The Reproducibility Debate in Science

Statisticians need to get involved

John Maindonald
Centre for Mathematics and Its Applications
Australian National University

25 November 2015

Focus

Focus is on comparative studies, e.g., compare speed of walking between group that has been primed to think about the disabilities of aged with unprimed group.

- Null hypothesis: Priming does not slow walking speed
- Alternative: Does slow walking speed, i.e., there is an effect

A small p-value (typically $p \leq \alpha$, with $\alpha = 0.05$) is commonly taken to justify rejection of the Null, implying a real effect.

- For medical research, focus is on pre-clinical studies
- Limited relevance to co-operative multi-disciplinary studies
- Carry-over from animal or in vitro models to humans is a separate discussion.

Aside: Fisher on *p*-values

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). ... A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level (0.05 or 0.02) of significance.[1]

... we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.[2]

1: Fisher (1926)
2: Fisher (1937)

Sources of evidence

- Ioannidis (2005)[1] — '... Most Published ... Findings Are False'
  - This paper put reproducibility issues "on the map"
- Direct evidence — results do not reproduce
  - Examples shortly, best evidence is in psychology
  - Most worrying evidence is in cancer studies
- Warning signals, from examination of papers

  - 'lack of +ve & -ve controls', faulty stats,
    'inappropriate use of key reagents', 'failure to repeat', ...
- Results may be unreplicible (check paper to see this)
  - Key information may be omitted or wrong

1: Ioannidis (2005), 'Why Most Published Research Findings Are False'

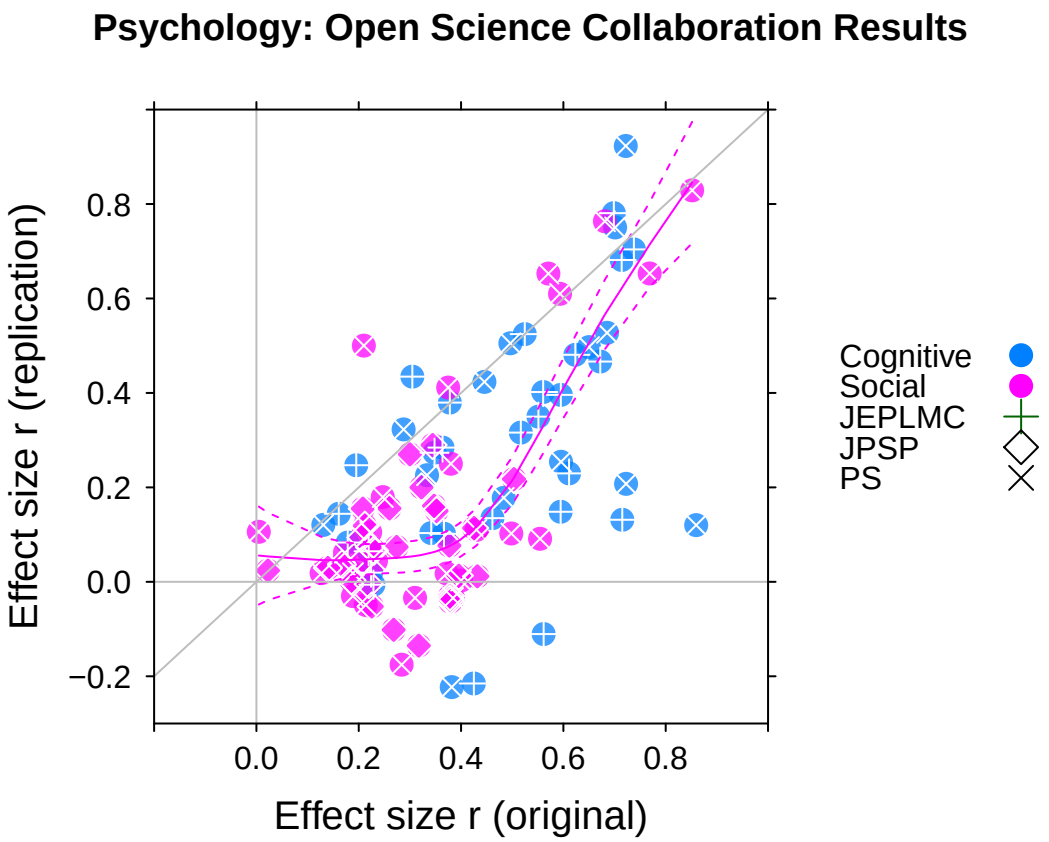
Selected Evidence

- Amgen: Reproduced 6 only of 53 'landmark' cancer studies.[1]
  - Begley (2013) notes issues with the studies that failed
- Bayer: Main results from 19 of 65 'seminal' drug studies
  - NB, journal impact factor was not a good predictor![2]
- fMRI studies: 57 of 134 papers (42%) had ≥ 1 case lacking check on separate test image. Another 14%, unclear ...[4]
- The Reproducibility:Psychology Project (~40% replicated)
  - Summary of results: 28 Aug 2015 issue of Science

1: Begley and Ellis (2012), 'Raise standards ...'; NB also Begley (2013)
2: Prinz, Schlange, and Asadullah (2011), 'Believe it or not ... drug targets'
3: Kriegeskorte et al. (2009), '... dangers of double dipping'
4: OSC (2015), 'Estimating the reproducibility of psychological science'

Begley's 6 red flags

- Blinding?
- Were basic expts repeated?
- Are all results given?
  (Cross-check images)
- +ve & -ve controls?
- Were reagents validated?
- Flawed statistical analysis



Collins & Tabak[1] — Factors include . . .

poor training . . . in experimental design

making provocative statements rather than presenting technical details

Crucial experimental design elements that are too frequently ignored include blinding, randomization, replication, sample-size calculation and the effect of sex differences

some scientists reputedly use a 'secret sauce' to make their experiments work — and withhold details . . . or describe them only vaguely . . .

Also: Deviations from stated protocol; errors in data; selective use of data; selection effects

1. Collins and Tabak (2014), '. . . NIH plans to enhance reproducibility'

What are the issues?

- Faulty design/reporting/execution (Begley, Collins & Tabak)
  - Repeating the same mistakes will not help
  - Some results are in principle unreplicable
- Selection effects – mostly there is no effect
  - More enlightened use of p-values will help

  - More crucially, the total process is not transparent

We do not know, certainly not with any certainty, where the balance lies between these two different types of issue. In this area, science lacks a scientific understanding of its own processes.

Is the criticism overblown?

- *The scientific process does finally identify the chaff*
  This contrasts with, e.g., alternative medicine.

Sure, but the process is far too protracted & tortuous.
Too many methodological failures go undetected.
Rewards systems encourage work that is poor quality.

- *One can never achieve 100% certainty*

Sure, but we can do a lot better than at present. Science should not be ignorant of its own processes.

- *Present processes are pretty much OK* (implied, not said)

We have the technology needed to do a vastly better job, but are not using it!

Journal & refereeing failures

- 'Publish' all; not just $p <= \alpha$ ($\alpha = 0.05$ or 0.01, or ...)
  - $p <= 0.01$ rather than $p <= 0.05$ is not an answer
- Referees & readers do not have the information needed
  - Know exactly what was done; check data, code
- Referees are at or beyond the limits of their expertise
  - Statistical analysis is an especial difficulty
- The system is not making good use of modern technology
  - plus, it interacts with rewards systems in malign ways
- Savvy critics are a huge untapped resource that is wasted
  - Other experts, in the area or in relevant areas
  - As, e.g., population projection, let the market choose.

Open Science's response: make all processes transparent

Commentary in *Science* (June 26 2015)

1. Self-correction in Science at work[1]

- Publish replications (PPS now has a section for this)
- Highlight & reward completeness of information
- Encourage publishing well (not often), ...
- Create a culture that is willing to admit mistakes, ...
- School scientists in research ethics

2. Promoting an open research culture[2]

- Transparency & Openness Promotion (TOP) Guidelines

1: Alberts and others (2015); 'Self-correction in science at work'
2: B. A. Nosek and others (2015); 'Promoting an open research culture'

TOP's 8 standards — 4 levels of each[1]

- (1) Citation standards (data, code, materials)
- (2)-(5) Transparency wrt data, analytic methods (code), research materials, design and analysis
- (6)-(7) Preregistration of studies, analysis plans

- (8) Replication

Levels for (8) Replication, as an example

- level 0: Discourages
- level 1: Encourages
- level 2: Encourages, & conducts blind review of results
- level 3: Encourages, with a protocol

1: B. A. Nosek and others (2015); 'Promoting an open research culture'

Scholarship: Beyond the paper

"Now we are witnessing the transition to yet another scholarly communication system — one that will harness the technology of the Web to vastly improve dissemination. ... The Web opens the workshop windows to disseminate scholarship as it happens, erasing the artificial distinction between process and product. ...

Today's publication silos will be replaced by a set of decentralized, interoperable services that are built on a core infrastructure of open data and evolving standards — like the Web itself .... This 'decoupled journal' publishes promiscuously, then subjects products to rigorous review through the aggregated judgements of expert communities, supporting both rapid, fine-grained filtering and consistent, meaningful evaluation."

Jason Priem: Nature 495, 437–440 (28 March 2013) doi:10.1038/495437a

An Open Source Model for Science

- Open Source Malaria — think "Linux for Malaria Research"[1]
  This follows a successful Schistosomiasis project.
- The Validation Science Exchange's Reproducibility Initiative[2]

- Cancer Studies — 50 "most impactful" from 2010-2012[3]

1: Todd and others (2015) (Matt Todd & others), OSBR (2015)
2: Iorns and others (2015) (Iorns & others) 3: Errington et al. (2014); Kaiser (2015), in June 26 2015 *Science*

Slides

Slides for this talk (pdf + R Markdown sources) will be posted at: http://maths-people.anu.edu.au/~johnm/stats-issues/

References

Alberts, Bruce, and others. 2015. "Self-Correction in Science at Work." *Science* 348 (6242): 1420–22.

Begley, C. Glenn. 2013. "Reproducibility: Six Red Flags for Suspect Work." *Nature* 497 (7450): 433–34. doi:10.1038/497433a.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483 (7391): 531–33. doi:10.1038/483531a.

Collins, Francis S., and Lawrence A. Tabak. 2014. "Policy: NIH Plans to Enhance Reproducibility." *Nature* 505 (7485): 612–13. doi:10.1038/505612a.

Errington, Timothy M, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, and Brian A Nosek. 2014. "An Open Investigation of the Reproducibility of Cancer Biology Research." *ELife* 3. doi:10.7554/elife.04333.

Fisher, Ronald Aylmer. 1926. "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture GB* 33: 503–13.

———. 1937. *The Design of Experiments*. 2nd ed. Oliver; Boyd.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *CHANCE* 18 (4): 40–47. doi:10.1080/09332480.2005.10722754.

Iorns, Elizabeth, and others. 2015. "Validation by Science Exchange - Identifying and Rewarding High-Quality Research." *Validation.scienceexchange.com*. http://validation.scienceexchange.com//#.

Kaiser, Jocelyn. 2015. "The Cancer Test." *Science* 348 (6242): 1411–13.

Kriegeskorte, Nikolaus, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. 2009. "Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping." *Nature Neuroscience* 12 (5): 535–40. doi:10.1038/nn.2303.

Nosek, B A, and others. 2015. "Promoting an Open Research Culture." *Science* 348 (3242): 1422–25.

OSBR. 2015. "The Synaptic Leap: Open Source Biomedical Research." http://www.thesynapticleap.org/.

OSC. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): 'aac4716–1'–'aac4716–7'. doi:10.1126/science.aac4716.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery* 10 (9): 712–12. doi:10.1038/nrd3439-c1.

Todd, Mat, and others. 2015. "OSM - Open Source Malaria." *Opensourcemalaria.org*. http://opensourcemalaria.org/.