Week Five

Open Science

Philip Leftwich

25.10.2021

Hello! 👋

- How are you doing today?
- Go to Slido.com #769388

Open Science

- A survey in 2015 estimated the % of published papers in STEM subjects that included accompanying raw data

Open Science

- A survey in 2015 estimated the % of published papers in STEM subjects that included accompanying raw data
- Only 13% of research articles included raw data

[Refs] 10.1371/journal.pone.0143460

"Reasons" not to share

- Fear of being "scooped"
- Patent applications
- Confidentiality - unclear about data protection
- Lack of technical ability

The Reproducibility crisis in Science

- In a 2011 study it was reported that only 25% of cancer research studies could be replicated or reproduced
- Inappropriate practices of science include:
  - HARKing
  - P-hacking
  - Selective reporting
  - Poor research design
  - Publication bias

[Refs]10.1038/nrd3439-c1

HARKing

HARKing (Hypothesizing After the Results are Known)

- Dismissing or neglecting to report on rejected *apriori* hypotheses
- Generating *new* hypotheses after collecting and analyzing data

Can increase Type 1 errors

Motivated by difficulty in publishing *null* results

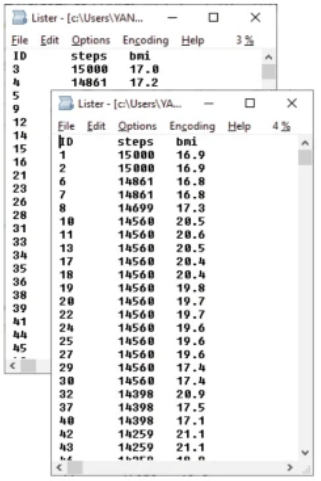
Confirmatory vs. Exploratory Science

Exploratory analysis is not the problem!

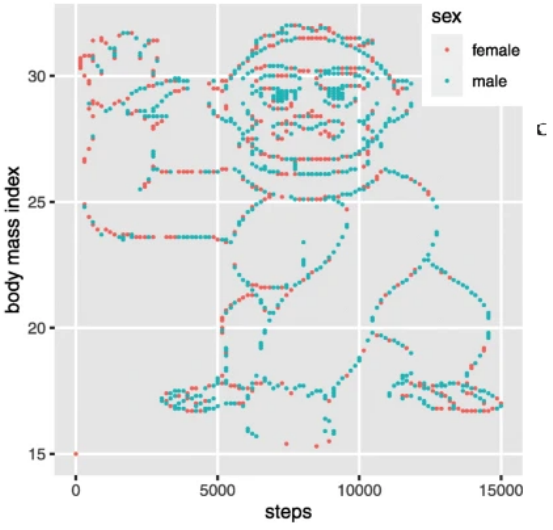
Confirmatory vs. Exploratory Science

Exploratory analysis is not the problem!

a



b



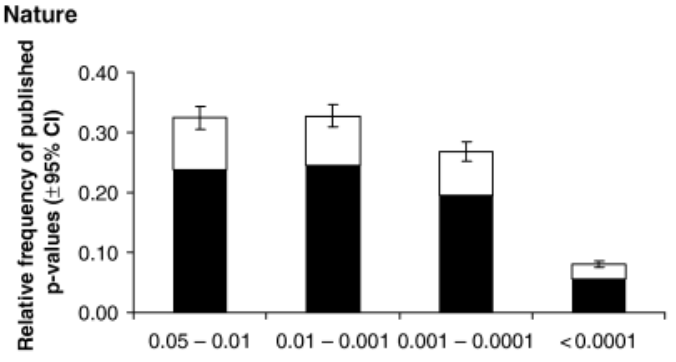
c

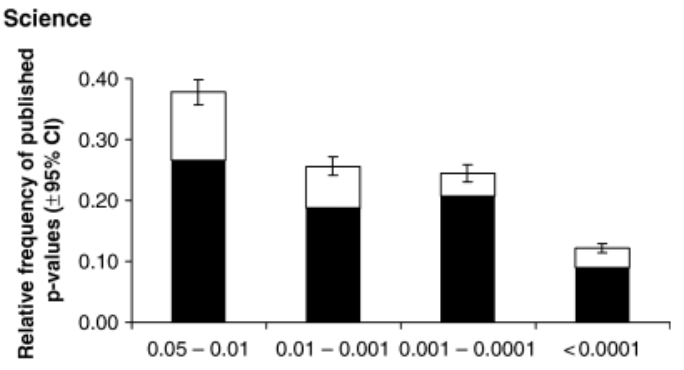
	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

[Refs]10.1186/s13059-020-02133-w

P-hacking

Performing many statistical tests on the data and only reporting those that come back with significant results

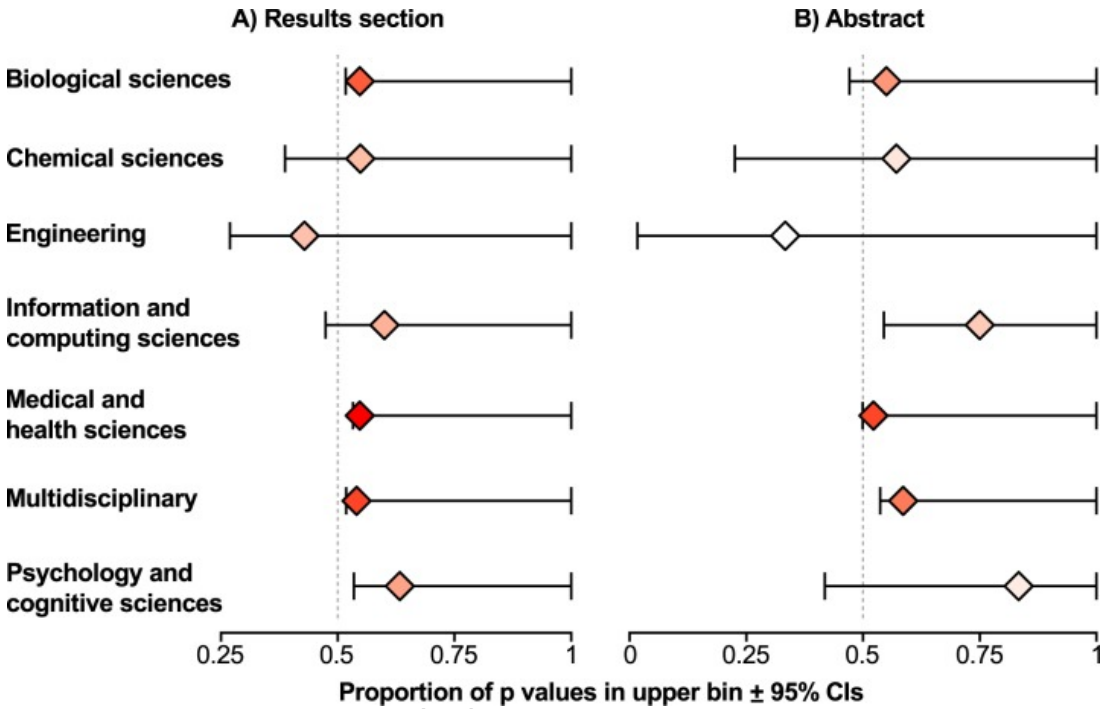




[Refs]10.1111/j.1420-9101.2006.01291.x

P-hacking

Widespread - but probably not but weak compared to the effect sizes being measured?



[Refs]10.1371/journal.pbio.1002106

Publication bias

- Significant results are more associate with *top* journals
- Novelty is prized
- Employers and funders count papers (weighted by journal impact) to assess performance

Mistakes vs. Fraud

- How prevalent is scientific misconduct?
  - 2% of scientists admit to having fabricated, falsified, or modified data at least once
  - Serial offenders thought to be primary drivers
- How can we prevent it & catch genuine mistakes or errors?

[]

https://ecoevoevoeco.blogspot.com/2020/02/integrityandtrust-idea-3-collaborate.html

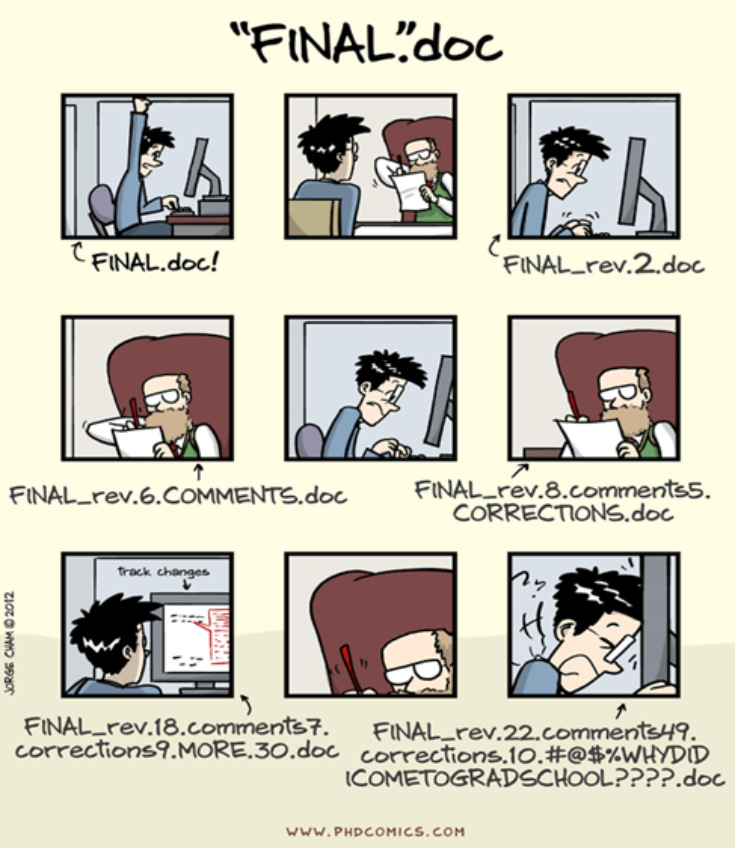
https://dynamicecology.wordpress.com/2020/02/17/some-data-and-historical-perspective-on-scientific-misconduct/

What can we do?

- Collaborative data analysis
- Share data *and* analyses
- Must be understandable

Version control

If you use Office 365 or Google Docs, you are already using version control



GitHub

- Git a free and open-source version control system
- GitHub a hosting system for Git that we can easily interact with using R and RStudio
- Set up a connection and your RStudio project can be Pushed to a repository on GitHub
- Fully collaborative, open, reproducible analyses
- Can be *archived* when ready to publish with a doi

Repositories

Specialist

- Genbank
- SRA
- UniProtKB
- The Cancer Imaging Archive

General

- Figshare
- Zenodo
- Dryad

https://www.nature.com/sdata/policies/repositories#nuc

Code readability

- What are some ways we can make our code readable?

Code readability

- What are some ways we can make our code readable?
- Frequent comments
- Pipes
- Use objects that make sense/relevance
- Organise directories and folders
- Provide READMEs

Thank you!

Next-time: Misleading figures!