Murdoch Childrens Research Institute
Healthier Kids. Healthier Future.

# Misplaced confidence in confidence intervals?

John Carlin

Clinical Epidemiology & Biostatistics Unit, Data Science,
Murdoch Childrens Research Institute;
School of Population & Global Health, University of Melbourne

*Biometrics by the Harbour, 30-Nov-15*

THE UNIVERSITY OF MELBOURNE

VICBiostat

# Outline of talk

- Example of a simple confidence interval
- What is a confidence interval? Is everyone clear?
  - Evidence that CIs not interpreted correctly
- The reproducibility crisis in science
  - “Statistical reform movement” giving weight to interval estimation cf. testing
  - But shouldn’t interval estimates be “credible”?
- Bayesian credible intervals
  - Sometimes but not always similar to confidence intervals
  - Example illustrates desirable shrinkage…

# Example: PPOIT trial

- A small randomised trial on treatment for peanut allergy in children:
  - Active treatment = probiotics + peanut oral immunotherapy (PPOIT)
  - Control = placebo
- Primary outcome = sustained unresponsiveness (2-5 weeks after treatment discontinuation)to peanut challenge
- 62 children randomised (31 each arm): outcome available for 56 (28 each arm)

Tang M. et al. (2015). *Journal of Allergy and Clinical Immunology*.

# PPOIT trial results

| | Success | n | (%) |
|---|---|---|---|
| Active | 23 | 28 | (82.1) |
| Control | 1 | 28 | (3.6) |

- Conventional reporting is as “risk” ratio or odds ratio (OR); we choose OR for illustration…
- Standard calculation gives

$$\widehat{\mathrm{OR}} = 124$$

with 95% confidence interval (CI): (14, 1140)

# How should the CI be interpreted?

- Our short course notes would suggest as follows:

  “With 95% confidence, the true population OR lies between 14 and 1140”

[after: Kirkwood & Sterne, *Essential Medical Statistics*, 2003]

- In teaching & texts this formulation commonly given after discussion of sampling variability:
  - In repeated sampling, 95% of intervals calculated this way will include the true value
- IS THERE A LOGICAL LINK between the two??

# Common difficulty in ‘service’ teaching

- E.g. from a popular textbook:

Utts & Heckard, *Mind on Statistics* (2nd ed, 2004)

Definition (accompanied by discussion of repeated sampling):

*“A confidence interval is an interval of values computed from sample data that is likely to include the true population value”*

… followed by example:

*“…poll finding was that 57% of the dating teens had been out with somebody of another race or ethnic group. […] We have 95% confidence that somewhere between 52.5% and 61.5% of all American teens…”*

# The Fundamental Confidence Fallacy

If the probability that a random interval contains the true value is *X*%, then the plausibility or probability that a particular observed interval contains the true value is also *X*%, or, alternatively, we can have *X*% confidence that the observed interval contains the true value.

- Key confusion between “pre-data” sampling probability and “post-data” inference
  - Neyman (1937, 1941) was very clear that post-data inference is not possible within his theory!

Morey et al, *Psychon Bull Rev*, 2015

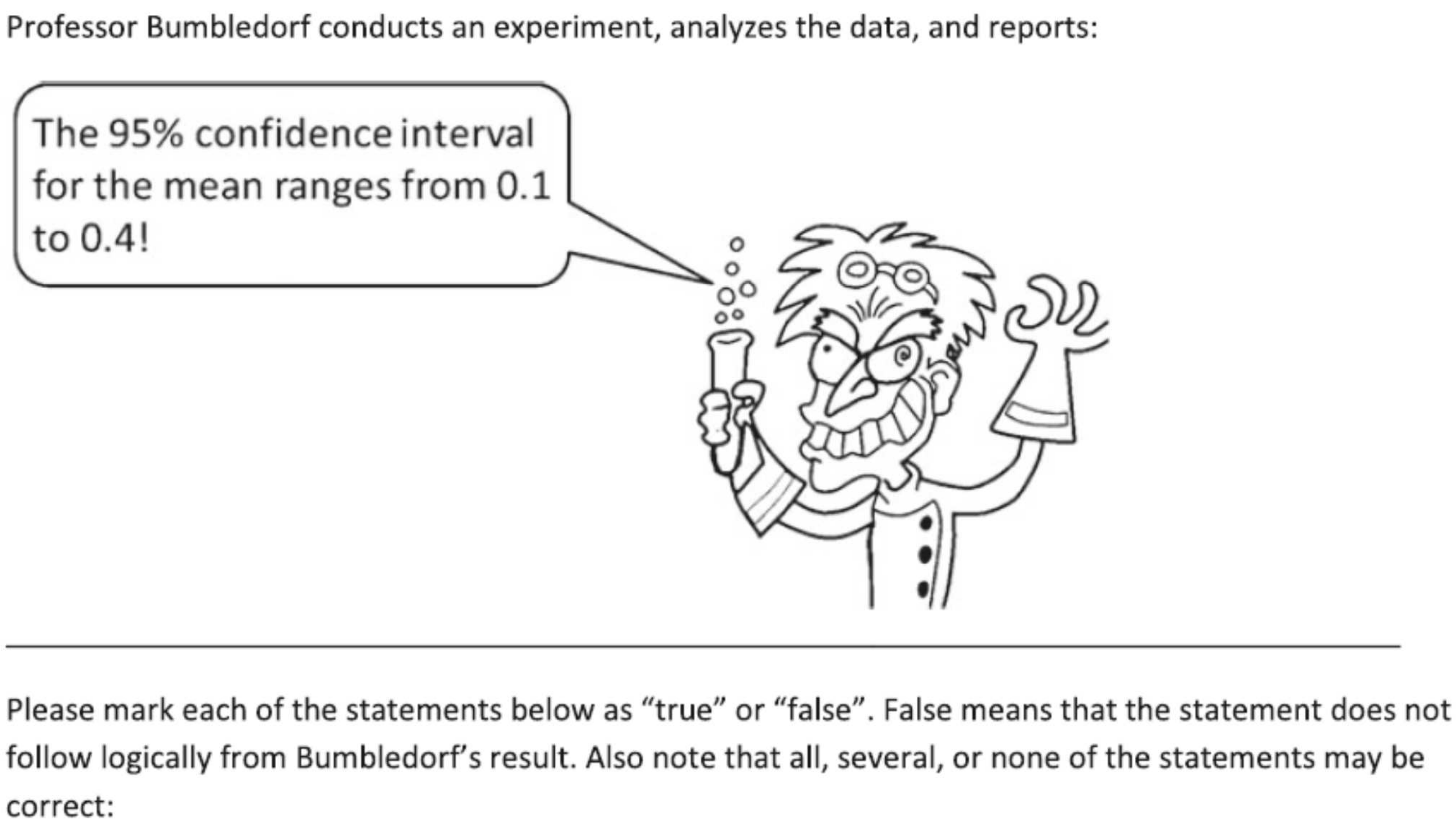
# I claim: the vast majority of CIs are interpreted as “post-data” inferences

Evidence?

- Introductory texts and courses invariably glide from the precise frequentist “pre-data” interpretation to a post-data version (as above)
- In actual practice, surely CIs are interpreted as having meaning for the particular data in hand
- Repeated empirical experiments demonstrate…

e.g. “Robust misinterpretation of confidence intervals” (Hoekstra et al, *Psychon Bull Rev*, 2014)

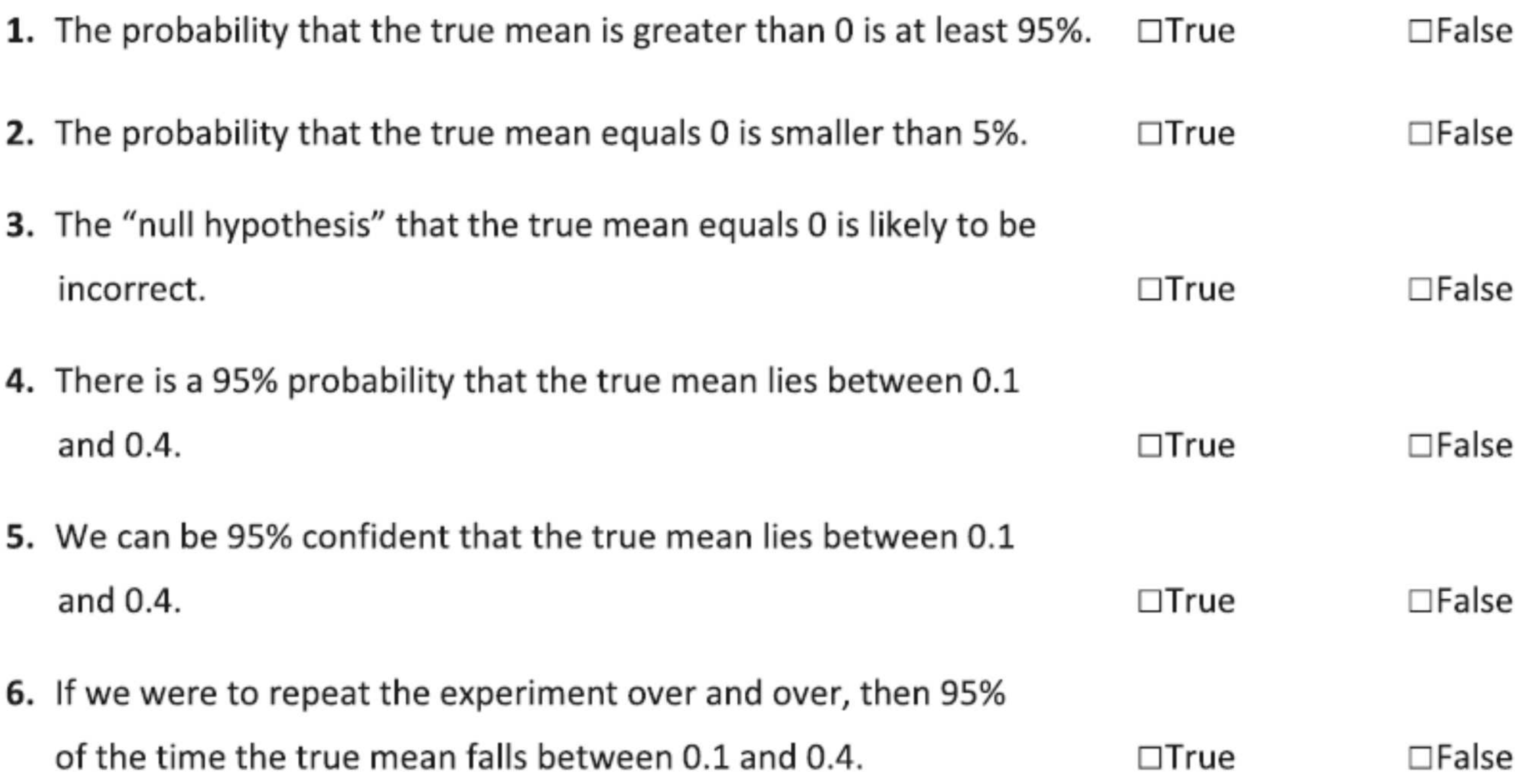
# Empirical evidence of the chaos



Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval for the mean ranges from 0.1 to 0.4!

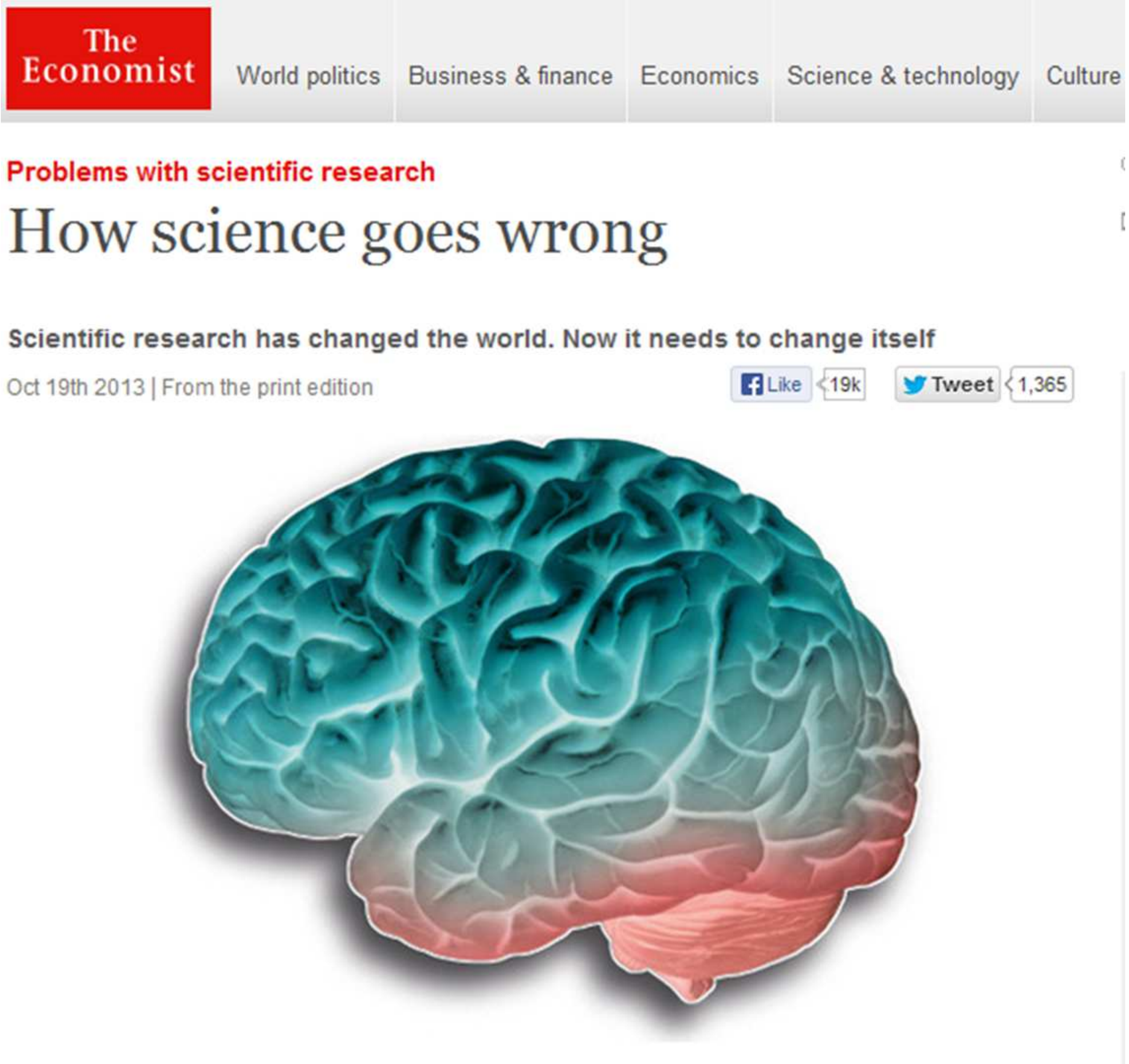
Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:



1. The probability that the true mean is greater than 0 is at least 95%. □True □False
2. The probability that the true mean equals 0 is smaller than 5%. □True □False
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect. □True □False
4. There is a 95% probability that the true mean lies between 0.1 and 0.4. □True □False
5. We can be 95% confident that the true mean lies between 0.1 and 0.4. □True □False
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4. □True □False

- All statements are false, but on average 3.5 were endorsed as true by respondents
  - Whether first-year students, Masters students or established researchers

# The reproducibility crisis in science



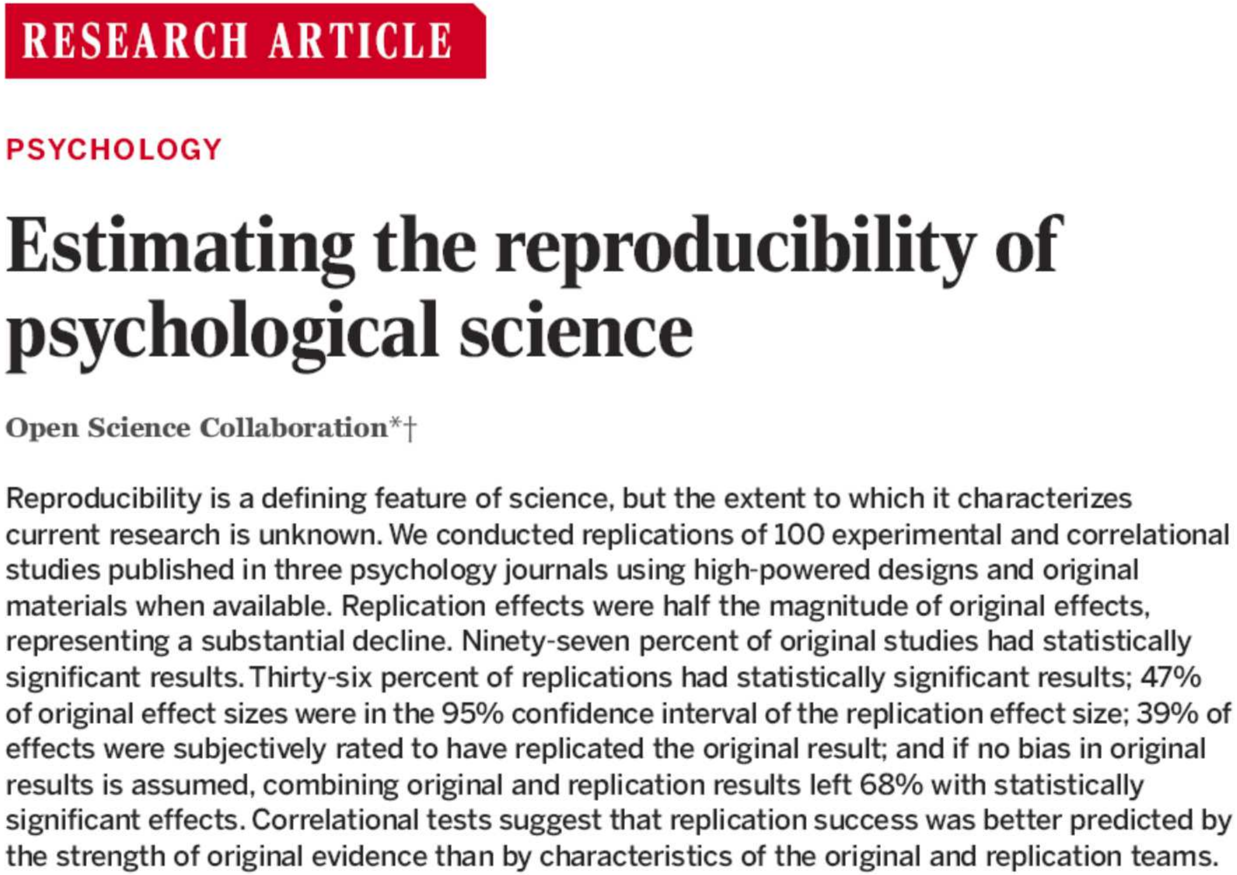
Problems with scientific research

## How science goes wrong

**Scientific research has changed the world. Now it needs to change itself**

Oct 19th 2013 | From the print edition

Replicating psychology experiments pp. 909 & 943
Why it pays to believe in a punitive god
Engineering cooperation among cells
Science
$10
28 AUGUST 2015
AAAS
Cellular proteins in live action
Pushing the resolution of structured illumination microscopy



RESEARCH ARTICLE

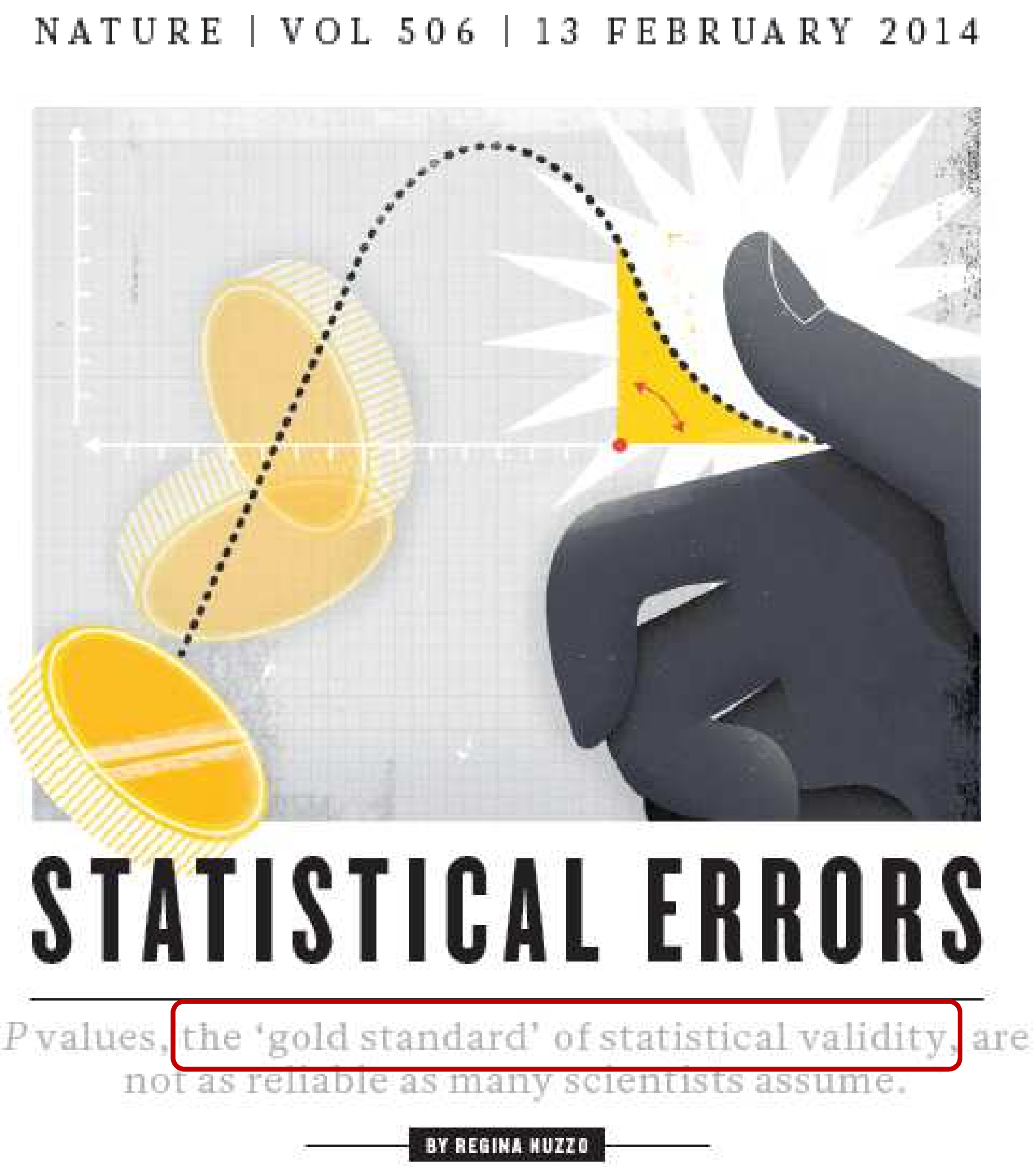
PSYCHOLOGY

# Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

28 August 2015



NATURE | VOL 506 | 13 FEBRUARY 2014

# STATISTICAL ERRORS

*P* values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

# The reproducibility crisis

Many scientific claims cannot be replicated

- Well documented examples from laboratory science (Begley, *Nature* 2012) & psychology

- WHY?
  - Pressures to publish, pressures to be first/ original/ novel
  - Peer review process imperfect
- Statistics done badly: in particular, significance tests widely misunderstood and misused
  - Confidence intervals widely promoted as preferred alternative…

# Much discussion of reform, especially in psychology…

Editorial, *Basic & Applied Social Psychology*, Feb-2015

- "… authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about 'significant' differences  or lack thereof, and so on)."
- "… confidence intervals also are banned from BASP"
- "with respect to Bayesian procedures…" (less clear!)
- Apparently only descriptive statistics allowed…

- Higher-profile changes at *Psychological Science*
  - Cumming (2014): strong emphasis on confidence intervals

# Priorities for statistical reform

- Multiple misinterpretations and misuses of *P*-values:
  - Main culprit: the false dichotomy of “statistically significant” (0.05 or other)
  - Null hypotheses themselves often represent false dichotomy
- Proponents of statistical reform see confidence intervals as providing a distinct alternative to “NHST”
  - But the actual theory is the same: a CI is no more than the set of parameter values “not rejected”

# BUT: I just remembered I was a Bayesian!

- Bayesian inference naturally produces "credible intervals"
- Can confidence intervals also be credible?
- Answer: yes, in many settings where there is a "pivotal quantity" e.g.
  - normal means (*t*-distribution)
  - approximate normal likelihood-based inference

$$\frac{\hat{\theta}-\theta}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}} \sim \mathrm{N}(0,1)$$

# Where “lazy Bayes” fails

Safe settings for above equivalence (“lazy Bayes”)?

- Estimating means & similar; large samples

Problem settings:

- Estimating parameters on bounded domains such as variances or their ratios
  - Boundaries of parameter space give difficulties (confidence interval width not reflecting precision of estimation), e.g. Morey et al (2015)
- Standard problems where *n* is small
  - Back to my example…

# PPOIT example: results

|  | Success | n | (%) |
|---|---|---|---|
| Active | 23 | 28 | (82.1) |
| Control | 1 | 28 | (3.6) |

- Standard calculation (likelihood approx$^{n}$)

$$\widehat{\mathrm{OR}} = 124, \text{ 95\% CI: } (14, 1140)$$

  - Probably not very valid in frequentist terms!
  - One alternative is so-called exact method:

$$\widehat{\mathrm{OR}} = 124, \text{ 95\% CI: } (13, 5290)$$

  - Even more obvious fail of common-sense test!

# Example: let's be Bayesian

- Treat problem in logistic regression framework

$$\mathrm{logit}(\Pr(\mathrm{success})) = \beta_0 + \beta_1 I[trt = \mathrm{active}]$$

- Need a prior distribution…
  - – Intercept parameter ($\beta_0$) – diffuse prior
  - – Log OR ($\beta_1$) – consider what's known/likely in this field: clinical optimist might think 10-fold OR plausible, so we set SD(log OR) = log(10) = 2.3 (with mean = 0)
- Exact Bayesian computation (Stata 14.1) gives

$\widehat{\mathrm{OR}} =$ 37, 95% credible interval: (9.8, 176)

# Summary of example

- The Bayesian credible interval depends on the prior distribution
  - So it should, to be credible! (the likelihood function is not sharply peaked)
- Even a modestly informative prior produces sensible shrinkage
  - "…the first randomized placebo-controlled trial evaluating the novel co-administration of a probiotic and peanut OIT and assessing sustained unresponsiveness in children with peanut allergy"
- Results from small studies could always use some shrinkage!

# Overall summary

- Statistical reasoning is the basis of many scientific claims to knowledge
- The frequentist theory of confidence intervals is counter-intuitive & arguably not helpful in practice
- True credible intervals require a Bayesian framework
  - In many problems a credible interval will be similar to a “standard” confidence interval
  - But when there’s a difference it can matter…
  - Teaching should at least acknowledge the issue
- BUT: beware Bayesian snake oil!

# Acknowledgements & References

- Andrew Gelman: www.andrewgelman.com
- ASA Working Party on “the meaning and use of p-values”
- Fellow teachers of short courses on introductory (bio)statistics

Cumming G. (2014). The New Statistics: Why and How. *Psychological Science* 25(1): 7-29.

Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ. (2014) Robust misinterpretation of confidence intervals. *Psychon Bull Rev.* 21(5):1157-64. doi: 10.3758/s13423-013-0572-3.

Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. (2015) The fallacy of placing confidence in confidence intervals. *Psychon Bull Rev.* Oct 8. [Epub ahead of print]

Nuzzo R. (2014) Scientific method: Statistical errors. *Nature*, 506:150–152.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

Tang MLK et al. (2015) Administration of a probiotic with peanut oral immunotherapy: A randomized trial. *Journal of Allergy and Clinical Immunology* 135(3): 737-744.