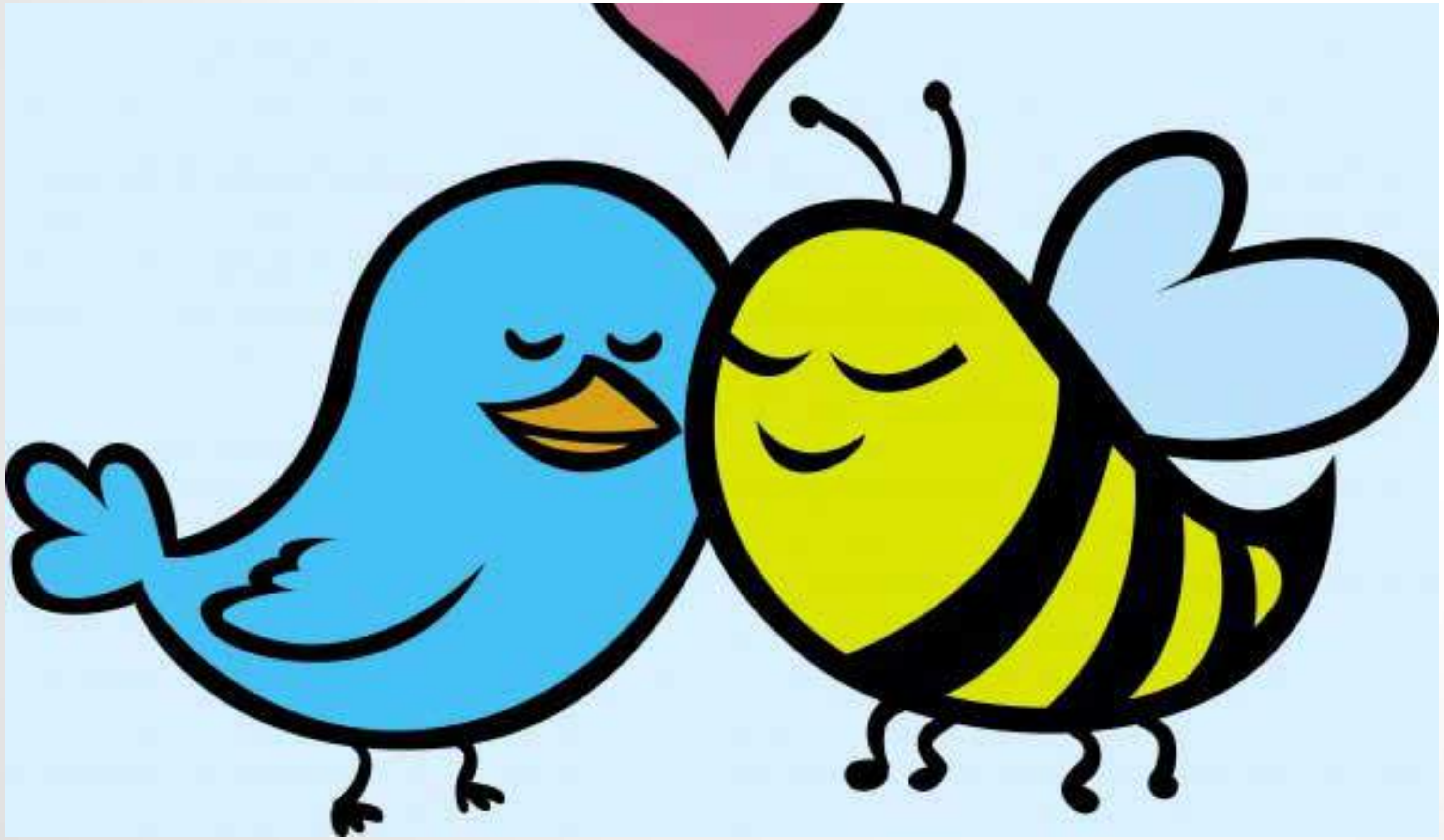
# P-Values Update

Allen Schirm

Ron Wasserstein

# Moving to a World Beyond $p < 0.05$

# The Talk



- They think they know all about it already, because they learned about it from others like them.
- It is not nearly as interesting as they thought it would be.
- They've stopped listening before you've stopped talking.
- Chances are, they now understand it even less.

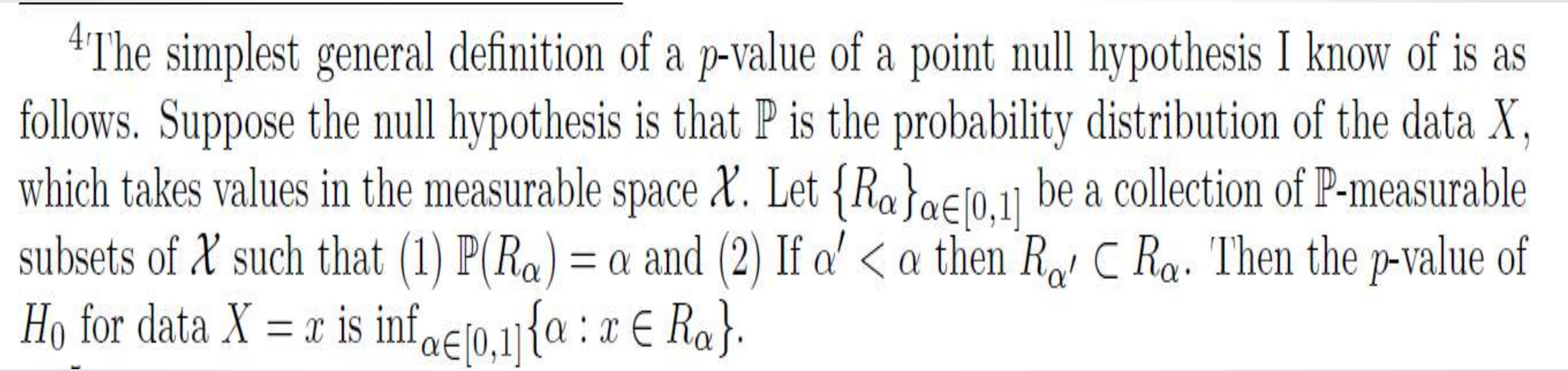
# P-value “clarified”
# (in the ASA Statement)

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

“That definition is about as clear as mud”

Christie Aschwanden, lead writer for science, *FiveThirtyEight*

# Perhaps this is clearer



[4] The simplest general definition of a $p$-value of a point null hypothesis I know of is as follows. Suppose the null hypothesis is that $\mathbb{P}$ is the probability distribution of the data $X$, which takes values in the measurable space $\mathcal{X}$. Let $\{R_\alpha\}_{\alpha \in [0,1]}$ be a collection of $\mathbb{P}$-measurable subsets of $\mathcal{X}$ such that (1) $\mathbb{P}(R_\alpha) = \alpha$ and (2) If $\alpha' < \alpha$ then $R_{\alpha'} \subset R_\alpha$. Then the $p$-value of $H_0$ for data $X = x$ is $\inf_{\alpha \in [0,1]}\{\alpha : x \in R_\alpha\}$.

(Stark, 2016)

# So, what is a p-value?

- We know some stuff
- We want to know some more
- We design an experiment to help us
- We collect data from the experiment
- We summarize the results
- Now what do we know?

# Summarizing the data

- We summarize the data into a number we call a “statistic”
- Compute a probability from that statistic – that’s the p-value
- If the p-value is small enough, we call it “statistically significant”
- What is small enough? Typically, 0.05.

# Example

- New treatment compared to placebo to improve TKV (a measure of health in kidney patients, in ml)
- After one year, difference in median TKV (treatment – placebo) = **96.8** (95% confidence interval (10.8 to 182.7))
- P-value computed to be **0.03** (0.027, but let's round off)

- **https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002777**

# Now home in on 0.03

- How did we get that number?

# To compute the p-value

- We assumed a bunch of stuff, including that there was no difference between the treatment and the placebo

# So what does the p-value mean?

- If there is no difference between the treatment and the placebo
- If everything else we assumed is also true
- Then the probability that we would observe the difference we found (96.8 ml), or one even larger, is 0.03

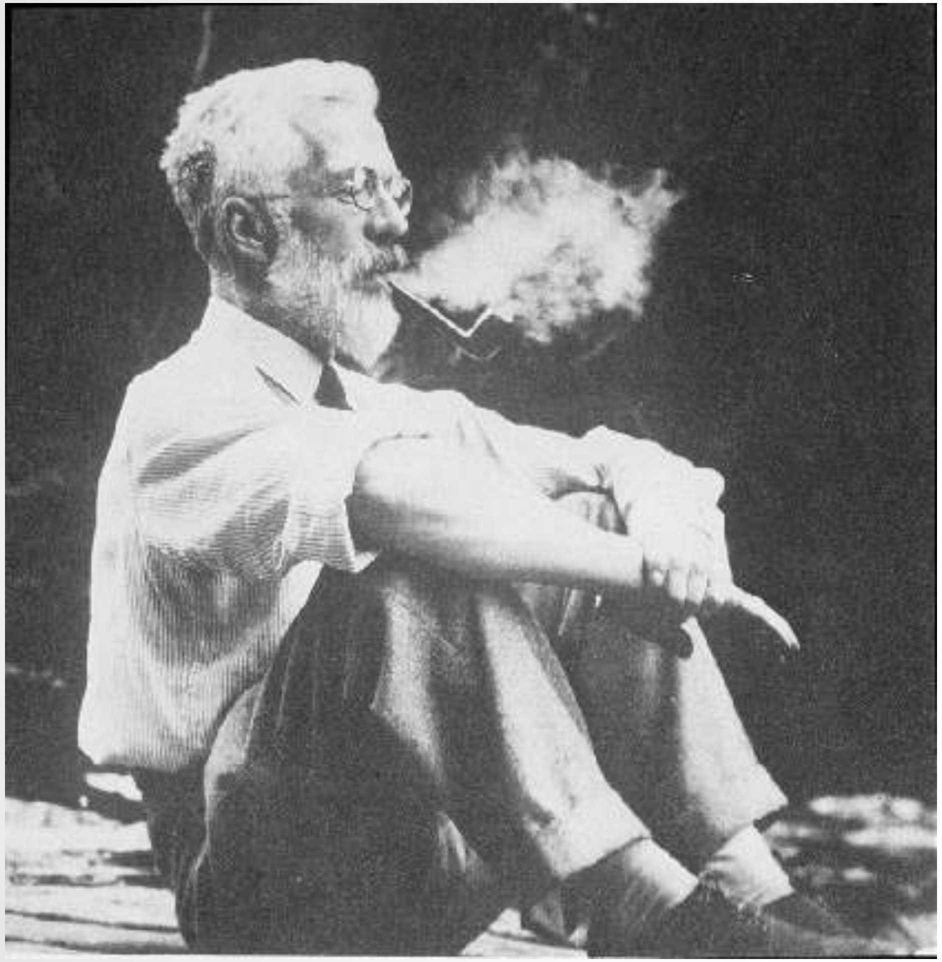
# Now, what can we conclude?

- We had really bad luck in our sample selection, or
- One or more of our assumptions was wrong
- Remember: one of those assumptions was that the treatment was no better than placebo

# So, our p-value was p=.03

- And that's smaller than .05
- And so it is…statistically significant

# Special Correspondent: Emily Latella

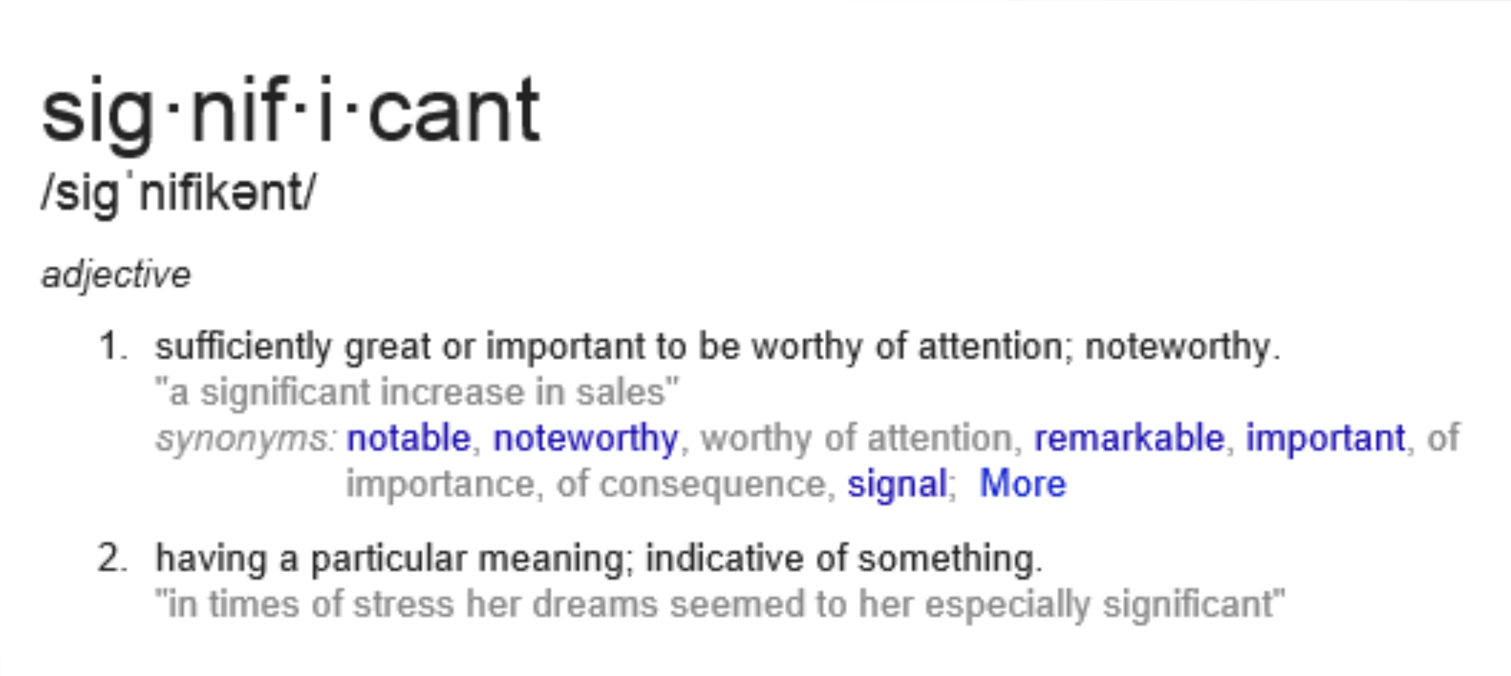


Returning to "small" p-values"

R. A. Fisher called such results "significant"

To Fisher, “statistical significance”

meant that the result was worth further scrutiny.

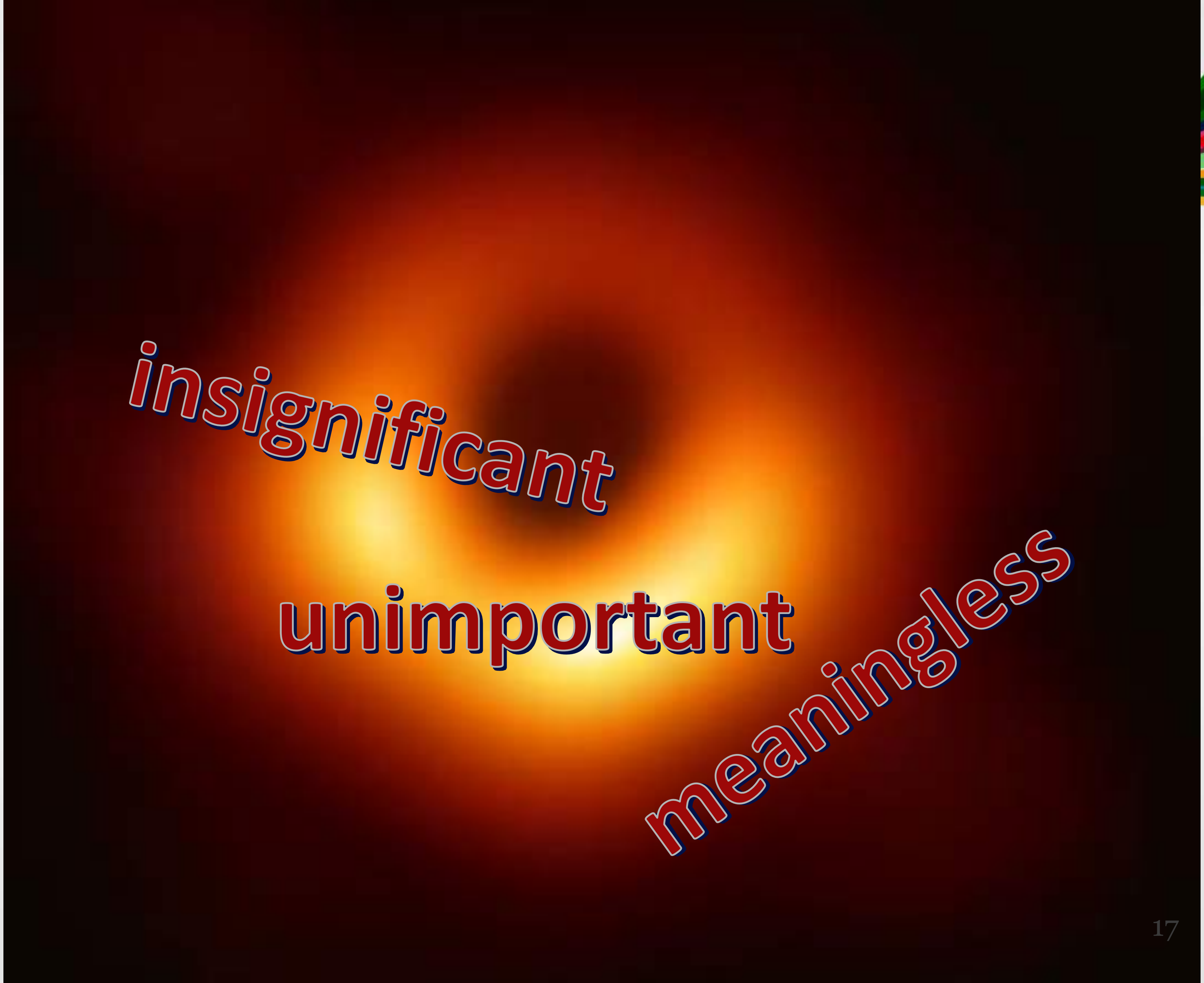


sig·nif·i·cant

/sigˈnifikənt/

*adjective*

1. sufficiently great or important to be worthy of attention; noteworthy.
   "a significant increase in sales"
   *synonyms:* notable, noteworthy, worthy of attention, remarkable, important, of importance, of consequence, signal; More
2. having a particular meaning; indicative of something.
   "in times of stress her dreams seemed to her especially significant"



insignificant

unimportant

meaningless

significant increase

significant event

significant other

# mole

The amount or sample of a chemical substance that contains as many constitutive particles, e.g., atoms, molecules, ions, electrons, or photons, as there are atoms in 12 grams of carbon-12

# Allan Rossman

“You keep using that word. I don’t think that it means what you think it means.” – Inigo Montoya

*“Just a Theory”: 7 Misused Scientific Words*, Scientific American, April 2, 2013 https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/

# Hypothesis

A proposed explanation that can be tested

Not an educated guess

“You keep using that word. I don’t think that it means what you think it means.” – Inigo Montoya

# Theory

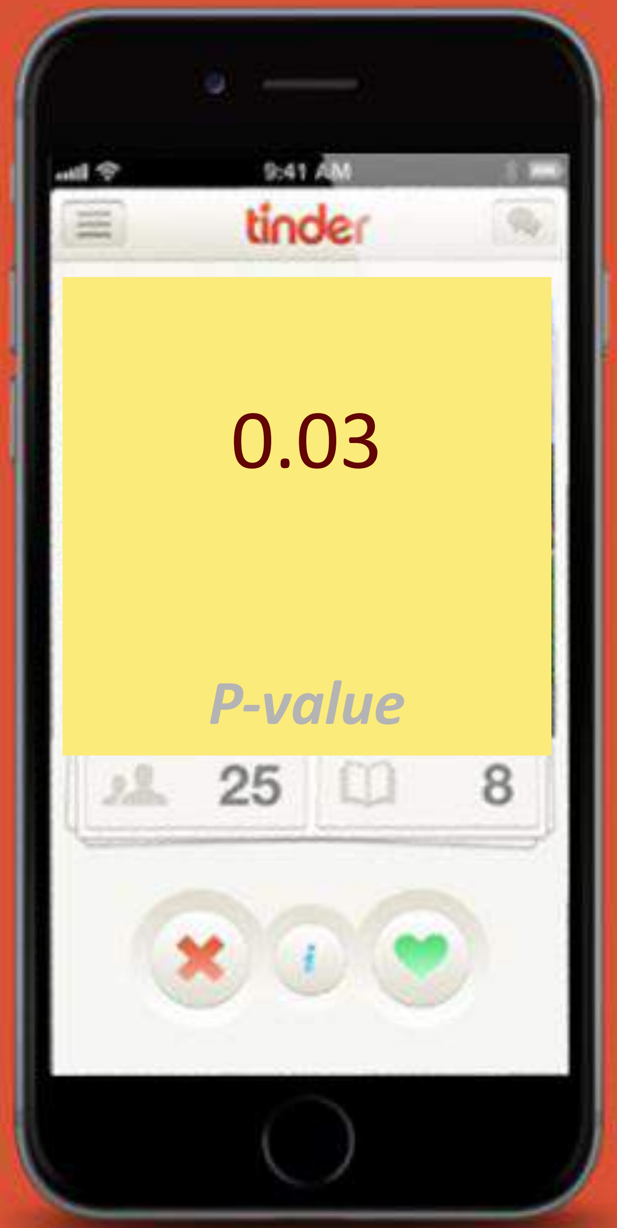
“You keep using that word. I don’t think that it means what you think it means.” – Inigo Montoya

An explanation of some aspect of the natural world that has been substantiated through repeated experiments or testing

# Word number 6:

“Significant”

tinder



0.03

*P-value*

SWIPE
RIGHT

My experimental results are interesting. I should spend more time with them, maybe repeat the experiment.  I may be on to something, but it will take time to be sure.

SWIPE
RIGHT

You tiny, beautiful p-value.  You are the result that I want to spent the rest of my life with. Let’s publish and get grants together.  I love you!

✓

# p equal or nearly equal to 0.06

- almost significant
- almost attained significance
- almost significant tendency
- almost became significant
- almost but not quite significant
- almost statistically significant
- almost reached statistical significance
- just barely below the level of significance
- just beyond significance

✓

# p equal or nearly equal to 0.08

- a certain trend toward significance
- a definite trend
- a slight tendency toward significance
- a strong trend toward significance
- a trend close to significance
- an expected trend
- approached our criteria of significance
- approaching borderline significance
- approaching, although not reaching, significance

# p close to but not less than 0.05

- hovered at nearly a significant level (p=0.058)
- hovers on the brink of significance (p=0.055)
- just about significant (p=0.051)
- just above the margin of significance (p=0.053)
- just at the conventional level of significance (p=0.05001)
- just barely statistically significant (p=0.054)
- just borderline significant (p=0.058)
- just escaped significance (p=0.057)
- just failed significance (p=0.057)

"... we only wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05.

Rosnow, R.L. and Rosenthal, R. 1989. Statistical procedures and the justification of knowledge and psychological science. *American Psychologist* 44: 1276-1284 https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

# Thanks to Matthew Hankins for these quotes

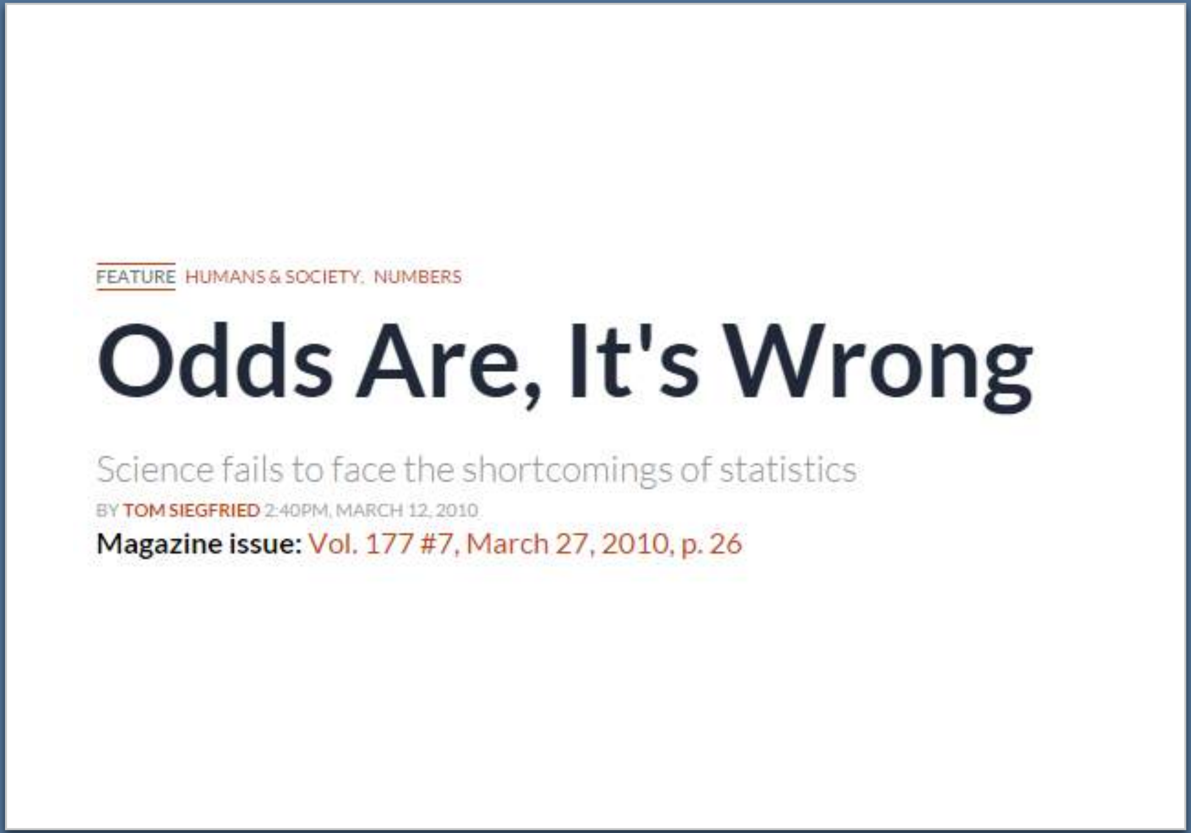
# Why the 2016 ASA statement?

- "It has been widely felt, probably for thirty years and more, that significance tests are overemphasized and often misused and that more emphasis should be put on estimation and prediction."
- Cox, D.R. 1986. Some general aspects of the theory of statistics. *International Statistical Review* 54: 117-126.
- A world of quotes illustrating the long history of concern about this can be viewed at David F. Parkhurst, School of Public and Environmental Affairs, Indiana University
- http://www.indiana.edu/~stigtsts/quotsagn.html

# “Let’s be clear. Nothing in the ASA statement is new.”

Statisticians and others have been sounding the alarm about these matters for decades, to little avail.

(Wasserstein and Lazar, 2016)



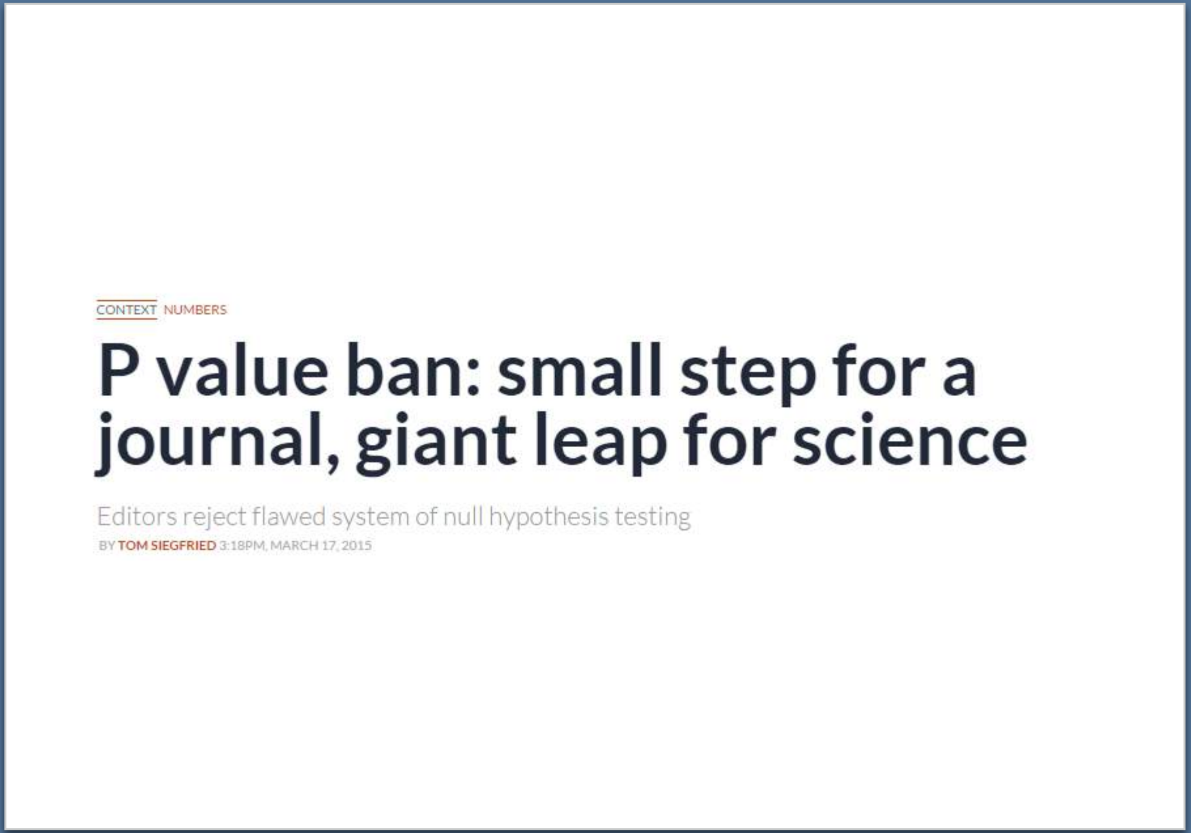
FEATURE HUMANS & SOCIETY, NUMBERS

Odds Are, It's Wrong

Science fails to face the shortcomings of statistics

BY TOM SIEGFRIED 2:40PM, MARCH 12, 2010

**Magazine issue:** Vol. 177 #7, March 27, 2010, p. 26



CONTEXT NUMBERS

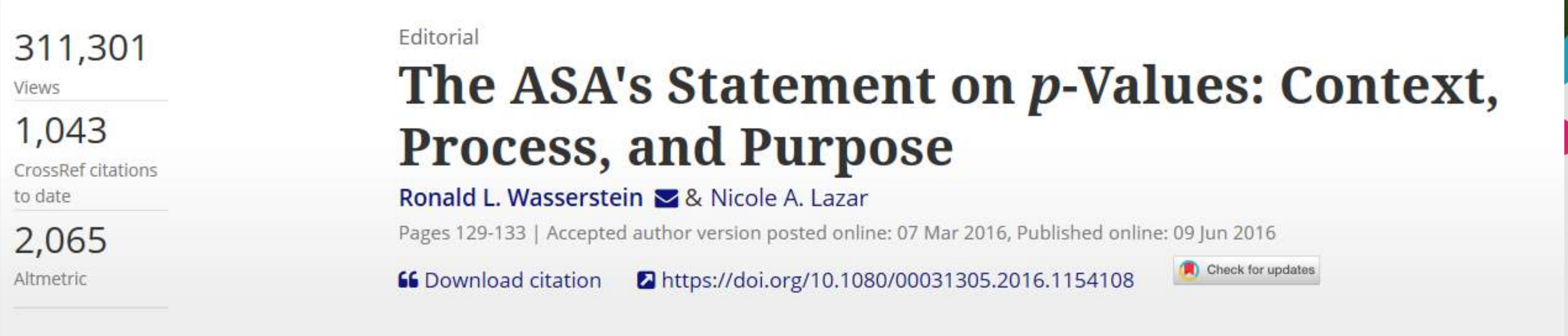
P value ban: small step for a journal, giant leap for science

Editors reject flawed system of null hypothesis testing

BY TOM SIEGFRIED 3:18PM, MARCH 17, 2015

# The ASA Statement on p-values and Statistical Significance

During the past century, though, a mutant form of math has deflected science’s heart from the modes of calculation that had long served so faithfully. Science was seduced by statistics, the math rooted in the same principles that guarantee profits for Las Vegas casinos. Supposedly, the proper use of statistics makes relying on scientific results a safe bet. But in practice, widespread misuse of statistical methods makes science more like a crapshoot.



311,301

Views

1,043

CrossRef citations to date

2,065

Altmetric

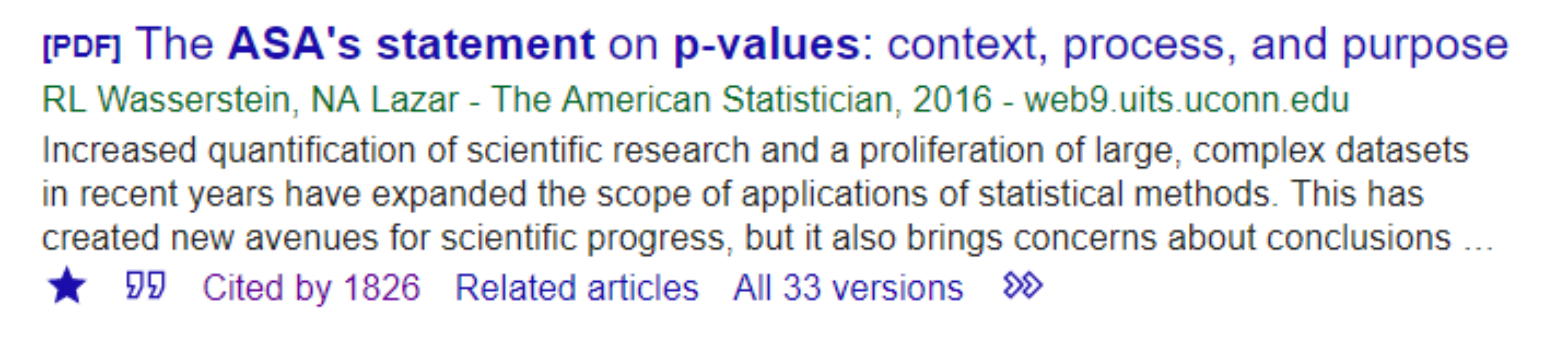
Editorial

# The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein ✉ & Nicole A. Lazar

Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online: 09 Jun 2016

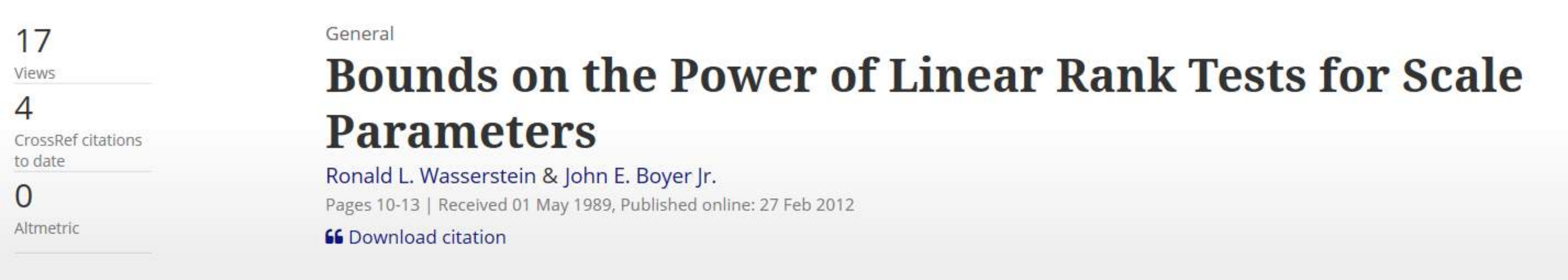
https://doi.org/10.1080/00031305.2016.1154108



[PDF] The **ASA's statement** on **p-values**: context, process, and purpose

RL Wasserstein, NA Lazar - The American Statistician, 2016 - web9.uits.uconn.edu

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions ...



17

Views

4

CrossRef citations to date

0

Altmetric

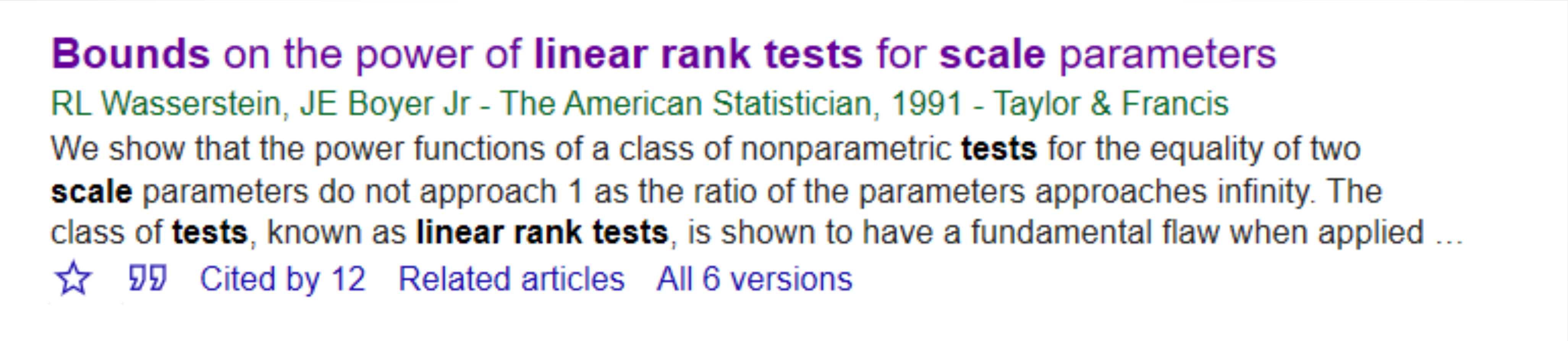
General

# Bounds on the Power of Linear Rank Tests for Scale Parameters

Ronald L. Wasserstein & John E. Boyer Jr.

Pages 10-13 | Received 01 May 1989, Published online: 27 Feb 2012

❝ Download citation

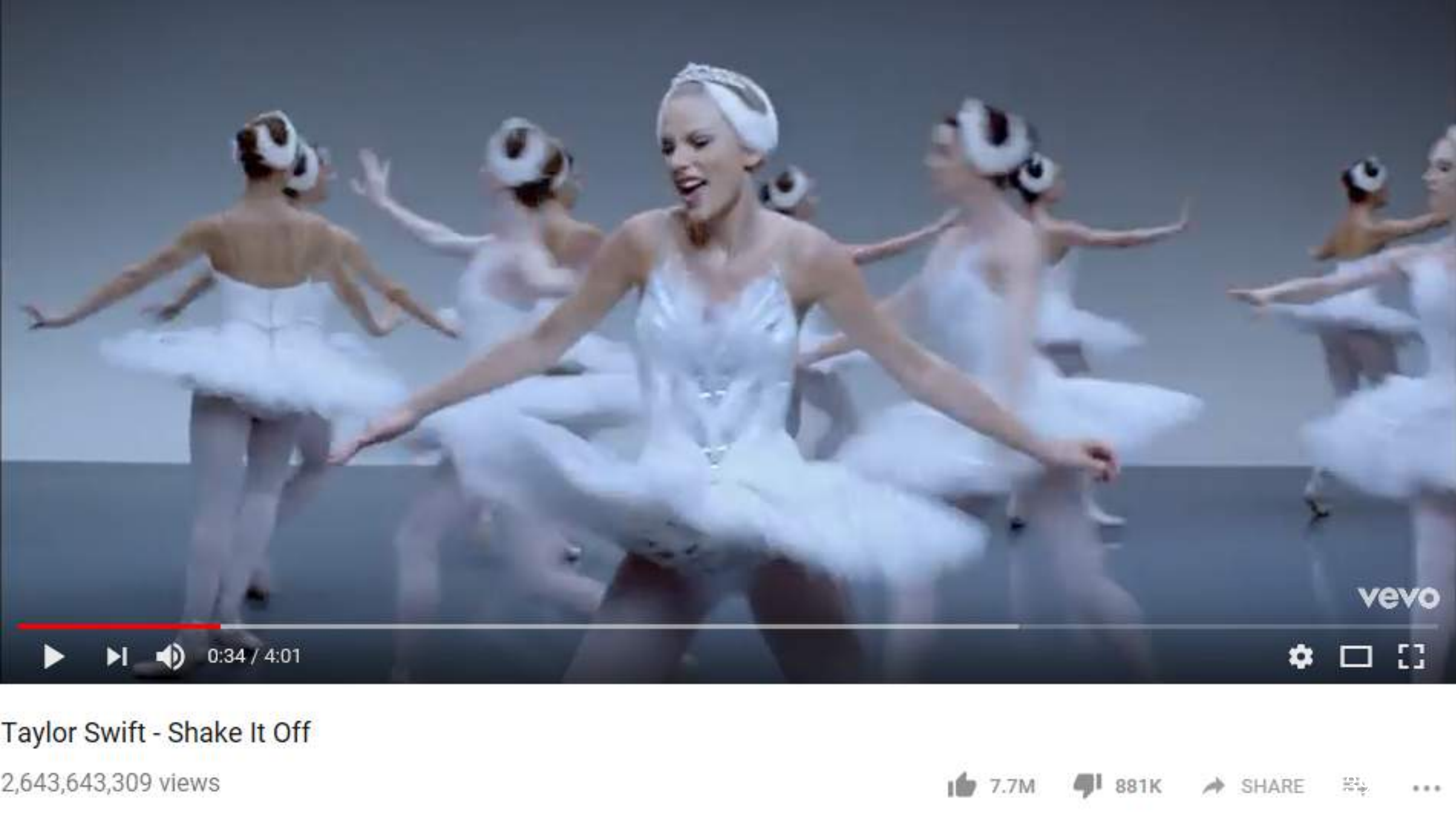


**Bounds** on the power of **linear rank tests** for **scale** parameters

RL Wasserstein, JE Boyer Jr - The American Statistician, 1991 - Taylor & Francis

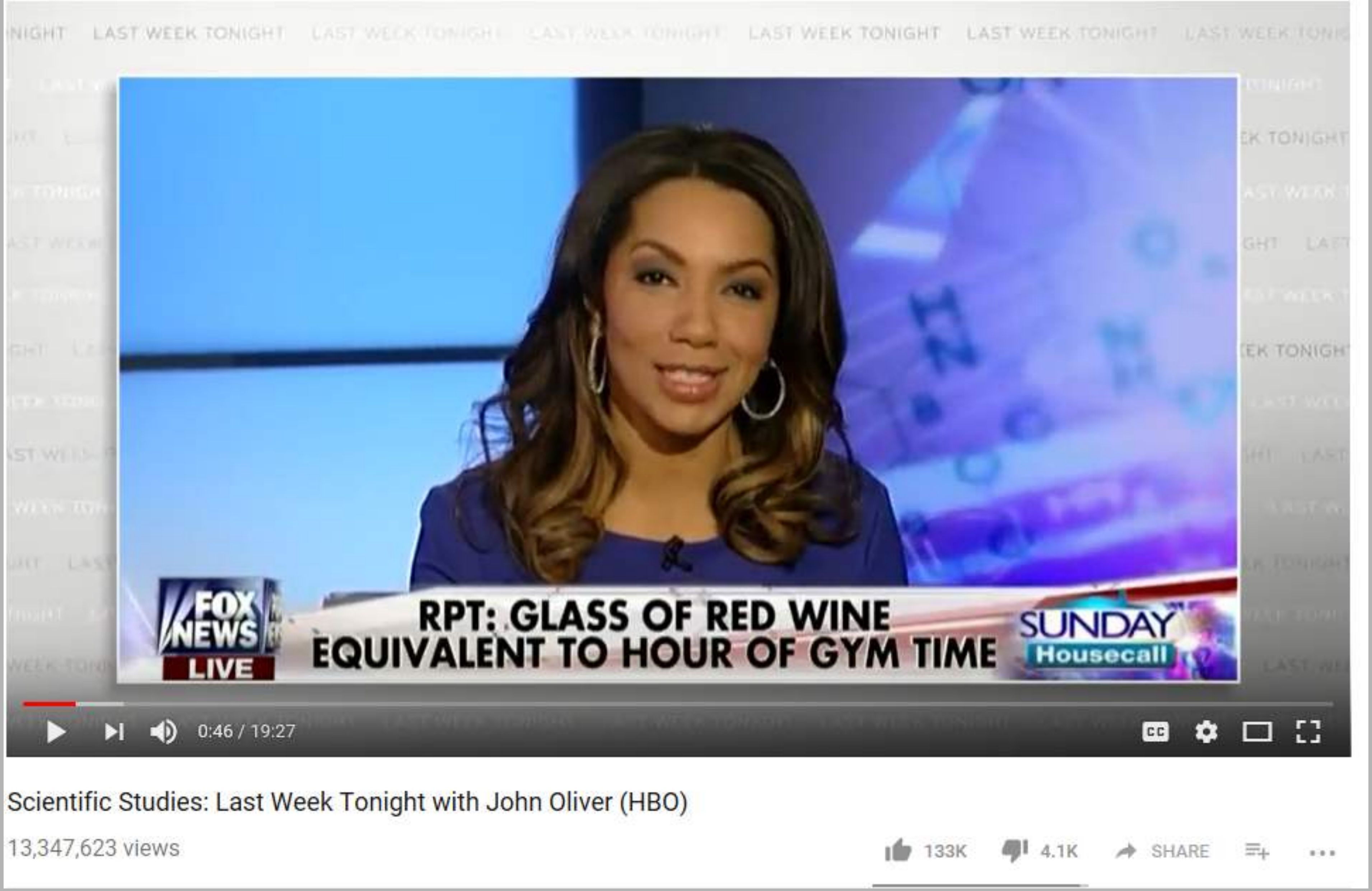
We show that the power functions of a class of nonparametric **tests** for the equality of two **scale** parameters do not approach 1 as the ratio of the parameters approaches infinity. The class of **tests**, known as **linear rank tests**, is shown to have a fundamental flaw when applied ...

☆ ❞ Cited by 12 Related articles All 6 versions



Taylor Swift - Shake It Off

2,643,643,309 views



Scientific Studies: Last Week Tonight with John Oliver (HBO)

13,347,623 views

# ASA statement articulated six principles

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**
6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

# Biggest takeaway message from the ASA statement

**Bright line thinking is bad for science**

"(S)cientists have embraced and even avidly **pursued meaningless differences** solely because they are statistically significant, and have **ignored important effects** because they failed to pass the screen of statistical significance…It is a safe bet that **people have suffered or died** because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results, and have consequently failed to identify the most beneficial courses of action."
(Rothman)

# P-value panel

- Naomi Altman
- Jim Berger
- Yoav Benjamini
- Don Berry
- Brad Carlin
- John Carlin
- George Cobb
- Marie Davidian
- Steve Fienberg
- Andrew Gelman
- Steve Goodman
- Sander Greenland
- Guido Imbens
- John Ioannidis
- Valen Johnson
- Michael Lavine
- Michael Lew
- Rod Little
- Deborah Mayo
- Chuck McCulloch
- Michele Millar
- Sally Morton
- Regina Nuzzo
- Hilary Parker
- Kenneth Rothman
- Don Rubin
- Stephen Senn
- Uri Simonsohn
- Dalene Stangl
- Philip Stark
- Steve Ziliak

SSI

ASA SYMPOSIUM ON

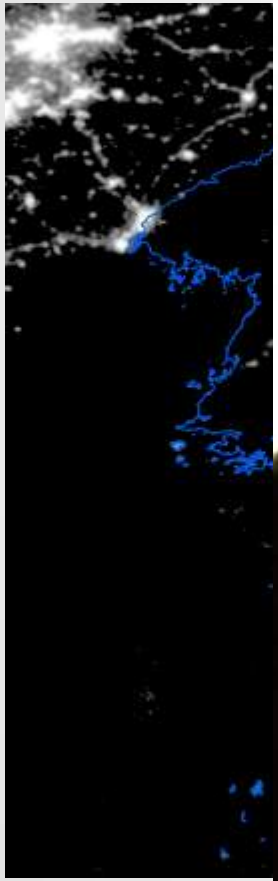
STATISTICAL

INFERENCE

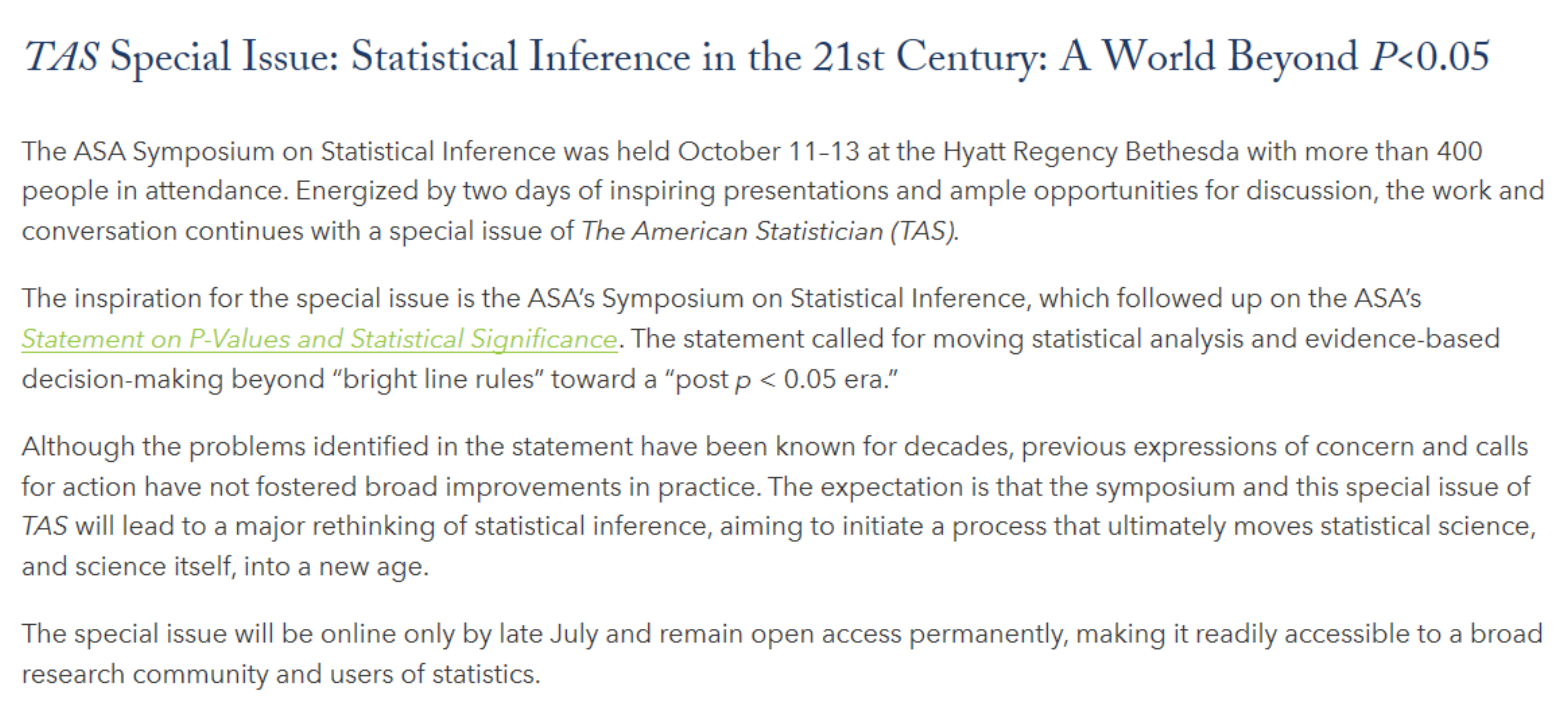
OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond $p < 0.05$

- “(D)rive change on the matters raised in the statement, providing necessary impetus for lasting improvements in science and society in the teaching of statistics, statistical practice, and the dissemination and many uses of statistical results”



NICKELBACK
WWW.ENTERTAINMENTWALLPAPER.COM



# *TAS* Special Issue: Statistical Inference in the 21st Century: A World Beyond *P*<0.05

The ASA Symposium on Statistical Inference was held October 11–13 at the Hyatt Regency Bethesda with more than 400 people in attendance. Energized by two days of inspiring presentations and ample opportunities for discussion, the work and conversation continues with a special issue of *The American Statistician (TAS)*.

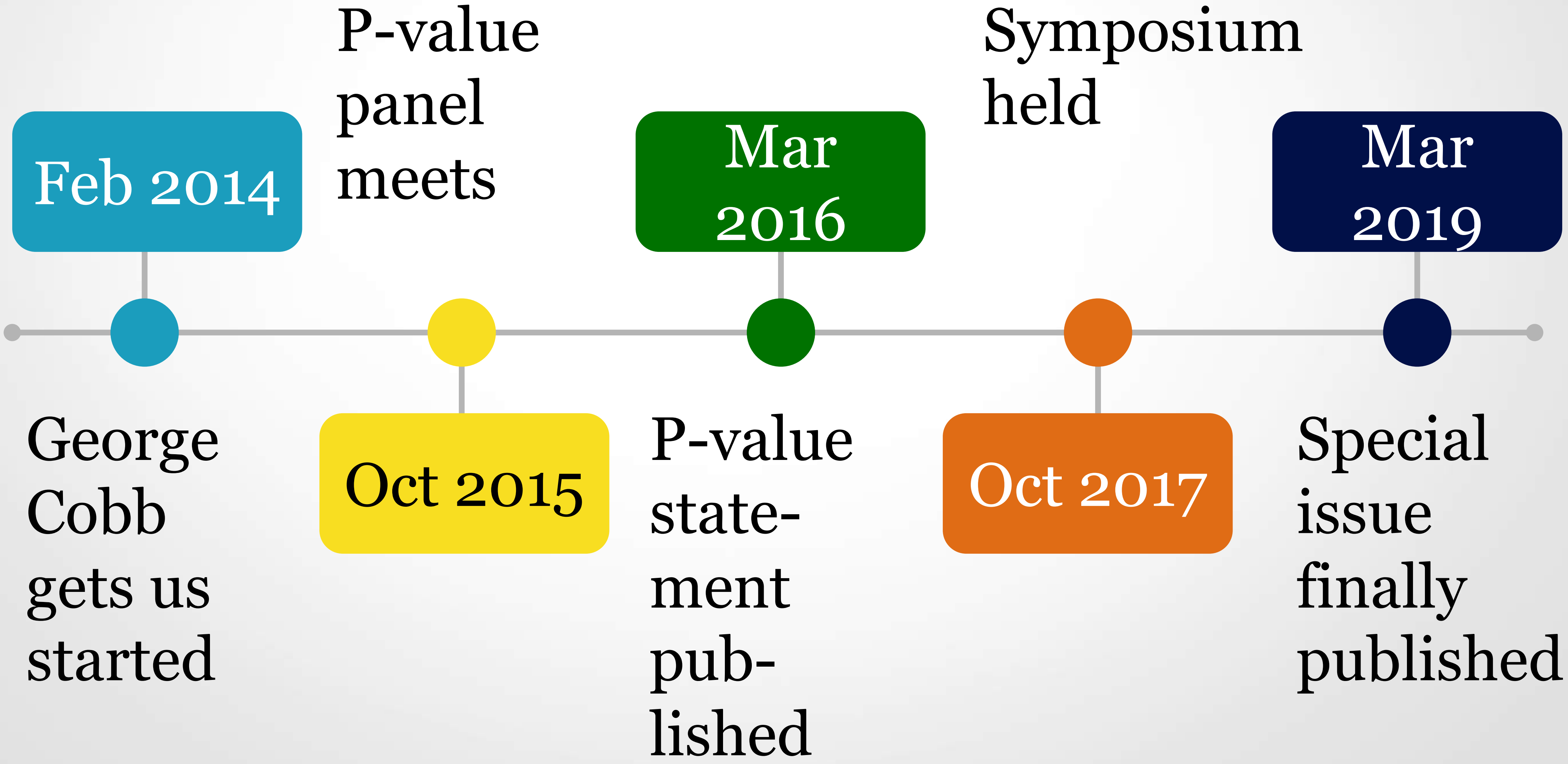
The inspiration for the special issue is the ASA's Symposium on Statistical Inference, which followed up on the ASA's *Statement on P-Values and Statistical Significance*. The statement called for moving statistical analysis and evidence-based decision-making beyond "bright line rules" toward a "post $p < 0.05$ era."

Although the problems identified in the statement have been known for decades, previous expressions of concern and calls for action have not fostered broad improvements in practice. The expectation is that the symposium and this special issue of *TAS* will lead to a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical science, and science itself, into a new age.

The special issue will be online only by late July and remain open access permanently, making it readily accessible to a broad research community and users of statistics.

# P-values timeline

Not to scale (so don't send us email)



Feb 2014
George Cobb gets us started

Oct 2015
P-value panel meets

Mar 2016
P-value statement published

Oct 2017
Symposium held

Mar 2019
Special issue finally published

# Associate editors

- Frank Bretz
- George Cobb
- Doug Hubbard
- Ray Hubbard
- Michael Lavine
- Fan Li
- Xihong Lin
- Tom Louis
- Regina Nuzzo
- Jane Pendergast
- Annie Qu
- Sherri Rose
- Steve Ziliak

# Anonymous reviewers

(redactions courtesy W. Barr and R. Rosenstein)

- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]

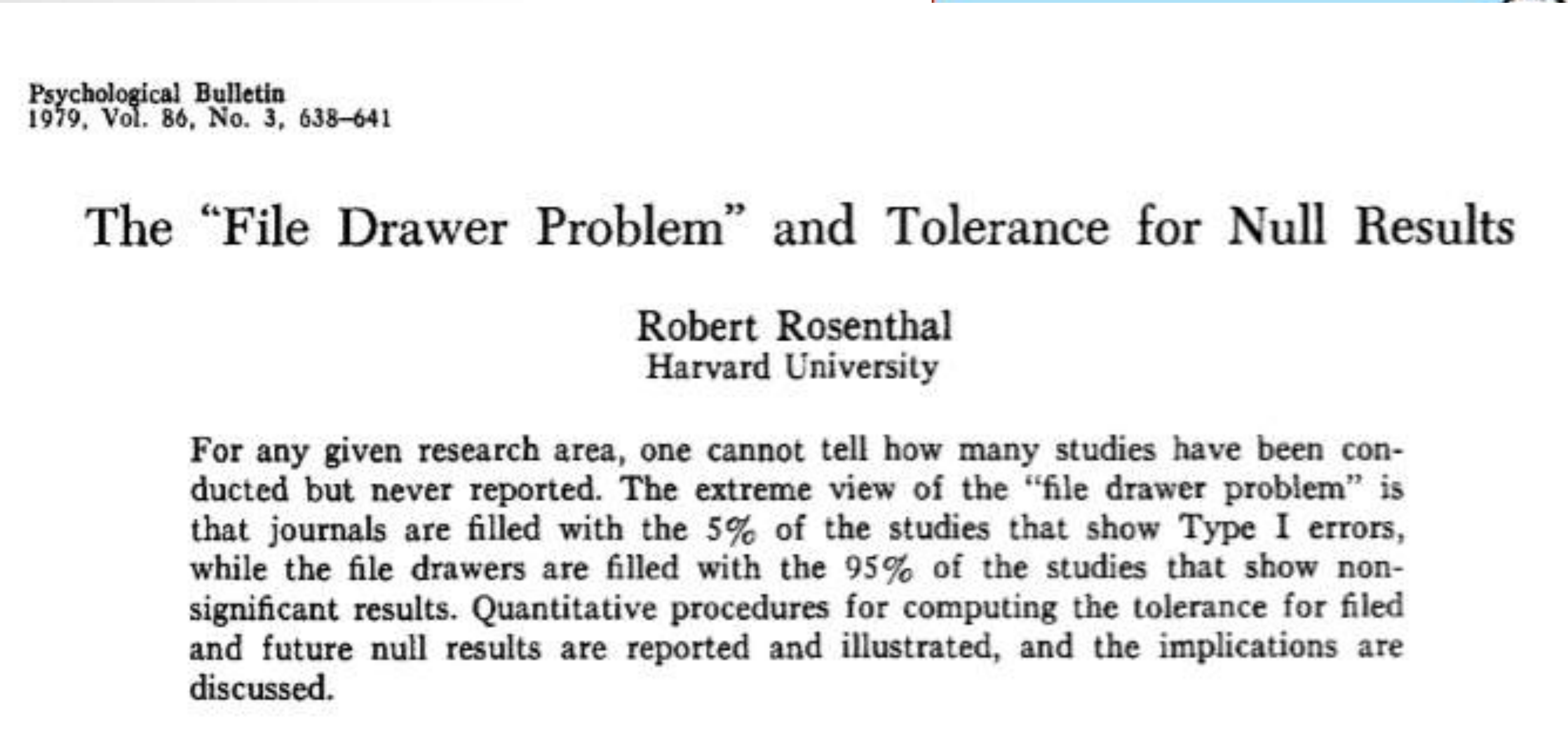
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]

- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]

- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]

# The big change

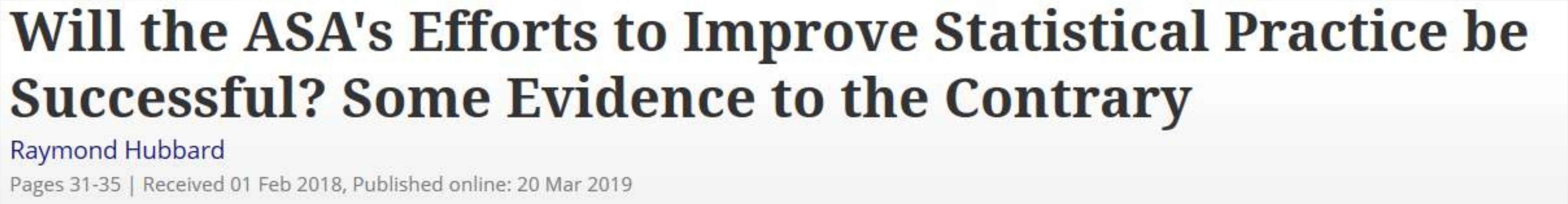
- Saying farewell to “statistically significant”


Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

# The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.



# Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary

Raymond Hubbard

Pages 31-35 | Received 01 Feb 2018, Published online: 20 Mar 2019

time to move on

# From the editorial

“We believe that a reasonable prerequisite for reporting any p-value is the ability to interpret it appropriately.”

# That p-value does not mean

- There is only a 3% chance the placebo was better than the treatment
- There is only a 3% chance of getting the result we did by chance alone
- The probability the null hypothesis is false is 97%
- …

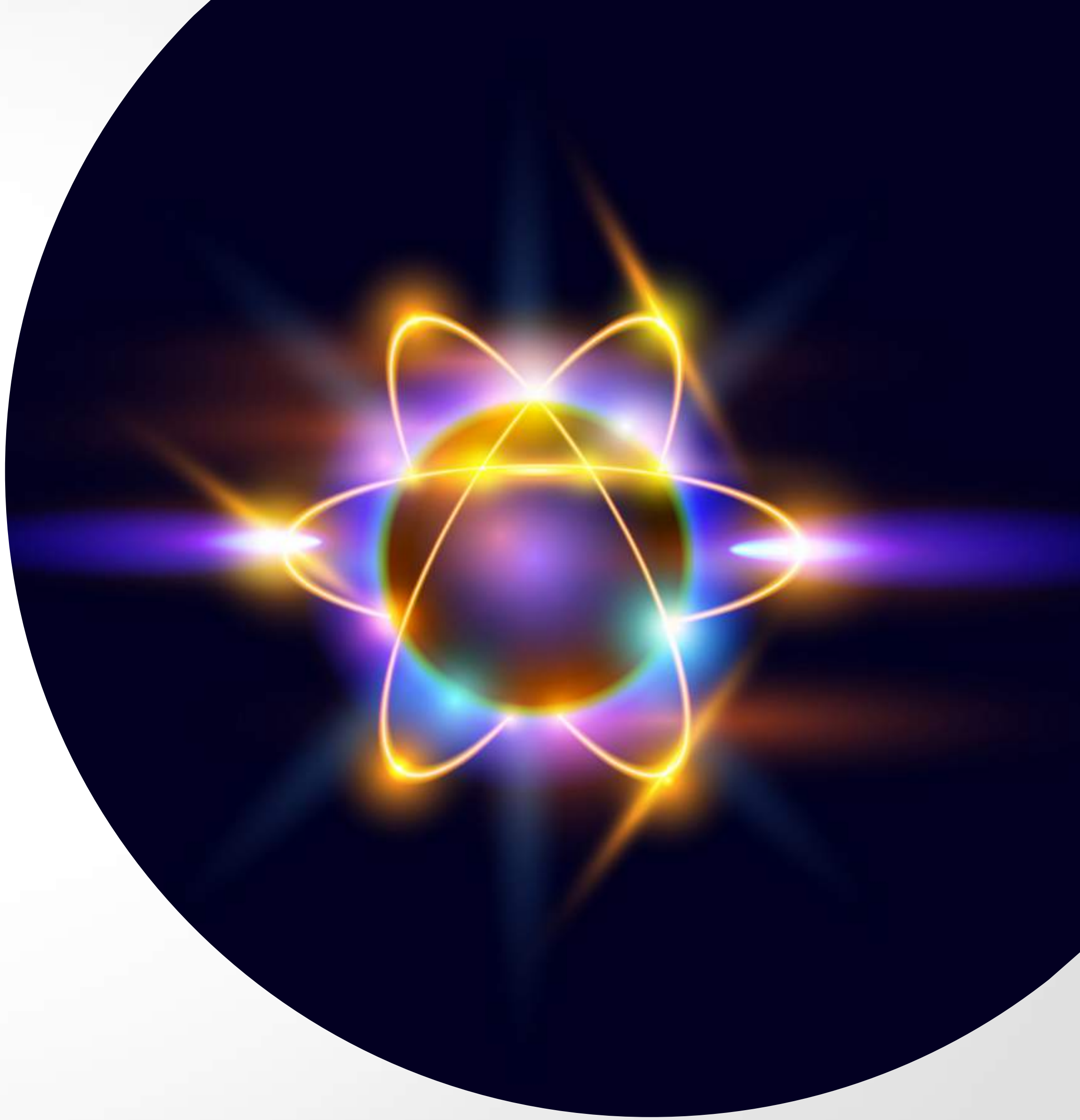
# It’s only about model incompatibility

- The null hypothesis is not the only assumption
  - But it is the only one that gets attention!
- Every other choice from design to statistic matters as well!

In a world where $p<0.05$ carried no meaning…

What would you have to do to get your paper published, your research grant funded, your drug approved, your policy or business recommendation accepted?

- Accept Uncertainty
- Be Thoughtful
- Be Open
- Be Modest



# Thoughtful research:

…looks ahead to prospective outcomes

(What magnitudes of differences, odds ratios, or other effect sizes are practically important?)

# Thoughtful research:

…considers “related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain…without giving priority to p-values or other purely statistical measures.”

# Thoughtful researchers:

…use a toolbox of statistical techniques

…consider multiple approaches for solving problems

# Alternatives

- Along with the standard p-value (null hypothesis), report some pre-specified other hypotheses. One example: instead of assuming no effect, assume the minimum meaningful effect size.

# Alternatives

- Transform $s = -\log_2(p)$

# Alternatives

- Analysis of credibility
- Second generation p-values
- False positive risk
- Bayes Factor Bound

# Be open

- Understand that subjectivity is involved in any statistical analysis.
- “(T)here is essentially no aspect of scientific investigation in which judgment is not required.”

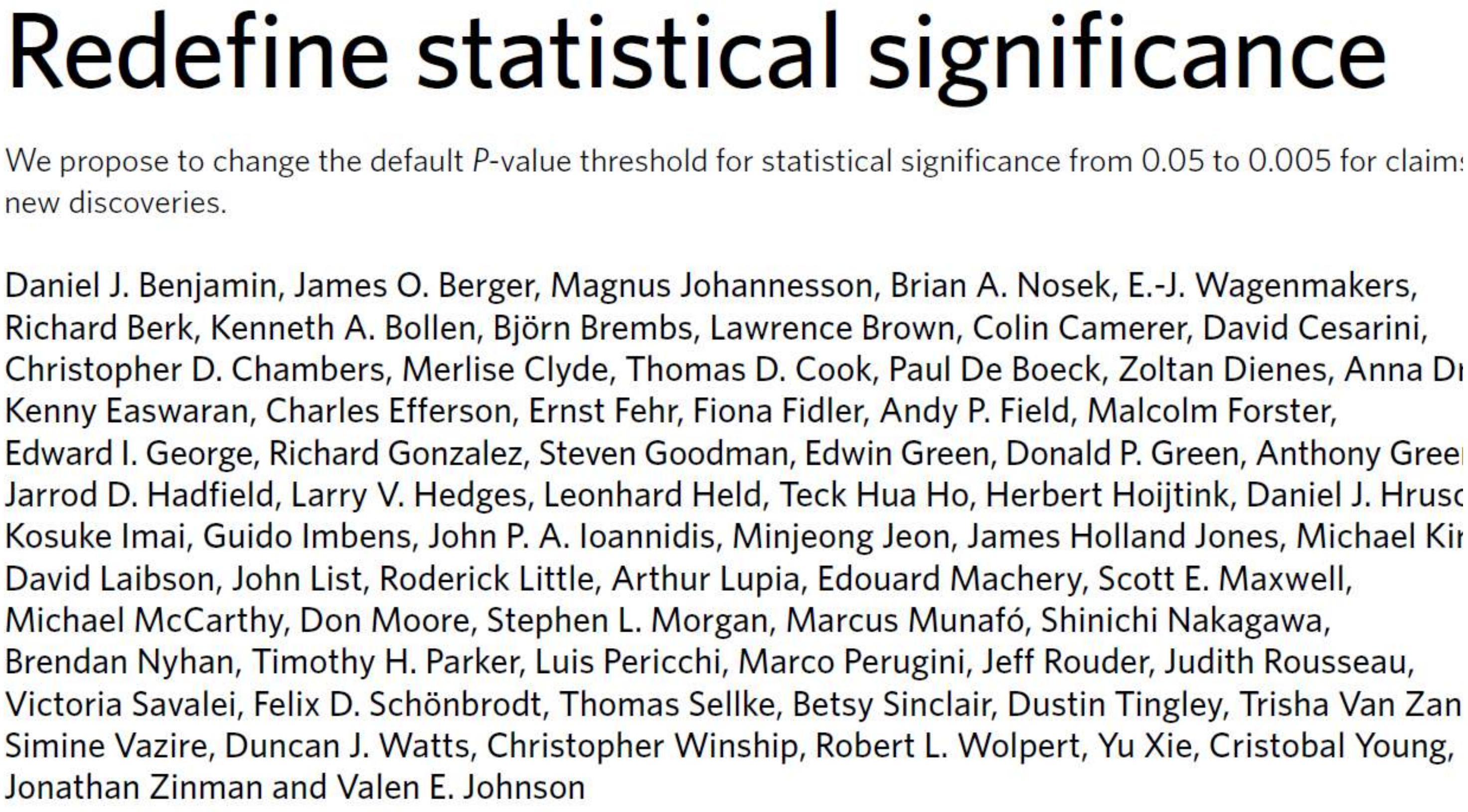
# Be open

Remember that one study is rarely enough. The words “a groundbreaking new study” might be loved by news writers but must be resisted by researchers. Breaking ground is only the first step in building a house. It will be suitable for habitation only after much more  hard work.

# Be modest

- P-values, confidence intervals, and other statistical measures are all uncertain.
- Encourage others to reproduce your work
- Statistical inference is (or should be) just one part of scientific inference

# $P<0.05$ versus $P<0.005$



# Redefine statistical significance

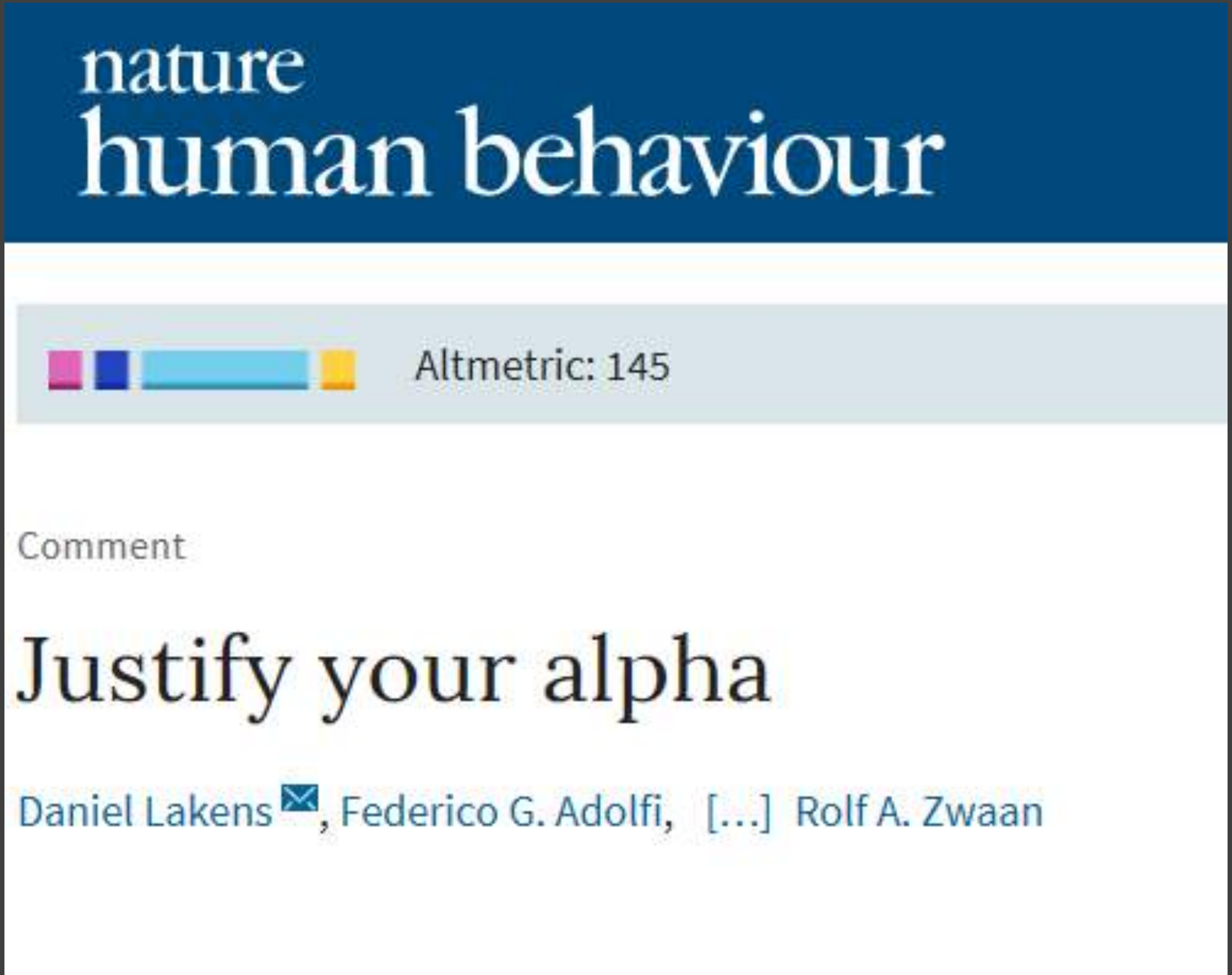
We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claim: new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dı Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Gree Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hrusc Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kiı David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zan Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

Nature Human Behavior
www.nature.com/nathumbehav
Sept 01 2017

DOI: 10.1038/s41562-017-0189-z

# A response



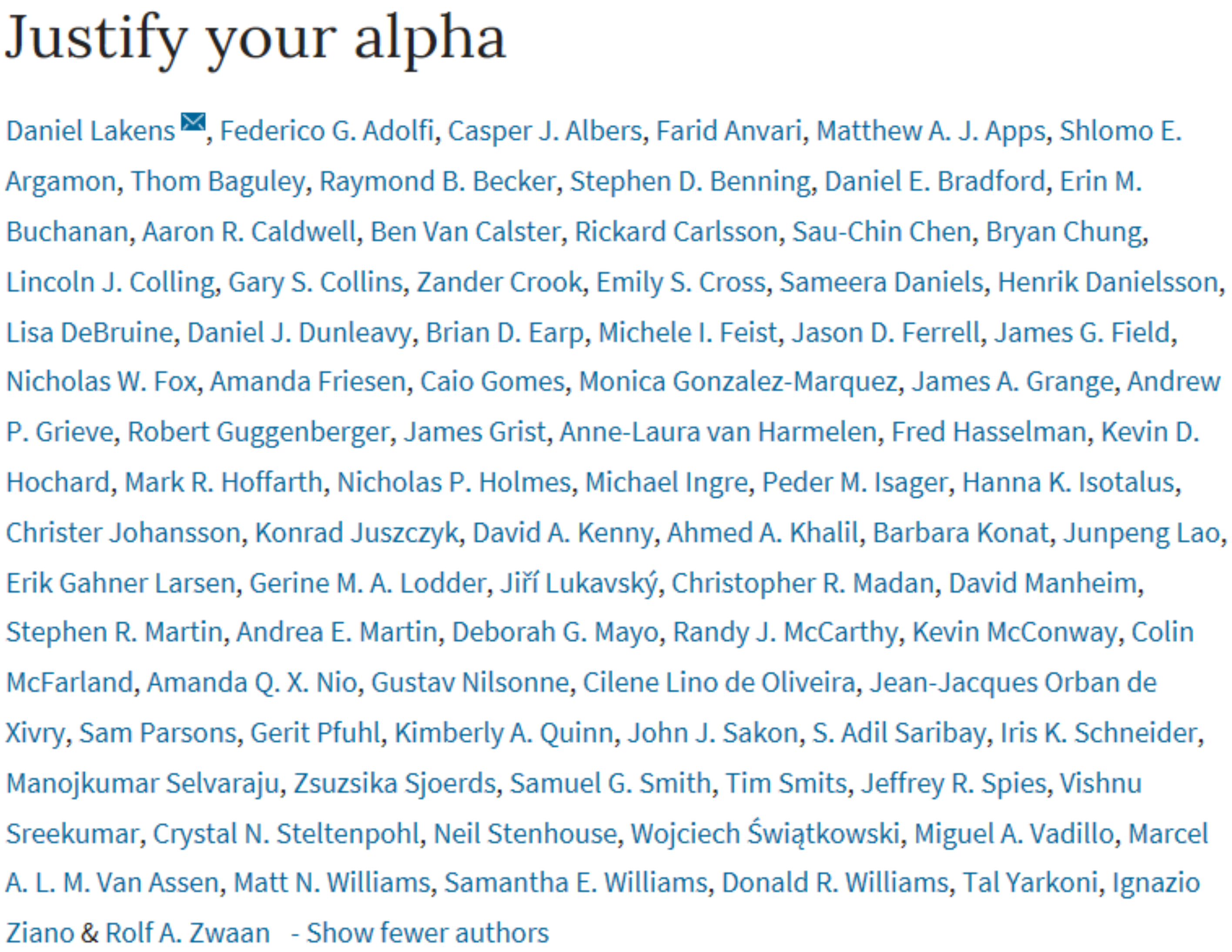
nature human behaviour

Altmetric: 145

Comment

Justify your alpha

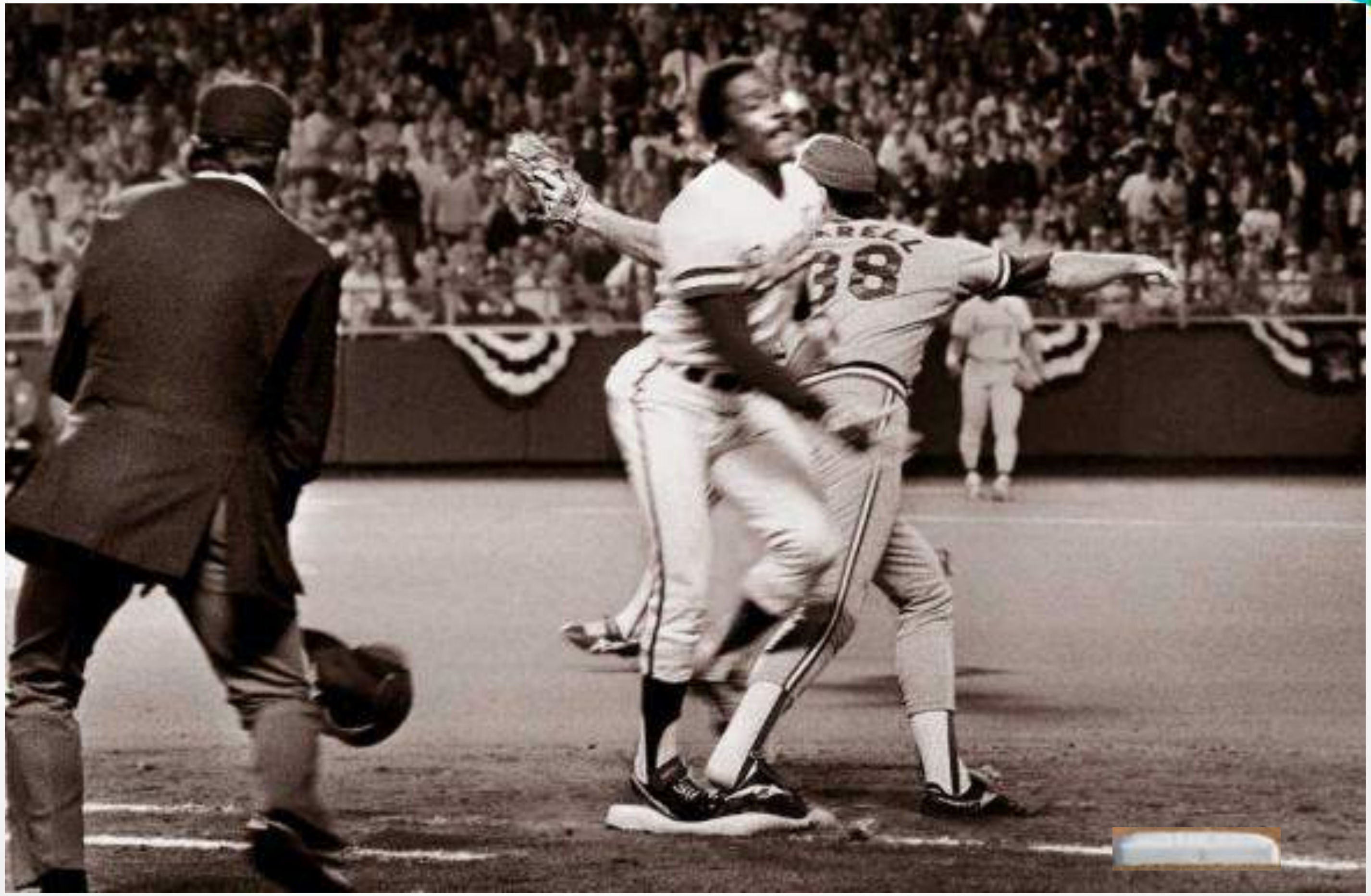
Daniel Lakens ✉, Federico G. Adolfi, […] Rolf A. Zwaan



# Justify your alpha

Daniel Lakens ✉, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau-Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa DeBruine, Daniel J. Dunleavy, Brian D. Earp, Michele I. Feist, Jason D. Ferrell, James G. Field, Nicholas W. Fox, Amanda Friesen, Caio Gomes, Monica Gonzalez-Marquez, James A. Grange, Andrew P. Grieve, Robert Guggenberger, James Grist, Anne-Laura van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark R. Hoffarth, Nicholas P. Holmes, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Christer Johansson, Konrad Juszczyk, David A. Kenny, Ahmed A. Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M. A. Lodder, Jiří Lukavský, Christopher R. Madan, David Manheim, Stephen R. Martin, Andrea E. Martin, Deborah G. Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Amanda Q. X. Nio, Gustav Nilsonne, Cilene Lino de Oliveira, Jean-Jacques Orban de Xivry, Sam Parsons, Gerit Pfuhl, Kimberly A. Quinn, John J. Sakon, S. Adil Saribay, Iris K. Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel G. Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal N. Steltenpohl, Neil Stenhouse, Wojciech Świątkowski, Miguel A. Vadillo, Marcel A. L. M. Van Assen, Matt N. Williams, Samantha E. Williams, Donald R. Williams, Tal Yarkoni, Ignazio Ziano & Rolf A. Zwaan - Show fewer authors

38



38

42
MIDWEST AIRLINES
State Farm
NOXZEMA
Sprint
PEPSI PARTY PORCH
ANKIEL
24
MECHE

# *Thank You, USCOTS!*

*ron@amstat.org, allenschirm@gmail.com*
*@Ron_Wasserstein*