431 Class 22

Thomas E. Love, Ph.D.

2022-11-29

Today's Agenda

- What I Taught for Many Years
- What is the problem?
- Do Confidence Intervals Solve the Problem?
- Borrowing from Bayesian Ideas
- Replicable Research and the Crisis in Science

What I Taught for Many Years

- Null hypothesis significance testing is here to stay.
  - Learn how to present your p value so it looks like what everyone else does
  - Think about "statistically detectable" rather than "statistically significant"
  - Don't accept a null hypothesis, just retain it.
- Use point **and** interval estimates
  - Try to get your statements about confidence intervals right (right = just like I said it)
- Use Bayesian approaches/simulation/hierarchical models when they seem appropriate or for "non-standard" designs
  - But look elsewhere for people to teach/do that stuff
- Power is basically a hurdle to overcome in a grant application

Conventions for Reporting *p* Values

1. Use an italicized, lower-case *p* to specify the *p* value. Don't use *p* for anything else.
2. For *p* values above 0.10, round to two decimal places, at most.
3. For *p* values near \(\alpha\), include only enough decimal places to clarify the reject/retain decision.
4. For very small *p* values, always report either $p < 0.0001$ or even just $p < 0.001$, rather than specifying the result in scientific notation, or, worse, as \(p = 0\) which is glaringly inappropriate.
5. Report *p* values above 0.99 as $p > 0.99$, rather than $p = 1$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if p < .05.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if p < .05,

which morphed into a

- **rule** for editors: reject the submitted article if p > .05.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if p < .05,

which morphed into a

- **rule** for editors: reject the submitted article if p > .05,

which morphed into a

- **rule** for journals: reject all articles that report p-values

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if p < .05, which morphed into a
- **rule** for editors: reject the submitted article if p > .05, which morphed into a
- **rule** for journals: reject all articles that report p-values.

Bottom line: **Reject rules. Ideas matter.**


P > 0.05
GAME OVER, TRY AGAIN
imgflip.com

American Statistical Association to the rescue!?!

The American Statistical Association

2016

- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

2019

- Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "p < 0.05", *The American Statistician*, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913.

Statistical Inference in the 21st Century

> … a world learning to venture beyond "p < 0.05"

> This is a world where researchers are free to treat "p = 0.051" and "p = 0.049" as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number.

Statistical Inference in the 21st Century

> In this world, where studies with "p < 0.05" and studies with "p > 0.05" are not automatically in conflict, researchers will see their results more easily replicated – and, even when not, they will better understand why.

> The 2016 ASA Statement on P-Values and Statistical Significance started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times-an average of about 11 citations per week since its release. Now we must go further.

The American Statistical Association Statement on P values and Statistical Significance

The ASA Statement (2016) was mostly about what **not** to do.

The 2019 effort represents an attempt to explain what to do.

ASA 2019 Statement

> Some of you exploring this special issue of The American Statistician might be wondering if it's a scolding from pedantic statisticians lecturing you about what not to dowith p-values, without offering any real ideas of what to do about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

"Don't" is not enough.

> If you're just arriving to the debate, here's a sampling of what not to do.

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p* value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.

"Don't" is not enough.

- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Problems with *p* Values

1. *P* values are inherently unstable
2. The *p* value, or statistical significance, does not measure the size of an effect or the importance of a result
3. Scientific conclusions should not be based only on whether a *p* value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. By itself, a *p* value does not provide a good measure of evidence regarding a model or hypothesis

http://jamanetwork.com/journals/jamaotolaryngology/fullartic

One More Don't...

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless. Made

> A label of statistical significance adds nothing to what is already conveyed by the value of *p*; in fact, this dichotomization of *p*-values makes matters worse.

Gelman on *p* values, 1

> The common practice of dividing data comparisons into categories based on significance levels is terrible, but it happens all the time…. so it's worth examining the prevalence of this error. Consider, for example, this division:

- "really significant" for $p < .01$,
- "significant" for $p < .05$,
- "marginally significant" for $p < .1$, and
- "not at all significant" otherwise.

Gelman on *p* values, 2

Now consider some typical *p*-values in these ranges: say, $p = .005$, $p = .03$, $p = .08$, and $p = .2$.

Translate these two-sided *p*-values back into z-scores…

Description	really sig.	sig.	marginally sig.	not at all sig.
p value	0.005	0.03	0.08	0.20
Z score	2.8	2.2	1.8	1.3

Gelman on *p* values, 3

The seemingly yawning gap in p-values comparing the not at all significant *p*-value of .2 to the really significant *p*-value of .005, is only a z score of 1.5.

If you had two independent experiments with z-scores of 2.8 and 1.3 and with equal standard errors and you wanted to compare them, you'd get a difference of 1.5 with a standard error of 1.4, which is completely consistent with noise.

Gelman on *p* values, 4

From a **statistical** point of view, the trouble with using the p-value as a data summary is that the p-value can only be interpreted in the context of the null hypothesis of zero effect, and (much of the time), nobody's interested in the null hypothesis.

Indeed, once you see comparisons between large, marginal, and small effects, the null hypothesis is irrelevant, as you want to be comparing effect sizes.

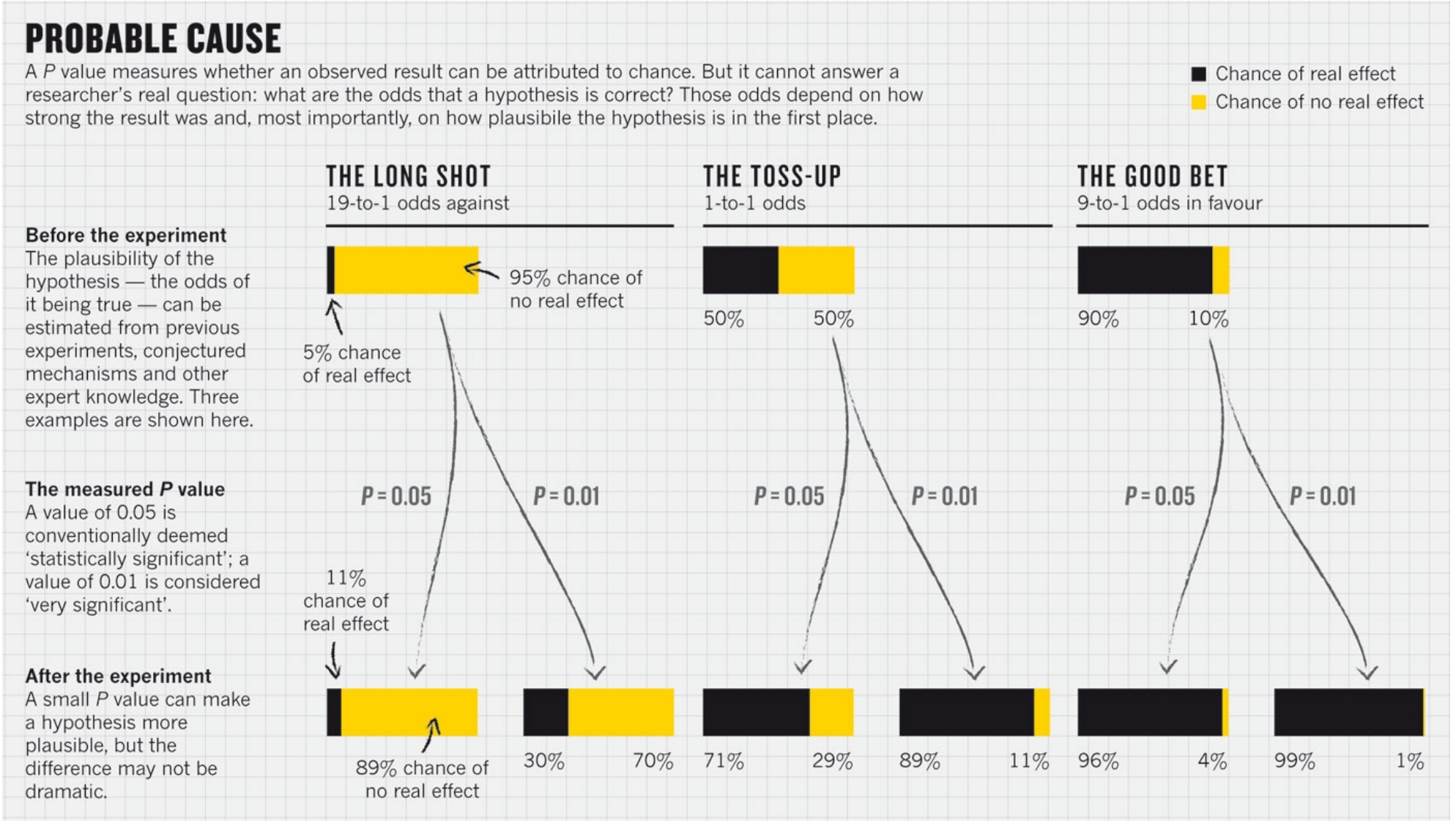
Gelman on *p* values, 5

From a **psychological** point of view, the trouble with using the p-value as a data summary is that this is a kind of deterministic thinking, an attempt to convert real uncertainty into firm statements that are just not possible (or, as we would say now, just not replicable).

The key point: The difference between statistically significant and NOT statistically significant is not, generally, statistically significant.

http://andrewgelman.com/2016/10/15/marginally-significant-effects-as-evidence-for-hypotheses-changing-attitudes-over-four-decades/

Regina Nuzzo: *Nature* Statistical Errors



Are P values all that bad?

Grumpy Old Health Stats Dude
@healthstatsdude

Following

"If you never use another p-value, you will have improved medicine."

-me, to clinicians

#statstwitter #medtwitter #epitwitter

12:36 AM - 4 Mar 2019

Grumpy Old Health Stats Dude
@healthstatsdude

Following

Replying to @healthstatsdude @EugeneDayDSc and 2 others

My main reason for being overtly/in public anti p-values is this:

P values
f overall
analyses
partly
atistical
even if
group,
wer, due
l distri-
fference
specific
ll death

mate of a 5% decrease in 10-year survival with watchful waiting, 750 men might have died prematurely as a result.

A mistake in the operating room can threaten the life of one patient; a mistake in statistical analysis or interpretation can lead to hundreds of early deaths. So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS® (SPSS, Chicago, IL) to almost anyone. Moreover, whilst only a surgeon would comment on surgical technique, it seems that anybody, regardless of statistical training,

day); a
that on
risk of t
in many
that the
event is

Comp
The aut
no com

7:59 PM - 19 Apr 2019

ASA Statement on *p* Values

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

"Not Even Scientists Can Easily Explain *p* Values" at fivethirtyeight.com

… Try to distill the p-value down to an intuitive concept and it loses all its nuances and complexity, said science journalist Regina Nuzzo, a statistics professor at Gallaudet University. "Then people get it wrong, and this is why statisticians are upset and scientists are confused." **You can get it right, or you can make it intuitive, but it's all but impossible to do both.**

"Statisticians found one thing they can agree on" at fivethirtyeight.com

A Few Comments on Significance

- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always "significant" even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?
- **A non-significant effect is not necessarily the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.
- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.
- "**Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**"

ASA 2016 statement on *p* values

p = 0.05?

> "For decades, the conventional p-value threshold has been 0.05," says Dr. Paul Wakim, chief of the biostatistics and clinical epidemiology service at the National Institutes of Health Clinical Center, "but it is extremely important to understand that this 0.05, there's nothing rigorous about it. It wasn't derived from statisticians who got together, calculated the best threshold, and then found that it is 0.05. No, it's Ronald Fisher, who basically said, 'Let's use 0.05,' and he admitted that it was arbitrary."

- NOVA "Rethinking Science's Magic Number" by Tiffany Dill 2018-02-28. See especially the video labeled "Science's most important (and controversial) number has its origins in a British experiment involving milk and tea."

More from Dr. Wakim...

> "People say, 'Ugh, it's above 0.05, I wasted my time.' No, you didn't waste your time." says Dr. Wakim. "If the research question is important, the result is important. Whatever it is."

- NOVA Season 45 Episode 6 2018-02-28.

p values don't trend...

Randy Sweis, MD
@RandySweisMD

Follow

If a P value of 0.06 trends toward statistical significance, then doesn't a P value of 0.04 trend toward non-significance?

9:47 AM - 12 Jan 2018

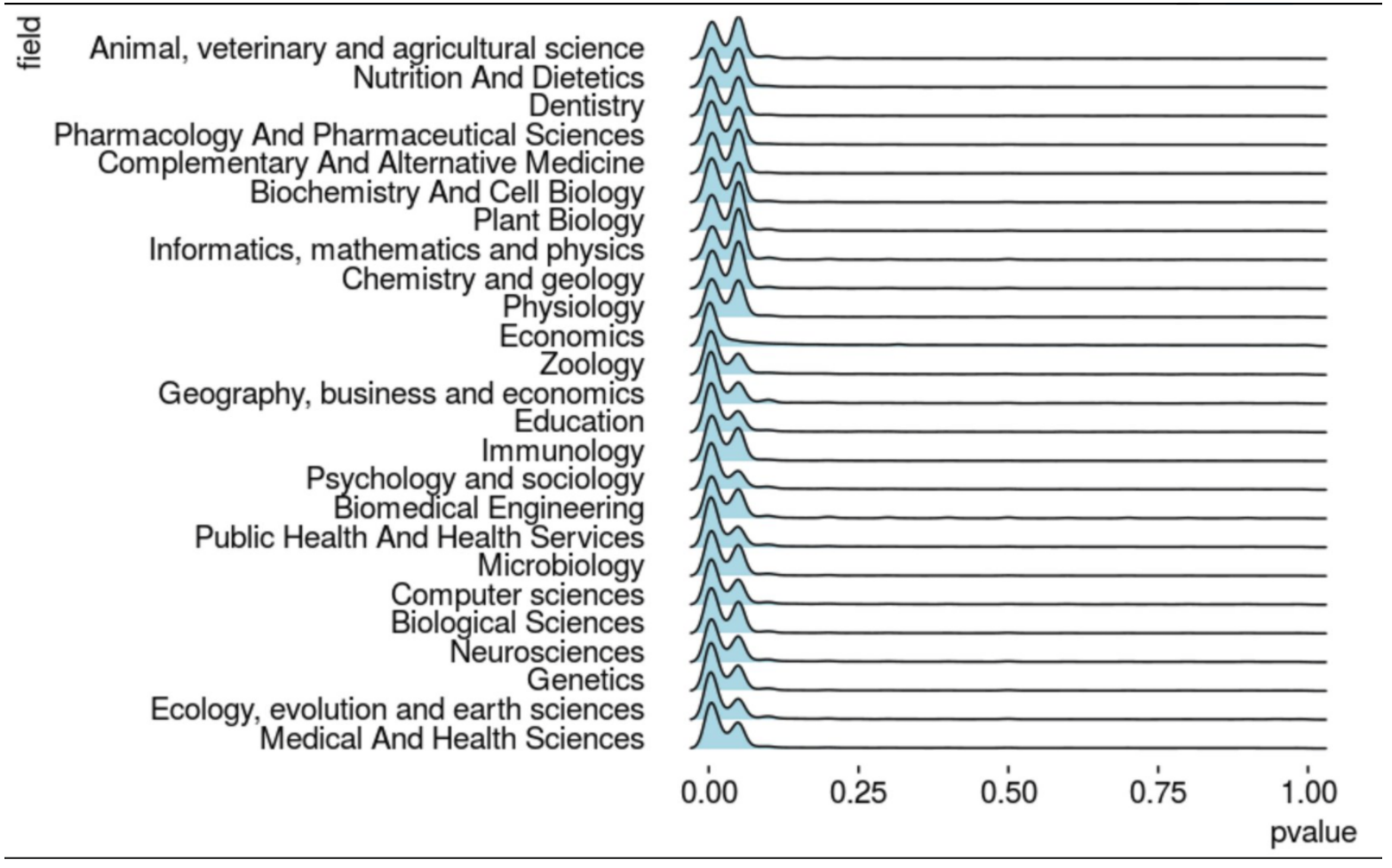
All the p values

> The p-value is the most widely-known statistic. P-values are reported in a large majority of scientific publications that measure and report data. R.A. Fisher is widely credited with inventing the p-value. If he was cited every time a p-value was reported his paper would have, at the very least, 3 million citations - making it the most highly cited paper of all time.

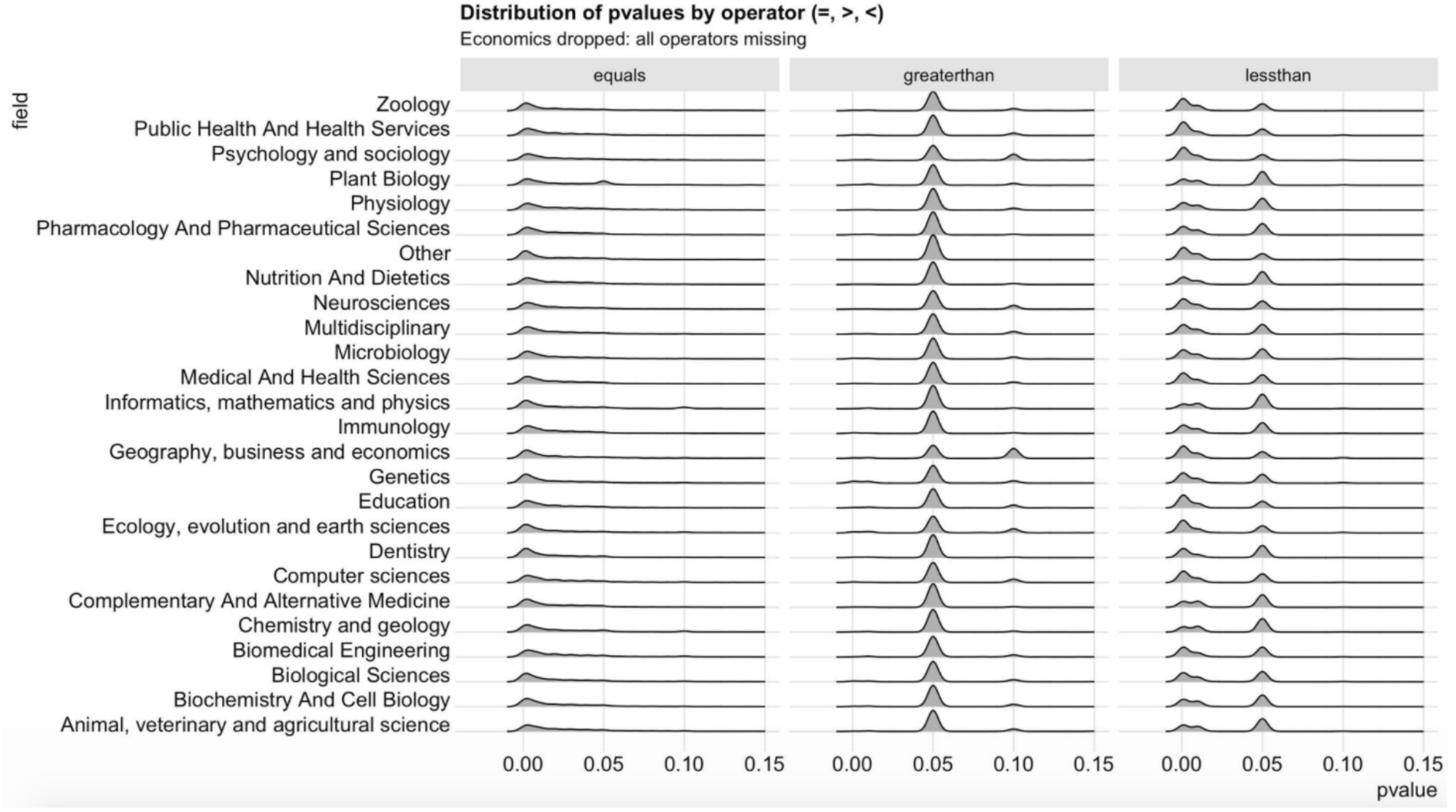
- Visit Jeff Leek's Github for tidypvals package
  - 2.5 million *p* values in 25 scientific fields

What do you suppose the distribution of those p values is going to look like?

2.5 million p values in 25 scientific fields: Jeff Leek



from Michael Lopez



Miguel Hernán ✓
@_MiguelHernan

Simple way for editors to improve science: If your journal still uses "statistical significance" in 2017, retire your statistical consultant

Practices that reduce scientific inference to mechanical "bright-line" rules (such as "p < 0.05") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making.

American Statistical Association, 2016

But many journals do present findings as "statistically significant" or "not statistically significant".

- How can an editor work with statistical consultants who ignore the ASA without publicly justifying their views?
- Would the editor work with a cardiology consultant who ignores the American Heart Association without providing any justification?

Unfortunately...

There are a lot of candidates for the most outrageous misuse of “statistical significance” out there.

Alvaro Alonso
@alonso_epi

More p-value silliness. HR 0.90, 95%CI 0.81-0.99--> 'effect'; HR 0.89, 95%CI 0.78-1.0009-->no 'effect' jaha.ahajournals.org/content/6/5/e0... @ken_rothman

Normalization of Testosterone Levels After Testosterone Replacement Therapy Is Associated With Decreased Incidence of Atrial Fibrillation

Rishi Sharma, MD, MHSA; Olurinde A. Oni, MBBS, MPH; Kamal Gupta, MD; Mukut Sharma, PhD; Ram Sharma, PhD; Vikas Singh, MD, MHSA; Deepak Parashara, MD; Surineni Kamalakar, MBBS, MPH; Buddhadeb Dawn, MD; Guoqing Chen, MD, PhD, MPH; John A. Ambrose, MD; Rajat S. Barua, MD, PhD

Background—Atrial fibrillation (AF) is the most common cardiac dysrhythmia associated with significant morbidity and mortality. Several small studies have reported that low serum total testosterone (TT) levels were associated with a higher incidence of AF. In contrast, it is also reported that anabolic steroid use is associated with an increase in the risk of AF. To date, no study has explored the effect of testosterone normalization on new incidence of AF after testosterone replacement therapy (TRT) in patients with low testosterone.

Methods and Results—Using data from the Veterans Administrations Corporate Data Warehouse, we identified a national cohort of 76 639 veterans with low TT levels and divided them into 3 groups. Group 1 had TRT resulting in normalization of TT levels (normalized TRT), group 2 had TRT without normalization of TT levels (nonnormalized TRT), and group 3 did not receive TRT (no TRT). Propensity score–weighted stabilized inverse probability of treatment weighting Cox proportional hazard methods were used for analysis of the data from these groups to determine the association between post-TRT levels of TT and the incidence of AF. Group 1 (40 856 patients, median age 66 years) had significantly lower risk of AF than group 2 (23 939 patients, median age 65 years; hazard ratio 0.90, 95% CI 0.81–0.99, *P*=0.0255) and group 3 (11 853 patients, median age 67 years; hazard ratio 0.79, 95% CI 0.70–0.89, *P*=0.0001). There was no statistical difference between groups 2 and 3 (hazard ratio 0.89, 95% CI 0.78–1.0009, *P*=0.0675) in incidence of AF.

Conclusions—These novel results suggest that normalization of TT levels after TRT is associated with a significant decrease in the incidence of AF. (*J Am Heart Assoc.* 2017;6:e004880. DOI: 10.1161/JAHA.116.004880.)

Key Words: atrial fibrillation • testosterone • testosterone replacement therapy

8:55 AM · May 12, 2017 · Twitter Web Client

Normalization of Testosterone Levels After Testosterone Replacement Therapy Is Associated With Decreased Incidence of Atrial Fibrillation

Rishi Sharma, MD, MHSA; Olurinde A. Oni, MBBS, MPH; Kamal Gupta, MD; Mukut Sharma, PhD; Ram Sharma, PhD; Vikas Singh, MD, MHSA; Deepak Parashara, MD; Surineni Kamalakar, MBBS, MPH; Buddhadeb Dawn, MD; Guoqing Chen, MD, PhD, MPH; John A. Ambrose, MD; Rajat S. Barua, MD, PhD

***Background*—**Atrial fibrillation (AF) is the most common cardiac dysrhythmia associated with significant morbidity and mortality. Several small studies have reported that low serum total testosterone (TT) levels were associated with a higher incidence of AF. In contrast, it is also reported that anabolic steroid use is associated with an increase in the risk of AF. To date, no study has explored the effect of testosterone normalization on new incidence of AF after testosterone replacement therapy (TRT) in patients with low testosterone.

***Methods and Results*—**Using data from the Veterans Administrations Corporate Data Warehouse, we identified a national cohort of 76 639 veterans with low TT levels and divided them into 3 groups. Group 1 had TRT resulting in normalization of TT levels (normalized TRT), group 2 had TRT without normalization of TT levels (nonnormalized TRT), and group 3 did not receive TRT (no TRT). Propensity score–weighted stabilized inverse probability of treatment weighting Cox proportional hazard methods were used for analysis of the data from these groups to determine the association between post-TRT levels of TT and the incidence of AF. Group 1 (40 856 patients, median age 66 years) had significantly lower risk of AF than group 2 (23 939 patients, median age 65 years; hazard ratio 0.90, 95% CI 0.81–0.99, P=0.0255) and group 3 (11 853 patients, median age 67 years; hazard ratio 0.79, 95% CI 0.70–0.89, P=0.0001). There was no statistical difference between groups 2 and 3 (hazard ratio 0.89, 95% CI 0.78– 1.0009, P=0.0675) in incidence of AF.

Conclusions*—**These novel results suggest that normalization of TT levels after TRT is associated with a significant decrease in the incidence of AF. (J Am Heart Assoc.*** **2017;6:e004880. DOI: 10.1161/JAHA.116.004880.)**

Key Words: atrial fibrillation • testosterone • testosterone replacement therapy

Mike Babyak
@mababyak

Replying to @_MiguelHernan

I've often tried to make a similar point to colleagues. They would never dream of ignoring medical consensus on the approach to an assay or dx procedure, but often cast statisticians as being "fussy" for trying to have them adhere to best statistical practice.

10:06 AM · Dec 31, 2017 · Twitter Web Client

Ken Rothman
@ken_rothman

Replying to @oncology_bg @pash22 and 8 others

.We shouldn't be "deciding" to reject or accept. We should be measuring effects. See, e.g.,

> **Alvaro Alonso** @alonso_epi · May 12, 2017
> More p-value silliness. HR 0.90, 95%CI 0.81-0.99--> 'effect'; HR 0.89, 95%CI 0.78-1.0009-->no 'effect' jaha.ahajournals.org/content/6/5/e0... @ken_rothman

George Cobb's Questions (with Answers)

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?

A: Because that's **still** what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

"Researcher Degrees of Freedom", 1

> [I]t is unacceptably easy to publish *statistically significant* evidence consistent with any hypothesis.

> The culprit is a construct we refer to as **researcher degrees of freedom**. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?

Simmons et al.

"Researcher Degrees of Freedom", 2

> … It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields statistical significance, and to then report only what worked. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

For more, see

- Gelman's blog 2012-11-01 "Researcher Degrees of Freedom",
- Simmons and others, defining the term.

And this is really hard to deal with...

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or p-hacking and the research hypothesis was posited ahead of time

> Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

- Gelman and Loken

Lakens et al. Justify Your Alpha

"In response to recommendations to redefine statistical significance to \(p \leq .005\), we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level." Visit link.

Abandon Statistical Significance

Gelman blog 2017-09-26 on "Abandon Statistical Significance"

"Measurement error and variation are concerns even if your estimate is more than 2 standard errors from zero. Indeed, if variation or measurement error are high, then you learn almost nothing from an estimate even if it happens to be 'statistically significant.'"

Read the whole paper here

JAMA 2018-04-10

VIEWPOINT

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Meta-Research Innovation Center at Stanford, Departments of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics, Stanford University, Stanford, California.

The Proposal to Lower *P* Value Thresholds to .005

P values and accompanying methods of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report *P* values in the abstract, full text, or both include some values of .05 or less.[1] However, many of the claims that these reports highlight are likely false.[2] Recognizing the major importance of the statistical significance conundrum, the American Statistical Association (ASA) published[3] a statement on *P* values in 2016. The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious. The contributors to the ASA statement also wrote 20 independent, accompanying commentaries focusing on different aspects and prioritizing different solutions. Another large coalition of 72 methodologists recently proposed[4] a specific, simple move: lowering the routine *P* value threshold for claiming statistical significance from .05 to .005 for new discoveries. The proposal met with strong endorsement in some circles and concerns in others.

P values are misinterpreted, overtrusted, and misused. The language of the ASA statement enables the dis-

fully considered how low a *P* value should be for a research finding to have a sufficiently high chance of being true. For example, adoption of genome-wide significance thresholds ($P < 5 \times 10^{-8}$) in population genomics has made discovered associations highly replicable and these associations also appear consistently when tested in new populations. The human genome is very complex, but the extent of multiplicity of significance testing involved is known, the analyses are systematic and transparent, and a requirement for $P < 5 \times 10^{-8}$ can be cogently arrived at.

However, for most other types of biomedical research, the multiplicity involved is unclear and the analyses are nonsystematic and nontransparent. For most observational exploratory research that lacks preregistered protocols and analysis plans, it is unclear how many analyses were performed and what various analytic paths were explored. Hidden multiplicity, nonsystematic exploration, and selective reporting may affect even experimental research and randomized trials. Even though it is now more common to have a preexisting protocol and statistical analysis plan and preregistration of

Blume et al. PLoS ONE (2018) 13(3): e0188299

RESEARCH ARTICLE

Second-generation *p*-values: Improved rigor, reproducibility, & transparency in statistical analyses

Jeffrey D. Blume[1]*, Lucy D'Agostino McGowan[2], William D. Dupont[3], Robert A. Greevy, Jr.[1]

Second-generation *p* values

Verifying that a statistically significant result is scientifically meaningful is not only good scientific practice, it is a natural way to control the Type I error rate. Here we introduce a novel extension of the *p*-value—a second-generation *p*-value (p_δ)–that formally accounts for scientific relevance and leverages this natural Type I Error control. The approach relies on a pre-specified interval null hypothesis that represents the collection of effect sizes that are scientifically uninteresting or are practically null. The second-generation *p*-value is the proportion of data-supported hypotheses that are also null hypotheses. As such, second-generation *p*-values indicate when the data are compatible with null hypotheses ($p_\delta = 1$), or with alternative hypotheses ($p_\delta = 0$), or when the data are inconclusive ($0 < p_\delta < 1$). Moreover, second-generation *p*-values provide a proper scientific adjustment for multiple comparisons and reduce false discovery rates. This is an advance for environments rich in data, where traditional *p*-value adjustments are needlessly punitive. Second-generation *p*-values promote transparency, rigor and reproducibility of scientific results by *a priori* specifying which candidate hypotheses are practically meaningful and by providing a more reliable statistical summary of when the data are compatible with alternative or null hypotheses.

Nature P values are just the tip of the iceberg!

COMMENT

P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

OK, so what SHOULD we do?

The American Statistician Volume 73, 2019, Supplement 1

Articles on:

1. Getting to a Post "$p < 0.05$" Era
2. Interpreting and Using *p*
3. Supplementing or Replacing *p*
4. Adopting more holistic approaches
5. Reforming Institutions: Changing Publication Policies and Statistical Education

- Note that there is an enormous list of "things to do" in Section 7 of the main editorial, too.

Statistical Inference in the 21st Century

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p<0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

To link to this article: https://doi.org/10.1080/00031305.2019.1583913

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

> We can make acceptance of uncertainty more natural to our thinking by accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate. Reporting and interpreting point and interval estimates should be routine.

> How will accepting uncertainty change anything? To begin, it will prompt us to seek better measures, more sensitive designs, and larger samples, all of which increase the rigor of research.

> It also helps us be modest … [and] leads us to be thoughtful.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

3.2. Be Thoughtful

What do we mean by this exhortation to "be thoughtful"? Researchers already clearly put much thought into their work. We are not accusing anyone of laziness. Rather, we are envisioning a sort of "statistical thoughtfulness." In this perspective, statistically **thoughtful researchers** begin above all else with clearly expressed objectives. They recognize when they are doing exploratory studies and when they are doing more rigidly pre-planned studies. They invest in producing solid data. They consider not one but a multitude of data analysis techniques. And they think about so much more.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research looks ahead to prospective outcomes in the context of theory and previous research. Researchers would do well to ask, *What do we already know, and how certain are we in what we know?* And building on that and on the field's theory, *what magnitudes of differences, odds ratios, or other effect sizes are practically important?* These questions would naturally lead a researcher, for example, to use existing evidence from a literature review to identify specifically the findings that would be practically important for the key outcomes under study.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research includes careful consideration of the definition of a meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed. Afterwards is just too late; it is dangerously easy to justify observed results after the fact and to overinterpret trivial effect sizes as being meaningful. Many authors in this special issue argue that consideration of the effect size and its "scientific meaningfulness" is essential for reliable inference (e.g., Blume et al. 2019; Betensky 2019). This concern is also addressed in the literature on equivalence testing (Wellek 2017).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

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" (McShane et al. 2019).

Thoughtful researchers "use a toolbox of statistical techniques, employ good judgment, and keep an eye on developments in statistical and data science," conclude Heck and Krueger (2019), who demonstrate how the *p*-value can be useful to researchers as a heuristic.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

In all instances, regardless of the value taken by p or any other statistic, consider what McShane et al. (2019) call the "currently subordinate factors"—the factors that should no longer be subordinate to "$p < 0.05$." These include relevant prior evidence, plausibility of mechanism, study design and data quality, and the real-world costs and benefits that determine what effects are scientifically important. The scientific context of your study matters, they say, and this should guide your interpretation.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

To **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 open by providing sufficient information so that other researchers can execute meaningful alternative analyses. van Dongen et al. (2019) provide an illustrative example of such alternative analyses by different groups attacking the same problem.

Being open goes hand in hand with **being modest.**

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Being modest requires a reality check (Amrhein, Trafimow, and Greenland 2019). "A core problem," they observe, "is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct and of the reality being studied.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Be modest in recognizing there is not a "true statistical model" underlying every problem, which is why it is wise to **thoughtfully** consider many possible models (Lavine 2019).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Be modest about the role of statistical inference in scientific inference. "Scientific inference is a far broader concept than statistical inference," says Hubbard, Haig, and Parsa (2019). "A major focus of scientific inference can be viewed as the pursuit of *significant sameness*, meaning replicable and empirically generalizable results among phenomena. Regrettably, the obsession with users of statistical inference to report *significant differences* in data sets actively thwarts cumulative knowledge development."

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

> The nexus of openness and modesty is to report everything while at the same time not concluding anything from a single study with unwarranted certainty. Because of the strong desire to inform and be informed, there is a relentless demand to state results with certainty. Again, accept uncertainty and embrace variation in associations and effects, because they are always there, like it or not. Understand that expressions of uncertainty are themselves uncertain. Accept that one study is rarely definitive, so encourage, sponsor, conduct, and publish replication studies.

> Be modest by encouraging others to reproduce your work. Of course, for it to be reproduced readily, you will necessarily have been thoughtful in conducting the research and open in presenting it.

What I Think I Think Now

- Null hypothesis significance testing is much harder than I thought.
  - The null hypothesis is almost never a real thing.
  - Rather than rejiggering the cutoff, I would largely abandon the *p* value as a summary
  - Replication is far more useful than I thought it was.
- Some hills aren't worth dying on.
  - Think about uncertainty intervals more than confidence or credible intervals

What I Think I Think Now

- Which method to use is far less important than finding better data
  - The biggest mistake I make regularly is throwing away useful data
  - I'm not the only one with this problem.
- The best thing I do most days is communicate more clearly.
  - When stuck in a design, I think about how to get better data.
  - When stuck in an analysis, I try to turn a table into a graph.
- I have A LOT to learn.