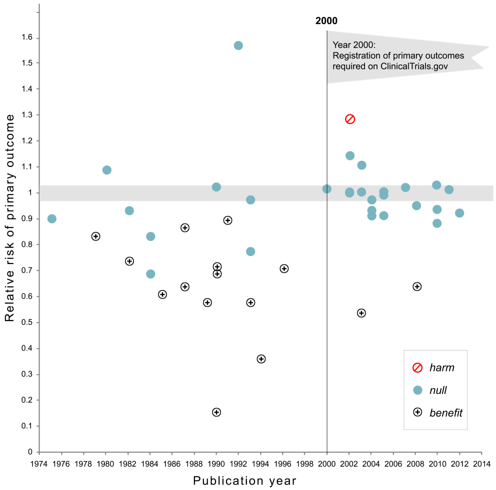
Open Science: Share **all** your research materials.

Preregistration: Publicly state what you will do *prior* to data collection.

Replicate: Results can be false positives. Try it again!

Meta-analyses: Combine the results of multiple studies.

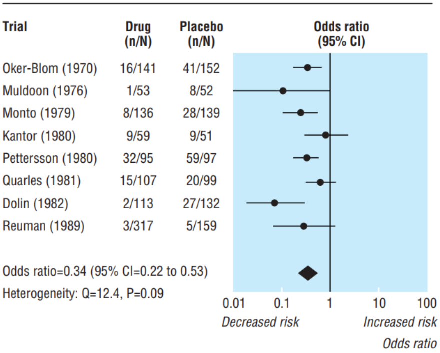
Pre-registration works



(Source: Kaplan & Irvin (2015), PLoS ONE.)

Meta-analyses

A meta-analysis is a structured way of combining findings from different studies.



Trial	Drug (n/N)	Placebo (n/N)
Oker-Blom (1970)	16/141	41/152
Muldoon (1976)	1/53	8/52
Monto (1979)	8/136	28/139
Kantor (1980)	9/59	9/51
Pettersson (1980)	32/95	59/97
Quarles (1981)	15/107	20/99
Dolin (1982)	2/113	27/132
Reuman (1989)	3/317	5/159

Odds ratio=0.34 (95% CI=0.22 to 0.53)

Heterogeneity: Q=12.4, P=0.09

Fig 1 Eight trials of amantadine for prevention of influenza.[11] Outcome is cases of influenza. Summary odds ratios calculated with random effects method

Source: Higgings et al. (2013) Measuring inconsistency in meta-analyses, BMJ.

Meta-analyses

- Each study has an effect size
- When there are no QRPs, roughly 95% of the 95% CI's should overlap.
- The observed effect sizes differ due to chance and QRPs:

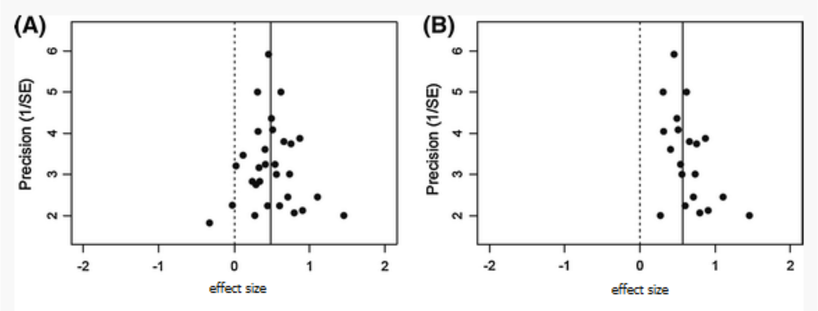
$$\boxed{\text{(Observed ES)} = \text{(True ES)} + \text{(Bias due to QRP)} + \text{(Random Variation)}}$$

- The smaller the sample size of a study, the larger the random variation

Note: For the exam, you should be able to interpret meta-analysis plots. You do not have to be able to reproduce them or do any of the specific meta-analytic calculations.

Meta-analyses

Funnel plots can be used to detect publication bias and other QRPs:



(Source: Nakagawa & Santos (2012). Evolutionary Ecology.)

Plot (A): ‘regular’ pyramid-shape: No sign of QRP; clear sign of an effect.
Plot (B): skewed pyramid-shape: Sign of QRP.

For the next lecture

Contents:

- Overview lecture

No new reading material