The Significance Delusion: Inconvenient Truths about *P*-values

Dorothy Dickson, M.Sc.
Vaccine Testing Center

Research Tapas
15 November 2018

The University of Vermont
LARNER COLLEGE OF MEDICINE

Overview

P-value definition

Conclusions from *P*-values – true/false

The Significance Game

P-values versus NHST

Is my finding real?

The Reproducibility Crisis

What exactly is a *P*-value?

A *P*-value obtained from an experiment:

Probability of obtaining data as extreme as, or more extreme than, that observed

given that the null hypothesis is true

$$Pr(X>=x|H_o)$$

Conclusions from *P*-values: Valid or not?

1. *P* >0.05: Conclusion - there is no effect

The larger the P-value, the more the null effect is consistent with the observed data.

A null effect is not necessarily the most likely effect.

The effect best supported by the data from a given experiment is always the observed effect, *regardless of its statistical significance*

Conclusions from *P*-values: Valid or not?

2. $P < 0.05$ - the finding is scientifically important

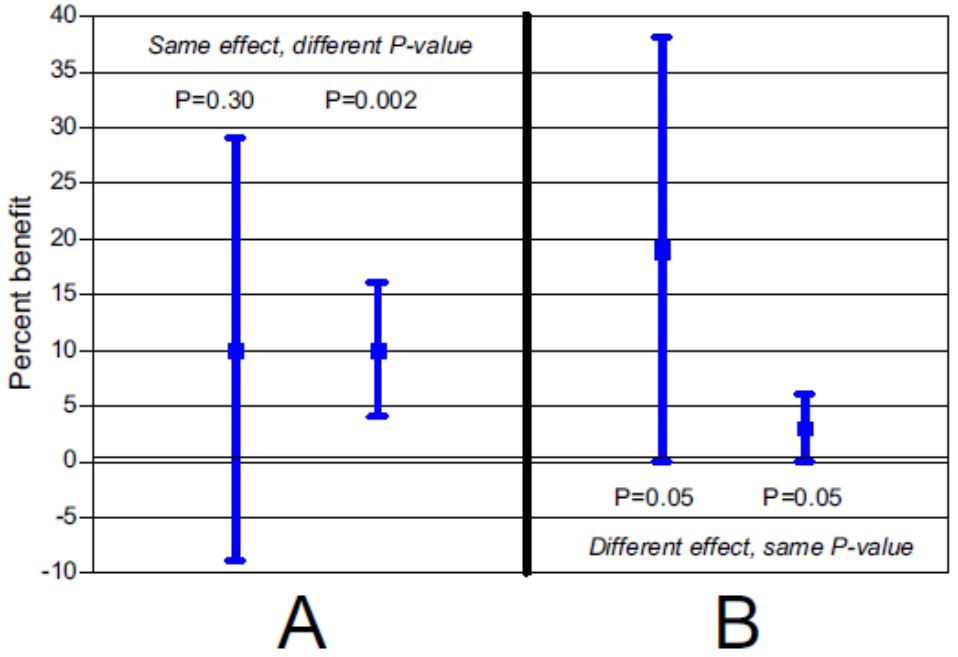
Especially when a study is large, very minor effects or small assumption violations can give rise to small *P*-values.

Do not confuse scientific and statistical significance – they are completely different things.

Conclusions from *P*-values: Valid or not?

3. Studies with *P*-values on opposite sides of 0.05 are conflicting

Even when effect sizes are identical, the *P*-values can differ enormously



Conclusions from *P*-values: Valid or not?

4. *P* = 0.02 - there is only a 2% probability the null hypothesis is true
5. *P* = 0.03 - there is a 3% probability my result is due to chance

The *P*-value says nothing about whether the null is true or false or due to chance because

By definition the *P*-value is a probability calculated GIVEN the null is true (chance only at play)

Conclusions from *P*-values: Valid or not?

6. $P < 0.05$ – the null hypothesis is false

A low *P*-value indicates that your data are unlikely assuming a true null, but it cannot evaluate which of two competing cases is more likely:

a) The null is true (but your sample was unusual)

b) The null is false

Conclusions from *P*-values: Valid or not?

7. *P* =0.01 - under the null hypothesis, these data would occur 1% of the time

1% probability of your data OR MORE EXTREME data occurring, under the null hypothesis

Conclusions from *P*-values: Valid or not?

8. P =0.05 - if you reject the null hypothesis, the probability of a Type I error (α) is 5%

P and α are incompatible (see later)

P-values: False Statements

1. When testing two groups and $P>0.05$ - there is no effect X
2. $P<0.05$ - the finding is scientifically important X
3. Studies with resulting *P*-values on opposite sides of 0.05 are conflicting X
4. $P=0.02$ - there is a 2% probability the null hypothesis is true X
5. $P=0.03$ - there is a 3% probability my result is due to chance X
6. $P<0.05$ – the null hypothesis is false X
7. $P=0.01$ - under the null hypothesis, these data would likely occur 1% of the time X
8. $P=0.05$ - if you reject the null hypothesis, the probability of a Type I error (α) is 5% X

P-value Misinterpretations

Most serious of all *P*-value misconceptions is the false belief that the

probability of a conclusion being in error

can be calculated from the data in a single experiment, without reference to external evidence or plausibility of any underlying mechanism

Theory versus Practice

In theory: A *P*-value is a data-dependent continuous measure of evidence from a single experiment.

In practice: A P-value is often used as a decision making tool for strong, weak, and no evidence against a null hypothesis labeled using cut-offs typically at 0.01, 0.05 and 0.10.

The Significance Game

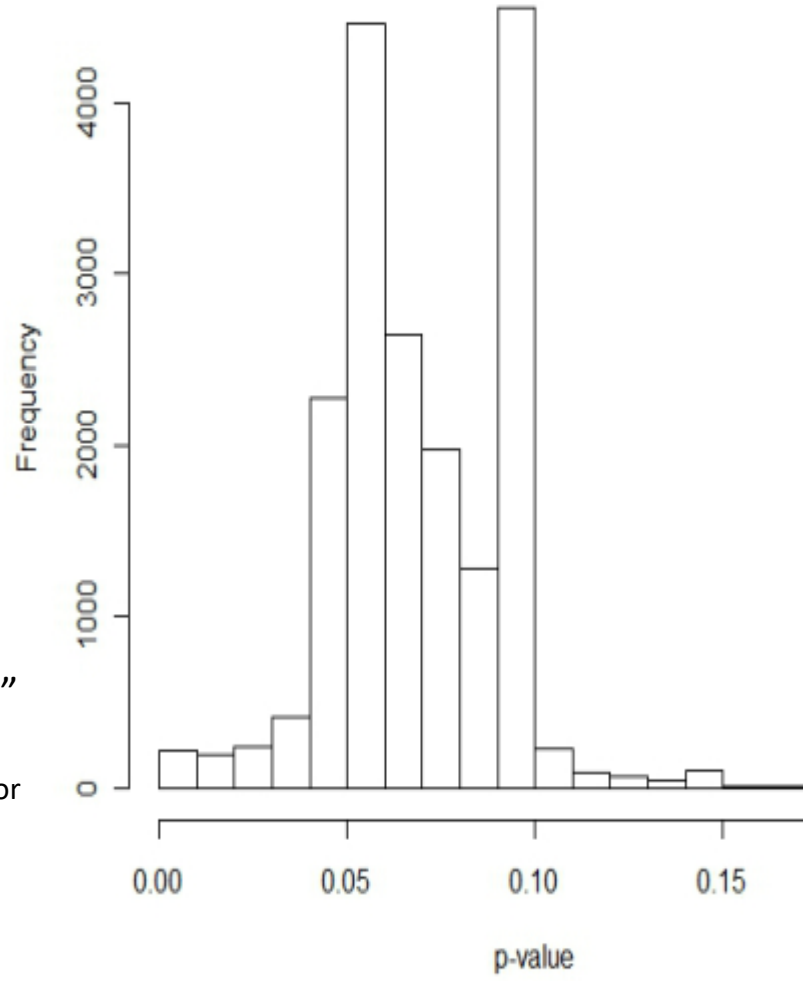
- marginally significant (p=0.056)
- almost significant (p=0.06)
- a suggestive trend (p=0.06)
- partially significant (p=0.08)
- borderline significant (p=0.09)
- fairly significant (p=0.09)
- hint of significance (p>0.05)
- approaching close to significance (p<0.1)

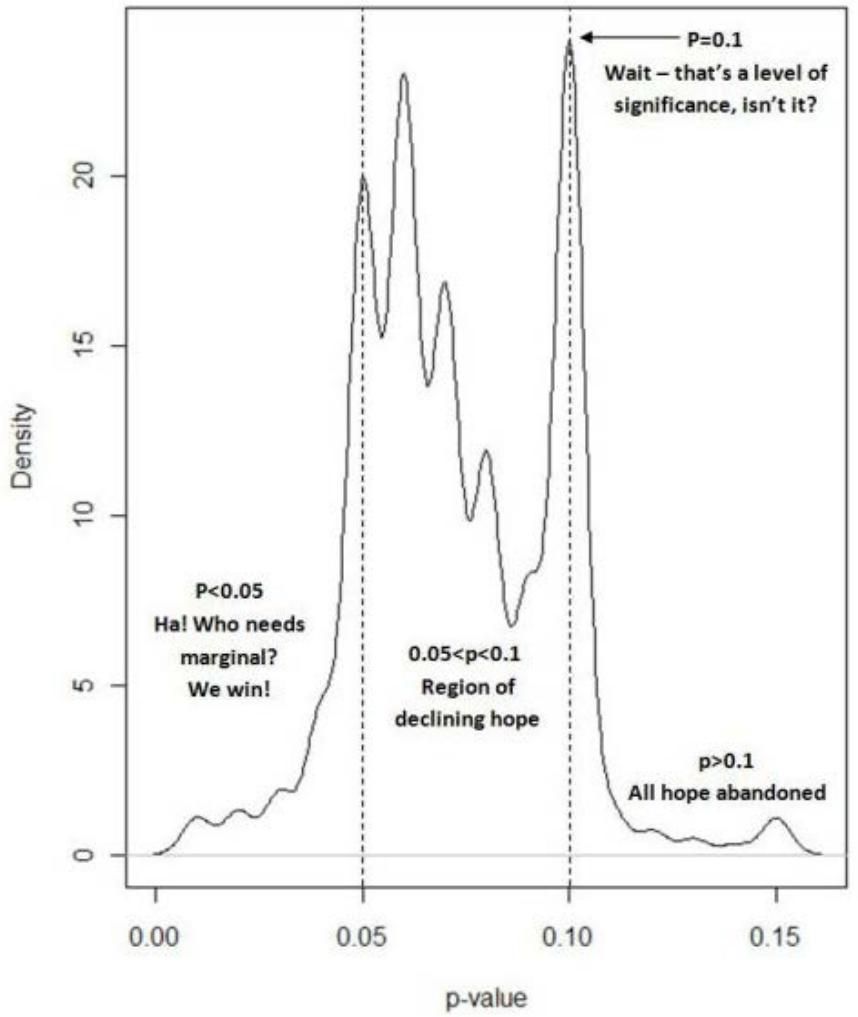
Everyone plays

“The Significance Game”

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

The Significance Game





Google Scholar search:
"Marginally significant (P=0.0*"
18,893 articles
https://mchankins.wordpress.com/author/mchankins/page/2/

The University of Vermont
LARNER COLLEGE OF MEDICINE

Blame the Statisticians

R. L. Fisher (1890 - 1962)

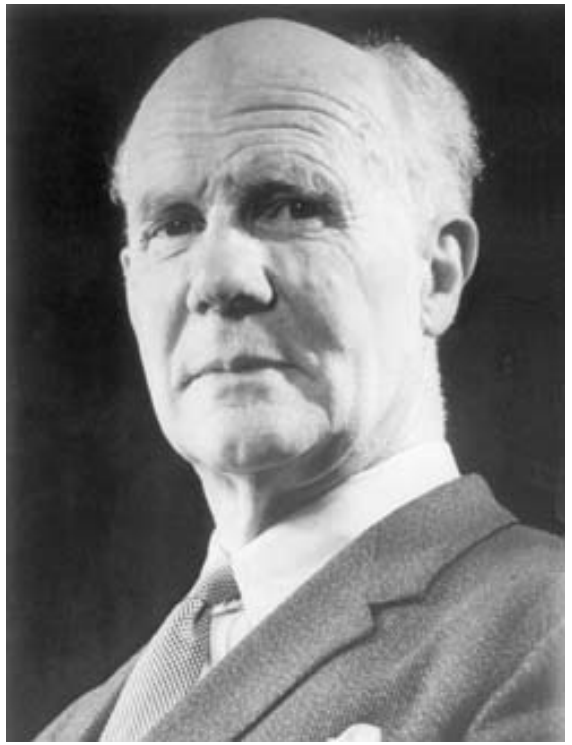
"Personally, the writer prefers to set a low standard of significance at the 5 percent point.
A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

Statistical Methods for Research Workers, 1925

P-value < 0.05 suggestive that the experiment is worthy of *a second look*

Blame the Statisticians

Jerzy Neyman (1894 - 1981)



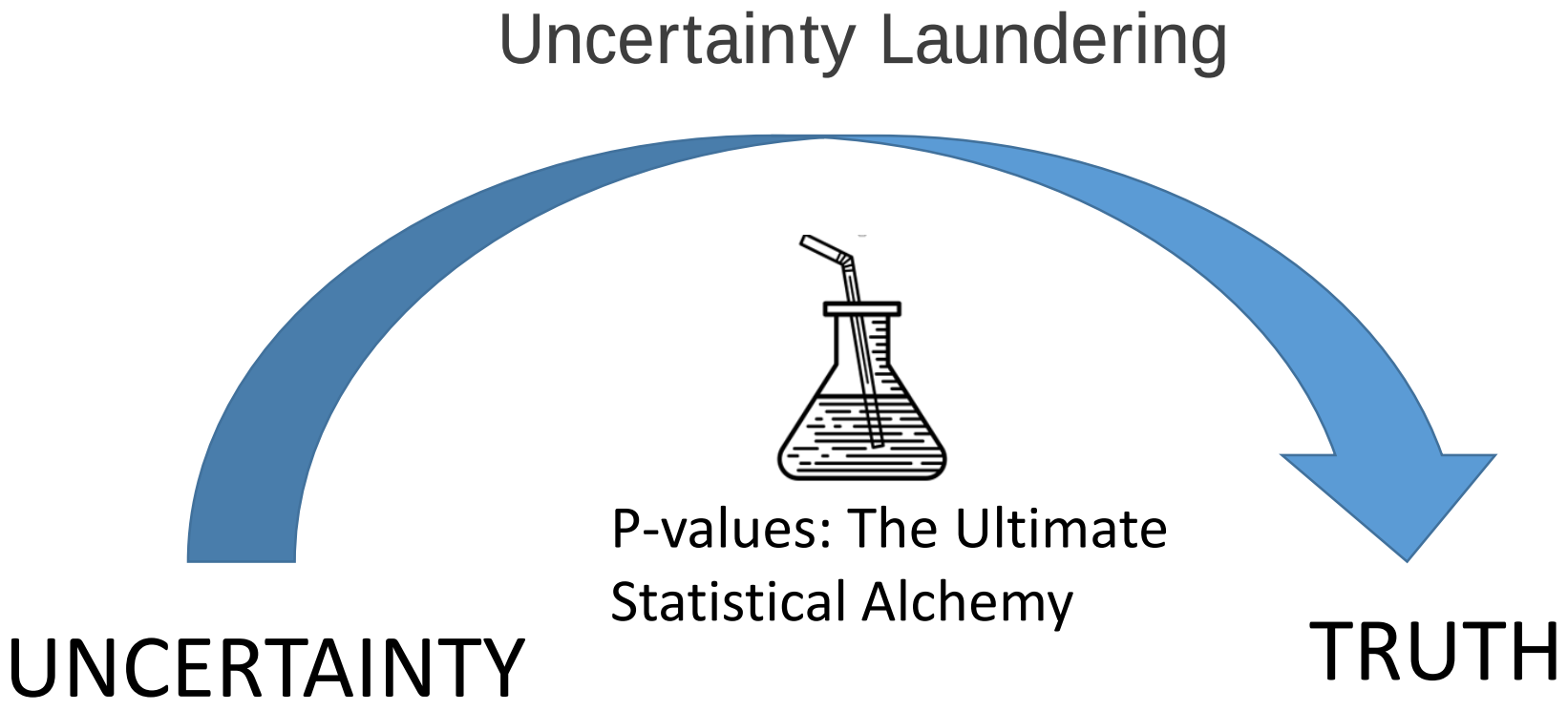
Egon S. Pearson (1895 - 1980)

Fisher vs. Neyman-Pearson

Fisher's p value	Hypothesis testing
Ronald Fisher	Jerzy Neyman and Egon Pearson
Significance test	Hypothesis test
p Value	α
The p value is a measure of the evidence against the null hypothesis	α and β levels provide rules to limit the proportion of errors
Computed a posteriori from the data observed	Determined a priori at some specified level
Applies to any single experiment	Applies in the long run through the repetition of experiments
Subjective decision	Objective behavior
Evidential, ie, based on the evidence observed	Nonevidential, ie, based on a rule of behavior

Note that Evidence *P* and Error α are **incompatible**

Statistical Alchemy



New Discovery

2010: A Psychology professor and his PhD student found evidence that Political Extremists perceive the world in black and white (figuratively and literally)

~2000 people
Moderates perceive shades of gray more accurately than those on the political Left or Right

P=0.01

Eureka!

Investigator suggests replicating the study

1300 participants
99% power to detect an effect of the original effect size at α = .05

P=0.59

Stupid reality!

Surprised?

The probability of replicating the original result was not 99%, as most might assume, but closer to 73%.

Or only 50%, if we wanted another 'very significant' ($P<0.01$) result.

Most Research Findings are False

The probability that a Hypothesis Test-based research finding is true or not depends on

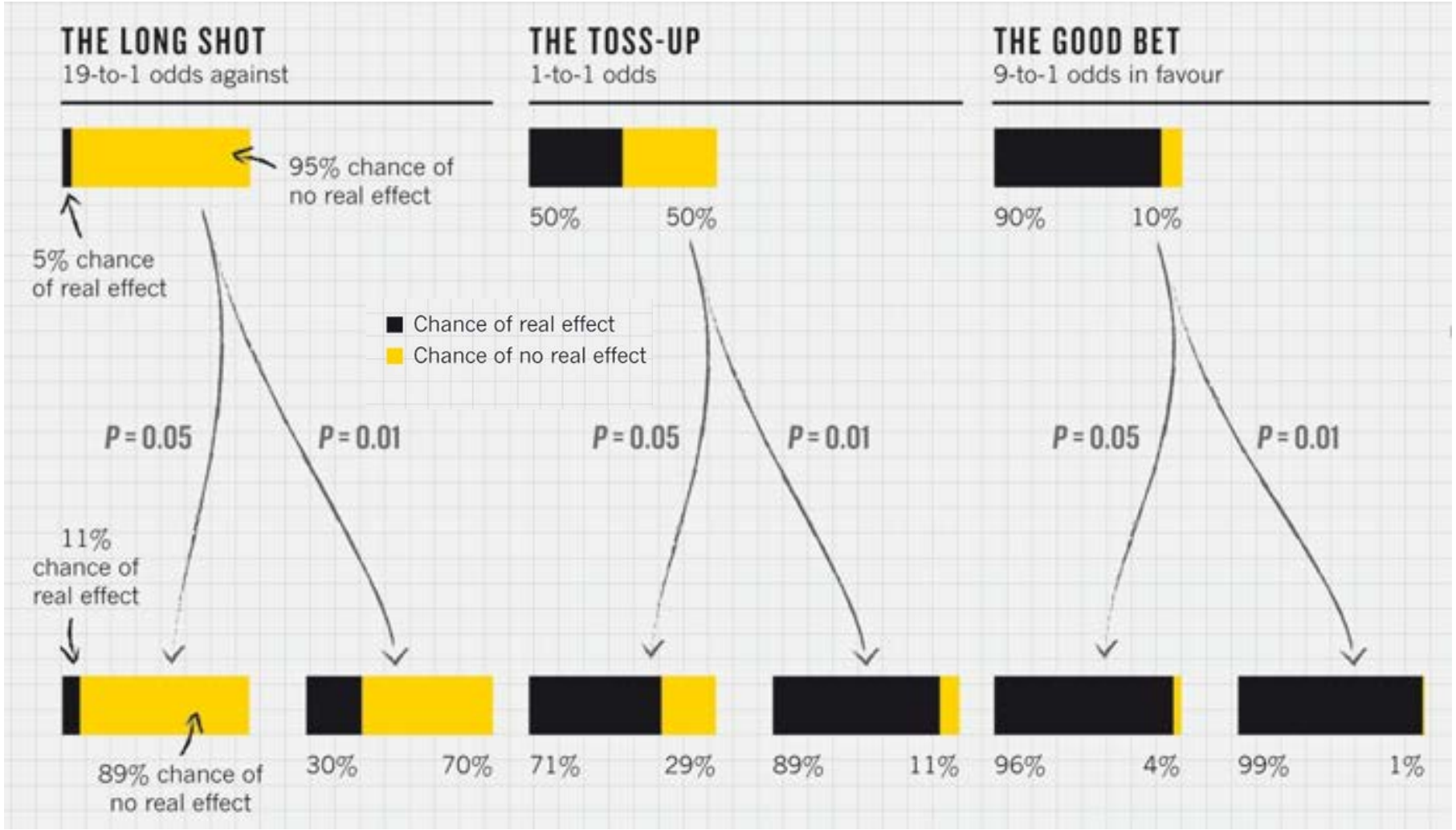
1. Statistical power
2. Level of statistical significance
3. PRIOR probability of it being true (before you even thought of your study)

What is the chance my finding is real?

Before the experiment

Measured *P*-value

After the experiment
(Bayesian interpretation)



R Nuzzo, Scientific method: statistical errors, Nature News 506 (7487), 150 (2014)

Why not report the false positive risk or FDR?

Researchers usually have no way of knowing what the prior probability is

Possible solutions:

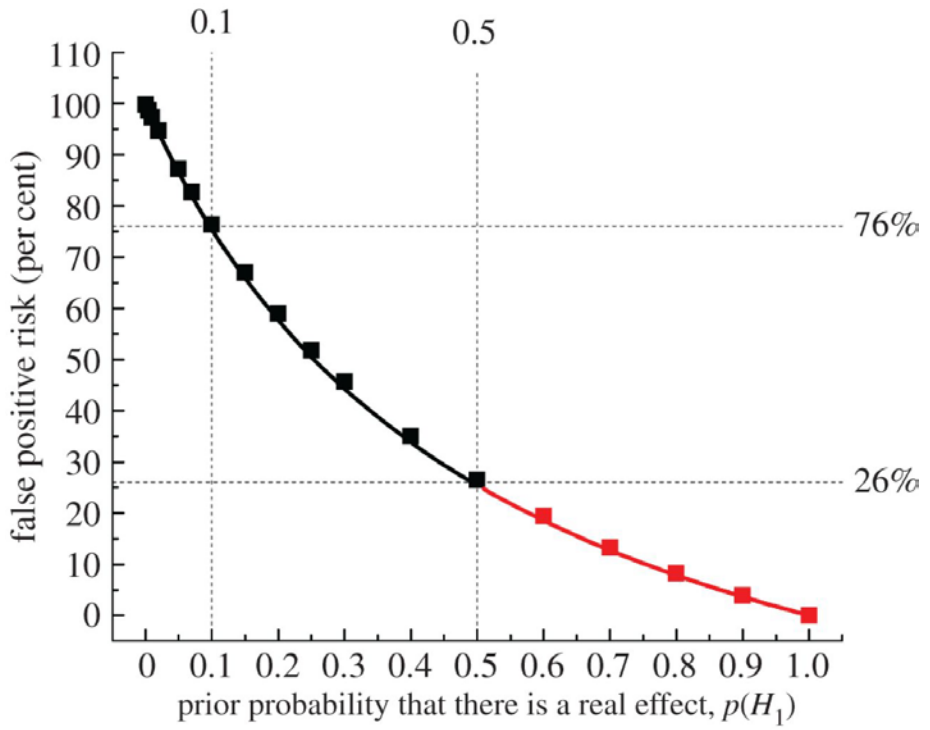
- Specify the prior probability needed in order to achieve an FDR of 5%, as well as providing the *P*-value and confidence interval
- Arbitrarily assume a prior probability of not more than 0.5 and calculate the minimum FDR for the observed *P*-value.

The False Positive Risk is always bigger than the *P*-value

How much bigger depends strongly on the plausibility of the hypothesis before the experiment

If the prior probability is low (10%), a *P* value close to 0.05 would have a FPR of 76%.

To lower that risk to 5% (what many folks falsely believe *P* = 0.05 means), the *P* value would need to be 0.00045.



The reproducibility of research and the misinterpretation of *p*-values
David Colquhoun; Royal Society Open Science 4:12 (2017)

Natural Selection of Bad Science

- Overreliance on p-values and significance testing in applied research has evolved into standard practice
- It often ignores magnitude of associations, estimation of precision, the consistency and pattern of results, possible bias arising from multiple sources, previous research findings, and foundational knowledge
- Many researchers have knowledge only to run statistical software that allows them to get their papers out quickly
- Researchers are rewarded primarily for publishing, so habits which promote publication are naturally selected and so are incentivized to increase publication (quantity over quality)
- Positive results are more likely to be published than negative results, particularly in high-impact journals
- False positives occur due to misinterpretations, poor theory, P-hacking, post-hoc hypotheses, biased selection of analyses, invalid assumptions
- The practice is misleading for inferences and intervention or policy decisions

Reproducibility Crisis in Science

P Values and Statistical Practice

Andrew Gelman

Reproducibility Crisis in Science

P Values and Statistical Practice

Andrew Gelman

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

Reproducibility Crisis in Science

P Values and Statistical Practice

Andrew Gelman

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing

Raymond HUBBARD and M. J. BAYARRI

Reproducibility Crisis in Science

P Values and Statistical Practice

Andrew Gelman

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. S[illegible][2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Good[illegible] · [illegible]las G. Altman[7]

Confusion Over Measures of Evidence (*p*'s) Versus Err[illegible] in Classical Statistical Testing

Raymond HUBBARD and M. J. BAYARRI

Common pitfalls in statistical analysis: "P" values, statistical significance and confidence intervals

Reproducibility Crisis in Science

P Values and Statistical Practice

Andrew Gelman

A Dirty Dozen: Twelve P-Value Misconceptions

Steven Goodman

...tistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

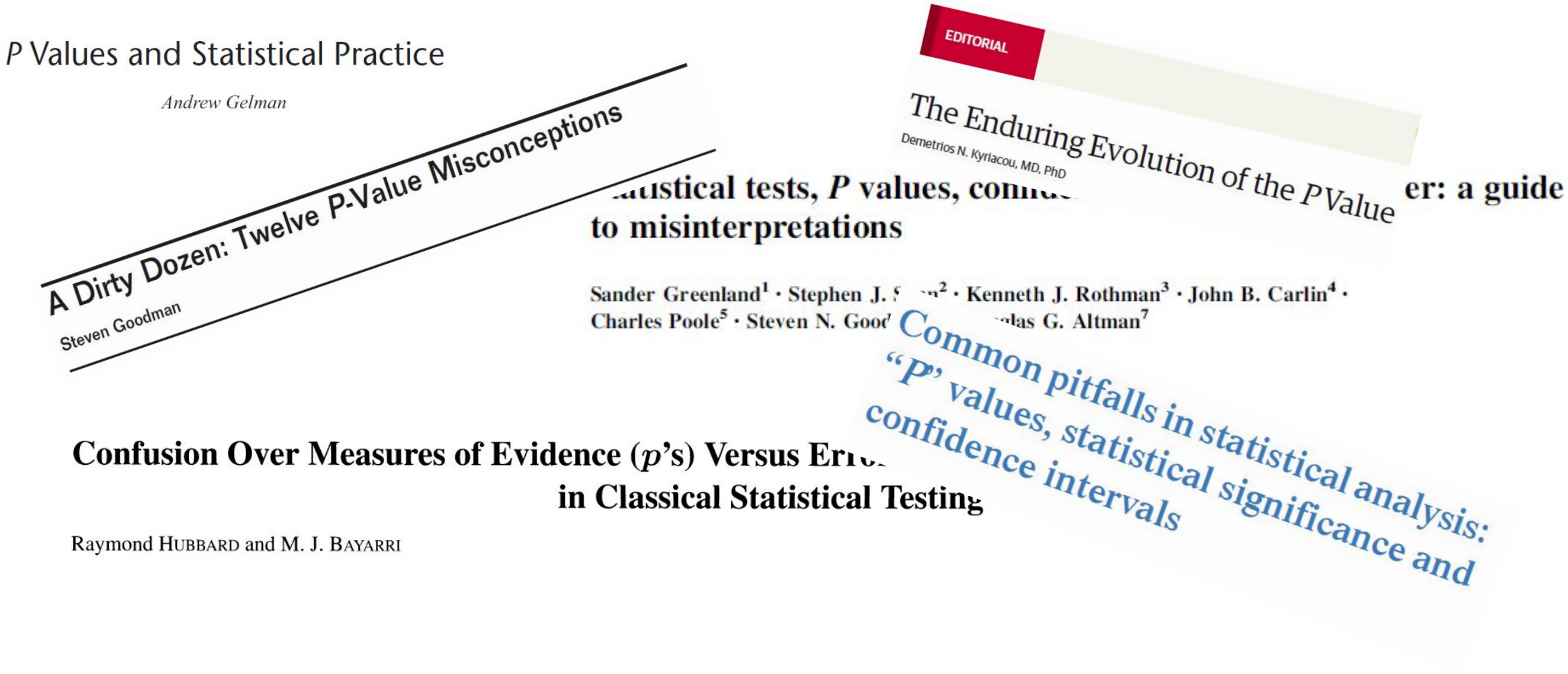
Sander Greenland[1] · Stephen J. S...[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Good... ...las G. Altman[7]

Common pitfalls in statistical analysis: "P" values, statistical significance and confidence intervals

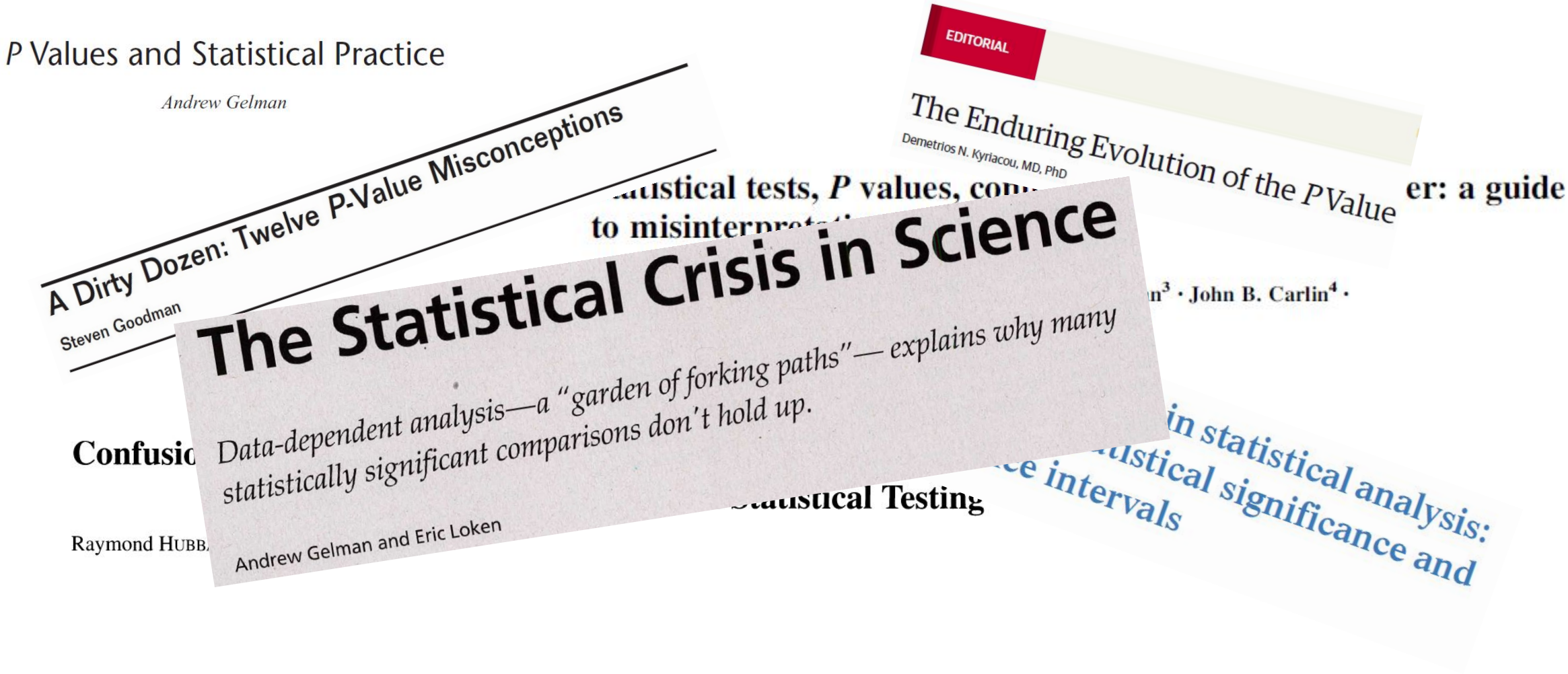
Confusion Over Measures of Evidence (*p*'s) Versus Err... in Classical Statistical Testing

Raymond HUBBARD and M. J. BAYARRI

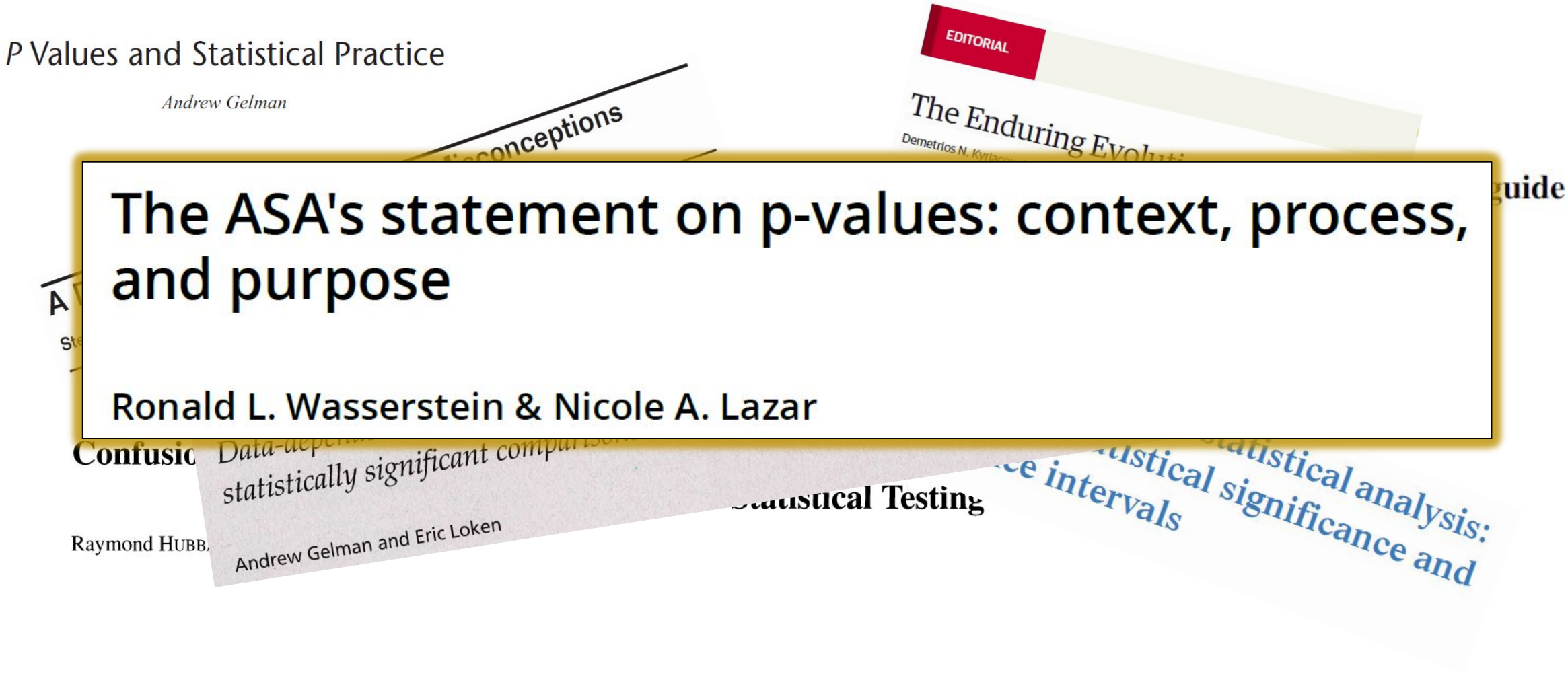
Reproducibility Crisis in Science



Reproducibility Crisis in Science



Reproducibility Crisis in Science



American Statistical Association Statement

Principles to Improve the Conduct and Interpretation of Quantitative Science

1. ***P*-values can indicate how incompatible the data are with a specified statistical model**
2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone**
3. **Scientific conclusions and business or policy decisions should not be based only on whether a *P*-value passes a specific threshold**
4. **Proper inference requires full reporting and transparency**
5. **A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result**
6. **By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis**

ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, Vol 70, Issue 2, 2016

Summary

- Hypothesis testing and *P*-values don't need to be banned, but they are limited tools that must be used and interpreted appropriately
- *P*-values and NHST should not used solely - more emphasis on complete reporting and estimation
- The *P*-value is not the false discovery probability
- The evidential strength of a result with a *P*-value of 0.05 or 0.01 is much weaker than the number suggests

Bibliography

1. Gelman A. *P* values and Statistical Practice, Epidemiology, 24 (1):69-72, 2013
2. Gelman, A Loken, E. The Statistical Crisis in Science, American Scientist 102:6 460-465, 2014
3. Greenland S, Senn S, Rothman K, Carlin J, Poole C, Goodman S, Altman D. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations, Eur J Epidemiol 31:337–350, 2016
4. Hubbard R, Bayarri MJ. Confusion over measures of evidence versus errors in classical statistical testing, The American Statistician, 57:3, 2003
5. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med 2(8): e124, 2005
6. Nuzzo R, Scientific method: statistical errors,Nature News 506 (7487), 150, 2014
7. Sellke T et. Al. Calibration of p Values for Testing Precise Null Hypotheses AM. STAT. 55, 62–71, 2001
8. Nosek BA., Spies, JR & Motyl M. Perspect. Psychol. Sci.7,615–631, 2012
9. Goodman S. A comment on replication, P values and Evidence, Stat. Med.11,875–879, 1992
10. Goodman S. Toward evidence-based medical statistics I. The P value fallacy Ann. Intern Med;130:995–1004, 1999
11. Goodman S. Towards evidence-based medical statistics. II. The Bayes Factor. Ann Intern Med; 130: 1005–1013, 1999
12. Goodman S. Of P-values and bayes: A modest proposal, Epidemiology12, 295–297, 2001
13. Goodman S. A dirty dozen: twelve p-value misconceptions, Semin Hematol, 45:3, 135-40, 2008
14. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values, Royal Society Open Science 1:3, 2014
15. Colquhoun D. The reproducibility of research and the misinterpretation of *p*-values ; Royal Society Open Science 4:12, 2017
16. Schervish, MJ. P values: What they are and what they are not, American Statistician 50:3, 1996
17. Kyriacou, DN. The Enduring Evolution of the P Value, JAMA, 315:11, 1113-5, 2016
18. Leek J et al. Five ways to fix statistics Nature **551**, 557-559, 2017
19. ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, Vol 70, Issue 2, 2016