



Doctor, It Hurts When I p

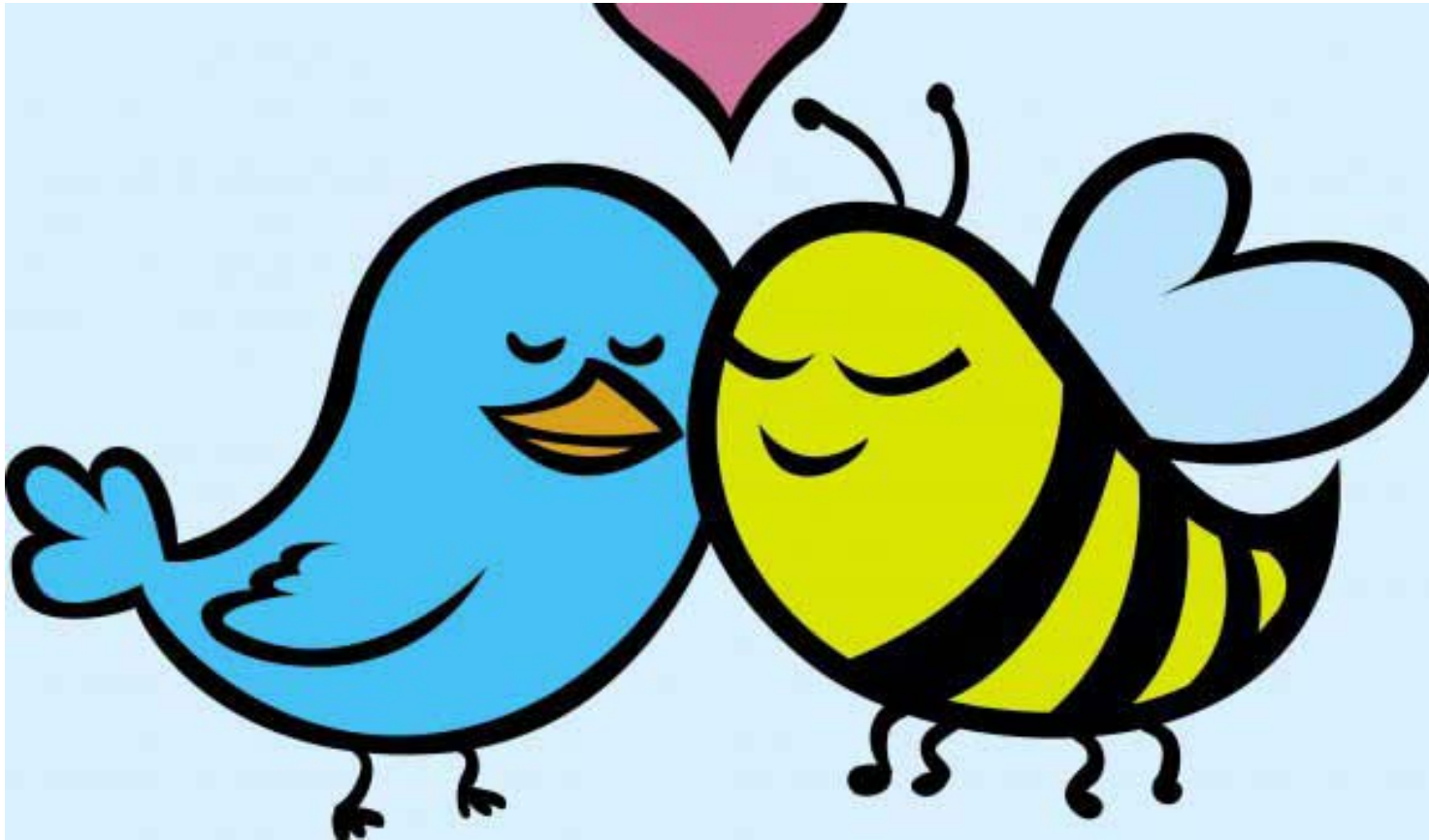
Ron Wasserstein

Executive Director

American Statistical Association

October 15, 2018





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.



Outline

- What are we talking about?
- The significance of “significant”
- Making a statement
- It’s worse than you think
- Cure worse than the disease?
- Aren’t we doing the wrong problem?
- The right way is hard work
- Some forthcoming suggestions

What is the
null
hypothesis
significance
testing
procedure?

- What do we already know?
- What do we want to know now?
- Experiment designed
- Data collected
- Data summarized
- Now what do we know?

Summarizing the data

- Compute a “statistic”
- Compute a probability called the “p-value”
- If the p-value is “small,” call the result “statistically significant”

What's the logic?
(With
oversimplifications)

- We assumed some stuff.
- We calculated a probability of observing the data that we did.
- If the probability is small, either
 - At least one assumption was wrong, or
 - We just had bad luck



R.A. Fisher called such results “significant”

sig·nif·i·cant

/sigˈnɪfɪkənt/

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"

To Fisher,
this meant
that the
result was
worth
further
scrutiny



insignificant

unimportant

meaningless

A silhouette of a person in mid-air, jumping over a gap between two dark, rectangular blocks. The background is a dramatic sky with large, billowing clouds in shades of blue and white. The person's arms are outstretched, and their legs are bent, capturing the peak of their jump.

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

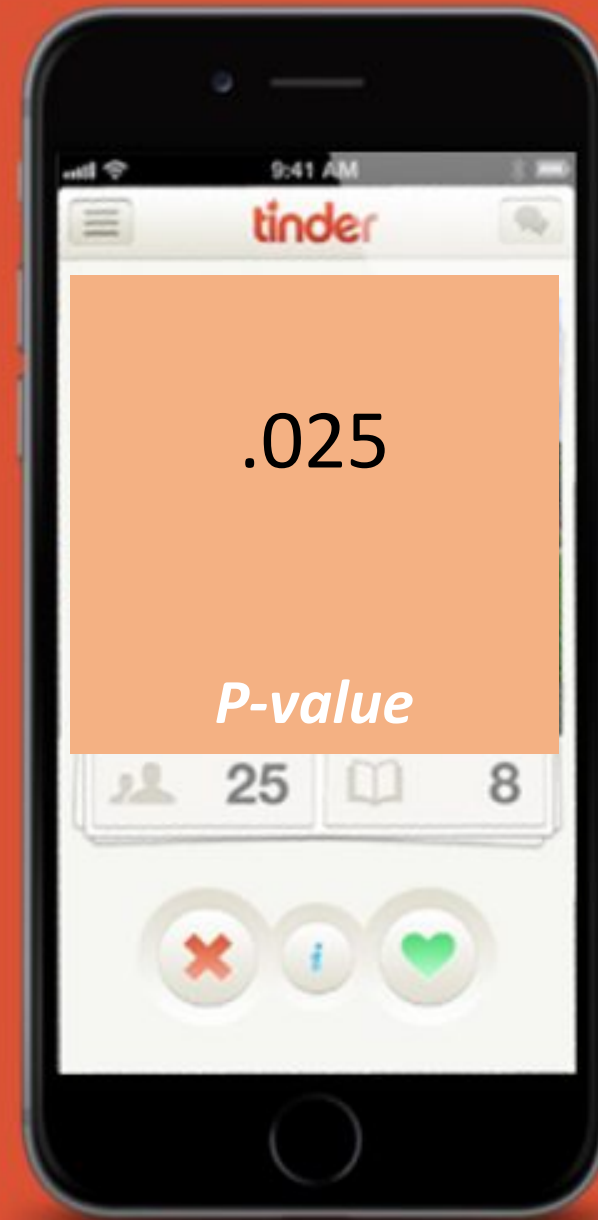


“You keep using that word. I don’t think that it means what you think it means.” – Inigo Montoya

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

Word number 6: “Significant”

tinder





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.



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!

The ASA Statement on p-values and Statistical Significance

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

270,748

Views

763

CrossRef citations
to date

2,017

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

 Download citation  <https://doi.org/10.1080/00031305.2016.1154108>

 Check for updates

[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 ...

  Cited by 1356 [Related articles](#) [All 35 versions](#) 

17

Views

4

CrossRef citations

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 - amstat.tandfonline.com

Abstract 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 to scale parameters, resulting in low power when the sample sizes are small.

  Cited by 12 [Related articles](#) [All 6 versions](#)



Taylor Swift - Shake It Off

2,643,643,309 views

7.7M 881K SHARE ...



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

13,347,623 views

133K

4.1K

SHARE



ASA statement
articulates 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.**

ASA statement
articulates six
principles

- 3. Scientific conclusions and business or policy decisions should not be based 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 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 ($p=0.057$)

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

Thanks to Matthew
Hankins for these
quotes

"... 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. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p ?"

Rosnow, R.L. and Rosenthal, R. 1989. Statistical procedures and the justification of knowledge and psychological science. *American Psychologist* 44: 1276-1284

Yes, dichotomizing evidence leads to strange behaviors!



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 Di 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 Kir 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

premise

...a leading cause of non-reproducibility has not yet been adequately addressed:

statistical standards of evidence for claiming new discoveries in many fields of science are simply too low.

premise

Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

premise

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields.



Why .005?

Essentially, because it approximates the level of certainty researchers mistakenly think they are getting with a .05 threshold.

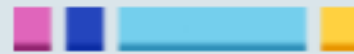
Naturally, this lowers the false positive rate.



Anticipated objections

- False negative rate becomes unacceptably high
- Does not address multiple hypothesis testing, P-hacking, publication bias, low power, or other biases
- Appropriate threshold for statistical significance should be different for different research communities
- Distracts from the real solution


nature
human behaviour



Altmetric: 145

Comment

Justify your alpha

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

A response

Justify your alpha

Daniel Lakens , Federico G. Adolphi, 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

premise

We do not think that redefining the threshold for statistical significance to the lower, but equally arbitrary threshold of $p \leq .005$ is advisable.

Arguments

There is insufficient evidence that the current standard for statistical significance is in fact a “leading cause of non-reproducibility”

Arguments

The arguments in favor of a blanket default of $p \leq .005$ are not strong enough to warrant the immediate and widespread implementation of such a policy

Arguments

A lower significance threshold will likely have positive and negative consequences, both of which should be carefully evaluated before any large-scale changes are proposed



Alternative proposal

When designing studies, they propose that authors transparently specify their design choices and justify these choices prior to collected data.



Alternative proposal

Instead of simple heuristics and an arbitrary blanket threshold, research should be guided by principles of rigorous science.

Their bottom line

Single studies, regardless of their p-value, are never enough to conclude that there is strong evidence for a *theory*.

We need to train researchers to recognize what cumulative evidence looks like and work towards an unbiased scientific literature.







A fundamental problem

We want $P(H | D)$ but p-values give $P(D | H)$

What is the probability of obtaining a dead person (D) given that the person was hanged (H); that is, in symbol form, what is $p(D|H)$?

Obviously, it will be very high, perhaps .97 or higher.

The problem illustrated (Carver 1978)

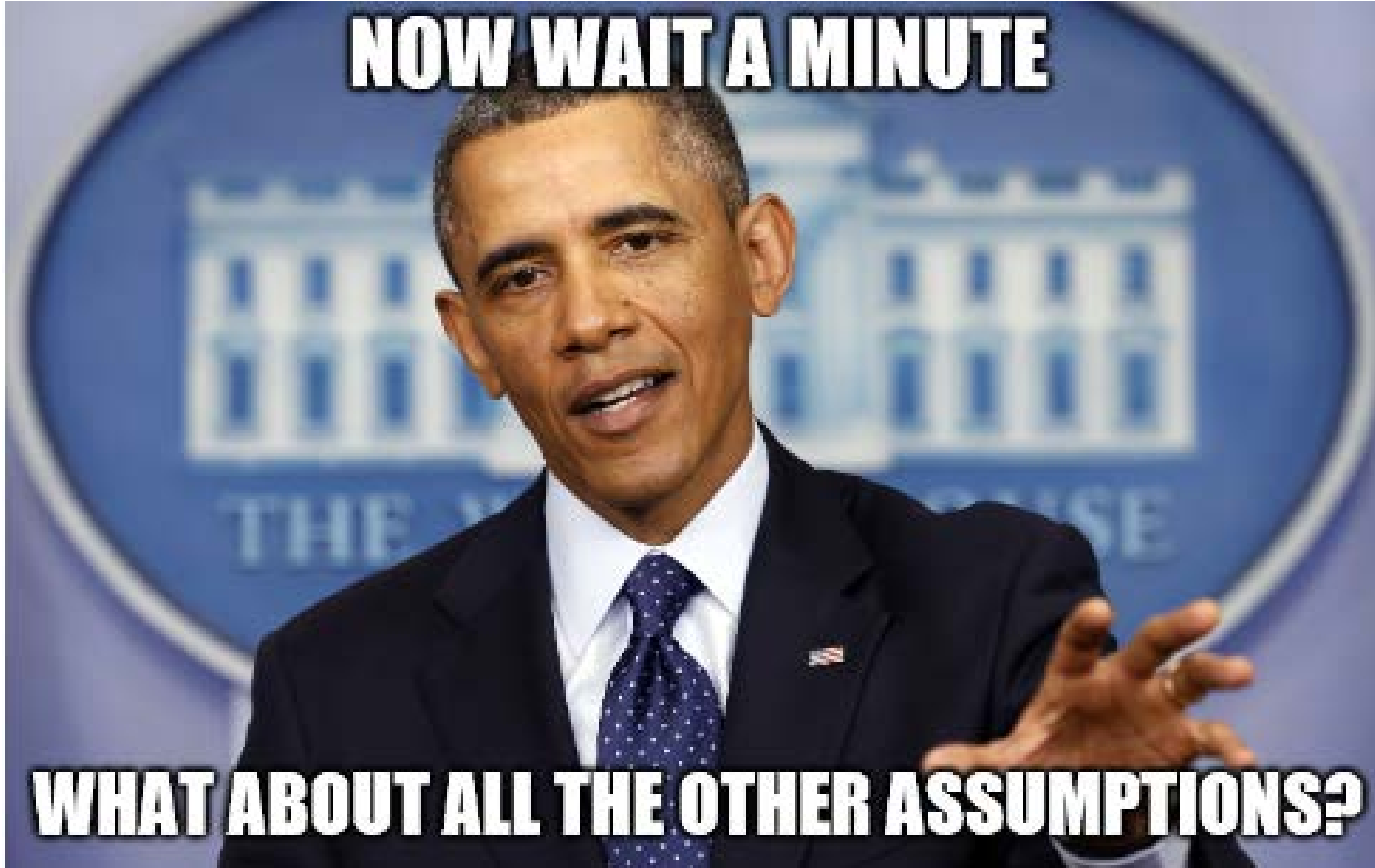
Now, let us reverse the question: What is the probability that a person has been hanged (H) given that the person is dead (D); that is, what is $p(H|D)$?

This time the probability will undoubtedly be very low, perhaps .01 or lower.

No one would be likely to make the mistake of substituting the first estimate (.97) for the second (.01); that is, to accept .97 as the probability that a person has been hanged given that the person is dead.

NOW WAIT A MINUTE

WHAT ABOUT ALL THE OTHER ASSUMPTIONS?



Simplistic (“cookbook”) rules and procedures are not a substitute for this hard work.

Cookbook + artificial threshold for significance = appearance of objectivity

So the ASA has
been asking
the question...

How would you conduct
research in a world where
 $p < 0.05$ (or 95% limits)
carried no meaning?

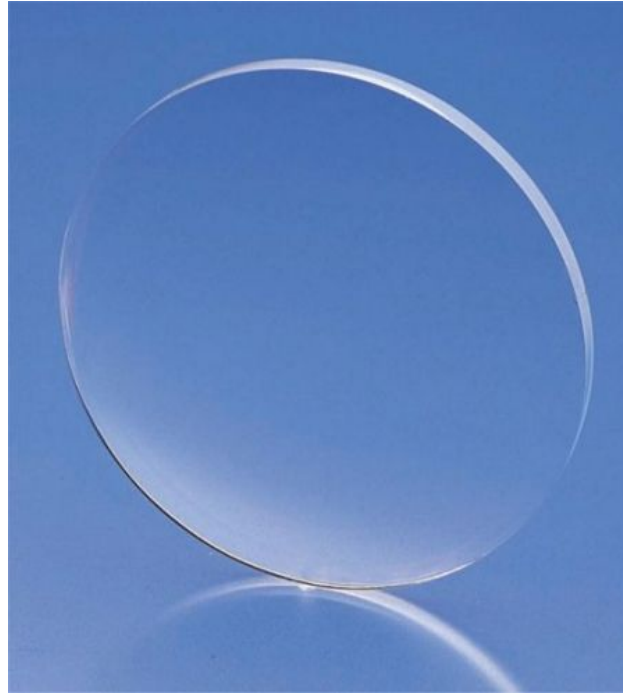


In a world where
 $p < 0.05$ carried no
meaning...

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

You wouldn't just do a hypothesis test

You'd check multiple models, review critical assumptions, use alternate methods of analysis



You would be relentlessly transparent

You wouldn't
need to

- P-hack
- HARK
- Cherry pick
- Use your “researcher degrees of freedom”

Converting “don’ts” to “do’s”

A sneak preview of the special issue of
The American Statistician

Use decision-theoretic approaches (Manski)

Treatment choice using statistical decision theory is not based at all on whether a p-value passes a threshold.

Statistical decision theory clearly distinguishes between the statistical and clinical significance of empirical estimates of treatment effects.

Abandon thresholds (McShane et.al)

...we propose that the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with currently subordinate factors (e.g., 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) as just one among many pieces of evidence.

Advise editors and reviewers (Trafimow)

- Give more consideration of the nature of the contribution
- Tolerate some ambiguity
- Emphasize thinking and execution, not results
- Replace NHST with *a priori* thinking
- Remember that the assumptions of random and independent sampling might be wrong

Introduce results-blind publishing (Locascio)

- Provide an initial provisional decision on a manuscript based exclusively on the judged importance of the research issues addressed by the study and the soundness of the reported methodology.
- Give no weight to the reported results of the study per se in the decision as to whether to publish or not.
- Commit to an initial decision regarding publication after having been provided with only the Introduction and Methods sections of a manuscript by the editor, not having seen the Abstract, Results, or Discussion.

Introduce results-blind publishing (Locascio)

- Emphasize the clinical and/or scientific importance of a study in the Introduction section of a manuscript
- Give a clear, explicit statement of the research questions being addressed and any hypotheses to be tested.
- Include a detailed statistical analysis sub-section in the Methods section
- Submit for publication reports of well-conducted studies on important research issues regardless of findings

Carefully elicit
expert
judgment
(Brownstein et
al.)

- Understand that subjective judgments are needed at all stages of a scientific study
- Ensure that all such judgments are made as carefully, rigorously and honestly as possible.
- Identify all judgments made, and measures applied to avoid bias whenever possible.
- Use protocol-guided elicitation of judgments.

Second generation p- values (SGPV) (Blume et. al.)

- Construct a composite null hypothesis by specifying the range of effects that are not scientifically meaningful
- Replace classical p-values with second-generation p-values (SGPV), which accommodate composite null hypotheses and encourage the proper communication of findings.
- Interpret the SGPV as a high-level summary of what the data say.
- Report an interval estimate of effect size (confidence interval, support interval, or credible interval) and note its proximity to the composite null hypothesis.

Address
thresholds a
different way
(Gannon et.
al)

- Blend Bayesian and classical tools to define optimal sample-size-dependent significance levels
- Procedure minimizes a linear combination of α and β while preserving the likelihood principle

Wrapping up...

Little p-value

what are you trying to say

of significance?

- Steve Ziliak

Haiku

ron@amstat.org

@RonWasserstein