431 Class 24

thomaselove.github.io/431

2021-11-18

Today's Agenda

- What I Taught for Many Years
- p-Hacking
- Do Confidence Intervals Solve the Problem?
- Borrowing from Bayesian Ideas
- Replicable Research and the Crisis in Science
- Retrospective Power and why most smart folks avoid it
  - Type S and Type M error: Saying something more useful

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 α, 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 $p = 1$.

American Statistical Association to the rescue!?!

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."

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

... 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.**

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

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 *statement* on *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.

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[1]

[1] http://www.nature.com/news/psychology-journal-bans-p-values-1.17001 describes the recent banning of null hypothesis significance testing by *Basic and Applied Psychology*.

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

$p = 0.05$?

"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."

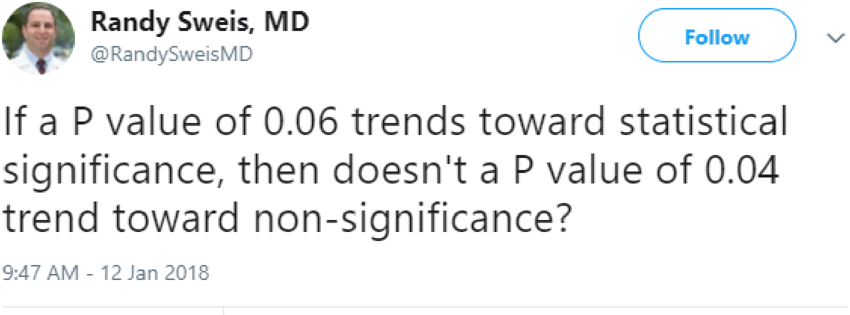
- NOVA "Rethinking Science's Magic Number" by Tiffany Dill 2018-02-28. See especially the video labeled "Science's most important (and controversial) number has its origins in a British experiment involving milk and tea."

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 Prediction by the Numbers 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

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.

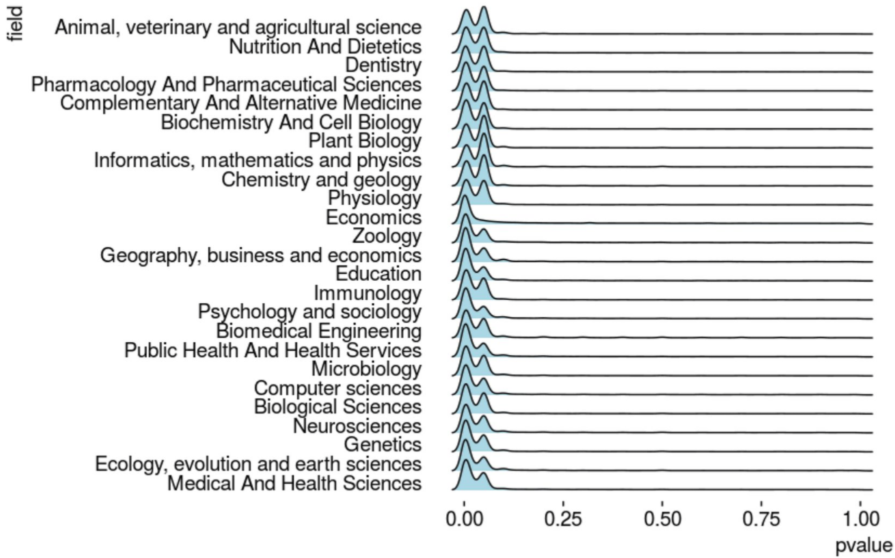
All the p values

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.

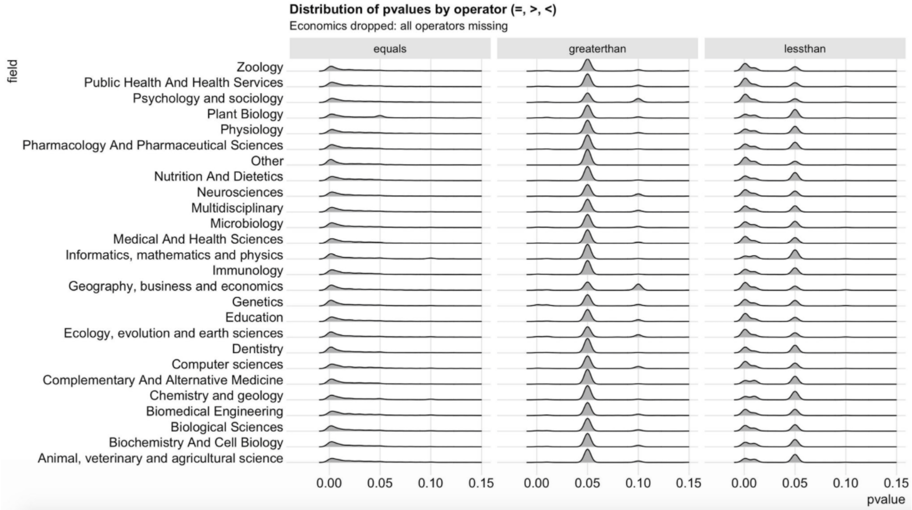
- Visit Jeff Leek's Github for tidypvals package
  - 2.5 million p values in 25 scientific fields

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;

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.,

p-Hacking

Hack Your Way To Scientific Glory (fivethirtyeight)

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [x] Governors
- [x] Senators
- [ ] Representatives

How do you want to measure economic performance?

- [ ] Employment
- [x] Inflation
- [x] GDP
- [x] Stock prices

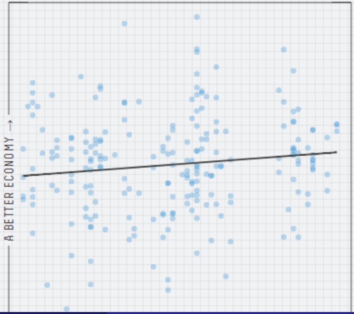
Other options

- [x] Factor in power

Weight more powerful positions more heavily

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.06** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy,

"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. *link*

"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",
- Paper by *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

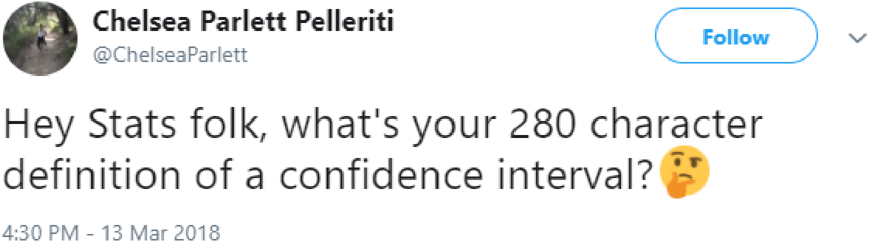
> *Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the 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.*

- *Link* to the paper from Gelman and Loken

Grim Reality

- Editorial, Educational and Other Institutional Practices Will Have to Change
- It Is Going to Take Work, and It Is Going to Take Time
- Why Will Change Finally Happen Now?

Confidence Intervals - do they solve our problem?



Confidence Intervals - do they solve our problem?

Thomas Leeper
@thosjleeper

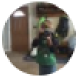
Follow

Replying to @ChelseaParlett

An interval drawn such that, were repeated, equal-sized samples of units drawn from the population of units using an identical sampling procedure and the same estimator was applied to each sample, 100*(1-α)% of those intervals would contain the population parameter of interest.

4:58 PM - 13 Mar 2018

Confidence Intervals - do they solve our problem?



Joran Elias
@joranelias

Follow

A confidence interval is a meaure of uncertainty such that all definitions of it elicit corrections from Bayesians.

(Didn't need all 280.)

Confidence Intervals - do they solve our problem?

Jenny Bryan
@JennyBryan

Following

Pedantry about the definition of a confidence interval ... why is this the hill statisticians choose to die on? Every time you feel the urge, go convert a table to a figure. It is likely to do more good.

Confidence Intervals - do they solve our problem?

Frank Harrell @f2harrell · 28 Dec 2017

Tables and figures are important but so is this. We need to get this right. Too many faulty conclusions being drawn with frequentist statistical analysis. If one is going to be a frequentist one should make exactly correct interpretations.

2 10

Jenny Bryan
@JennyBryan

Following

Replying to @f2harrell

I just feel like the people we're often trying to reach aren't making informed comparisons of frequentist vs Bayesian methods, they're still struggling with decision making under uncertainty

Using Bayesian Ideas: Confidence Intervals

> *My current favorite (hypothetical) example is an epidemiology study of some small effect where the point estimate of the odds ratio is 3.0 with a 95% conf interval of [1.1, 8.2].*
>
> *As a 95% conf interval, this is fine (assuming the underlying assumptions regarding sampling, causal identification, etc. are valid).*

(but on some level you need to deal with the fact that. . .)

> *. . . real-world odds ratios are much more likely to be near 1.1 than to be near 8.2.*

See Gelman 2014-12-11.

Uncertainty intervals?

I've (Gelman) become increasingly uncomfortable with the term "confidence interval" for several reasons:

- The well-known difficulties in interpretation (officially the confidence statement can be interpreted only on average, but people typically implicitly give the Bayesian interpretation to each case.)
- The ambiguity between confidence intervals and predictive intervals.
- The awkwardness of explaining that confidence intervals are big in noisy situations where you have less confidence, and confidence intervals are small when you have more confidence.

  So here's my proposal. Let's use the term "uncertainty interval" instead. The uncertainty interval tells you how much uncertainty you have.

See Gelman 2010-12-21.

Some Noisy Recent Suggestions

Benjamin et al 2017 Redefine Statistical Significance

We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

Motivations:

- links to Bayes Factor interpretation
- 0.005 is stringent enough to "break" the current system - makes it very difficult for researchers to reach threshold with noisy, useless studies.

Visit the main *article*. Visit an explanatory piece in *Science*.

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

Opinion

VIEWPOINT

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

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.

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.

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.

- Statistical methods do not rid data of their uncertainty.

Statistical methods do not rid data of their uncertainty. "Statistics," Gelman (2016) says, "is often sold as a sort of alchemy that transmutes randomness into certainty, an 'uncertainty laundering' that begins with data and concludes with success as measured by statistical significance." To accept uncertainty requires that we "treat statistical results as being much more incomplete and uncertain than is currently the norm" (Amrhein, Trafimow, and Greenland 2019). We must "countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error" (Calin-Jageman and Cumming 2019).

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
  - Retrospective calculations about Type S (sign) and Type M (magnitude) errors can help me illustrate ideas.
- 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.