ReproducibiliTea

The Garden of Forking Paths (Gelman and Loken, 2013)

Clement Lee

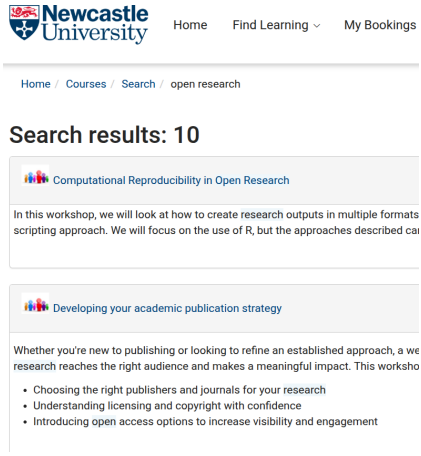
2025-10-16 (Thu)

Welcome

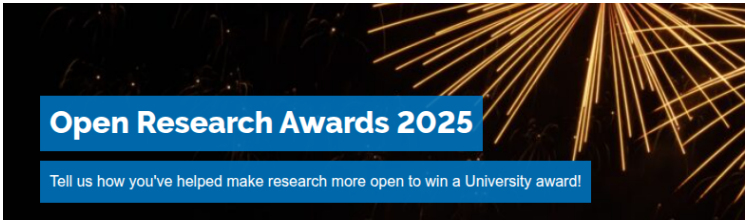
- A blend of:
  - Open Research
  - https://reproducibilitea.org
  - UK Reproducibility Network (UKRN)
- Monthly informal discussion
- Practices sharing
- Precursor of training

Training

- Research Software Engineering:
  - https://rse.ncldata.dev/events
- Open Research workshops:
  1. Go to https://elements.ncl.ac.uk
  2. "Find Learning" → "Courses"
  3. Search "Open Research"
- Email openresearch@ncl.ac.uk



Other events

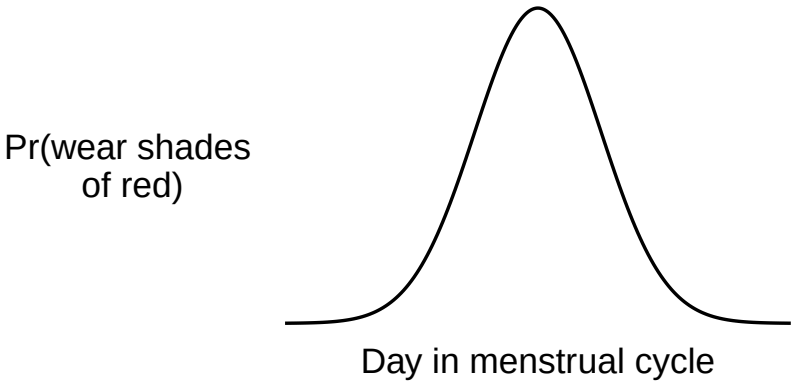


- Open Research Awards (Deadine Nov 7th):
  - https://www.ncl.ac.uk/open-research/awards/
- Statistics & Data Science Clinic:
  - https://www.ncl.ac.uk/maths-physics/engagement/nusolve/
- (External) Replication Games
  - We plan to form a Newcastle team if there's enough interest

The Garden of Forking Paths

- 1941 short story by Argentine writer and poet Jorge Luis Borges
- On the surface: story of a spy in World War I
- Foreshadows the many-worlds interpretation of quantum mechanics
- An inspiration of hypertext fiction

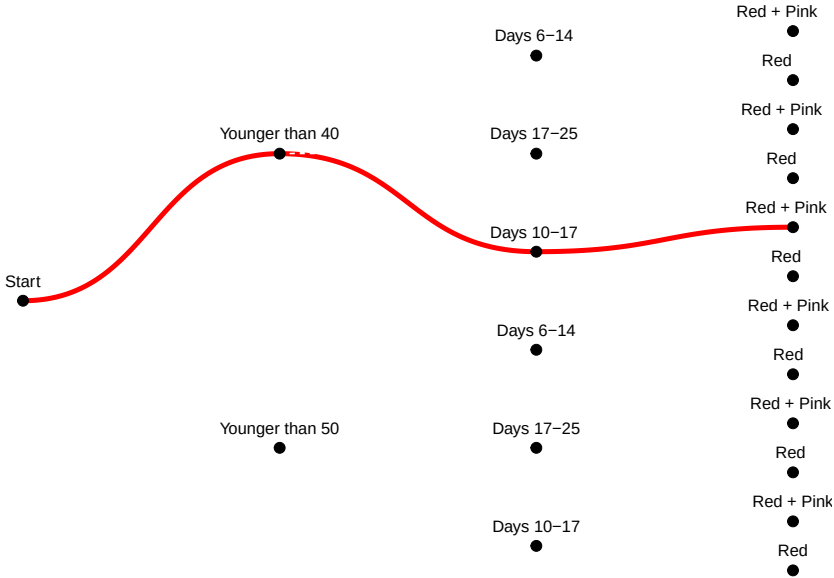
Gelman and Loken (2013) on Beall and Tracy (2013)



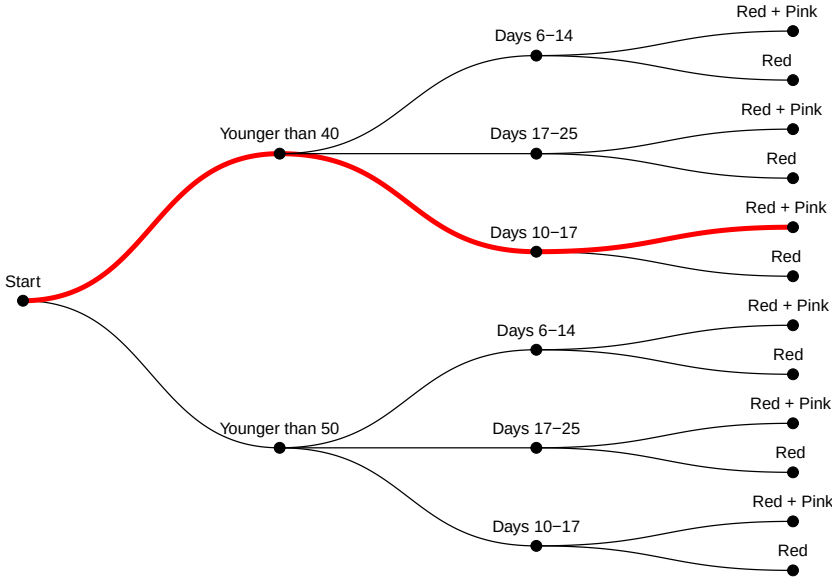
Good practice

- Specify the model & criteria before you collect & analyse the data
- Inclusion: Women younger than 40
- Peak: Days 10–17
- Colors: Red + pink

Illustration



Fishing / *p*-hacking



Rebuttal

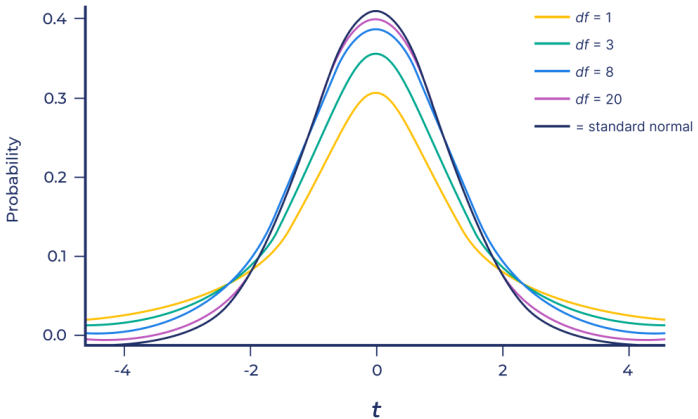
Tracy and Beall (2013)

- In a blog post that doesn't exist anymore
- Claimed they were not fishing
- Only one single analysis on one single data set

Gelman and Loken (2013)

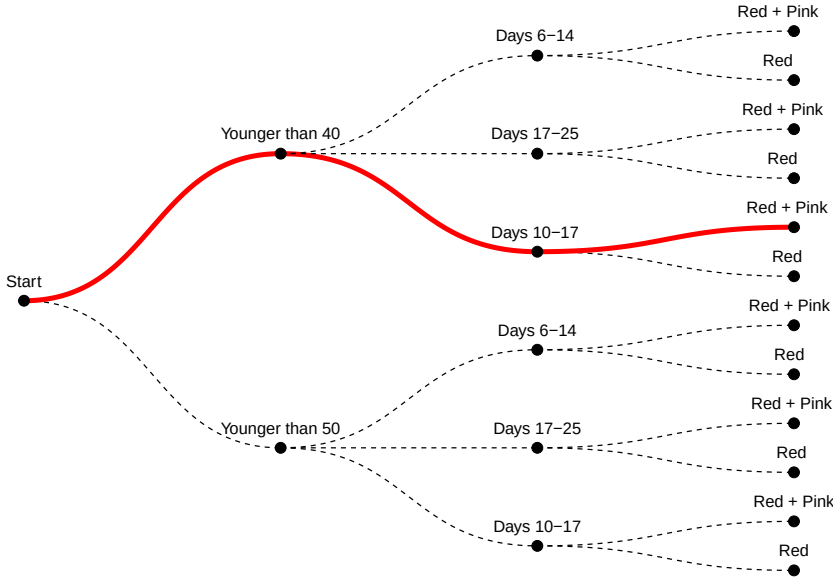
- Even if they were not fishing, they didn't follow the good practice either
- Pointed to **researcher** degrees of freedom

~~Researcher~~ degrees of freedom

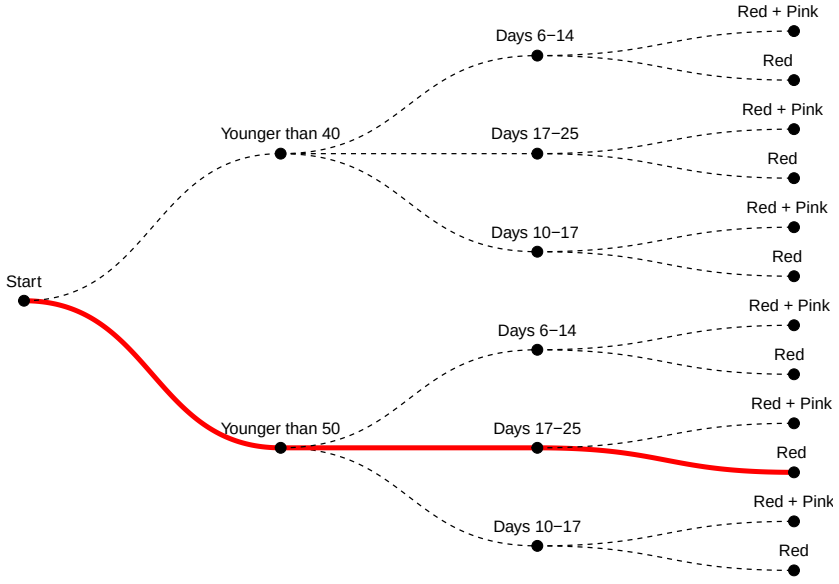


Source: https://www.scribbr.co.uk/stats/degrees-of-freedom-df/

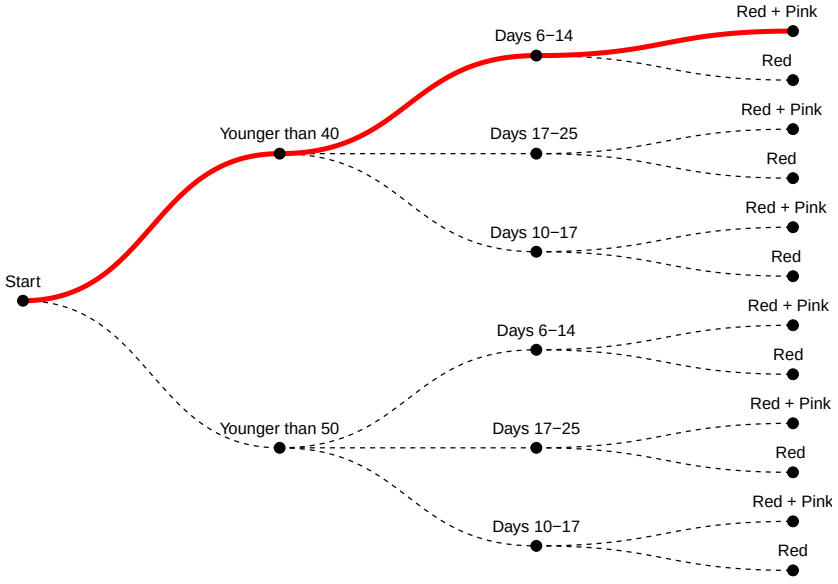
Researcher degrees of freedom



Had the data been different



Or this


Start
Younger than 40
Younger than 50
Days 6−14
Days 17−25
Days 10−17
Days 6−14
Days 17−25
Days 10−17
Red + Pink
Red
Red + Pink
Red
Red + Pink
Red
Red + Pink
Red
Red + Pink
Red
Red + Pink
Red

Potential multiple comparison

- Had the data been different, you would end up with a different model
- Analysis contingent on data $\rightarrow$ p-value argument disappears

"One can no longer argue that, if nothing were going on, that something as extreme as what was observed would occur less than 5% of the time."

Q: How to resolve this?

Preregistration

This Article Is Licensed Under CC0 For Maximum Reuse.

Preregistration is the practice of **documenting your research plan** at the beginning of your study and storing that plan in a **read-only public repository** such as OSF Registries or the National Library of Medicine's Clinical Trials Registry.

- Open Science Framework (OSF):
  - https://help.osf.io/article/145-preregistration
- As Predicted:
  - https://aspredicted.org/

Q: What about null results?

Registered Reports

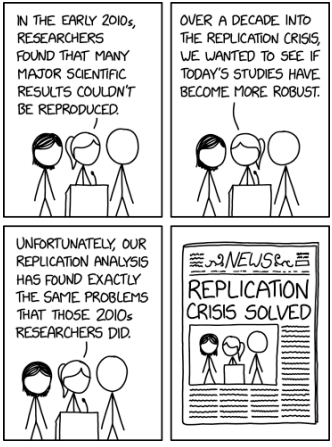
Author and reviewer guidelines

A Registered Report is a form of empirical article offered at *Nature* for papers in selected areas. As of 22 February 2023 these include cognitive neuroscience and the behavioural and social sciences. It is a format in which the methods and proposed analyses are pre-registered and peer reviewed prior to research being conducted. Submissions that pass peer review are provisionally accepted for publication before data collection and/or analyses are done. This format is designed to minimize publication bias and research bias in hypothesis-driven research, while also allowing the flexibility to conduct exploratory (unregistered) analyses and report serendipitous findings.

- Training!
  - Experts on preregistration & registered reports within University

Q: But many hypotheses were not well formed before

- Don't treat the finding as definitive truth
- More like exploratory analysis
- Reproduce/replicate it!
  - Experiment, collect data & analyse again



Source: https://xkcd.com/3117/

Quotes from discussion

- *"... evidence in these research papers is not as strong as stated."*
- *"... watch out for a system that allows any hunch to be ratched up to a level of statistical significance that is then taken as scientific proof."*
- *"... it is ironic that, by way of data-dependent analyses, p-values are often used to lend credence to noisy claims based on small samples."*

Why psychology?

- Small & noisy data
  - Small effect sizes & sample sizes
  - Large measurement errors & high variation
- More common in social sciences
- Much higher threshold in other fields
  - In physics: p-values $< 3 \times 10^{-7}$ required

Some more context

- Original article unpublished
  - Citation count > 1000
- A similar version in Gelman and Loken (2014)

Source: https://www.americanscientist.org/article/the-statistical-crisis-in-science

Some more context

- Fighting bad papers on his blog: https://statmodeling.stat.columbia.edu

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

Female hurricanes are deadlier than male hurricanes

Kiju Jung, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe Authors Info & Affiliations

Edited* by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 14, 2014 (received for review February 13, 2014)

June 2, 2014 | 111 (24) 8782-8787 | https://doi.org/10.1073/pnas.1402786111

References

Beall, Alec T, and Jessica L Tracy. 2013. “Women Are More Likely to Wear Red or Pink at Peak Fertibility.” *Psychological Science* 24 (9).

Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "P-Hacking" and the Research Hypothesis Was Posited Ahead of Time.”

———. 2014. “The Statistical Crisis in Science.” *American Scientist* 102 (6): 460.