OSC. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716–16. doi:10.1126/science.aac4716 (http://dx.doi.org/10.1126/science.aac4716).

Pentland, Alex. 2014. *Social Physics*. Scribe.

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 (http://dx.doi.org/10.1038/nrd3439-c1).

Scherr, George. 1983. *The Best of the Journal of Irreproducible Results: Improbable Investigations and Unfounded Findings*. Workman Publishing Company.

Sellke, Thomas, M. J Bayarri, and James O Berger. 2001. “Calibration of P-Values for Testing Precise Null Hypotheses.” *The American Statistician* 55 (1): 62–71. doi:10.1198/000313001300339950 (http://dx.doi.org/10.1198/000313001300339950).

Todd, Mat, and others. 2015. “OSM - Open Source Malaria.” *Opensourcemalaria.org*. http://opensourcemalaria.org/ (http://opensourcemalaria.org/).

Tukey, John. 1997. “More Honest Foundations for Data Analysis.” *Journal of Statistical Planning and Inference* 57 (1): 21–28. doi:10.1016/s0378-3758(96)00032-8 (http://dx.doi.org/10.1016/s0378-3758(96)00032-8).

Yong, Ed. 2012. “Replication Studies: Bad Copy.” *Nature* 485 (7398): 298–300. doi:10.1038/485298a (http://dx.doi.org/10.1038/485298a).

# Reproducibility in Science — Rethink, or Crisis?

## A search for a more scientific science

John Maindonald
Centre for Mathematics and Its Applications
Australian National University

04 November, 2015

## Focus

Focus is on comparative studies, e.g., compare a cocktail thought to slow aging in mice with a placebo

- Null hypothesis: Cocktail does not slow aging
- Alternative: Does slow aging, 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 large co-operative multi-disciplinary studies
- Carry-over from animal or in vitro models to humans is a separate discussion.

# Overview

- Evidence of a problem —
  - Direct evidence — results do not reproduce
  - Warning signals — 'lack of +ve & -ve controls', faulty stats, 'inappropriate use of key reagents', 'failure to repeat', …
- Use/misuse of $p$-values — maybe mostly no effect, but …
  - always, expect $\geq$ 5% of studies to show p $\leq$ 0.05
- Initiatives (NB: articles in 'Science' for June 26 2015)
  - Pointers to a radical reshaping of research & publication?
- What needs to happen?



# Matching Ideal to Reality

**If only!**
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)
2: Economist (2013), 'Unreliable research. Trouble at the lab', October 19



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

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

King, Gary. 2013. “Restructuring the Social Sciences: Reflections from Harvard’s Institute for Quantitative Social Science.” *PS: Political Science & Politics* 47 (01): 165–72. doi:10.1017/s1049096513001534 (http://dx.doi.org/10.1017/s1049096513001534).

Klein, Richard A., and others. 2014. “Investigating Variation in Replicability.” *Social Psychology* 45 (3): 142–52. doi:10.1027/1864-9335/a000178 (http://dx.doi.org/10.1027/1864-9335/a000178).

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 (http://dx.doi.org/10.1038/nn.2303).

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

Nuzzo, Regina. 2014. “Scientific Method: Statistical Errors.” *Nature* 506 (7487): 150–52. doi:10.1038/506150a (http://dx.doi.org/10.1038/506150a).

OSBR. 2015. “The Synaptic Leap: Open Source Biomedical Research.” http://www.thesynapticleap.org/ (http://www.thesynapticleap.org/).

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

COS. 2015. “Center for Open Science (COS).” *Cos.io*. https://cos.io (https://cos.io).

Economist. 2013. “Unreliable Research. Trouble at the Lab.” *Economist*. http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong (http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong).

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 (http://dx.doi.org/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.

———. 1956. *Statistical Methods and Scientific Inference*. Hafner.

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

## 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]
- Failed attempts to reproduce widely cited priming studies[3]
  (Thinking about Grandpa makes one walk slowly?)
- fMRI studies: 57 of 134 papers (42%) had $\geq$ 1 case lacking check on separate test image. Another 14%, unclear …[4]

1: Begley and Ellis (2012), 'Raise standards …'; NB also Begley (2013)
2: Prinz, Schlange, and Asadullah (2011), 'Believe it or not … drug targets'
3: Yong (2012), 'Replication studies: Bad copy'
4: Kriegeskorte et al. (2009), '… dangers of double dipping'



## Collins & Tabak lay it on![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'

# $p$ is a relative measure of evidence

If no real effects, expect $\alpha$ = 5% of cases to show effect

- What then if the probability of a real effect is very small?
  - Might explain the Bayer 19/65 result; unlikely for 6/53

For what follows, note that $\alpha$ and power are tightly linked

- Power = Probability that a real effect will be detected
- Decrease $\alpha$ and, for a given design, power decreases.
- Increasing power (better or bigger expt) shifts the distribution of $p$-values — more fall under threshold (0.05, or …)
- If there are no real effects, power is irrelevant!



# 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)



# In conclusion

## Acknowledgements

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

## Slides

Slides for this talk (pdf + R Markdown sources) can be found at: http://maths-people.anu.edu.au/~johnm/stats-issues/ (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 (http://dx.doi.org/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 (http://dx.doi.org/10.1038/483531a).

# Further Comment (1)

- A finding of 'no detectable effect' is important evidence
  - It is essential context for 'successes'
  - It may warn later experimenters against a blind alley.
- The publishing model needs to change. But how?
  - Publish everything somehow, 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!
- Cooperate in larger more definitive studies; look for a dose-effect pattern.



# 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 people skills, not page counts!
- In major areas, (social science, public health, …) new technologies offer rich new data sources and opportunities for new types of research.[1,2]

1: c.f. King (2013) 'Restructuring the Social Sciences'
2: Pentland (2014) 'Social Physics'



# "… no … fixed level" — If only!

. . . 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 … may be much less than the frequency specifying the level of significance.

Fisher was not especially consistent. See http://www.jerrydallal.com/lhsp/p05.htm (http://www.jerrydallal.com/lhsp/p05.htm)

1: Fisher (1956)



# Thought Experiment (950 attempts)

Ratio of real to no relationship: 50:900 (1:18)
$\alpha$ (Prob[detect if No]) = 0.05; Power (Prob[detect if Yes]) = 0.5

| ***From 950 total*** | ***900 True No*** | ***50 True Yes*** | **Total = 950** |
|---|---|---|---|
| **Yes result** | 900 $\times$ 0.05 = 45 | 50 $\times$ 0.5 = 25 | 70 |
| ***True*:*Total* ratio** | | | $\frac{25}{70}$ = 0.36 |

For 'Yes', $p$ 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.

The *True*:*Total* ratio has the name Positive Predictive Value (PPV).
False Discovery Rate (FDR) = 1 - PPV

P-Value Demonstration (shinyPPV.Rmd)

# The Ioannidis results — Observations[1]

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

All the above have the proviso "other factors being equal".

1: More details, including a model for the effect of bias; see Ioannidis (2005), 'Why Most Published Research Findings Are False'



# A Defendable Use of $p$-values

$p$-values are relative measures of the weight of evidence

**Repeated** $p \leq 0.05$ results establish a result

- How many repeats? It surely depends on the prior odds.
- Each new $p \leq 0.05$ (or whatever) makes the null less likely!

Scenarios that might explain PPV = 37%

(e.g., take Bayer 19/65 as 80% of ~ $\frac{24}{65} \simeq 37\%$ with real effect)

- A ~1:27 prior odds for an effect, with power=0.8
- Or ~1:17 prior odds, with power 0.5



# Initiatives — Cancer and Other

**Cancer studies**

- The Reproducibility Project — Cancer Biology (noted above)[1]
  - Replicate 50 "most impactful" studies from 2010-2012.
  - Substantial progress, no published reports yet
  - Raw datasets and data analyses will be publicly available

**Applying the Open Source Model to Science**

- Open Source Malaria — think “Linux for Malaria Research"[2]
  This follows a successful Schistosomiasis project.

**General**

- The Validation Science Exchange's Reproducibility Initiative[3]

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



# Challenges for Statistics Education

- Get design issues back on the agenda
- Use the Nature checklist[1] to monitor learning
  - Tick each item off as its import is mastered.
- Drive home: $p$-value $\neq$ probability of a real effect!
- Analyses must be robust against any reasonable challenge
  - Challenge science, design, execution, use of diagnostics.
- This is an era of post-publication review. Involve students.

1: http://www.nature.com/authors/policies/checklist.pdf (http://www.nature.com/authors/policies/checklist.pdf)

# Center for Open Science Projects[1]

- Many Labs — reproduce 13 classical psych studies[2]
  - Of 13 studies — 10: successful, 1: weakly, 2: no!
  - Plots show scatter across the 36 participating teams
- Reproducibility: Psychology — 100 studies (3 journals, 2008)[3]
  - 1 replicate only of each study
  - Subjectively, 39% replicated original results[3]
- Cancer Studies — 50 "most impactful" from 2010-2012[4]

1: COS (2015); 'COS (Center for Open Science)'; https://cos.io/
2: Klein and others (2014); 'Many Labs'
3: OSC (2015); 'Estimating the replicability …'
4: Errington et al. (2014); 'An open investigation … cancer biology research'



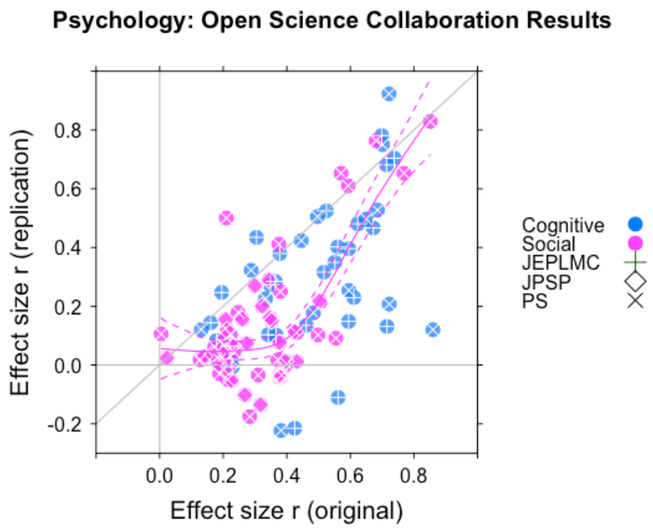
# $p \leq 0.05$ vs $p = 0.05$; prior odds 1:17

For a given $p$, posterior odds = $\frac{-1}{e\,p\,log(p)} \times$ prior odds[1]

| $p$-value | Xplier | $\frac{1}{17} \times$ Xplier | Prob (PPV) |
|---|---|---|---|
| **0.05** | 2.5 | $\frac{2.5}{17}$ | $\frac{2.5}{2.5+18}$ = 0.13 |
| **0.01** | 8.0 | $\frac{8}{17}$ | 0.32 |
| **0.001** | 53.2 | $\frac{53.2}{17}$ | 0.76 |

For $p \leq 0.05$, power=0.5, Xply by 10.0 — much as for $p \simeq 0.0075$
NB that $p \leq 0.05$ is much stronger evidence than $p = 0.05$








# Tukey on Data Analysis Foundations[1]

**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 inferences.

Accept inferences that have survived diverse challenges.

- Challenges to all aspects, not just the statistics

1: Tukey (1997)

# Types of Challenge

1. For experiments, critique/criticise the design.
2. Look for biases in processes that generated the data.
3. Look for inadequacies in laboratory procedure.
4. Examine any available model diagnostics (NB plots).
5. Check model on test data (Test data chosen how?)
6. Repeat the exercise (a training/test approach)
    - But how robust is the replication?

Also, critique the science underlying interpretation of 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 savvy 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 wastes effort** as other researchers follow false leads, or go down what had been identified (but not advertised) as blind alleys.



# 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
- Publish 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'