Reproducibility in Science - Rethink, or Crisis?

How believable is published laboratory science?

John Maindonald, MSI, ANU

revised Oct 23 2015

All is Not Well at the Lab

Myth? “The glorious endeavour that we know today as science has grown out of the murk of sorcery, religious ritual, and cooking. But while witches, priests and chefs were developing taller and taller hats, scientists worked out a method for determining the validity of their results: they learned to ask”Are they reproducible¿‘[1]

Reality “Scientists like to think of science as self-correcting. To an alarming degree, it is not.”[2]

[1] Scherr 1983, in "The Best of the Journal of Irreproducible Results".

[2] Economist article, Oct 19 2013

Selected Evidence

- 9 attempts to reproduce a 1998 priming study failed (Does thinking about a professor improve test performance?)
- Amgen scientists: reproduced results from 6 only of 53 'landmark' cancer studies. (but no details are available)
- Bayer scientists: main results in 19 of 65 'seminal' drug studies
  - NB, journal impact factor was not a good predictor!
  - Limited summary information, scant detail
- Of 134 fMRI papers examined by Kriegeskorte et al (2000), 42% (57) had at least one case where results were checked on the images used to find an effect. Another 14%, ?

$p < 0.05$ has implied a low ($< 25\%$) probability of a real effect! Many/most were fishing experiments, plus other factors also?

See also references in: Landis et al (2012): "A call for transparent reporting to optimize the predictive value of preclinical research." Nature 490: 187–191.

The Wide Reach of Study Quality Issues

- Collins & Tabak (2014): **"Policy: NIH plans to enhance reproducibility"**. Nature 505, 612–613. {doi:10.1038/505612a}
- Yong (2012): **"Nobel laureate challenges psychologists to clean up their act"**. {Nature doi:10.1038/nature.2012.11535}
- Chen (2013): **"Hidden depths: Brain science is drowning in uncertainty."** New Scientist, issue 2939, 17 October.
- Everitt, J I (2015): **"The Future of Preclinical Animal Models in Pharmaceutical Discovery and Development"**. Toxicologic Pathology 43:70-77. {doi:10.1177/0192623314555162}
- Leek et al (2010): **"Tackling the widespread and critical impact of batch effects in high-throughput data."** Nat. Rev. Genet. 2010 Oct;11(10):733-739.
- Suter & Cormier (2014): **"The Problem of Biased Data & Potential Solutions for Health & Environmental Assessments"**. Human and Ecological Risk Assessment, to appear. {doi:10.1080/10807039.2014.974499}

Collins & Tabak (Nature, 27 Jan 2014) lay it on!

“**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 . . .

"

NB also: Mismatch between stated and actual protocol; selective use of data; data mistakes.

A Thought Experiment (1100 relationships probed)

Ratio of no to true relationship; here, 1000:100
Power (Prob[detect true relationship]); here, 0.8
Type 1 error rate (set in advance); here, 0.05

From 1100 total	*1000 True No*	*100 True Yes*
Yes result	$1000 \times 0.05 = 50$	$100 \times 0.8 = 80$
Fraction of total		$\text{PPV}^{3} = \frac{80}{130} = 0.615$

For the 'Yes' results, p is from a distribution in the range $p \leq 0.05$; thus $p = 0.01$, $p = 0.005$, and $p = 0.05$ all add just 1 to the count.

$p \leq 0.05$ may be roughly equivalent to an individual $p \simeq 0.0038$

[3] PPV = Positive Predictive Value; FDR (False Discovery Rate) = 1 - PPV

More detailed analysis[4]

Ionnidis gives formulae that account for:

- u = Proportion of claimed results that in truth reflect bias
- $n = \#$ of independent teams addressing the same questions

Note that:

- Increasing the power shifts the distribution of observed *p*-values so that more fall under the threshold.
- Power is a function of α. Decrease α and, for a given experiment, power decreases.

[4] Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

The Ionnidis results — Observations

1. Large studies are more likely to yield true results (NB, caveats)
2. The smaller the effect size, the less likely that findings are true
3. Many relationships, less testing => low PPV
4. Flexibility in designs, definitions, outcomes & analysis ...
5. Financial & other interests & prejudices ...
6. "Truth"" is less likely in fields that are "hot" (justifiably?)

$p \leq 0.05$ versus $p = 0.05$ (with prior odds 1:10)[5]

With prior odds = 1:10 (e.g.), what does $p = 0.05, \ldots,$ imply?

Slightly (?) optimistically, multiply prior odds by by $\frac{-1}{e\,p\,log(p)}$

p-value	Xplier	$\frac{1}{10} \times$ Xplier	Prob
0.05	2.5	2.5:10	$\frac{2.5}{2.5+10} = 0.2$
0.01	8.0	8:10	0.44
0.001	53.2	53.2:10	0.84

Xplier as for $p \leq 0.05$ with power=0.8, requires $p \simeq 0.0038$

$p = 0.05$ is weaker evidence than a lump all $p \leq 0.05$ together.
(but ideally, look at the individual p-value.)

[5] For a semi-popular exposition ("*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume."), see:
Nuzzo, R. Scientific method: Statistical errors. Nature 506, 150–152 (13 February 2014) doi:10.1038/506150a (uses result from Sellke et al 2001)

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 togive this level [0.05 or 0.02] of significance.
[The arrangement of field experiments. J Minist Agric GB 33: 503–513, 1926.]

. . . 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.
[The Design of Experiments, 2nd Edition, 1937, p.16]

So then, repeated $p \leq 0.05$ results are needed to establish a result?
[How many repeats? It surely depends on the prior odds.]

Hence a way to shape *p*-value discourse to make it defendable.
Each new $p \leq 0.05$ (or whatever) makes the null less likely!

Fisher in a very accommodating mode

. . . the calculation is absurdly academic, for . . . no scientific worker has a fixed level of significance at which . . . in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which . . . is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance. [Fisher RA. Statistical Methods and Scientific Inference. London: Oliver and Boyd/Longman Group, 1956.]

Fisher was not especially consistent.[6]

[6] See http://www.jerrydallal.com/lhsp/p05.htm

Data Analysis Foundations (cf Tukey)

Distinguish Model Development from Inference:

- Models aim to give real world descriptions NB: Simulation extends the range of useful models
- Inference (or Generalization) Use diverse challenges to build (or destroy!) confidence in model-based inferences.

Tukey's warnings:

- Do not assume "that we always know what in fact we never know – the exact probability structure . . . "
- No data set is large enough [and has sufficient background information] to tell us how it should be analysed.

Tukey: **More honest foundations for data analysis.** Journal of Statistical Planning and Inference, vol. 57, no. 1, pp. 21-28, 1997

Types of Challenge

1. For experiments, is the design open to criticism?
2. Look for biases in processes that generated the data
3. Look for inadequacies in laboratory procedure
4. Model diagnostics (NB plots) & other critiques
5. Check model on test data (Test data chosen how?)
6. Repeat the exercise (the ultimate in training/test?)
   - But how robust is the replication result?

Add also, more general criticisms of the science that underlies interpretation or results.

Current Practice relies mostly on:

- A hope that researchers will keep challenges 1 – 4 in mind
    - Are they equipped to think about such challenges?
- Challenges from referees.
    - Statistical issues require statistically expert referees
    - Do referees have all necessary details (Typically, no!)
- Some high profile papers attract post-publication critique
    - How can the system use/accommodate this?

Current practice results in huge wasted effort as other researchers follow false leads, or go down what others had earlier identified (but not advertised) as blind alleys.

Hence, increasingly, a demand to **“Repeat the study”**

The Science of Doing Science

- There is a science of the doing of science. Statistics is a large part of it. So also is human psychology.
- Since 2012, these issues have gained increasing momentum
- Independent replication has become the surest way to show that results are replicable. Initiatives are afoot to build it into scientific processes.
  - The checks must however be done to a high standard
  - How useful is a further $p < 0.05$ result?
  - Independendent replication is an 'easy' add-on
- Genuinely scientific approach: Understand the system
  - All results should be reported, 'successes' & failures This provides a needed context for judging 'successes'
  - This is not an easy add-on to current procedures
  - New Researchers — do not believe all you read!

Reproducibility Initiatives — Psychology

- The Center for Open Science Many Labs project
  - reproduce 13 classical psych studies
  - 10 successful, 1 weakly reproduced, 2 not reproduced
  - Plots show the scatter across the 36 participating teams https://osf.io/ebmf8/
- Replication Project: Psychology[7]
  - 270 authors participated in replications — 100 studies
  - Studies were sampled from 2008 issues of 3 psych journals
  - Initial results: 39 reproduced + 24 "moderately similar" + 37 failures[8]

[7] https://osf.io/ezcuj/wiki/home/

[8] http://www.nature.com/news/first-results-from-psychology-s-largest-reproducibility-test-1.17433

Initiatives — Cancer and Other

Cancer studies * The Center for Open Science has a $1.3 million grant from the Laura and John Arnold Foundation, to fund replication of the 50 “most impactful” cancer biology studies from 2010-2012. + Substantial progress, no published reports yet + Raw datasets and data analyses will be publicly available + Individual replication reports, plus final summary report + See as an example https://osf.io/tcauf/

General * The Reproducibility Initiative: http://validation.scienceexchange.com/#

Funding Agencies and Journals

- Collins & Tabak: ". . . NIH plans to enhance reproducibility", Nature, 27 Jan 2014
- Nature: **Statistics & general methods** checklist [9]
  Match experimental designs against criteria — demands for randomisation, replication, local control (or blocking), etc. — that parallel Fisher's criteria for field experiments
- REMARK: REporting recommendations for tumor MARKer prognostic studies [10]
- ARRIVE guidelines for animal experimentation:[11] (maybe, really, for a different list. NB that ethics committees need to be aware of the statistical issues for efforts to reduce numbers of animals used.)

[9] http://www.nature.com/authors/policies/checklist.pdf

[10] http://www.nature.com/nrclinonc/journal/v2/n8/full/ncponc0252.html

[11] http://www.nc3rs.org.uk/arrive-guidelines

Challenges for Statistics Education

- Get design issues back on the agenda
- Hammer hard: $P(H_0 \mid p <= 0.05) \neq Pr(p <= 0.05 \mid H_0)$
- What does $p <= 0.05$ really imply?
  - Not much, unless the prior "no effect" proportion is known
  - A prize for the best short demonstration (paper or film)?
- Analyses must be robust against any reasonable challenge
  - Challenge science, design, execution, apply model diagnostics
- Familiarize students with the Nature style of checklist
  - Its items are a checklist for their learning. Tick them off as each idea is mastered
- This is an era of post-publication review. Get students involved.

The case for better understanding and better standards is now widely recognised.

Further Comment (1)

- In a context where 29% of positives are true positives
  - $\sim$40:1 prior odds for the NULL, with power=0.8, might explain the $\sim$29%
  - How much else is at play – bad design, faulty analyses, . . .
- A finding of 'no detectable effect' is important evidence
  - It provides essential context for a $p <= 0.05$ or $p <= 0.01$
  - It helps warn later experimenters against blind alleys
- The publishing model needs to change. But how?
  - Publish everything, one way or another, including "failures".
  - Report on a plan of research, not the individual study.
  - Post-publication review has a large & useful role.
  - Bloggers may sometimes fill in for deficiencies in scientific processes!
- Foster cooperation in large more definitive studies; ask how effect may change with dose.

Further Comment (2)

- Warnings for science funding & management systems
  - Avoid measures that corrupt what they presume to measure.
  - Funding regimes are prone to reward a "muddle through ourselves" mentality, with a devaluing of key skills (esp statistics?)
- High quality work requires high level skills from all relevant specialisms. Focus on the people skills, not the page counts!
- In social science and areas such as public health, new technologies offer rich new data sources and opportunities for new types of research. [cf Gary King (2014): "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science ." PS: Political Science & Politics, 47, 165-172. `doi:10.1017/S1049096513001534`.]

Ionnidis (2005) on what might be done

- Target large-scale evidence, where pre-study probability is high
  - Watch however for bias, as much in large studies as small
- . . . many teams, . . . what is the totality of the evidence?
  - upfront registration.
  - develop & adhere to a protocol
- Do not chase statistical significance
  - Assess, as far as possible, the pre-study odds (R)
  - Check findings that are judged relatively established
- Most new discoveries: from research with low pre-study odds
  - A single study will often give a very partial picture
  - Look for insight on the extent of pre-study data dredging?

Context: **Careful study design + accurate & full reporting.**

References & Selected Further Papers

Begley and Ellis (2012): **Raise standards for preclinical cancer research.** Nature, vol 483, pp 531-533.

Klein et al (2014). **Investigating Variation in Replicability. A "Many Labs" Replication Project.** Social Psychology 2014; Vol. 45(3):142–152.

Kriegeskorte, Simmons, Bellgowan & Baker (2010): **Circular analysis in systems neuroscience: the dangers of double dipping.** Nature Neuroscience 12, 535-540.

Prinz, Schlange and Asadullah (2011). **Believe it or not: how much can we rely on published data on potential drug targets?** Nature Reviews Drug Discovery 10, p712.

Sellke et al. **Calibration of p Values for Testing Precise Null Hypotheses**. American Statistician, Vol 55: pp.62-71, 2001

Further Papers

Couzin-Frankel (2013): **“When Mice Mislead.”** Science 342: 922-925. {doi:10.1126/science.342.6161.922}

Pfeiffer, Bertram, Ioannidis (2011). **Quantifying Selective Reporting and the Proteus Phenomenon for Multiple Datasets with Similar Bias.** PLoS ONE 6(3): e18362. doi:10.1371/journal.pone.0018362

Acknowledgements [12]

Dr Terry Neeman (ANU Statistical Consulting Unit) and ANU SCU clients — plied me with several of the key papers that I referenced.

[12] Slides for this talk (pdf + R Markdown sources; this + older version) can be found at: http://maths-people.anu.edu.au/ johnm/stats-issues/

More general issues – Herricanes vs Himmicanes

Claim: Female hurricanes are deadlier than male hurricanes

The suggestion is that they are taken less seriously?.[13]

log(E[`deaths`]) was regressed on damage in 2013 US$ (graph that follows has 2014 US$), barometric pressure at landfall, femaleness of the name, & interactions.

More realistically, transform the damage variable:

With a log transformation, the femaleness effect vanishes.

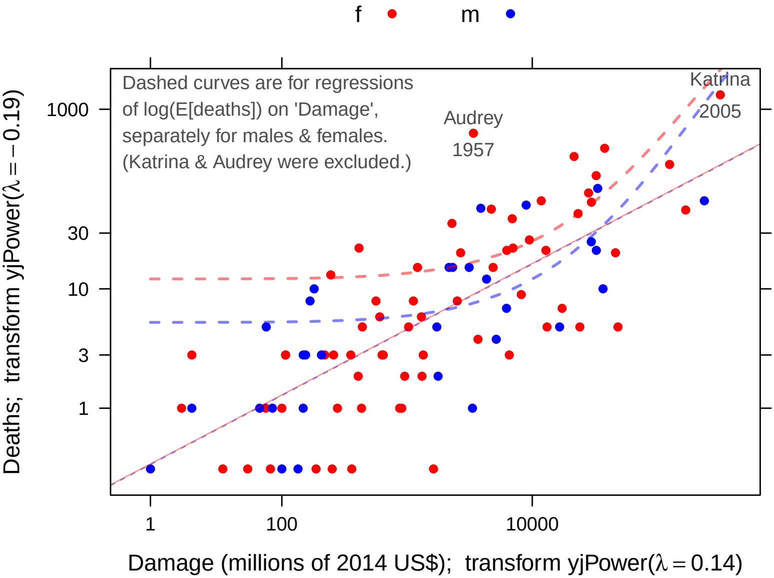
Even better, use the Yeo & Johnson (2000) extension of Box-Cox power transform methodology.[14] Use transforms thus: $\lambda = -0.19$ for `deaths+1`; $\lambda = 0.14$ for the damage variable.}

A storm of reaction on the blogosphere was mostly himmicane.

[13] Jung et al `http://www.pnas.org/cgi/doi/10.1073/pnas.1402786111`

[14] "A new family of power transformations to improve normality or symmetry". Biometrika, 87, 954-959. From R, type `help(yjPower, package='car')`

Hurricanes – Deaths vs Damage in 2014 $US


f ●
m ●
Dashed curves are for regressions
of log(E[deaths]) on 'Damage',
separately for males & females.
(Katrina & Audrey were excluded.)
Audrey
Katrina
Deaths; transform yjPower(λ = −0.19)
Damage (millions of 2014 US$); transform yjPower(λ = 0.14)