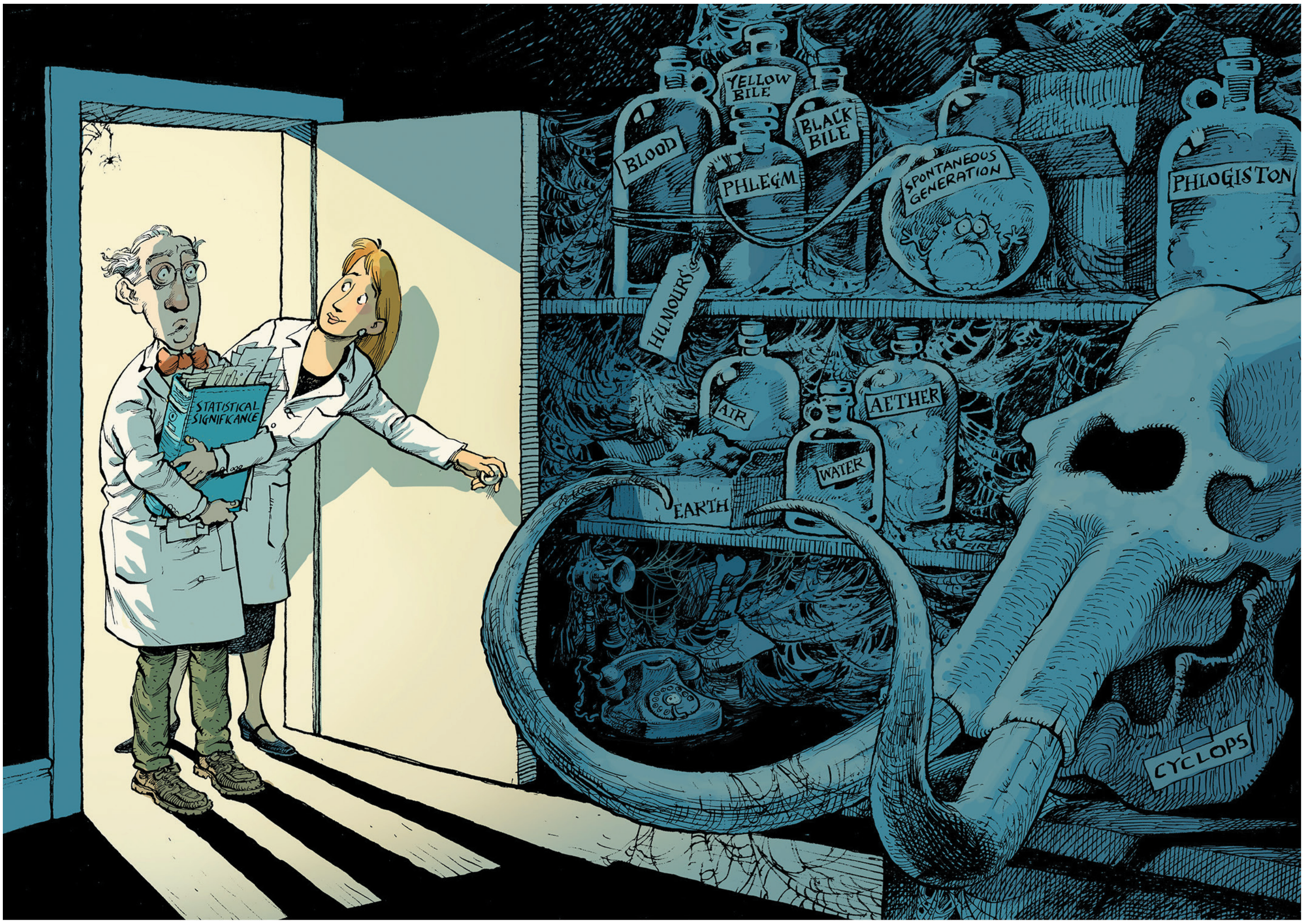
COMMENT

ILLUSTRATION BY DAVID PARKINS



Retire statistical significance

Valentin Amrhein, **Sander Greenland**, **Blake McShane** and more than 800 signatories

call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)[1]. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

PERVASIVE PROBLEM

Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a *P* value is larger than a threshold such as 0.05 ▶

© 2019 Springer Nature Limited. All rights reserved.

▶ or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.

For example, consider a series of analyses of unintended effects of anti-inflammatory drugs[2]. Because their results were statistically non-significant, one set of researchers concluded that exposure to the drugs was "not associated" with new-onset atrial fibrillation (the most common disturbance to heart rhythm) and that the results stood in contrast to those from an earlier study with a statistically significant outcome.

Now, let's look at the actual data. The researchers describing their statistically non-significant results found a risk ratio of 1.2 (that is, a 20% greater risk in exposed patients relative to unexposed ones). They also found a 95% confidence interval that spanned everything from a trifling risk decrease of 3% to a considerable risk increase of 48% ($P = 0.091$; our calculation). The researchers from the earlier, statistically significant, study found the exact same risk ratio of 1.2. That study was simply more precise, with an interval spanning from 9% to 33% greater risk ($P = 0.0003$; our calculation).

It is ludicrous to conclude that the statistically non-significant results showed "no association", when the interval estimate included serious risk increases; it is equally absurd to claim these results were in contrast with the earlier results showing an identical observed effect. Yet these common practices show how reliance on thresholds of statistical significance can mislead us (see 'Beware false conclusions').

These and similar errors are widespread. Surveys of hundreds of articles have found that statistically non-significant results are interpreted as indicating 'no difference' or 'no effect' in around half (see 'Wrong interpretations' and Supplementary Information).

In 2016, the American Statistical Association released a statement in *The American Statistician* warning against the misuse of statistical significance and *P* values. The issue also included many commentaries on the subject. This month, a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on 'Statistical inference in the 21st century: a world beyond $P < 0.05$'. The editors introduce the collection with the caution "don't say 'statistically significant'"[3]. Another article[4] with dozens of signatories also calls on authors and journal editors to disavow those terms.

We agree, and call for the entire concept of statistical significance to be abandoned.

"Eradicating categorization will help to halt overconfident claims, unwarranted declarations of 'no difference' and absurd statements about 'replication failure'."

We are far from alone. When we invited others to read a draft of this comment and sign their names if they concurred with our message, 250 did so within the first 24 hours. A week later, we had more than 800 signatories — all checked for an academic affiliation or other indication of present or past work in a field that depends on statistical modelling (see the list and final count of signatories in the Supplementary Information). These include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica. One advocate called it a "surgical strike against thoughtless testing of statistical significance" and "an opportunity to register your voice in favour of better scientific practices".

We are not calling for a ban on *P* values. Nor are we saying they cannot be used as a decision criterion in certain specialized applications (such as determining whether a manufacturing process meets some quality-control standard). And we are also not advocating for an anything-goes situation, in which weak evidence suddenly becomes credible. Rather, and in line with many others over the decades, we are calling for a stop to the use of *P* values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis[5].

QUIT CATEGORIZING

The trouble is human and cognitive more than it is statistical: bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different[6–8]. The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.

Unfortunately, the false belief that crossing the threshold of statistical significance is enough to show that a result is 'real' has led scientists and journal editors to privilege such results, thereby distorting the literature. Statistically significant estimates are biased upwards in magnitude and potentially to a large degree, whereas statistically non-significant estimates are biased downwards in magnitude. Consequently, any discussion that focuses on estimates chosen for their significance will be biased. On top of this, the rigid focus on statistical significance encourages researchers to choose data and methods that yield statistical significance for some desired (or simply publishable) result, or that yield statistical non-significance for an undesired result, such as potential side effects of drugs — thereby invalidating conclusions.

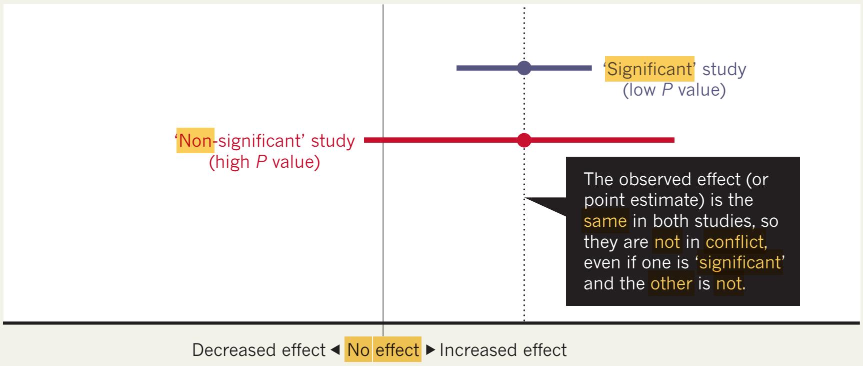
The pre-registration of studies and a commitment to publish all results of all analyses can do much to mitigate these issues. However, even results from pre-registered studies can be biased by decisions invariably left open in the analysis plan[9]. This occurs even with the best of intentions.

Again, we are not advocating a ban on *P* values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.

One reason to avoid such 'dichotomania' is that all statistics, including *P* values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree. In fact, random variation alone can easily lead to large disparities in *P* values, far beyond falling just to either side of the 0.05 threshold. For example, even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving $P < 0.05$, it would not be very surprising for one to obtain $P < 0.01$ and the other $P > 0.30$.

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.



SOURCE: V. AMRHEIN *ET AL.*

© 2019 Springer Nature Limited. All rights reserved.

Whether a *P* value is small or large, caution is warranted.

We must learn to embrace uncertainty. One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence. Specifically, we recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits. In doing so, they should remember that all the values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval[7,10]. Therefore, singling out one particular value (such as the null value) in the interval as 'shown' makes no sense.

We're frankly sick of seeing such nonsensical 'proofs of the null' and claims of non-association in presentations, research articles, reviews and instructional materials. An interval that contains the null value will often also contain non-null values of high practical importance. That said, if you deem all of the values inside the interval to be practically unimportant, you might then be able to say something like 'our results are most compatible with no important effect'.

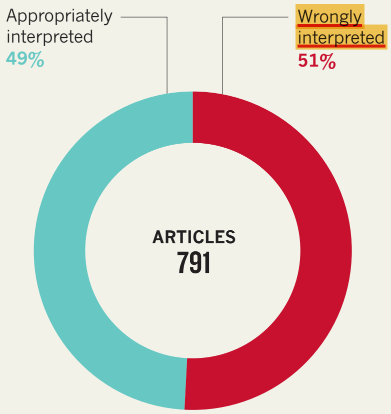
When talking about compatibility intervals, bear in mind four things. First, just because the interval gives the values most compatible with the data, given the assumptions, it doesn't mean values outside it are incompatible; they are just less compatible. In fact, values just outside the interval do not differ substantively from those just inside the interval. It is thus wrong to claim that an interval shows all possible values.

Second, not all values inside are equally compatible with the data, given the assumptions. The point estimate is the most compatible, and values near it are more compatible than those near the limits. This is why we urge authors to discuss the point estimate, even when they have a large *P* value or a wide interval, as well as discussing the limits of that interval. For example, the authors above could have written: 'Like a previous study, our results suggest a 20% increase in risk of new-onset atrial fibrillation in patients given the anti-inflammatory drugs. Nonetheless, a risk difference ranging from a 3% decrease, a small negative association, to a 48% increase, a substantial positive association, is also reasonably compatible with our data, given our assumptions.' Interpreting the point estimate, while acknowledging its uncertainty, will keep you from making false declarations of 'no difference', and from making overconfident claims.

Third, like the 0.05 threshold from which it came, the default 95% used to compute intervals is itself an arbitrary convention. It is based on the false idea that there is a 95% chance that the computed interval itself contains the true value, coupled with the vague feeling that this is a basis for a confident decision. A different level can be justified, depending on the application. And, as in the anti-inflammatory-drugs example, interval estimates can perpetuate the problems of statistical significance when the dichotomization they impose is treated as a scientific standard.

WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.



*Data taken from: P. Schatz *et al. Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* **33**, 1–15 (2017).

SOURCE: V. AMRHEIN *ET AL.*

Last, and most important of all, be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval. In practice, these assumptions are at best subject to considerable uncertainty[7,8,10]. Make these assumptions as clear as possible and test the ones you can, for example by plotting your data and by fitting alternative models, and then reporting all results.

Whatever the statistics show, it is fine to suggest reasons for your results, but discuss a range of potential explanations, not just favoured ones. Inferences should be scientific, and that goes far beyond the merely statistical. Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often more important than statistical measures such as *P* values or intervals.

The objection we hear most against retiring statistical significance is that it is needed to make yes-or-no decisions. But for the choices often required in regulatory, policy and business environments, decisions based on the costs, benefits and likelihoods of all potential consequences always beat those made based solely on statistical significance. Moreover, for decisions about whether to pursue a research idea further, there is no simple connection between a *P* value and the probable results of subsequent studies.

What will retiring statistical significance look like? We hope that methods sections and data tabulation will be more detailed and nuanced. Authors will emphasize their estimates and the uncertainty in them — for example, by explicitly discussing the lower and upper limits of their intervals. They will not rely on significance tests. When *P* values are reported, they will be given with sensible precision (for example, $P=0.021$ or $P=0.13$) — without adornments such as stars or letters to denote statistical significance and not as binary inequalities ($P<0.05$ or $P>0.05$). Decisions to interpret or to publish results will not be based on statistical thresholds. People will spend less time with statistical software, and more time thinking.

Our call to retire statistical significance and to use confidence intervals as compatibility intervals is not a panacea. Although it will eliminate many bad practices, it could well introduce new ones. Thus, monitoring the literature for statistical abuses should be an ongoing priority for the scientific community. But eradicating categorization will help to halt overconfident claims, unwarranted declarations of 'no difference' and absurd statements about 'replication failure' when the results from the original and replication studies are highly compatible. The misuse of statistical significance has done much harm to the scientific community and those who rely on scientific advice. *P* values, intervals and other statistical measures all have their place, but it's time for statistical significance to go. ■

Valentin Amrhein *is a professor of zoology at the University of Basel, Switzerland.* **Sander Greenland** *is a professor of epidemiology and statistics at the University of California, Los Angeles.* **Blake McShane** *is a statistical methodologist and professor of marketing at Northwestern University in Evanston, Illinois. For a full list of co-signatories, see Supplementary Information.*
e-mail: v.amrhein@unibas.ch

1. Fisher, R. A. *Nature* **136**, 474 (1935).
2. Schmidt, M. & Rothman, K. J. *Int. J. Cardiol.* **177**, 1089–1090 (2014).
3. Wasserstein, R. L., Schirm, A. & Lazar, N. A. *Am. Stat.* https://doi.org/10.1080/00031305.2019.1583913 (2019).
4. Hurlbert, S. H., Levine, R. A. & Utts, J. *Am. Stat.* https://doi.org/10.1080/00031305.2018.1543616 (2019).
5. Lehmann, E. L. *Testing Statistical Hypotheses* 2nd edn 70–71 (Springer, 1986).
6. Gigerenzer, G. *Adv. Meth. Pract. Psychol. Sci.* **1**, 198–218 (2018).
7. Greenland, S. *Am. J. Epidemiol.* **186**, 639–645 (2017).
8. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. *Am. Stat.* https://doi.org/10.1080/00031305.2018.1527253 (2019).
9. Gelman, A. & Loken, E. *Am. Sci.* **102**, 460–465 (2014).
10. Amrhein, V., Trafimow, D. & Greenland, S. *Am. Stat.* https://doi.org/10.1080/00031305.2018.1543137 (2019).

Supplementary information accompanies this article; see go.nature.com/2tc5nkm

© 2019 Springer Nature Limited. All rights reserved.

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 73 • NUMBER S1 MARCH 2019

The American Statistician

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

What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

To cite this article: John P. A. Ioannidis (2019) What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?, The American Statistician, 73:sup1, 20-25, DOI: 10.1080/00031305.2018.1447512

To link to this article: https://doi.org/10.1080/00031305.2018.1447512

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 2592

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utas20

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 20–25: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1447512

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University and Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA

ABSTRACT

P values linked to null hypothesis significance testing (NHST) is the most widely (mis)used method of statistical inference. Empirical data suggest that across the biomedical literature (1990–2015), when abstracts use *P* values 96% of them have *P* values of 0.05 or less. The same percentage (96%) applies for full-text articles. Among 100 articles in PubMed, 55 report *P* values, while only 4 present confidence intervals for all the reported effect sizes, none use Bayesian methods and none use false-discovery rate. Over 25 years (1990–2015), use of *P* values in abstracts has doubled for all PubMed, and tripled for meta-analyses, while for some types of designs such as randomized trials the majority of abstracts report *P* values. There is major selective reporting for *P* values. Abstracts tend to highlight most favorable *P* values and inferences use even further spin to reach exaggerated, unreliable conclusions. The availability of large-scale data on *P* values from many papers has allowed the development and applications of methods that try to detect and model selection biases, for example, p-hacking, that cause patterns of excess significance. Inferences need to be cautious as they depend on the assumptions made by these models and can be affected by the presence of other biases (e.g., confounding in observational studies). While much of the unreliability of past and present research is driven by small, underpowered studies, NHST with *P* values may be also particularly problematic in the era of overpowered big data. NHST and *P* values are optimal only in a minority of current research. Using a more stringent threshold, as in the recently proposed shift from $P < 0.05$ to $P < 0.005$, is a temporizing measure to contain the flood and death-by-significance. NHST and *P* values may be replaced in many fields by other, more fit-for-purpose, inferential methods. However, curtailing selection biases requires additional measures, beyond changes in inferential methods, and in particular reproducible research practices.

ARTICLE HISTORY
Received November 2017
Revised February 2018

KEYWORDS
Bias; P-value; Statistical significance

1. Introduction

Null hypothesis significance testing (NHST) and *P* value thresholds such as 0.05 have long been a mainstay of empirical work in the sciences. Increasingly, however, statisticians and other scientists concerned with learning from data have come to recognize major shortcomings in the way these methods are used. This paper, based on an invited plenary address to a recent ASA-sponsored workshop on statistical inference, summarizes recent empirical work on the use and misuse of *P* values and places in context what we have learnt towards solving this conundrum.

In what follows, Section 2 summarizes empirical results from a database of 13 million abstracts and 844 thousand full articles taken from PubMed Central between 1990 and 2015. Section 3 discusses how bias emerges from a multilayered selection process that leads to specific reported *P* values. Section 4 describes and discusses a variety of proposed remedies intended to address the problem of selection bias. These include: (4.1) alternative approaches to inference (effect sizes and confidence intervals, Bayesian methods, changing the *P* value threshold); (4.2) attempts to model the selection process (the *P* value curve and meta-analysis of publication selection); (4.3) examples of alternatives based on context and goals; and, finally (4.4) how reproducible research practices might offer the best solution. A concluding section offers some final thoughts.

2. Empirical Results: NHST is Widespread and Reliance on *P* Values Increases Over Time

There are over 100 million published articles in the scientific literature (Khabsa and Giles 2014), and a substantial proportion of them use data. Among those that use data an increasing proportion use also some tools of statistical inference beyond simple description. Different scientific fields use different statistical tools by tradition, but their traditions are not necessarily justified or fit-for-purpose. Convenience, inertia, poor quantitative and statistical training of scientists, and lack of initiative from journals and funding agencies may perpetuate the use and misuse of those tools (Szucs and Ioannidis 2017a).

In particular, the use and misuse of *P* values is, arguably, the most widely perpetrated misdeed of statistical inference across all of science (Chavalarias et al. 2016). NHST coupled with the use of *P* value thresholds dominates most fields in the biomedical and life sciences, social sciences, and physical sciences. Most fields use *P* value thresholds of 0.05 to differentiate in black-and-white fashion between "significant" and "nonsignificant" results. Exceptions do occur, for example, the use of *P* value thresholds of 3×10^{-7} (5 sigma) in high-energy physics of 5×10^{-8} (genome-wide significance) in genome epidemiology, but they are relatively uncommon when the scientific literature is seen in its total volume.

CONTACT John P. A. Ioannidis jioannid@stanford.edu Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University, 1265 Welch Road, Medical School Office Building Room X306, Stanford CA 94305.

The paper is based on the opening keynote on the same topic at the American Statistical Association Symposium on Statistical Inference, Bethesda, Maryland, October 2017.

© 2019 The Author. Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

The remainder of this section first describes a database of published biomedical literature, then summarizes results obtained from text mining of this database (Chavalarias et al. 2016).

- *The database from PubMed and PubMed Central*

The results presented in this section are based on a survey of the entire biomedical literature published during the quarter-century from 1990 to 2015. Text mining was used to assess the presence of *P* values in the abstracts of 16.2 million items (13.0 million of which had an abstract). Similar text mining was performed in PubMed Central (PMC) for 844,000 full-text articles. For details, see ref. 3. Across this large corpus of biomedical literature, when abstracts use *P* values 96% of them have *P* values of 0.05 or less. The same percentage (96%) applies for full-text articles. This is too good to be true. Dissecting the use and misuse or *P* values may explain why.

- *The use of null hypothesis significance testing and P value thresholds is widespread*

The proportion of PMC papers with *P* values in their abstract or text is 51.1% for all papers. However, this figure is an underestimate, because the text mining could not capture most *P* values embedded in tables and figures. Manual evaluation of the full articles (including tables and figures) in a sample of 100 randomly selected articles from PubMed, found 55 that report *P* values. The use of other tools of statistical inference is rare or nonexistent: of the 100 papers, four report confidence intervals for all the reported effect sizes, none uses Bayesian methods, and none uses methods based on false discovery rates.

- *The rate is higher in clinical journals and in meta-analyses*

Although the use of *P* values is widespread, there are categories of studies for which the percentage reporting p-values is substantially higher than the overall rate. Among these categories are:

Overall (all papers)	51.1%
Articles published in core clinical journals	78.4%
Meta-analyses	82.8%
Randomized controlled trials	76.0%
Other clinical trials (excluding randomized controlled trials)	75.7%

To the extent that *P* values are misused, or are used in place of other more suitable methods, these high percentages are particularly concerning.

- *Reliance on P values is increasing over time*

The same survey (Chavalarias et al. 2016) suggests that the use of *P* values has increased over the 25 years covered by the sample. For all Pubmed abstracts, the percentage of abstracts reporting *P* values doubled from 8% in 1990 to 17% in 2015; the rate tripled for meta-analyses. For some types of designs such as randomized controlled trials, about 60% of articles currently have some *P* value(s) in their abstracts. Of note, the proportion of articles that have *P* values in the full-text is much larger (as shown above) than the proportion of those that have *P* values in the abstract. Abstracts are, of course, highly prominent as they represent the façade of articles in terms of what they communicate.

- *The "typical size" of reported P values is 0.01, more or less*

Most of the reported *P* values are modest. An exception is the tiny fraction (0.4%) of those presented with exponents of 10 (e.g., 2 × 10–8) or "EXP" of "E" notation, for which the mean –log(*P* value) is around 9, the other *P* values have an average –log(*P* value) of 2, corresponding to $P = 0.01$.

Why should the widespread and growing reliance on *P* values be such a concern? A main reason is selection bias. As described in the next section, selection operates at many levels, and the resulting bias substantially inflates the rate of false alarms in the published literature.

3. Selection Effects: The Typical Direction is Toward Claiming Greater Significance

"Selection" refers here to the collection of choices that lead from the planning of a study to the reporting of *P* values. The premise of this section is that such selection occurs in many ways, at many steps in the analysis of a data set or study. At each step, there are choices to be made, and with each choice, there is an opportunity to shape the presentation of results. Section 3.1 offers simple empirical evidence of selection bias, namely, a tendency to choose smaller (more significant) *P* values for inclusion in a paper's abstract. Expanding on this theme, Section 3.2 describes four expanding sets of choices, four layers of selection, as a frame for thinking about sources of bias in the analysis of data. These sets of choices are rarely reported in full transparency, and so remain hidden from the reader. However, the placement of *P* values within the sections of an article (3.3) provides a way to track some of the selection bias. Cherry-picking is more pronounced in the most competitive journals (3.4). Big data sets, which offer greater scope for pattern searching, are correspondingly at greater risk for false positives (3.5).

3.1. P Values in the Abstracts are More "Significant" than P Values in the Full Text

Abstracts offer authors the best opportunity to say they have something important to present. If there is a selection bias at work in the choice of which *P* values to highlight, we would expect to find that bias to show up in a comparison of *P* values in the abstract with *P* values in the full text. Specifically, we would expect an author to select for the abstract some of the more impressive of the *P* values reported in the full text. As a measure of the selection effect, we use the ratio of papers/abstracts reporting *P* values at 0.05 to those reporting *P* values ≤ 0.001. For the papers in the full text sample (Chavalarias et al. 2016), the number of *P* values at 0.05 exceeds by 11% the number of *P* values at 0.001 or less. However, the opposite in seen in the abstracts of these papers, where *P* values of 0.05 are 41% fewer than *P* values of 0.001 or less. Clearly, there is conscious or subconscious selection of more impressive *P* values in the abstracts.

The selection gradient is more steep in the Core Clinical Journals category where in the abstracts *P* values of 0.05 are 73% fewer than the *P* values of = <0.001, while in the full text they are only 16% fewer.

The comparisons here are based on observable data, but they are merely the visible manifestation of a multilayered selection process.

3.2. Layers of Selection for P Values: Which P Values Get Reported

One can think of layers of selection that are applied in the presentation and highlighting of results through *P* values.

- The universe of all *P* values obtained in all the analyses conducted during a scientific process. Unless everything is rigorously prespecified (an uncommon scenario) there can be many trail-and-error efforts at different analyses, a "garden of forking paths" as Andrew Gelman has characterized the process (Gelman 2014). With few exceptions, these iterations and forking paths are not yet documented anywhere. The next layer:
- All the *P* values that other analysts that other authors may obtain, if the original authors can make the data and script/code for their analysis available. If only the data are available without the guidance of a specific analysis plan, one can explore still more options:
- All possible analyses that might be run, for example, using different modeling choices, different adjustments for multivariable models, or different definitions for variables of interest. This can give a sense of the magnitude of the "vibration of effects," that is, by how much results can vary depending on the endorsed exact analytical choices (Patel et al. 15). This layer of variability is of course conditional on the data made available, and does not take into account whatever tailoring the authors may have done to arrive at the version of the data made available to others. Thus, there is a final layer:
- All possible results from pre-processing of the data, for example, trying multiple covariates but only making available in public those included in the "nicest" model (the one presented in the paper). In the absence of full preregistration (Chambers 2013), there is no obstacle to such an approach.

Empirical evaluations (Patel et al. 2015) of the vibration of effects (obtained with different analyses) has shown that if there is sufficient data and degrees of freedom for choices of models almost any result can be obtained. This results in the "Janus phenomenon" where totally opposite results are possible to obtain routinely provided there are sufficient degrees of freedom (Patel et al. 2015).

Most of the time data and script/code are not available, so these selection dilemmas are hidden from an outsider examining a report of a study. However, what can still be visible is the extent of selection within different sections of a published paper.

3.3. Selection Within Sections of a Paper

Section 3.1 compared the set of *P* values reported in the full text of an article with the set of *P* values reported in the abstract. The main finding was that *P* values chosen for the abstract tended to show greater significance than those reported in the text, and that the gradient was more pronounced in some types of journals and types of designs. It is useful to extend this approach by defining a hierarchy of prominence for the location of reported *P* values within an article.

(a) Tables and figures tend to offer the most comprehensive recording of results, although even these may be cherry-picked from among several analyses of the data. Unfortunately, as noted before, text mining is not consistently able to recognize *P* values reported in tables and figures.

(b) *P* values chosen for discussion in the text constitute a subset of all *P* values, and are typically chosen for discussion either because of interest, but occasionally because of some anomaly.

(c) *P* values chosen for the abstract are even more likely to be chosen for impressive significance (Chavalarias et al. 2016) and are sometimes accompanied by implausible effect sizes (Gøtzsche 2006).

(d) Finally, at the top of the hierarchy, are the *P* values taken most seriously by the authors in reaching their conclusions. Conclusions have been documented empirically to depend very often on "spin" (Boutron et al. 2014). With spin, results that fail to register formally as statistically significant can still be taken as "significant."

The typical direction of bias is towards claiming more significance as one moves through these selection steps. Of course, there can be exceptions to this rule, as in some cases nonsignificant *P* values are more attractive, for example, in noninferiority studies, but these tend to be the minority. Also, the exact use of and selection bias on *P* values at these different steps may depend on the type of discipline and the journal where research is published.

3.4. Cherry Picking in the More Competitive Basic Science Journals

For most journals, the tabulated results are likely to be more complete and less selective than the one or few *P* values highlighted in the abstract. However, for extremely competitive basic science journals such as Nature, Science, and PNAS, we have observed (Cristea and Ioannidis 2018) that when authors use *P* values in a Figure or Table (something they do increasingly over time) these *P* values are almost uniformly statistically significant. This uniformity of significant *P* values suggests that cherry-picking has already happened at the step where results are tabulated. It is reasonable to infer that "artificial scarcity"—the availability of very limited print space in prestigious journals—creates pressure to report impressive results (Young et al. 2008).

3.5. P Values, Big Data, and False Positives

An emerging compounding problem is the increasing availability of massive databases that can be analyzed in many scientific fields. While the typical challenge for most scientific work to date has been the conduct of underpowered studies (Szucs and Ioannidis 2017b), "big data" is bringing the reverse challenge of overpowered studies. Massive data sets expand the number of analyses that can be performed, and the multiplicity of possible analyses combines with lenient *P* value thresholds like 0.05 to generate vast potential for false positives. As just one extreme example, an analysis of the entire Swedish population might conclude—if results are taken at face value using lenient *P* value thresholds—that three quarters of medication classes are associated with cancer risk: obviously an impossible result (Patel et al. 2016).

Table 1. Is NHST a good choice for various research applications?

Research application	Is NHST a good choice?
Developing a prognostic score for CVD?	No, selection of variables should not use NHST or use very lenient Type 1 error
Assessing a diagnostic test for depression?	No, absolute magnitude of improvement in diagnostic performance matters more
Evaluating medical therapies in randomized trials?	Mostly no, the "2 trials with $P < 0.05$" rule has modest discriminating performance
Mining electronic health records?	No, specificity of $P < 0.05$ in searching for genuine effects is very low
Mining big data from metabolomics?	No, except for screening, and only with multiplicity-corrected *P* values thresholds
Assessing whether to exclude women athletes with high testosterone should from the Olympics?	No, magnitude of the competitive advantage Is what matters

Unfortunately, it is easier to document the problem than to offer a simple, effective solution. We next consider some proposed remedies.

4. Some Proposed Remedies

This section summarizes four sets of proposed remedies: alternative approaches to inference (4.1); examples of fit-to-purpose measures (4.2); attempts to model the selection process (4.3); and standards for reproducible research (4.4).

4.1. Alternatives Approaches to Inference and Complements to P Values

We mention here well-known alternatives to null hypothesis testing via *P* values at the 5% level: effect sizes, confidence intervals, methods based on false discovery rates, Bayesian methods, and a change to far more stringent thresholds for *P* values.

Within the frequentist framework, some have proposed more extensive and routine use of effect sizes and confidence intervals as alternatives or complements to null hypothesis testing using *P* values. Surely, such a change could help many papers become more understandable and less often misleading for both experts and users, especially for decision-making in medical applications. Across the biomedical literature the proportion of abstracts (11%) that report at least one effect size is a bit less but roughly the same as the proportion of abstracts (12.5%) that report at least one *P* value (Chavalarias et al. 2016). For large-scale inference, methods based on the false discovery rate may be more appropriate in many if not most papers that currently use *P* values, and, of course, Bayesian methods offer yet another approach. However, it is not clear that greater use of these alternative approaches would substantially diminish bias from selective reporting of the sort described in Section 3, because similar biases can be present regardless of the approach to inference used.

Meanwhile the widespread and expanding use of *P* values suggests the urgency of our need for change. The recent proposal (Benjamin et al. 2017) to lower the traditional threshold for declaring significance from 0.05 to 0.005 should be seen mostly as a temporizing measure, a dam to contain the flood. Many caveats exists for such an approach, most of them raised in the original paper (Benjamin et al. 2017); their discussion is beyond the scope of the current paper. If applied across the biomedical literature of 1990–2015 surveyed in (Chavalarias et al. 2016), the proposed threshold of 0.005 will change the characterization of about one third of the *P* values that are reported in the abstract and considered statistically significant. To the extent that the large majority of these *P* values reflect spuriously significant associations due to selection bias, a change to the more stringent threshold is likely to do more good than harm. The benefit may apply both to the (generally more appropriate) interpretation of past literature and the generation and reporting of new studies (Ioannidis 2018). All the same, changing the threshold cannot directly address the threat of selection bias.

So far, we have considered only broad-brush changes: greater use of effect sizes and confidence intervals, methods based on false discovery rates, Bayesian methods, and more stringent thresholds for declaring a result significant. For a great many applied problems there is a fit-to-purpose measure that is more suitable than the observed significance level for NHST.

4.2. Specific Examples Based on Context and Goals

In most fields and with most types of study designs, NHST should not be the default choice for analysis. Table 1 lists a (nonrandom) sample of some common questions that arise in biomedical research.

None of these applications seems to be a good fit to for NHST:

- A prognostic score should be developed either without using statistical significance for choosing variables to include, or else using a very lenient Type I error rate such as alpha = 0.2 or even higher rather than 0.05.
- For estimating diagnostic performance metrics, *P* values from testing against the null are not meaningful. The magnitude of the improvement in sensitivity and specificity is what matters, not whether the null hypothesis of no improvement can be rejected.
- Randomized trials have used *P* values routinely, but simulations suggest (van Ravenzwaaij and Ioannidis 2017) that they are suboptimal and the rule of "Two trials with $P < 0.05$" for licensing is problematic.
- In the big data environment of electronic health records of omics, $P < 0.05$ makes no sense: It has negligible specificity and can lead to myriad false positive results.
- Finally, a recent consultation concerned the question, "Should women athletes with high testosterone be excluded from the Olympics?" The proponents of this exclusion use results from a paper (Bermon and Garnier 2018) that shows a barely significant difference between women with high and low testosterone (details omitted). However, the magnitude of the difference is tiny. Taking 99% confidence intervals into account, the possibility of a 10% advantage (the disqualifying limit) can be clearly excluded.
- Neither the broad-brush changes of 4.1 nor the more narrowly tailored statistical inference tools address directly the threat of selection bias. We turn next to proposals for modeling the selection process.

4.3. Attempts to Model the Selection Process

The large-scale availability of *P* values that can be readily extracted has led to interesting models that try to differentiate (a) distributions of *P* values in a body of literature commensurate with bias (e.g., p-hacking for passing traditional thresholds of statistical significance (Szucs 2016) from (b) those distributions that are most compatible with genuine discoveries of non-null associations and effects in the absence of such bias. The assumptions of each of these modeling approaches need to be carefully considered. An increasingly popular approach are P-curves, curves that plot the distribution of *P* values in a set of studies. It is speculated that a P-curve analysis (Simonsohn et a. 2014) may differentiate between bias and genuine discoveries. Such P-curves (Simonsohn et al. 2014) may indeed work quite well for randomized experimental studies with no other sources of bias. However, P-curves are sensitive to even tiny bias, corresponding to distortions of the effect sizes by 0.01 standard deviations in a setting of observational studies with confounding. Such a bias, through tiny, can generate a spurious P-curve that resembles genuine discoveries when in fact it is a mere artifact of confounding and omitted variable bias (Bruns and Ioannidis 2016).

Other approaches that can benefit from large-scale availability of *P* value data aim to model the publication selection process over time. When data from large collections of studies or from hundreds or thousands of meta-analyses are available one can assess the average pattern of selection. Consider, for example, the potential strength of publication bias for initial studies, early replication, and later replications (Pfeiffer et al. 2011). The selection forces may depend on the circumstances and the availability of prior evidence on the same question, as, for example, in the "Proteus phenomenon" (Ioannidis and Trikalinos 2005), where once a highly significant result is prominently published, there is a window of opportunity in the next year or two to publish a result that is totally opposite to the original. Furthermore, one can model average biases in sets of multiple studies, but, unfortunately, it is not possible to apply the averages to correct the results of any one particular study.

Meta-analyses can fix only a part of the problem of selective reporting. Sometimes different selection effects will have opposite directions and may cancel out, but more frequently, they may become more prominent. In this sense, meta-analyses may be useful in that a large body of literature can show the bias in sharper relief (Fanelli, Costas, and Ioannidis, 2017).

None of the proposed changes considered so far in this section can offer a head-on challenge to the threat of selection bias. In long term, the only direct protection must come from standards for reproducible research.

4.4. Reproducible Research is Key to Addressing Selection Effects Head-On

Given the shortcomings of proposed remedies and the vulnerability to selection bias present in all approaches to inference, that bias cannot be prevented even by requiring authors to make available both their data and the script or code used for the analysis. Unless this script and code were preregistered (Chambers 2013) it is not possible to tell whether the analysis plan was pre-specified or that it represents the final step of an extreme data exploration that remains unshared. Selection biases may be manageable mostly with improvements in reproducible research practices, such as better transparency, pre-registration, availability of all raw data and software code, greater collaboration and openness among scientists, and the adoption of rewards and incentives that can facilitate such behavior (Munafò et al. 2017).

5. Concluding Thoughts

In conclusion, the use of *P* values has become an epidemic affecting the majority of scientific disciplines. Decisive action is needed both from the statistical and wider scientific community (Wasserstein and Lazar 2016). Strong selection biases can make almost everything (seem) statistically significant and it is very likely that these biases do operate in many, probably most scientific fields that use *P* values, especially with lenient $P < 0.05$ thresholds for claiming success. Implausibly, 96% of the biomedical literature that uses *P* values in the abstract or in the full text claims statistically significant results (Chavalarias et al. 2016). Empirical data combined with plausible argument show that selection effects occur at multiple steps in the process of analyzing data and presenting the results, and that these strongly bias the selection of *P* values in the direction of greater significance. It has even been argued (Fanelli 2010) that fields with the highest proportion of significant claims may be least reliable, and that this ecological relationship can serve as the basis for a hierarchy of scientific fields.

NHST and *P* values are inherently most suitable/optimal for only a minority of current research. Using a more stringent threshold is a temporizing measure to avoid death-by-significance. NHST and *P* values may be replaced in many fields by other inferential methods that will be more fit for reading the results, understanding what they mean, and (when needed) acting on them. However, curtailing selection biases will still require additional drastic measures rather than just a change in inferential method. Changes in the choice of inferential methods do not necessarily address the threat of selection bias head-on. The only direct protection against selection bias is to embrace reproducible research practices, including careful choice and layout of study design and hypotheses with specified and registered in advance methods and analyses, whenever appropriate.

References

Benjamin, D. J., et al. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [23]

Bermon, S., and Garnier, P. Y. (2018), "Serum Androgen Levels and their Relation to Performance in Track and Field: Mass Spectrometry Results from 2127 Observations in Male and Female Elite Athletes," *British Journal of Sports Medicine*, 51, 1309–1314. [23]

Boutron, I., Altman, D. G., Hopewell, S., Vera-Badillo, F., Tannock, I., and Ravaud, P. (2014), "Impact of Spin in the Abstracts of Articles Reporting Results of Randomized Controlled Trials in the Field of Cancer: the SPIIN Randomized Controlled Trial," *Journal of Clinical Oncology*, 32, 4120–4126. [22]

Bruns, S. B., and Ioannidis, J.P (2016), "p-Curve and p-Hacking in Observational Research," *PLoS One*, 11, e0149144. [24]

Chambers, C. D. (2013), "Registered Reports: a New Publishing Initiative at Cortex," *Cortex*, 49, 609–610. [22,24]

Chavalarias, D., Wallach, J. D., Li, A. H., and Ioannidis, J. P. (2016), "Evolution of Reporting P Values in the Biomedical Literature, 1990–2015," *JAMA*, 315, 1141–1148. [20,21,22,23,24]

Cristea, I. A., and Ioannidis, J. P. (2018), "P-values in Display Items are Ubiquitous and Almost Invariably Significant: A Survey of Top Science Journals," *PLoS ONE*, 13, e0197440. [22]

Fanelli, D. (2010), ""Positive" Results Increase Down the Hierarchy of the Sciences," *PLoS One*, 5, e10068. [24]

Fanelli, D., Costas, R., and Ioannidis, J. P. (2017), "Meta-assessment of Bias in Science," *Proceedings of the National Academy of Sciences U S A*, 114, 3714–3719. [24]

Gelman, A. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–65. [22]

Gøtzsche, P. C. (2006), "Believability of Relative Risks and Odds Ratios in Abstracts: Cross Sectional Study," *BMJ*, 333, 231–234. [22]

Ioannidis, J. P. (2018), "The Proposal to Lower P-value Thresholds to .005," *JAMA*, 319, 1429–1430. [23]

Ioannidis, J. P., and Trikalinos, T. A. (2005), "Early Extreme Contradictory Estimates may Appear in Published Research: the Proteus Phenomenon in Molecular Genetics Research and Randomized Trials," *Journal of Clinical Epidemiology*, 58, 543–549. [24]

Khabsa, M., and Giles, C. L. (2014), "The Number of Scholarly Documents on the Public Web," *PLoS One*, 9, e93949. [20]

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017), "A Manifesto for Reproducible Science," *Nature Human Behaviour*, 1, 0021. [24]

Patel, C. J., Burford, B., and Ioannidis, J. P. (2015), "Assessment of Vibration of Effects due to Model Specification can Demonstrate the Instability of Observational Associations," *Journal of Clinical Epidemiology*, 68, 1046–1058. [22]

Patel, C. J., Ji, J., Sundquist, J., Ioannidis, J. P., and Sundquist, K. (2016), "Systematic Assessment of Pharmaceutical Prescriptions in Association with Cancer Risk: a Method to Conduct a Population-wide Medication-wide Longitudinal Study," *Scientific Reports*, 10, 31308. [22]

Pfeiffer, T., Bertram, L., and Ioannidis, J. P. (2011), "Quantifying Selective Reporting and the Proteus Phenomenon for Multiple Datasets with Similar Bias," *PLoS One*, 6, e18362. [24]

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014), "P-curve: a Key to the File-drawer," *Journal of Experimental Psychology General*, 143, 534–547. [24]

Szucs, D., and Ioannidis, J. P. (2017a), "When Null Hypothesis Significance Testing is Unsuitable for Research: a Reassessment," *Frontiers in Human Neuroscience*, 11, 390. [20]

Szucs, D., and Ioannidis, J. P. (2017b), "Empirical assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature," *PLoS Biology*, 15, e2000797. [22]

Szucs, D. (2016), "A Tutorial on Hunting Statistical Significance by chasing N," *Frontiers in Psychology*, 7, 1444. [24]

van Ravenzwaaij, D., and Ioannidis, J. P. (2017), "A Simulation Study of the Strength of Evidence in the Recommendation of Medications Based on Two Trials with Statistically Significant Results," *PLoS One*, 12, e0173184. [23]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [24]

Young, N. S., Ioannidis, J. P., and Al-Ubaydli, O. (2008), "Why Current Publication Practices may Distort Science," *PLoS Medicine*, 5, e201. [22]

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 73 • NUMBER S1 MARCH 2019

The American Statistician

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

The *p*-Value Requires Context, Not a Threshold

Rebecca A. Betensky

To cite this article: Rebecca A. Betensky (2019) The *p*-Value Requires Context, Not a Threshold, The American Statistician, 73:sup1, 115-117, DOI: 10.1080/00031305.2018.1529624

To link to this article: https://doi.org/10.1080/00031305.2018.1529624

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 1191

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utas20

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 115–117: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529624

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The *p*-Value Requires Context, Not a Threshold

Rebecca A. Betensky

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

ABSTRACT

It is widely recognized by statisticians, though not as widely by other researchers, that the *p*-value cannot be interpreted in isolation, but rather must be considered in the context of certain features of the design and substantive application, such as sample size and meaningful effect size. I consider the setting of the normal mean and highlight the information contained in the *p*-value in conjunction with the sample size and meaningful effect size. The *p*-value and sample size jointly yield 95% confidence bounds for the effect of interest, which can be compared to the predetermined meaningful effect size to make inferences about the true effect. I provide simple examples to demonstrate that although the *p*-value is calculated under the null hypothesis, and thus seemingly may be divorced from the features of the study from which it arises, its interpretation as a measure of evidence requires its contextualization within the study. This implies that any proposal for improved use of the *p*-value as a measure of the strength of evidence cannot simply be a change to the threshold for significance.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Effect size; Sample size; Statistical significance.

1. Introduction

Seventy-two prominent researchers proposed changing the default *p*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries (Benjamin et al. 2018). They were motivated to address a leading cause of nonreproducibility of scientific studies; standards of evidence such as the usual rule for formal inference, "reject if $p < 0.05$," are too low. This is not a new concern and was raised several years ago by Ionnidis (2005), who presented data and analysis in support of his title claim that most published research findings are false. The solution proposed by Benjamin et al. (2018) is simple, but surprisingly does not implement some of the basic tenets put forth in the recent statement published by the American Statistical Association (ASA) on the topic of statistical significance and *p*-values (Wasserstein and Lazar 2016). The third principle listed in the ASA statement asserts that scientific conclusions should not be based on whether a *p*-value passes a threshold. The fifth principle acknowledges that the *p*-value, in isolation, does not measure the effect size or the importance of a result. A better solution is not to change the threshold, as suggested by Benjamin et al. (2018), but to find an alternative to exclusive reliance on threshold alone. Context matters.

In this article, I use the terms design and context to refer to characteristics of experiments such as sample size and substantively meaningful effect size, which impact the interpretation of a *p*-value and the conclusions that are drawn. I then rely on examples from the simple setting of a single normal sample with variance one to articulate and illustrate two informal principles for interpreting *p*-values. A first pair of examples (Section 2) shows how data leading to a *p*-value of 0.005 as in Benjamin et al. (2018) can lead to different inferences depending on the combination of sample size and context-specific magnitude of an interesting or important effect size. These examples rely on a functional relationship between the observed *p*-value and sample size, and the lower endpoint of a one-sided confidence interval. A second pair of examples (Section 3) shows how data leading to a large *p*-value can also lead to different inferences, also depending on the combination of sample size and context-specific magnitude of an uninteresting or inconsequential effect size. These examples similarly rely on a relationship between the observed *p*-value and sample size, and the upper endpoint of a one-sided confidence interval. Although I rely on a simple setting for examples, the informal principles for interpreting *p*-values extend in a natural way to more general settings. The article concludes (Section 4) with a summary and discussion.

2. Interpreting a Small *p*-Value

The Benjamin et al. (2018) proposal calls for reducing the *p*-value threshold from 0.05 to 0.005 as a solution to nonreproducibility in science. This section presents two examples, both with $p = 0.005$. For the first example, $p = 0.005$ is too stringent of a threshold for detecting a meaningful signal. For the second example, $p = 0.005$ is not stringent enough. What distinguishes the two examples is the context, namely, the combination of sample size, n, and size, d, of the effect judged to be meaningful. Here, I assume that d has been identified; in practice, the identification of clinically or substantively meaningful effects is complicated and may not be consistent across the various stakeholders, including patients, clinicians, regulators,

CONTACT Rebecca A. Betensky betensky@hsph.harvard.edu Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA 02115

© 2019 The Author. Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

investors, and payers (Keefe et al. 2013; Rosnow and Rosenthal 2003). By definition, d is nonzero.

It is well-known that there is a duality relating hypothesis tests and confidence intervals: we reject the null hypothesis, H_0 at level α if and only if the null value of the parameter lies outside the corresponding $1-\alpha$ level confidence interval. I rely here on a different correspondence, that between the p-value calculated for a one-sided test of $H_0 : \mu = 0$ versus the one-sided alternative hypothesis, $H_1 : \mu > 0$, and the endpoint μ_* of a one-sided confidence interval for μ of the form (μ_*, ∞). The extension to two-sided tests and confidence intervals is straightforward. Browne (2010) also elucidated the relationship between the p-value and the observed effect. He did not, however, relate the interpretation of the 95% confidence interval to a predetermined meaningful effect size as the basis for inference.

For the simple case of a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with known variance 1 and unknown mean μ, the value of the sample mean $\bar{x}$ determines both the p-value, denoted by p, and the lower endpoint of the interval, μ_*. For example, X might be the measured standardized change in systolic blood pressure from baseline to one year after some treatment and μ is the expectation of X. Of interest is the test of the hypothesis that the drug has a positive effect on blood pressure: $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. The p-value is given by $p = P_{\mu=0}(Z > \sqrt{n}\bar{x})$, where Z is standard normal, $\mu_* = \bar{x} - Z_{1-\alpha}/\sqrt{n}$, and $Z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. Inverting the equation that defines the p-value and solving for $\bar{x}$ yields that $\mu_* = (Z_{1-p} - Z_{1-\alpha})/\sqrt{n}$. Note that smaller p are associated with larger μ_*, and thus there is a threshold p^* such that p-values that are below p^* provide evidence that $\mu > d$, that is, of a non-null meaningful effect. In the systolic blood pressure example, d might be $10/\sigma$, where σ is the standard deviation of the change in blood pressure from baseline to one year. (These calculations follow from the fact that the sample mean $\bar{X}$ is normally distributed with mean μ and variance $1/n$ and so $\sqrt{n}(\bar{X} - \mu)$ is standard normal). In the setting of a two-sided test, with positive $\bar{x}$, $\mu_* = (Z_{1-p/2} - Z_{1-\alpha/2})/\sqrt{n}$ and a p-value threshold can likewise be derived.

Given the observed p-value, it is possible to calculate the lower endpoint μ_* of a one-sided $1-\alpha$% confidence interval for μ. In particular, if $p = 0.005$, the corresponding 0.995 quantile of a standard normal is $Z_{0.995} = 2.576$, and the lower endpoint of a 95% one-sided interval is $\mu_* = (2.576 - Z_{0.95})/\sqrt{n}$, where $Z_{0.95} = 1.645$. Thus, $\mu_* = (2.576 - 1.645)/\sqrt{n}$.

Now consider two examples, both with $p = 0.005$. I take as context the combination of sample size, n, and the meaningful effect size, d, defined as the smallest value of $\mu > 0$ judged to be meaningful. Note that the sample size might have been selected to attain a certain level of power to detect a particular value of μ. If it is important to fix the power at a certain level, for practical considerations such as availability of subjects and cost of the trial, this value of μ is frequently larger than d, the smallest *meaningful* value of μ. However, if it is important to design the study to detect the meaningful effect, d, it may be underpowered given the constraints of subject availability and cost.

Example 1(a): Suppose that an effect size of $d = 0.10$ is considered meaningful, and that the sample size is $n = 50$. Given that $p = 0.005$, the lower endpoint of the one-sided 95% confidence interval is equal to $\mu_* = (2.576 - 1.645)/\sqrt{50} = 0.1317$ (Table 1). Thus, with 95% confidence, $p = 0.005$ excludes values of μ that are less than or equal to 0.1317, and thus certainly those that are less than 0.10. In this context, $p = 0.005$ identifies meaningful signals, but potentially misses some signals (i.e., those between 0.10 and 0.1317). The optimal (i.e., maximum) p-value threshold corresponding to $\mu_* = 0.1$ in this context is 0.0093.

Example 1(b): For a contrasting example, I increase the sample size to $n = 200$ and maintain the same effect size of $d = 0.10$. The same p-value yields a lower 95% confidence limit of 0.0658, which includes values of μ less than d. Here, $p = 0.005$ is not a useful threshold relative to the meaningful effect size as it admits values of μ less than 0.10. In this example, the optimal threshold is 0.0011.

Generalizing from these examples suggests a strategy for finding the p-value threshold for concluding a meaningful effect for any given sample size:

1. Based on substantive knowledge about the applied context, select a value d for the smallest effect size considered meaningful. While ideally this is the value that is used to design the study to achieve a fixed power, practical considerations often do not permit this.

2. Take as the upper p-value threshold that value p^*, for which $\mu_* = d$. That is, reject H_0 if and only if $p < p^*$, or equivalently, the 95% confidence interval for μ, (μ_*, ∞) excludes d.

The principle: Reject the null in favor of a meaningful effect if and only if the lower 95% confidence bound exceeds the smallest effect size considered meaningful. Thus, rejecting the null means we can be 95% confident that the true effect size is at least as large as the size considered to be clinically meaningful.

As an example, consider the 1993 GUSTO-I study of streptokinase plus intravenous heparin versus rt-PA (recombinant tissue plasminogen activator) plus intravenous heparinthrombolytic drugs for acute myocardial infarction, as discussed by Lesaffre (2008). The primary endpoint was 30-day mortality. There were approximately 10,300 subjects in each of these treatment arms, and the observed percentages of 30-day mortality were 7.4% and 6.3%. The two-sided p-value testing the equality of the percentages was 0.0028, with a 95% confidence interval for the difference of (0.36%, 1.73%). The conclusion was that there was a significant reduction in 30-day mortality advantage for the rt-PA group versus the streptokinase plus intravenous heparin group. This conclusion implies that a difference as small as 0.36% is considered to be clinically meaningful (i.e., $d < 0.0036$). If this is not the case, and $d > 0.0036$, then even the small p-value of 0.0028 does not provide strong evidence of a meaningful effect.

3. Interpreting a Large *p*-Value

In the previous section, I illustrated that a small p-value relative to a fixed threshold has different meanings depending on the context. I now consider what can be learned from large p-values. Students of introductory statistics courses are taught that no conclusions can be drawn from large p-values. This maxim was reiterated in the ASA statement (Wasserstein and Lazar

2016). In this section, I illustrate that large p-values relative to a fixed threshold also have different meanings depending on the context.

Just as the lower confidence limit for the normal mean has a direct relationship with the p-value for the same one-sided test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$, so does the upper limit of a one-sided confidence interval $(-\infty, \mu^*)$. In particular, simple algebra yields that $\mu^* = (Z_{1-p} + Z_{1-\alpha})/\sqrt{n}$. Note that larger p are associated with smaller μ^*, and thus there is a threshold p_* such that p-values that exceed p_* provide evidence that $\mu < d$, that is, of a nonmeaningful effect. In the setting of a two-sided test, with positive $\bar{x}$, $\mu^* = (Z_{1-p/2} + Z_{1-\alpha/2})/\sqrt{n}$ and a p-value threshold can likewise be derived.

Now consider two examples, both with $p = 0.6286$. Again, I take as context the combination of sample size, n, and the effect size, d, defined as the smallest value of $\mu > 0$ judged to be meaningful.

Example 2(a): Suppose that an effect size of $d = 0.10$ is considered meaningful, and that the sample size is $n = 50$. Given that $p = 0.6286$ and $Z_{1-p} = -0.328$, the calculations above yield that the upper endpoint of the one-sided 95% confidence interval is equal to $\mu_* = (-0.328 + 1.645)/\sqrt{50} = 0.1862$. Thus, with 95% confidence, $p = 0.6286$ excludes values of μ that are greater than or equal to 0.1862, but is uninformative about whether μ is less than $d = 0.10$ or not. In this example, a lower p-value threshold of $p_* = 0.8259$ would provide evidence of a nonmeaningful effect (i.e., $\mu < d$).

Example 2(b): For a contrasting example, I increase the sample size to $n = 200$ and maintain the same effect size of $d = 0.10$. The same p-value yields an upper 95% confidence limit of 0.0931, which excludes values of μ greater than $d = 0.10$. Here, the large p-value of 0.6286 is useful in providing evidence against a meaningful effect. In this example, a lower p-value threshold of $p_* = 0.5913$ would be sufficient to provide evidence of a nonmeaningful effect.

Generalizing from these examples suggests a strategy for finding the p-value threshold for concluding a nonmeaningful effect for any given sample size:

1. Based on substantive knowledge about the applied context, select a value d for the smallest effect size considered meaningful. While ideally this is the value that is used to design the study to achieve a fixed power, practical considerations often do not permit this.
2. Take as a lower p-value threshold that value, p_*, for which $\mu^* = d$. That is, accept H_0, that is, conclude no meaningful effect, if and only if $p > p_*$, or equivalently, the 95% confidence interval for μ, $(-\infty, \mu^*)$ excludes d.

The principle: Accept the null with respect to a prespecified d if and only if the upper 95% confidence bound falls below the smallest effect size considered meaningful. Thus, accepting the null means we can be 95% confident that the true effect size is no larger than the minimal size considered to be clinically meaningful.

As an example in the different context of a two-sided test of a relative risk, the RE-LY trial of atrial fibrillation compared dabigratran to warfarin with respect to risk of stroke or systemic embolism (Connolly et al. 2009). A relative risk of 1.46 was identified as the clinically meaningful threshold for noninferiority of dabigatran relative to warfarin; that is, if the upper two-sided 95% confidence limit (i.e., the upper one-sided 97.5% limit) for the relative risk fell below 1.46, noninferiority could be declared. The upper one-sided 97.5% limit was used to account for the two dabigatran dose groups that were tested versus warfarin and because superiority was tested, as well. The relative risk for the 6015 subjects in the 110 mg dabigatran group versus the 6022 subjects in the warfarin group was 0.91, with a 95% confidence interval of (0.74,1.11) and a p-value of 0.34. Because the upper limit of 1.11 is below 1.46, this dose group of dabigatran could be concluded to be noninferior to warfarin. In this setting, the large p-value of 0.34 (and associated confidence interval) is large enough to declare noninferiority of dabigatran.

4. Summary

In conjunction with the design and context of the study, such as sample size and the minimum meaningful effect size, which are inputs to the calculation of confidence limits for measures of effect, the p-value may indeed be informative about the effect of interest and/or about the null. However, absolute thresholds for the p-value do not render it meaningful with regard to a positive or null effect; the thresholds depend on n and d. This understanding expands on the ASA statement (Wasserstein and Lazar 2016), which enumerates truisms about the p-value, but does not provide guidance regarding best uses of the p-value, and provides nuance to the simple stringent threshold suggested by Benjamin et al. (2018). In summary, I have elucidated the importance of contextualizing the p-value within the salient features of the study when formal hypothesis testing is undertaken. When this is done, it can be a useful measure of evidence for the truth.

Funding

This work was supported in part by NIH Grant UL1TR001102.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., and Cesarini, D. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6. [115,117]

Browne, R. H. (2010), "The *t*-Test *p* Value and Its Relationship to the Effect Size and $P(X > Y)$," *The American Statistician*, 64, 30–33. [116]

Connolly, S. J., Ezekowitz, M. D., Yusuf, S., Eikelboom, J., Oldgren, J., Parekh, A., Pogue, J., Reilly, P. A., Themeles, E., Varrone, J., and Wang, S. (2009) "Dabigatran Versus Warfarin in Patients With Atrial Fibrillation," *New England Journal of Medicine*, 361, 1139–1151. [Erratum, N Engl J Med 2010;363:1877.] [117]

Ionnidis, J. P. A. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124. [115]

Keefe, R. S. E., Kraemer, H. C., Epstein, R. S., Frank, E., Haynes, G., Laughren, T. P., and Leon, A. C. (2013), "Defining a Clinically Meaningful Effect for the Design and Interpretation of Randomized Controlled Trials," *Innovations in Clinical Neuroscience*, 10, 4S–19S. [116]

Lesaffre, E. (2008), "Superiority, Equivalence, and Non-inferiority Trials," *Bulletin of the NYU Hospital for Joint Diseases*, 66, 150–154. [116]

Rosnow, R. L., and Rosenthal, R. (2003), "Effect Sizes for Experimenting Psychologists," *Canadian Journal of Experimental Psychology*, 57, 221–237. [116]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process and Purpose," *The American Statistician*, 70, 129–133. [115,117]

Taylor & Francis
Taylor & Francis Group

THE AMERICAN STATISTICIAN
A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION
VOLUME 73 • NUMBER S1 MARCH 2019

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

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 30728

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utas20

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

Taylor & Francis
Taylor & Francis Group

EDITORIAL

OPEN ACCESS

Check for updates

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

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 do with 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.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of p-values and significance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). 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 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).

Don't. Don't. Just…don't. Yes, we talk a lot about don'ts. The *ASA Statement on p-Values and Statistical Significance* (Wasserstein and Lazar 2016) was developed primarily because after decades, warnings about the don'ts had gone mostly unheeded. The statement was about what not to do, because there is widespread agreement about the don'ts.

Knowing what not to do with p-values is indeed necessary, but it does not suffice. It is as though statisticians were asking users of statistics to tear out the beams and struts holding up the edifice of modern scientific research without offering solid construction materials to replace them. Pointing out old, rotting timbers was a good start, but now we need more.

Recognizing this, in October 2017, the American Statistical Association (ASA) held the Symposium on Statistical Inference, a two-day gathering that laid the foundations for this special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today (Mayo 2018). At other times you'll hear melodies wrapping in a rich counterpoint that may herald an increasingly harmonious new era of statistics. To us, these are all the sounds of statistical inference in the 21st century, the sounds of 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. 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*. As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies. Researchers will be free to communicate all their findings in all their glorious uncertainty, knowing their work is to be judged by the quality and effective communication of their science, and not by their p-values. As "statistical significance" is used less, statistical thinking will be used more.

The *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. That's what this special issue of *The American Statistician* sets out to do.

To get to the do's, though, we must begin with one more don't.

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

2. Don't Say "Statistically Significant"

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 broadly known by Fisher's use of the phrase (1925), Edgeworth's (1885) original intention for statistical significance was simply as a tool to indicate when a result warrants further scrutiny. But that idea has been irretrievably lost. Statistical significance was never meant to imply scientific importance, and the confusion of the two was decried soon after its widespread use (Boring 1919). Yet a full century later the confusion persists.

And so the tool has become the tyrant. The problem is not simply use of the word "significant," although the statistical and ordinary language meanings of the word are indeed now hopelessly confused (Ghose 2013); the term should be avoided for that reason alone. The problem is a larger one, however: using bright-line rules for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making (ASA statement, Principle 3). 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.

For example, no p-value can reveal the plausibility, presence, truth, or importance of an association or effect. Therefore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant. Yet the dichotomization into "significant" and "not significant" is taken as an imprimatur of authority on these characteristics. In a world without bright lines, on the other hand, it becomes untenable to assert dramatic differences in interpretation from inconsequential differences in estimates. As Gelman and Stern (2006) famously observed, the difference between "significant" and "not significant" is not itself statistically significant.

Furthermore, this false split into "worthy" and "unworthy" results leads to the selective reporting and publishing of results based on their statistical significance—the so-called "file drawer problem" (Rosenthal 1979). And the dichotomized reporting problem extends beyond just publication, notes Amrhein, Trafimow, and Greenland (2019): when authors use p-value thresholds to select which findings to discuss in their papers, "their conclusions and what is reported in subsequent news and reviews will be biased…Such selective attention based on study outcomes will therefore not only distort the literature but will slant published descriptions of study results—biasing the summary descriptions reported to practicing professionals and the general public." For the integrity of scientific publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.

To be clear, the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).

Despite the limitations of p-values (as noted in Principles 5 and 6 of the ASA statement), however, we are not recommending that the calculation and use of continuous p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (e.g., $p = 0.08$). They should also be described in language stating what the value means in the scientific context. We believe that a reasonable prerequisite for reporting any p-value is the ability to interpret it appropriately. We say more about this in Section 3.3.

To move forward to a world beyond "$p < 0.05$," we must recognize afresh that statistical inference is not—and never has been—equivalent to scientific inference (Hubbard, Haig, and Parsa 2019; Ziliak 2019). However, looking to statistical significance for a marker of scientific observations' credibility has created a guise of equivalency. Moving beyond "statistical significance" opens researchers to the real significance of statistics, which is "the science of learning from data, and of measuring, controlling, and communicating uncertainty" (Davidian and Louis 2012).

In sum, "statistically significant"—don't say it and don't use it.

3. There Are Many Do's

With the don'ts out of the way, we can finally discuss ideas for specific, positive, constructive actions. We have a massive list of them in the seventh section of this editorial! In that section, the authors of all the articles in this special issue each provide their own short set of do's. Those lists, and the rest of this editorial, will help you navigate the substantial collection of articles that follows.

Because of the size of this collection, we take the liberty here of distilling our readings of the articles into a summary of what can be done to move beyond "$p < 0.05$." You will find the rich details in the articles themselves.

What you will NOT find in this issue is one solution that majestically replaces the outsized role that statistical significance has come to play. The statistical community has not yet converged on a simple paradigm for the use of statistical inference in scientific research—and in fact it may never do so. A one-size-fits-all approach to statistical inference is an inappropriate expectation, even after the dust settles from our current remodeling of statistical practice (Tong 2019). Yet solid principles for the use of statistics do exist, and they are well explained in this special issue.

We summarize our recommendations in two sentences totaling seven words: "**A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest." Remember "ATOM."

3.1. Accept Uncertainty

Uncertainty exists everywhere in research. And, just like with the frigid weather in a Wisconsin winter, there are those who will flee from it, trying to hide in warmer havens elsewhere. Others, however, accept and even delight in the omnipresent cold; these are the ones who buy the right gear and bravely take full advantage of all the wonders of a challenging climate. Significance tests and dichotomized *p*-values have turned many researchers into scientific snowbirds, trying to avoid dealing with uncertainty by escaping to a "happy place" where results are either statistically significant or not. In the real world, data provide a noisy signal. Variation, one of the causes of uncertainty, is everywhere. Exact replication is difficult to achieve. So it is time to get the right (statistical) gear and "move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016).

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).

"Accept uncertainty and embrace variation in effects," advise McShane et al. in Section 7 of this editorial. "[W]e can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being 'an effect' or 'no effect'—based on some *p*-value or other statistical threshold being attained."

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. However, simplistic use of confidence intervals as a measurement of uncertainty leads to the same bad outcomes as use of statistical significance (especially, a focus on whether such intervals include or exclude the "null hypothesis value"). Instead, Greenland (2019) and Amrhein, Trafimow, and Greenland (2019) encourage thinking of confidence intervals as "compatibility intervals," which use *p*-values to show the effect sizes that are most compatible with the data under the given model.

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** (the fourth of our four principles, on which we will expand in Section 3.4) and encourages "meta-analytic thinking" (Cumming 2014). Accepting uncertainty as inevitable is a natural antidote to the seductive certainty falsely promised by statistical significance. With this new outlook, we will naturally seek out replications and the integration of evidence through meta-analyses, which usually requires point and interval estimates from contributing studies. This will in turn give us more precise overall estimates for our effects and associations. And this is what will lead to the best research-based guidance for practical decisions.

Accepting uncertainty leads us to **be thoughtful**, the second of our four principles.

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.

3.2.1. Thoughtfulness in the Big Picture

"[M]ost scientific research is exploratory in nature," Tong (2019) contends. "[T]he design, conduct, and analysis of a study are necessarily flexible, and must be open to the discovery of unexpected patterns that prompt new questions and hypotheses. In this context, statistical modeling can be exceedingly useful for elucidating patterns in the data, and researcher degrees of freedom can be helpful and even essential, though they still carry the risk of overfitting. The price of allowing this flexibility is that the validity of any resulting statistical inferences is undermined."

Calin-Jageman and Cumming (2019) caution that "in practice the dividing line between planned and exploratory research can be difficult to maintain. Indeed, exploratory findings have a slippery way of 'transforming' into planned findings as the research process progresses." At the bottom of that slippery slope one often finds results that don't reproduce.

Anderson (2019) proposes three questions **thoughtful researchers** asked thoughtful researchers evaluating research results: What are the practical implications of the estimate? How precise is the estimate? And is the model correctly specified? The latter question leads naturally to three more: Are the modeling assumptions understood? Are these assumptions valid? And do the key results hold up when other modeling choices are made? Anderson further notes, "Modeling assumptions (including all the choices from model specification to sample selection and the handling of data issues) should be sufficiently documented so independent parties can critique, and replicate, the work."

Drawing on archival research done at the Guinness Archives in Dublin, Ziliak (2019) emerges with ten "*G*-values" he believes we all wish to maximize in research. That is, we want large *G*-values, not small *p*-values. The ten principles of Ziliak's "Guinnessometrics" are derived primarily from his examination of experiments conducted by statistician William Sealy Gosset while working as Head Brewer for Guinness. Gosset took an economic approach to the logic of uncertainty, preferring balanced designs over random ones and estimation of gambles over bright-line "testing." Take, for example, Ziliak's *G*-value 10: "Consider purpose of the inquiry, and compare with best practice," in the spirit of what farmers and brewers must do. The purpose is generally NOT to falsify a null hypothesis, says Ziliak. Ask what is at stake, he advises, and determine what magnitudes of change are humanly or scientifically meaningful in context.

Pogrow (2019) offers an approach based on practical benefit rather than statistical or practical significance. This approach is especially useful, he says, for assessing whether interventions in complex organizations (such as hospitals and schools) are effective, and also for increasing the likelihood that the observed benefits will replicate in subsequent research and in clinical practice. In this approach, "practical benefit" recognizes that reliance on small effect sizes can be as problematic as relying on p-values.

Thoughtful research prioritizes sound data production by putting energy into the careful planning, design, and execution of the study (Tong 2019).

Locascio (2019) urges researchers to be prepared for a new publishing model that evaluates their research based on the importance of the questions being asked and the methods used to answer them, rather than the outcomes obtained.

3.2.2. Thoughtfulness Through Context and Prior Knowledge

Thoughtful research considers the scientific context and prior evidence. In this regard, a declaration of statistical significance is the antithesis of thoughtfulness: it says nothing about practical importance, and it ignores what previous studies have contributed to our knowledge.

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.

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).

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.

3.2.3. Thoughtful Alternatives and Complements to P-Values

Thoughtful research considers multiple approaches for solving problems. This special issue includes some ideas for supplementing or replacing p-values. Here is a short summary of some of them, with a few technical details:

Amrhein, Trafimow, and Greenland (2019) and Greenland (2019) advise that null p-values should be supplemented with a p-value from a test of a pre-specified alternative (such as a minimal important effect size). To reduce confusion with posterior probabilities and better portray evidential value, they further advise that p-values be transformed into s-values (Shannon information, surprisal, or binary logworth) $s = -\log_2(p)$. This measure of evidence affirms other arguments that the evidence against a hypothesis contained in the p-value is not nearly as strong as is believed by many researchers. The change of scale also moves users away from probability misinterpretations of the p-value.

Blume et al. (2019) offer a "second generation p-value (SGPV)," the characteristics of which mimic or improve upon those of p-values but take practical significance into account. The null hypothesis from which an SGPV is computed is a composite hypothesis representing a range of differences that would be practically or scientifically inconsequential, as in equivalence testing (Wellek 2017). This range is determined in advance by the experimenters. When the SGPV is 1, the data only support null hypotheses; when the SGPV is 0, the data are incompatible with any of the null hypotheses. SGPVs between 0 and 1 are inconclusive at varying levels (maximally inconclusive at or near SGPV = 0.5.) Blume et al. illustrate how the SGPV provides a straightforward and useful descriptive summary of the data. They argue that it eliminates the problem of how classical statistical significance does not imply scientific relevance, it lowers false discovery rates, and its conclusions are more likely to reproduce in subsequent studies.

The "analysis of credibility"(AnCred) is promoted by Matthews (2019). This approach takes account of both the width of the confidence interval and the location of its bounds when assessing weight of evidence. AnCred assesses the credibility of inferences based on the confidence interval by determining the level of prior evidence needed for a new finding to provide credible evidence for a nonzero effect. If this required level of prior evidence is supported by current knowledge and insight, Matthews calls the new result "credible evidence for a non-zero effect," irrespective of its statistical significance/nonsignificance.

Colquhoun (2019) proposes continuing the use of continuous p-values, but only in conjunction with the "false positive risk (FPR)." The FPR answers the question, "If you observe a 'significant' p-value after doing a single unbiased experiment, what is the probability that your result is a false positive?" It tells you what most people mistakenly still think the p-value does, Colquhoun says. The problem, however, is that to calculate the FPR you need to specify the prior probability that an effect is real, and it's rare to know this. Colquhoun suggests that the FPR could be calculated with a prior probability of 0.5, the largest value reasonable to assume in the absence of hard prior data. The FPR found this way is in a sense the minimum false positive risk (mFPR); less plausible hypotheses (prior probabilities below 0.5) would give even bigger FPRs, Colquhoun says, but the mFPR would be a big improvement on reporting a p-value alone. He points out that p-values near 0.05 are, under a variety of assumptions, associated with minimum false positive risks of 20–30%, which should stop a researcher from making too big a claim about the "statistical significance" of such a result.

Benjamin and Berger (2019) propose a different supplement to the null p-value. The Bayes factor bound (BFB)—which under typically plausible assumptions is the value $1/(-ep \ln p)$—represents the upper bound of the ratio of data-based odds of the alternative hypothesis to the null hypothesis. Benjamin and Berger advise that the BFB should be reported along with the continuous p-value. This is an incomplete step toward revising practice, they argue, but one that at least confronts the researcher with the maximum possible odds that the alternative hypothesis is true—which is what researchers often think they are getting with a p-value. The BFB, like the FPR, often clarifies that the evidence against the null hypothesis contained in the p-value is not nearly as strong as is believed by many researchers.

Goodman, Spruill, and Komaroff (2019) propose a two-stage approach to inference, requiring both a small p-value below a pre-specified level and a pre-specified sufficiently large effect size before declaring a result "significant." They argue that this method has improved performance relative to use of dichotomized p-values alone.

Gannon, Pereira, and Polpo (2019) have developed a testing procedure combining frequentist and Bayesian tools to provide a significance level that is a function of sample size.

Manski (2019) and Manski and Tetenov (2019) urge a return to the use of statistical decision theory, which they say has largely been forgotten. Statistical decision theory is not based on p-value thresholds and readily distinguishes between statistical and clinical significance.

Billheimer (2019) suggests abandoning inference about parameters, which are frequently hypothetical quantities used to idealize a problem. Instead, he proposes focusing on the prediction of future observables, and their associated uncertainty, as a means to improving science and decision-making.

3.2.4. Thoughtful Communication of Confidence

Be thoughtful and clear about the level of confidence or credibility that is present in statistical results.

Amrhein, Trafimow, and Greenland (2019) and Greenland (2019) argue that the use of words like "significance" in conjunction with p-values and "confidence" with interval estimates misleads users into overconfident claims. They propose that researchers think of p-values as measuring the compatibility between hypotheses and data, and interpret interval estimates as "compatibility intervals."

In what may be a controversial proposal, Goodman (2018) suggests requiring "that any researcher making a claim in a study accompany it with their estimate of the chance that the claim is true." Goodman calls this the confidence index. For example, along with stating "This drug is associated with elevated risk of a heart attack, relative risk (RR) = 2.4, $p = 0.03$," Goodman says investigators might add a statement such as "There is an 80% chance that this drug raises the risk, and a 60% chance that the risk is at least doubled." Goodman acknowledges, "Although simple on paper, requiring a confidence index would entail a profound overhaul of scientific and statistical practice."

In a similar vein, Hubbard and Carriquiry (2019) urge that researchers prominently display the probability the hypothesis is true or a probability distribution of an effect size, or provide sufficient information for future researchers and policy makers to compute it. The authors further describe why such a probability is necessary for decision making, how it could be estimated by using historical rates of reproduction of findings, and how this same process can be part of continuous "quality control" for science.

Being thoughtful in our approach to research will lead us to **be open** in our design, conduct, and presentation of it as well.

3.3. Be Open

We envision **openness** as embracing certain positive practices in the development and presentation of research work.

3.3.1. Openness to Transparency and to the Role of Expert Judgment

First, we repeat oft-repeated advice: **Be open** to "open science" practices. Calin-Jageman and Cumming (2019), Locascio (2019), and others in this special issue urge adherence to practices such as public pre-registration of methods, transparency and completeness in reporting, shared data and code, and even pre-registered ("results-blind") review. Completeness in reporting, for example, requires not only describing all analyses performed but also presenting all findings obtained, without regard to statistical significance or any such criterion.

Openness also includes understanding and accepting the role of expert judgment, which enters the practice of statistical inference and decision-making in numerous ways (O'Hagan 2019). "Indeed, there is essentially no aspect of scientific investigation in which judgment is not required," O'Hagan observes. "Judgment is necessarily subjective, but should be made as carefully, as objectively, and as scientifically as possible."

Subjectivity is involved in any statistical analysis, Bayesian or frequentist. Gelman and Hennig (2017) observe, "Personal decision making cannot be avoided in statistical data analysis and, for want of approaches to justify such decisions, the pursuit of objectivity degenerates easily to a pursuit to merely *appear* objective." One might say that subjectivity is not a problem; it is part of the solution.

Acknowledging this, Brownstein et al. (2019) point out that expert judgment and knowledge are required in all stages of the scientific method. They examine the roles of expert judgment throughout the scientific process, especially regarding the integration of statistical and content expertise. "All researchers, irrespective of their philosophy or practice, use expert judgment in developing models and interpreting results," say Brownstein et al. "We must accept that there is subjectivity in every stage of scientific inquiry, but objectivity is nevertheless the fundamental goal. Therefore, we should base judgments on evidence and careful reasoning, and seek wherever possible to eliminate potential sources of bias."

How does one rigorously elicit expert knowledge and judgment in an effective, unbiased, and transparent way? O'Hagan (2019) addresses this, discussing protocols to elicit expert knowledge in an unbiased and as scientifically sound was as possible. It is also important for such elicited knowledge to be examined critically, comparing it to actual study results being an important diagnostic step.

3.3.2. Openness in Communication

Be open in your reporting. Report p-values as continuous, descriptive statistics, as we explain in Section 2. We realize that this leaves researchers without their familiar bright line anchors. Yet if we were to propose a universal template for presenting and interpreting continuous p-values we would violate our own principles! Rather, we believe that the thoughtful use and interpretation of p-values will never adhere to a rigid rulebook, and will instead inevitably vary from study to study. Despite these caveats, we can offer recommendations for sound practices, as described below.

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.

When using p-values, remember not only Principle 5 of the ASA statement: "A p-value…does not measure the size of an effect or the importance of a result" but also Principle 6: "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis." Despite these limitations, if you present p-values, do so for more than one hypothesized value of your variable of interest (Fraser 2019; Greenland 2019), such as 0 and at least one plausible, relevant alternative, such as the minimum practically important effect size (which should be determined before analyzing the data).

Betensky (2019) also reminds us to interpret the p-value in the context of sample size and meaningful effect size.

Instead of p, you might consider presenting the s-value (Greenland 2019), which is described in Section 3.2. As noted in Section 3.1, you might present a confidence interval. Sound practices in the interpretation of confidence intervals include (1) discussing both the upper and lower limits and whether they have different practical implications, (2) paying no particular attention to whether the interval includes the null value, and (3) remembering that an interval is itself an estimate subject to error and generally provides only a rough indication of uncertainty given that all of the assumptions used to create it are correct and, thus, for example, does not "rule out" values outside the interval. Amrhein, Trafimow, and Greenland (2019) suggest that interval estimates be interpreted as "compatibility" intervals rather than as "confidence" intervals, showing the values that are most compatible with the data, under the model used to compute the interval. They argue that such an interpretation and the practices outlined here can help guard against overconfidence.

It is worth noting that Tong (2019) disagrees with using p-values as descriptive statistics. "Divorced from the probability claims attached to such quantities (confidence levels, nominal Type I errors, and so on), there is no longer any reason to privilege such quantities over descriptive statistics that more directly characterize the data at hand." He further states, "Methods with alleged generality, such as the p-value or Bayes factor, should be avoided in favor of discipline- and problem-specific solutions that can be designed to be fit for purpose."

Failing to **be open** in reporting leads to publication bias. Ioannidis (2019) notes the high level of selection bias prevalent in biomedical journals. He defines "selection" as "the collection of choices that lead from the planning of a study to the reporting of p-values." As an illustration of one form of selection bias, Ioannidis compared "the set of p-values reported in the full text of an article with the set of p-values reported in the abstract." The main finding, he says, "was that p-values chosen for the abstract tended to show greater significance than those reported in the text, and that the gradient was more pronounced in some types of journals and types of designs." Ioannidis notes, however, that selection bias "can be present regardless of the approach to inference used." He argues that in the long run, "the only direct protection must come from standards for reproducible research."

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.**

3.4. Be Modest

Researchers of any ilk may rarely advertise their personal modesty. Yet the most successful ones cultivate a practice of **being modest** throughout their research, by understanding and clearly expressing the limitations of their work.

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. Statistical results must eventually mislead us when they are used and communicated as if they present this complex reality, rather than a model for it. This is not a problem of our statistical methods. It is a problem of interpretation and communication of results."

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). Rougier (2019) calls on researchers to "recognize that behind every choice of null distribution and test statistic, there lurks a plausible family of alternative hypotheses, which can provide more insight into the null distribution."

p-values, confidence intervals, and other statistical measures are all uncertain. Treating them otherwise is immodest overconfidence.

Remember that statistical tools have their limitations. Rose and McGuire (2019) show how use of stepwise regression in health care settings can lead to policies that are unfair.

Remember also that the amount of evidence for or against a hypothesis provided by *p*-values near the ubiquitous $p < 0.05$ threshold (Johnson 2019) is usually much less than you think (Benjamin and Berger 2019; Colquhoun 2019; Greenland 2019).

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."

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. Then, use meta-analysis, evidence reviews, and Bayesian methods to synthesize evidence across studies.

Resist the urge to overreach in the generalizability of claims, Watch out for pressure to embellish the abstract or the press release. If the study's limitations are expressed in the paper but not in the abstract, they may never be read.

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.

Hubbard and Carriquiry (see their "do list" in Section 7) suggest encouraging reproduction of research by giving "a byline status for researchers who reproduce studies." They would like to see digital versions of papers dynamically updated to display "Reproduced by…." below original research authors' names or "not yet reproduced" until it is reproduced.

Indeed, when it comes to reproducibility, Amrhein, Trafimow, and Greenland (2019) demand that we **be modest** in our expectations. "An important role for statistics in research is the summary and accumulation of information," they say. "If replications do not find the same results, this is not necessarily a crisis, but is part of a natural process by which science evolves. The goal of scientific methodology should be to direct this evolution toward ever more accurate descriptions of the world and how it works, not toward ever more publication of inferences, conclusions, or decisions."

Referring to replication studies in psychology, McShane et al. (2019) recommend that future large-scale replication projects "should follow the 'one phenomenon, many studies' approach of the Many Labs project and Registered Replication Reports rather than the 'many phenomena, one study' approach of the Open Science Collaboration project. In doing so, they should systematically vary method factors across the laboratories involved in the project." This approach helps achieve the goals of Amrhein, Trafimow, and Greenland (2019) by increasing understanding of why and when results replicate or fail to do so, yielding more accurate descriptions of the world and how it works. It also speaks to significant sameness versus significant difference a la Hubbard, Haig, and Parsa (2019).

Kennedy-Shaffer's (2019) historical perspective on statistical significance reminds us to **be modest**, by prompting us to recall how the current state of affairs in *p*-values has come to be.

Finally, **be modest** by recognizing that different readers may have very different stakes on the results of your analysis, which means you should try to take the role of a neutral judge rather than an advocate for any hypothesis. This can be done, for example, by pairing every null *p*-value with a *p*-value testing an equally reasonable alternative, and by discussing the endpoints of every interval estimate (not only whether it contains the null).

Accept that both scientific inference and statistical inference are hard, and understand that no knowledge will be efficiently advanced using simplistic, mechanical rules and procedures. Accept also that pure objectivity is an unattainable goal—no matter how laudable—and that both subjectivity and expert judgment are intrinsic to the conduct of science and statistics. Accept that there will always be uncertainty, and be **t**houghtful, **o**pen, and **m**odest. ATOM.

And to push this acronym further, we argue in the next section that **i**nstitutional **c**hange is needed, so we put forward that change is needed at the ATOMIC level. Let's go.

4. Editorial, Educational and Other Institutional Practices Will Have to Change

Institutional reform is necessary for moving beyond statistical significance in any context—whether journals, education, academic incentive systems, or others. Several papers in this special issue focus on reform.

Goodman (2019) notes considerable social change is needed in academic institutions, in journals, and among funding and regulatory agencies. He suggests (see Section 7) partnering "with science reform movements and reformers within disciplines, journals, funding agencies and regulators to promote and reward 'reproducible' science and diminish the impact of statistical significance on publication, funding and promotion." Similarly, Colquhoun (2019) says, "In the end, the only way to solve the problem of reproducibility is to do more replication and to reduce the incentives that are imposed on scientists to produce unreliable work. The publish-or-perish culture has damaged science, as has the judgment of their work by silly metrics."

Trafimow (2019), who added energy to the discussion of *p*-values a few years ago by banning them from the journal he edits (Fricker et al. 2019), suggests five "nonobvious changes" to editorial practice. These suggestions, which demand reevaluating traditional practices in editorial policy, will not be trivial to implement but would result in massive change in some journals.

Locascio (2017, 2019) suggests that evaluation of manuscripts for publication should be "results-blind." That is, manuscripts should be assessed for suitability for publication based on the substantive importance of the research without regard to their reported results. Kmetz (2019) supports this approach as well and says that it would be a huge benefit for reviewers, "freeing [them] from their often thankless present jobs and instead allowing them to review research designs for their potential to provide useful knowledge." (See also "registered reports" from the Center for Open Science (*https://cos.io/rr/?_ga=2.184185454.979594832.1547755516-1193527346.1457026171*) and "registered replication reports" from the Association for Psychological Science (*https://www.psychologicalscience.org/publications/replication*) in relation to this concept.)

Amrhein, Trafimow, and Greenland (2019) ask if results-blind publishing means that anything goes, and then answer affirmatively: "Everything should be published in some form if whatever we measured made sense *before we obtained the data* because it was connected in a potentially useful way to some research question." Journal editors, they say, "should be proud about [their] exhaustive methods sections" and base their decisions about the suitability of a study for publication "on the quality of its materials and methods rather than on results and conclusions; the quality of the presentation of the latter is only judged after it is determined that the study is valuable based on its materials and methods."

A "variation on this theme is *pre-registered replication*, where a *replication* study, rather than the original study, is subject to strict pre-registration (e.g., Gelman 2015)," says Tong (2019). "A broader vision of this idea (Mogil and Macleod 2017) is to carry out a whole series of exploratory experiments *without* any formal statistical inference, and summarize the results by descriptive statistics (including graphics) or even just disclosure of the raw data. When results from this series of experiments converge to a single working hypothesis, it can *then* be subjected to a pre-registered, randomized and blinded, appropriately powered confirmatory experiment, carried out by another laboratory, in which valid statistical inference may be made."

Hurlbert, Levine, and Utts (2019) urge abandoning the use of "statistically significant" in all its forms and encourage journals to provide instructions to authors along these lines: "There is now wide agreement among many statisticians who have studied the issue that for reporting of statistical tests yielding *p*-values it is illogical and inappropriate to dichotomize the *p*-scale and describe results as 'significant' and 'nonsignificant.' Authors are strongly discouraged from continuing this never justified practice that originated from confusions in the early history of modern statistics."

Hurlbert, Levine, and Utts (2019) also urge that the *ASA Statement on P-Values and Statistical Significance* "be sent to the editor-in-chief of every journal in the natural, behavioral and social sciences for forwarding to their respective editorial boards and stables of manuscript reviewers. That would be a good way to quickly improve statistical understanding and practice." Kmetz (2019) suggests referring to the ASA statement whenever submitting a paper or revision to any editor, peer reviewer, or prospective reader. Hurlbert et al. encourage a "community grassroots effort" to encourage change in journal procedures.

Campbell and Gustafson (2019) propose a statistical model for evaluating publication policies in terms of weighing novelty of studies (and the likelihood of those studies subsequently being found false) against pre-specified study power. They observe that "no publication policy will be perfect. Science is inherently challenging and we must always be willing to accept that a certain proportion of research is potentially false."

Statistics education will require major changes at all levels to move to a post "$p < 0.05$" world. Two papers in this special issue make a specific start in that direction (Maurer et al. 2019; Steel, Liermann, and Guttorp 2019), but we hope that volumes will be written on this topic in other venues. We are excited that, with support from the ASA, the US Conference on Teaching Statistics (USCOTS) will focus its 2019 meeting on teaching inference.

The change that needs to happen demands change to editorial practice, to the teaching of statistics at every level where inference is taught, and to much more. However…

5. It Is Going to Take Work, and It Is Going to Take Time

If it were easy, it would have already been done, because as we have noted, this is nowhere near the first time the alarm has been sounded.

Why is eliminating the use of *p*-values as a truth arbiter so hard? "The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them," says Goodman (2019). "It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and *p*-values for knowledge claims, publication, funding, and promotion. It doesn't matter if the *p*-value doesn't mean what people think it means; it becomes valuable because of what it buys."

Goodman observes that statisticians alone cannot address the problem, and that "any approach involving only statisticians will not succeed." He calls on statisticians to ally themselves "both with scientists in other fields and with broader based, multidisciplinary scientific reform movements. What statisticians can do within our own discipline is important, but to effectively disseminate or implement virtually any method or policy, we need partners."

"The loci of influence," Goodman says, "include journals, scientific lay and professional media (including social media), research funders, healthcare payors, technology assessors, regulators, academic institutions, the private sector, and professional societies. They also can include policy or informational entities like the National Academies…as well as various other science advisory bodies across the government. Increasingly, they are also including non-traditional science reform organizations comprised both of scientists and of the science literate lay public…and a broad base of health or science advocacy groups…"

It is no wonder, then, that the problem has persisted for so long. And persist it has! Hubbard (2019) looked at citation-count data on twenty-five articles and books severely critical of the effect of null hypothesis significance testing (NHST) on good science. Though issues were well known, Hubbard says, this did nothing to stem NHST usage over time.

Greenland (personal communication, January 25, 2019) notes that cognitive biases and perverse incentives to offer firm conclusions where none are warranted can warp the use of any method. "The core human and systemic problems are not addressed by shifting blame to *p*-values and pushing alternatives as magic cures—especially alternatives that have been subject to little or no comparative evaluation in either classrooms or practice," Greenland said. "What we need now is to move beyond debating only our methods and their interpretations, to concrete proposals for elimination of systemic problems such as pressure to produce noteworthy findings rather than to produce reliable studies and analyses. Review and provisional acceptance of reports before their results are given to the journal (Locascio 2019) is one way to address that pressure, but more ideas are needed since review of promotions and funding applications cannot be so blinded. The challenges of how to deal with human biases and incentives may be the most difficult we must face." Supporting this view is McShane and Gal's (2016, 2017) empirical demonstration of cognitive dichotomization errors among biomedical and social science researchers—and even among statisticians.

Challenges for editors and reviewers are many. Here's an example: Fricker et al. (2019) observed that when *p*-values were suspended from the journal *Basic and Applied Social Psychology* authors tended to overstate conclusions.

With all the challenges, how do we get from here to there, from a "$p < 0.05$" world to a post "$p < 0.05$" world?

Matthews (2019) notes that "Any proposal encouraging changes in inferential practice must accept the ubiquity of NHST.…Pragmatism suggests, therefore, that the best hope of achieving a change in practice lies in offering inferential tools that can be used alongside the concepts of NHST, adding value to them while mitigating their most egregious features."

Benjamin and Berger (2019) propose three practices to help researchers during the transition away from use of statistical significance. "…[O]ur goal is to suggest minimal changes that would require little effort for the scientific community to implement," they say. "Motivating this goal are our hope that easy (but impactful) changes might be adopted and our worry that more complicated changes could be resisted simply because they are perceived to be too difficult for routine implementation."

Yet there is also concern that progress will stop after a small step or two. Even some proponents of small steps are clear that those small steps still carry us far short of the destination.

For example, Matthews (2019) says that his proposed methodology "is not a panacea for the inferential ills of the research community." But that doesn't make it useless. It may "encourage researchers to move beyond NHST and explore the statistical armamentarium now available to answer the central question of research: what does our study tell us?" he says. It "provides a bridge between the dominant but flawed NHST paradigm and the less familiar but more informative methods of Bayesian estimation."

Likewise, Benjamin and Berger (2019) observe, "In research communities that are deeply attached to reliance on '$p < 0.05$,' our recommendations will serve as initial steps away from this attachment. We emphasize that our recommendations are intended merely as initial, temporary steps and that many further steps will need to be taken to reach the ultimate destination: a holistic interpretation of statistical evidence that fully conforms to the principles laid out in the ASA Statement…"

Yet, like the authors of this editorial, not all authors in this special issue support gradual approaches with transitional methods.

Some (e.g., Amrhein, Trafimow, and Greenland 2019; Hurlbert, Levine, and Utts 2019; McShane et al. 2019) prefer to rip off the bandage and abandon use of statistical significance altogether. In short, no more dichotomizing *p*-values into categories of "significance." Notably, these authors do not suggest banning the use of *p*-values, but rather suggest using them descriptively, treating them as continuous, and assessing their weight or import with nuanced thinking, clear language, and full understanding of their properties.

So even when there is agreement on the destination, there is disagreement about what road to take. The questions around reform need consideration and debate. It might turn out that different fields take different roads.

The catalyst for change may well come from those people who fund, use, or depend on scientific research, say Calin-Jageman and Cumming (2019). They believe this change has not yet happened to the desired level because of "the cognitive opacity of the NHST approach: the counter-intuitive *p*-value (it's good when it is small), the mysterious null hypothesis (you want it to be false), and the eminently confusable Type I and Type II errors."

Reviewers of this editorial asked, as some readers of it will, is a *p*-value threshold ever okay to use? We asked some of the authors of articles in the special issue that question as well. Authors identified four general instances. Some allowed that, while *p*-value thresholds should not be used for inference, they might still be useful for applications such as industrial quality control, in which a highly automated decision rule is needed and the costs of erroneous decisions can be carefully weighed when specifying the threshold. Other authors suggested that such dichotomized use of *p*-values was acceptable in model-fitting and variable selection strategies, again as automated tools, this time for sorting through large numbers of potential models or variables. Still others pointed out that *p*-values with very low thresholds are used in fields such as physics, genomics, and imaging as a filter for massive numbers of tests. The fourth instance can be described as "confirmatory setting[s] where the study design and statistical analysis plan are specified prior to data collection, and then adhered to during and after it" (Tong 2019). Tong argues these are the only proper settings for formal statistical inference. And Wellek (2017) says at present it is essential in these settings. "[B]inary decision making is indispensable in medicine and related fields," he says. "[A] radical rejection of the classical principles of statistical inference…is of virtually no help as long as no conclusively substantiated alternative can be offered."

Eliminating the declaration of "statistical significance" based on $p < 0.05$ or other arbitrary thresholds will be easier in some venues than others. Most journals, if they are willing, could fairly rapidly implement editorial policies to effect these changes. Suggestions for how to do that are in this special issue of *The American Statistician*. However, regulatory agencies might require longer timelines for making changes. The U.S. Food and

Drug Administration (FDA), for example, has long established drug review procedures that involve comparing *p*-values to significance thresholds for Phase III drug trials. Many factors demand consideration, not the least of which is how to avoid turning every drug decision into a court battle. Goodman (2019) cautions that, even as we seek change, "we must respect the reason why the statistical procedures are there in the first place." Perhaps the ASA could convene a panel of experts, internal and external to FDA, to provide a workable new paradigm. (See Ruberg et al. 2019, who argue for a Bayesian approach that employs data from other trials as a "prior" for Phase 3 trials.)

Change is needed. Change has been needed for decades. Change has been called for by others for quite a while. So…

6. Why Will Change Finally Happen Now?

In 1991, a confluence of weather events created a monster storm that came to be known as "the perfect storm," entering popular culture through a book (Junger 1997) and a 2000 movie starring George Clooney. Concerns about reproducible science, falling public confidence in science, and the initial impact of the ASA statement in heightening awareness of long-known problems created a perfect storm, in this case, a good storm of motivation to make lasting change. Indeed, such change was the intent of the ASA statement, and we expect this special issue of TAS will inject enough additional energy to the storm to make its impact widely felt.

We are not alone in this view. "60+ years of incisive criticism has not yet dethroned NHST as the dominant approach to inference in many fields of science," note Calin-Jageman and Cumming (2019). "Momentum, though, seems to finally be on the side of reform."

Goodman (2019) agrees: "The initial slow speed of progress should not be discouraging; that is how all broad-based social movements move forward and we should be playing the long game. But the ball is rolling downhill, the current generation is inspired and impatient to carry this forward."

So, let's do it. Let's move beyond "statistically significant," even if upheaval and disruption are inevitable for the time being. It's worth it. In a world beyond "$p < 0.05$," by breaking free from the bonds of statistical significance, statistics in science and policy will become more significant than ever.

7. Authors' Suggestions

The editors of this special TAS issue on statistical inference asked all the contact authors to help us summarize the guidance they provided in their papers by providing us a short list of do's. We asked them to be specific but concise and to be active—start each with a verb. Here is the complete list of the authors' responses, ordered as the papers appear in this special issue.

7.1. Getting to a Post "p < 0.05" Era

Ioannidis, J., What Have We (Not) Learnt From Millions of Scientific Papers With p-Values?

1. Do not use *p*-values, unless you have clearly thought about the need to use them and they still seem the best choice.
2. Do not favor "statistically significant" results.
3. Do be highly skeptical about "statistically significant" results at the 0.05 level.

Goodman, S., Why Is Getting Rid of p-Values So Hard? Musings on Science and Statistics

1. Partner with science reform movements and reformers within disciplines, journals, funding agencies and regulators to promote and reward reproducible science and diminish the impact of statistical significance on publication, funding and promotion.
2. Speak to and write for the multifarious array of scientific disciplines, showing how statistical uncertainty and reasoning can be conveyed in non-"bright-line" ways both with conventional and alternative approaches. This should be done not just in didactic articles, but also in original or reanalyzed research, to demonstrate that it is publishable.
3. Promote, teach and conduct meta-research within many individual scientific disciplines to demonstrate the adverse effects in each of over-reliance on and misinterpretation of *p*-values and significance verdicts in individual studies and the benefits of emphasizing estimation and cumulative evidence.
4. Require reporting a quantitative measure of certainty—a "confidence index"—that an observed relationship, or claim, is true. Change analysis goal from achieving significance to appropriately estimating this confidence.
5. Develop and share teaching materials, software, and published case examples to help with all of the do's above, and to spread progress in one discipline to others.

Hubbard, R., Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary

This list applies to the ASA and to the professional statistics community more generally.

1. Specify, where/if possible, those situations in which the *p*-value plays a clearly valuable role in data analysis and interpretation.
2. Contemplate issuing a statement abandoning the use of *p*-values in null hypothesis significance testing.

Kmetz, J., Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p-Values

1. Refer to the ASA statement on *p*-values whenever submitting a paper or revision to any editor, peer reviewer, or prospective reader. Many in the field do not know of this statement, and having the support of a prestigious organization when authoring any research document will help stop corrupt research from becoming even more dominant than it is.
2. Train graduate students and future researchers by having them reanalyze published studies and post their findings to appropriate websites or weblogs. This practice will benefit not only the students, but will benefit the professions, by increasing the amount of replicated (or nonreplicated) research available and readily accessible, and as well as reformer organizations that support replication.
3. Join one or more of the reformer organizations formed or forming in many research fields, and support and publicize their efforts to improve the quality of research practices.

4. Challenge editors and reviewers when they assert that incorrect practices and interpretations of research, consistent with existing null hypothesis significance testing and beliefs regarding p-values, should be followed in papers submitted to their journals. Point out that new submissions have been prepared to be consistent with the ASA statement on p-values.
5. Promote emphasis on research quality rather than research quantity in universities and other institutions where professional advancement depends heavily on research "productivity," by following the practices recommended in this special journal edition. This recommendation will fall most heavily on those who have already achieved success in their fields, perhaps by following an approach quite different from that which led to their success; whatever the merits of that approach may have been, one objectionable outcome of it has been the production of voluminous corrupt research and creation of an environment that promotes and protects it. We must do better.

Hubbard, D., and Carriquiry, A., Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness and Relevance

1. Compute and prominently display the probability the hypothesis is true (or a probability distribution of an effect size) or provide sufficient information for future researchers and policy makers to compute it.
2. Promote publicly displayed quality control metrics within your field—in particular, support tracking of reproduction studies and computing the "level 1" and even "level 2" priors as required for #1 above.
3. Promote a byline status for researchers who reproduce studies: Digital versions are dynamically updated to display "Reproduced by…." below original research authors' names or "Not yet reproduced" until it is reproduced.

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J., The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making

1. Staff the study team with members who have the necessary knowledge, skills and experience—statistically, scientifically, and otherwise.
2. Include key members of the research team, including statisticians, in all scientific and administrative meetings.
3. Understand that subjective judgments are needed in all stages of a study.
4. Make all judgments as carefully and rigorously as possible and document each decision and rationale for transparency and reproducibility.
5. Use protocol-guided elicitation of judgments.
6. Statisticians specifically should:

   - Refine oral and written communication skills.
   - Understand their multiple roles and obligations as collaborators.
   - Take an active leadership role as a member of the scientific team; contribute throughout all phases of the study.
   - Co-own the subject matter—understand a sufficient amount about the relevant science/policy to meld statistical and subject-area expertise.
   - Promote the expectation that your collaborators co-own statistical issues.
   - Write a statistical analysis plan for all analyses and track any changes to that plan over time.
   - Promote co-responsibility for data quality, security, and documentation.
   - Reduce unplanned and uncontrolled modeling/testing (HARK-ing, p-hacking); document all analyses.

O'Hagan, A., Expert Knowledge Elicitation: Subjective but Scientific

1. Elicit expert knowledge when data relating to a parameter of interest is weak, ambiguous or indirect.
2. Use a well-designed protocol, such as SHELF, to ensure expert knowledge is elicited in as scientific and unbiased a way as possible.

Kennedy-Shaffer, L., Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing

1. Ensure that inference methods match intuitive understandings of statistical reasoning.
2. Reduce the computational burden for nonstatisticians using statistical methods.
3. Consider changing conditions of statistical and scientific inference in developing statistical methods.
4. Address uncertainty quantitatively and in ways that reward increased precision.

Hubbard, R., Haig, B. D., and Parsa, R. A., The Limited Role of Formal Statistical Inference in Scientific Inference

1. Teach readers that although deemed equivalent in the social, management, and biomedical sciences, formal methods of statistical inference and scientific inference are very different animals.
2. Show these readers that formal methods of statistical inference play only a restricted role in scientific inference.
3. Instruct researchers to pursue significant *sameness* (i.e., replicable and empirically generalizable results) rather than significant *differences* in results.
4. Demonstrate how the pursuit of significant differences actively impedes cumulative knowledge development.

McShane, B., Tackett, J., Böckenholt, U., and Gelman, A., Large Scale Replication Projects in Contemporary Psychological Research

1. When planning a replication study of a given psychological phenomenon, bear in mind that replication is complicated in psychological research because studies can never be direct or exact replications of one another, and thus heterogeneity—effect sizes that vary from one study of the phenomenon to the next—cannot be avoided.
2. Future large scale replication projects should follow the "one phenomenon, many studies" approach of the Many Labs project and Registered Replication Reports rather than the

"many phenomena, one study" approach of the Open Science Collaboration project. In doing so, they should systematically vary method factors across the laboratories involved in the project.

3. Researchers analyzing the data resulting from large scale replication projects should do so via a hierarchical (or multilevel) model fit to the totality of the individual-level observations. In doing so, all theoretical moderators should be modeled via covariates while all other potential moderators—that is, method factors—should induce variation (i.e., heterogeneity).
4. Assessments of replicability should not depend solely on estimates of effects, or worse, significance tests based on them. Heterogeneity must also be an important consideration in assessing replicability.

7.2. Interpreting and Using p

Greenland, S., Valid p-Values Behave Exactly as They Should: Some Misleading Criticisms of p-Values and Their Resolution With s-Values

1. Replace any statements about statistical significance of a result with the p-value from the test, and present the p-value as an equality, not an inequality. For example, if $p = 0.03$ then "…was statistically significant" would be replaced by "…had $p = 0.03$," and "$p < 0.05$" would be replaced by "$p = 0.03$." (An exception: If p is so small that the accuracy becomes very poor then an inequality reflecting that limit is appropriate; e.g., depending on the sample size, p-values from normal or χ^2 approximations to discrete data often lack even 1-digit accuracy when $p < 0.0001$.) In parallel, if $p = 0.25$ then "…was not statistically significant" would be replaced by "…had $p = 0.25$," and "$p > 0.05$" would be replaced by "$p = 0.25$."
2. Present p-values for more than one possibility when testing a targeted parameter. For example, if you discuss the p-value from a test of a null hypothesis, also discuss alongside this null p-value another p-value for a plausible alternative parameter possibility (ideally the one used to calculate power in the study proposal). As another example: if you do an equivalence test, present the p-values for both the lower and upper bounds of the equivalence interval (which are used for equivalence tests based on two one-sided tests).
3. Show confidence intervals for targeted study parameters, but also supplement them with p-values for testing relevant hypotheses (e.g., the p-values for both the null and the alternative hypotheses used for the study design or proposal, as in #2). Confidence intervals only show clearly what is in or out of the interval (i.e., a 95% interval only shows clearly what has $p > 0.05$ or $p \leq 0.05$), but more detail is often desirable for key hypotheses under contention.
4. Compare groups and studies directly by showing p-values and interval estimates for their differences, not by comparing p-values or interval estimates from the two groups or studies. For example, seeing $p = 0.03$ in males and $p = 0.12$ in females does ***not*** mean that different associations were seen in males and females; instead, one needs a p-value and confidence interval for the difference in the sex-specific associations to examine the between-sex difference. Similarly, if an early study reported a confidence interval which excluded the null and then a subsequent study reported a confidence interval which included the null, that does not mean the studies gave conflicting results or that the second study failed to replicate the first study; instead, one needs a p-value and confidence interval for the difference in the study-specific associations to examine the between-study difference. In all cases, differences-between-differences must be analyzed directly by statistics for that purpose.
5. Supplement a focal p-value p with its Shannon information transform (s-value or surprisal) $s = -\log_2(p)$. This measures the amount of information supplied by the test against the tested hypothesis (or model): Rounded off, the s-value s shows the number of heads in a row one would need to see when tossing a coin to get the same amount of information against the tosses being "fair" (independent with "heads" probability of $1/2$) instead of being loaded for heads. For example, if $p = 0.03$, this represents $-\log_2(0.03) = 5$ bits of information against the hypothesis (like getting 5 heads in a trial of "fairness" with 5 coin tosses); and if $p = 0.25$, this represents only $-\log_2(0.25) = 2$ bits of information against the hypothesis (like getting 2 heads in a trial of "fairness" with only 2 coin tosses).

Betensky, R., The p-Value Requires Context, Not a Threshold

1. Interpret the p-value in light of its context of sample size and meaningful effect size.
2. Incorporate the sample size and meaningful effect size into a decision to reject the null hypothesis.

Anderson, A., Assessing Statistical Results: Magnitude, Precision and Model Uncertainty

1. Evaluate the importance of statistical results based on their practical implications.
2. Evaluate the strength of empirical evidence based on the precision of the estimates and the plausibility of the modeling choices.
3. Seek out subject matter expertise when evaluating the importance and the strength of empirical evidence.

Heck, P., and Krueger, J., Putting the p-Value in Its Place

1. Use the p-value as a heuristic, that is, as the base for a tentative inference regarding the presence or absence of evidence against the tested hypothesis.
2. Supplement the p-value with other, conceptually distinct methods and practices, such as effect size estimates, likelihood ratios, or graphical representations.
3. Strive to embed statistical hypothesis testing within strong *a priori* theory and a context of relevant prior empirical evidence.

Johnson, V., Evidence From Marginally Significant t-Statistics

1. Be transparent in the number of outcome variables that were analyzed.
2. Report the number (and values) of all test statistics that were calculated.
3. Provide access to protocols for studies involving human or animal subjects.

4. Clearly describe data values that were excluded from analysis and the justification for doing so.
5. Provide sufficient details on experimental design so that other researchers can replicate the experiment.
6. Describe only p-values less than 0.005 as being "statistically significant."

Fraser, D., The p-Value Function and Statistical Inference

1. Determine a primary variable for assessing the hypothesis at issue.
2. Calculate its well defined distribution function, respecting continuity.
3. Substitute the observed data value to obtain the "p-value function."
4. Extract the available well defined confidence bounds, confidence intervals, and median estimate.
5. Know that you don't have an intellectual basis for decisions.

Rougier, J., p-Values, Bayes Factors, and Sufficiency

1. Recognize that behind every choice of null distribution and test statistic, there lurks a plausible family of alternative hypotheses, which can provide more insight into the null distribution.

Rose, S., and McGuire, T., Limitations of p-Values and R-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

1. Formulate a clear objective for variable inclusion in regression procedures.
2. Assess all relevant evaluation metrics.
3. Incorporate algorithmic fairness considerations.

7.3. Supplementing or Replacing p

Blume, J., Greevy, R., Welty, V., Smith, J., and DuPont, W., An Introduction to Second Generation p-Values

1. Construct a composite null hypothesis by specifying the range of effects that are not scientifically meaningful (do this before looking at the data). Why: Eliminating the conflict between scientific significance and statistical significance has numerous statistical and scientific benefits.
2. Replace classical p-values with second-generation p-values (SGPV). Why: SGPVs accommodate composite null hypotheses and encourage the proper communication of findings.
3. Interpret the SGPV as a high-level summary of what the data say. Why: Science needs a simple indicator of when the data support only meaningful effects (SGPV = 0), when the data support only trivially null effects (SGPV = 1), or when the data are inconclusive ($0 < \text{SGPV} < 1$).
4. Report an interval estimate of effect size (confidence interval, support interval, or credible interval) and note its proximity to the composite null hypothesis. Why: This is a more detailed description of study findings.
5. Consider reporting false discovery rates with SGPVs of 0 or 1. Why: FDRs gauge the chance that an inference is incorrect under assumptions about the data generating process and prior knowledge.

Goodman, W., Spruill, S., and Komaroff, E., A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting Its Use

1. Determine how far the true parameter's value would have to be, in your research context, from exactly equaling the conventional, point null hypothesis to consider that the distance is meaningfully large or practically significant.
2. Combine the conventional p-value criterion with a minimum effect size criterion to generate a two-criteria inference-indicator signal, which provides heuristic, but nondefinitive evidence, for inferring the parameter's true location.
3. Document the intended criteria for your inference procedures, such as a p-value cut-point and a minimum practically significant effect size, prior to undertaking the procedure.
4. Ensure that you use the appropriate inference method for the data that are obtainable and for the inference that is intended.
5. Acknowledge that every study is fraught with limitations from unknowns regarding true data distributions and other conditions that one's method assumes.

Benjamin, D., and Berger, J., Three Recommendations for Improving the Use of p-Values

1. Replace the 0.05 "statistical significance" threshold for claims of novel discoveries with a 0.005 threshold and refer to p-values between 0.05 and 0.005 as "suggestive."
2. Report the data-based odds of the alternative hypothesis to the null hypothesis. If the data-based odds cannot be calculated, then use the p-value to report an upper bound on the data-based odds: $1/(-ep \ln p)$.
3. Report your prior odds and posterior odds (prior odds * data-based odds) of the alternative hypothesis to the null hypothesis. If the data-based odds cannot be calculated, then use your prior odds and the p-value to report an upper bound on your posterior odds: (prior odds) * $(1/(-ep \ln p))$.

Colquhoun, D., The False Positive Risk: A Proposal Concerning What to Do About p-Values

1. Continue to provide p-values and confidence intervals. Although widely misinterpreted, people know how to calculate them and they aren't entirely useless. Just don't ever use the terms "statistically significant" or "nonsignificant."
2. Provide in addition an indication of false positive risk (FPR). This is the probability that the claim of a real effect on the basis of the p-value is in fact false. The FPR (not the p-value) is the probability that your result occurred by chance. For example, the fact that, under plausible assumptions, observation of a p-value close to 0.05 corresponds to an FPR of at least 0.2–0.3 shows clearly the weakness of the conventional criterion for "statistical significance."
3. Alternatively, specify the prior probability of there being a real effect that one would need to be able to justify in order to achieve an FPR of, say, 0.05.

Notes:

There are many ways to calculate the FPR. One, based on a point null and simple alternative can be calculated with the web calculator at *http://fpr-calc.ucl.ac.uk/*. However other approaches to the calculation of FPR, based on different assumptions, give results that are similar (Table 1 in Colquhoun 2019).

To calculate FPR it is necessary to specify a prior probability and this is rarely known. My recommendation 2 is based on giving the FPR for a prior probability of 0.5. Any higher prior probability of there being a real effect is not justifiable in the absence of hard data. In this sense, the calculated FPR is the minimum that can be expected. More implausible hypotheses would make the problem worse. For example, if the prior probability of there being a real effect were only 0.1, then observation of $p = 0.05$ would imply a disastrously high FPR = 0.76, and in order to achieve an FPR of 0.05, you'd need to observe $p = 0.00045$. Others (especially Goodman) have advocated giving likelihood ratios (LRs) in place of p-values. The FPR for a prior of 0.5 is simply 1/(1 + LR), so to give the FPR for a prior of 0.5 is simply a more-easily-comprehensible way of specifying the LR, and so should be acceptable to frequentists and Bayesians.

Matthews, R., Moving Toward the Post $p < 0.05$ Era via the Analysis of Credibility

1. Report the outcome of studies as effect sizes summarized by confidence intervals (CIs) along with their point estimates.
2. Make full use of the point estimate and width and location of the CI relative to the null effect line when interpreting findings. The point estimate is generally the effect size best supported by the study, irrespective of its statistical significance/nonsignificance. Similarly, tight CIs located far from the null effect line generally represent more compelling evidence for a nonzero effect than wide CIs lying close to that line.
3. Use the analysis of credibility (AnCred) to assess quantitatively the credibility of inferences based on the CI. AnCred determines the level of prior evidence needed for a new finding to provide credible evidence for a nonzero effect.
4. Establish whether this required level of prior evidence is supported by current knowledge and insight. If it is, the new result provides credible evidence for a nonzero effect, irrespective of its statistical significance/nonsignificance.

Gannon, M., Pereira, C., and Polpo, A., Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels

1. Retain the useful concept of statistical significance and the same operational procedures as currently used for hypothesis tests, whether frequentist (Neyman–Pearson p-value tests) or Bayesian (Bayes-factor tests).
2. Use tests with a sample-size-dependent significance level—ours is optimal in the sense of the generalized Neyman–Pearson lemma.
3. Use a testing scheme that allows tests of any kind of hypothesis, without restrictions on the dimensionalities of the parameter space or the hypothesis. Note that this should include "sharp" hypotheses, which correspond to subsets of lower dimensionality than the full parameter space.
4. Use hypothesis tests that are compatible with the likelihood principle (LP). They can be easier to interpret consistently than tests that are not LP-compliant.
5. Use numerical methods to handle hypothesis-testing problems with high-dimensional sample spaces or parameter spaces.

Pogrow, S., How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings

1. Switch from reliance on statistical or practical significance to the more stringent statistical criterion of practical benefit for (a) assessing whether applied research findings indicate that an intervention is effective and should be adopted and scaled—particularly in complex organizations such as schools and hospitals and (b) determining whether relationships are sufficiently strong and explanatory to be used as a basis for setting policy or practice recommendations. Practical benefit increases the likelihood that observed benefits will replicate in subsequent research and in clinical practice by avoiding the problems associated with relying on small effect sizes.
2. Reform statistics courses in applied disciplines to include the principles of practical benefit, and have students review influential applied research articles in the discipline to determine which findings demonstrate practical benefit.
3. Recognize the need to develop different inferential statistical criteria for assessing the importance of applied research findings as compared to assessing basic research findings.
4. Consider consistent, noticeable improvements across contexts using the quick prototyping methods of improvement science as a preferable methodology for identifying effective practices rather than on relying on RCT methods.
5. Require that applied research reveal the actual unadjusted means/medians of results for all groups and subgroups, and that review panels take such data into account—as opposed to only reporting relative differences between adjusted means/medians. This will help preliminarily identify whether there appear to be clear benefits for an intervention.

7.4. Adopting More Holistic Approaches

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J., Abandon Statistical Significance

1. Treat p-values (and other purely statistical measures like confidence intervals and Bayes factors) continuously rather than in a dichotomous or thresholded manner. In doing so, bear in mind that it seldom makes sense to calibrate evidence as a function of p-values or other purely statistical measures because they are, among other things, typically defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error.
2. Give consideration to 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. Do this always—not just once some p-value or other statistical threshold has been attained—and do this without giving priority to p-values or other purely statistical measures.

3. Analyze and report all of the data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold.
4. Conduct a decision analysis: p-value and other statistical threshold-based rules implicitly express a particular tradeoff between Type I and Type II error, but in reality this tradeoff should depend on the costs, benefits, and probabilities of all outcomes.
5. Accept uncertainty and embrace variation in effects: we can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect"—based on some p-value or other statistical threshold being attained.
6. Obtain more precise individual-level measurements, use within-person or longitudinal designs more often, and give increased consideration to models that use informative priors, that feature varying treatment effects, and that are multilevel or meta-analytic in nature.

Tong, C., Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

1. Prioritize effort for sound data production: the planning, design, and execution of the study.
2. Build scientific arguments with many sets of data and multiple lines of evidence.
3. Recognize the difference between exploratory and confirmatory objectives and use distinct statistical strategies for each.
4. Use flexible descriptive methodology, including disciplined data exploration, enlightened data display, and regularized, robust, and nonparametric models, for exploratory research.
5. Restrict statistical inferences to confirmatory analyses for which the study design and statistical analysis plan are prespecified prior to, and strictly adhered to during, data acquisition.

Amrhein, V., Trafimow, D., and Greenland, S., Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication

1. Do not dichotomize, but embrace variation.

   (a) Report and interpret inferential statistics like the p-value in a continuous fashion; do not use the word "significant."

   (b) Interpret interval estimates as "compatibility intervals," showing effect sizes most compatible with the data, under the model used to compute the interval; do not focus on whether such intervals include or exclude zero.

   (c) Treat inferential statistics as highly unstable local descriptions of relations between models and the obtained data.

      (i) Free your "negative results" by allowing them to be potentially positive. Most studies with large p-values or interval estimates that include the null should be considered "positive," in the sense that they usually leave open the possibility of important effects (e.g., the effect sizes within the interval estimates).

      (ii) Free your "positive results" by allowing them to be different. Most studies with small p-values or interval estimates that are not near the null should be considered provisional, because in replication studies the p-values could be large and the interval estimates could show very different effect sizes.

      (iii) There is no replication crisis if we don't expect replication. Honestly reported results *must* vary from replication to replication because of varying assumption violations and random variation; excessive agreement itself would suggest deeper problems such as failure to publish results in conflict with group expectations.

Calin-Jageman, R., and Cumming, G., The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known

1. Ask quantitative questions and give quantitative answers.
2. Countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error.
3. Seek replication, and use quantitative methods to synthesize across data sets as a matter of course.
4. Use Open Science practices to enhance the trustworthiness of research results.
5. Avoid, wherever possible, any use of p-values or NHST.

Ziliak, S., How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough

- *G-10 Consider the Purpose of the Inquiry, and Compare with Best Practice.* Falsification of a null hypothesis is not the main purpose of the experiment or observational study. Making money or beer or medicine—ideally more and better than the competition and best practice—is. Estimating the importance of your coefficient relative to results reported by others, is. To repeat, as the 2016 ASA Statement makes clear, merely falsifying a null hypothesis with a qualitative yes/no, exists/does not exist, significant/not significant answer, is not itself significant science, and should be eschewed.
- *G-9 Estimate the Stakes (Or Eat Them).* Estimation of magnitudes of effects, and demonstrations of their substantive meaning, should be the center of most inquiries. Failure to specify the stakes of a hypothesis is the first step toward eating them (gulp).
- *G-8 Study Correlated Data: ABBA, Take a Chance on Me.* Most regression models assume "iid" error terms—independently and identically distributed—yet most data in the social and life sciences are correlated by systematic, nonrandom effects—and are thus not independent. Gosset solved the problem of correlated soil plots with the "ABBA" layout, maximizing the correlation of paired differences between the As and Bs with a perfectly balanced chiasmic arrangement.
- *G-7 Minimize "Real Error" with the 3 R's: Represent, Replicate, Reproduce.* A test of significance on a single set of data is nearly valueless. Fisher's p, Student's t, and other tests should only be used when there is actual repetition of the experi-

ment. "One and done" is scientism, not scientific. Random error is not equal to real error, and is usually smaller and less important than the sum of nonrandom errors. Measurement error, confounding, specification error, and bias of the auspices are frequently larger in all the testing sciences, agronomy to medicine. Guinnessometrics minimizes real error by repeating trials on stratified and balanced yet independent experimental units, controlling as much as possible for local fixed effects.

- *G-6 Economize with "Less is More": Small Samples of Independent Experiments*. Small sample analysis and distribution theory has an economic origin and foundation: changing inputs to the beer on the large scale (for Guinness, enormous global scale) is risky, with more than money at stake. But smaller samples, as Gosset showed in decades of barley and hops experimentation, does not mean "less than," and Big Data is in any case not the solution for many problems.
- *G-5 Keep Your Eyes on the Size Matters/How Much? Question*. There will be distractions but the expected loss and profit functions rule, or should. Are regression coefficients or differences between means large or small? Compared to what? How do you know?
- *G-4 Visualize*. Parameter uncertainty is not the same thing as model uncertainty. Does the result hit you between the eyes? Does the study show magnitudes of effects across the entire distribution? Advances in visualization software continue to outstrip advances in statistical modeling, making more visualization a no brainer.
- *G-3 Consider Posteriors and Priors too ("It pays to go Bayes")*. The sample on hand is rarely the only thing that is "known." Subject matter expertise is an important prior input to statistical design and affects analysis of "posterior" results. For example, Gosset at Guinness was wise to keep quality assurance metrics and bottom line profit at the center of his inquiry. How does prior information fit into the story and evidence? Advances in Bayesian computing software make it easier and easier to do a Bayesian analysis, merging prior and posterior information, values, and knowledge.
- *G-2 Cooperate Up, Down, and Across (Networks and Value Chains)*. For example, where would brewers be today without the continued cooperation of farmers? Perhaps back on the farm and not at the brewery making beer. Statistical science is social, and cooperation helps. Guinness financed a large share of modern statistical theory, and not only by supporting Gosset and other brewers with academic sabbaticals (Ziliak and McCloskey 2008).
- *G-1 Answer the Brewer's Original Question ("How should you set the odds?")*. No bright-line rule of statistical significance can answer the brewer's question. As Gosset said way back in 1904, how you set the odds depends on "the importance of the issues at stake" (e.g., the expected benefit and cost) together with the cost of obtaining new material.

Billheimer, D., Predictive Inference and Scientific Reproducibility

1. Predict observable events or quantities that you care about.
2. Quantify the uncertainty of your predictions.

Manski, C., Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing

1. Statisticians should relearn statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end.
2. Statistical decision theory should supplant hypothesis testing when statisticians study treatment choice with trial data.
3. Statisticians should use statistical decision theory when analyzing decision making with sample data more generally.

Manski, C., and Tetenov, A., Trial Size for Near Optimal Choice between Surveillance and Aggressive Treatment: Reconsidering MSLT-II

1. Statisticians should relearn statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end.
2. Statistical decision theory should supplant hypothesis testing when statisticians study treatment choice with trial data.
3. Statisticians should use statistical decision theory when analyzing decision making with sample data more generally.

Lavine, M., Frequentist, Bayes, or Other?

1. Look for and present results from many models that fit the data well.
2. Evaluate models, not just procedures.

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C., Inference and Decision-Making for 21st Century Drug Development and Approval

1. Apply Bayesian paradigm as a framework for improving statistical inference and regulatory decision making by using probability assertions about the magnitude of a treatment effect.
2. Incorporate prior data and available information formally into the analysis of the confirmatory trials.
3. Justify and pre-specify how priors are derived and perform sensitivity analysis for a better understanding of the impact of the choice of prior distribution.
4. Employ quantitative utility functions to reflect key considerations from all stakeholders for optimal decisions via a probability-based evaluation of the treatment effects.
5. Intensify training in Bayesian approaches, particularly for decision makers and clinical trialists (e.g., physician scientists in FDA, industry and academia).

van Dongen, N., Wagenmakers, E.J., van Doorn, J., Gronau, Q., van Ravenzwaaij, D., Hoekstra, R., Haucke, M., Lakens, D., Hennig, C., Morey, R., Homer, S., Gelman, A., and Sprenger, J., Multiple Perspectives on Inference for Two Simple Statistical Scenarios

1. Clarify your statistical goals explicitly and unambiguously.
2. Consider the question of interest and choose a statistical approach accordingly.
3. Acknowledge the uncertainty in your statistical conclusions.
4. Explore the robustness of your conclusions by executing several different analyses.
5. Provide enough background information such that other researchers can interpret your results and possibly execute meaningful alternative analyses.

7.5. Reforming Institutions: Changing Publication Policies and Statistical Education

Trafimow, D., Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post P < 0.05 Universe

1. Tolerate ambiguity.
2. Replace significance testing with a priori thinking.
3. Consider the nature of the contribution, on multiple levels.
4. Emphasize thinking and execution, not results.
5. Consider that the assumption of random and independent sampling might be wrong.

Locascio, J., The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration

For journal reviewers

1. Provide an initial provisional decision regarding acceptance for publication of a journal manuscript based exclusively on the judged importance of the research issues addressed by the study and the soundness of the reported methodology. (The latter would include appropriateness of data analysis methods.) Give no weight to the reported results of the study per se in the decision as to whether to publish or not.
2. To ensure #1 above is accomplished, 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. (The latter would be reviewed only if and after a generally irrevocable decision to publish has already been made.)

For investigators/manuscript authors

1. Obtain consultation and collaboration from statistical consultant(s) and research methodologist(s) early in the development and conduct of a research study.
2. Emphasize the clinical and scientific importance of a study in the Introduction section of a manuscript, and give a clear, explicit statement of the research questions being addressed and any hypotheses to be tested.
3. Include a detailed statistical analysis subsection in the Methods section, which would contain, among other things, a justification of the adequacy of the sample size and the reasons various statistical methods were employed. For example, if null hypothesis significance testing and p-values are used, presumably supplemental to other methods, justify why those methods apply and will provide useful additional information in this particular study.
4. Submit for publication reports of well-conducted studies on important research issues regardless of findings, for example, even if only null effects were obtained, hypotheses were not confirmed, mere replication of previous results were found, or results were inconsistent with established theories.

Hurlbert, S., Levine, R., and Utts, J., Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

1. Encourage journal editorial boards to disallow use of the phrase "statistically significant," or even "significant," in manuscripts they will accept for review.
2. Give primary emphasis in abstracts to the magnitudes of those effects most conclusively demonstrated and of greatest import to the subject matter.
3. Report precise p-values or other indices of evidence against null hypotheses as continuous variables not requiring any labeling.
4. Understand the meaning of and rationale for neoFisherian significance assessment (NFSA).

Campbell, H., and Gustafson, P., The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication

1. *Consider the meta-research implications of implementing new publication/funding policies.* Journal editors and research funders should attempt to model the impact of proposed policy changes before any implementation. In this way, we can anticipate the policy impacts (both positive and negative) on the types of studies researchers pursue and the types of scientific articles that ultimately end up published in the literature.

Fricker, R., Burke, K., Han, X., and Woodall, W., Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban

1. Use measures of statistical significance combined with measures of practical significance, such as confidence intervals on effect sizes, in assessing research results.
2. Classify research results as either exploratory or confirmatory and appropriately describe them as such in all published documentation.
3. Define precisely the population of interest in research studies and carefully assess whether the data being analyzed are representative of the population.
4. Understand the limitations of inferential methods applied to observational, convenience, or other nonprobabilistically sampled data.

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer J., Content Audit for p-Value Principles in Introductory Statistics

1. Evaluate the coverage of p-value principles in the introductory statistics course using rubrics or other systematic assessment guidelines.
2. Discuss and deploy improvements to curriculum coverage of p-value principles.
3. Meet with representatives from other departments, who have majors taking your statistics courses, to make sure that inference is being taught in a way that fits the needs of their disciplines.
4. Ensure that the correct interpretation of p-value principles is a point of emphasis for all faculty members and embedded within all courses of instruction.

Steel, A., Liermann, M., and Guttorp, P., Beyond Calculations: A Course in Statistical Thinking

1. Design curricula to teach students how statistical analyses are embedded within a larger science life-cycle, including steps such as project formulation, exploratory graphing, peer review, and communication beyond scientists.
2. Teach the p-value as only one aspect of a complete data analysis.

3. Prioritize helping students build a strong understanding of what testing and estimation can tell you over teaching statistical procedures.
4. Explicitly teach statistical communication. Effective communication requires that students clearly formulate the benefits and limitations of statistical results.
5. Force students to struggle with poorly defined questions and real, messy data in statistics classes.

    5. Encourage students to match the mathematical metric (or data summary) to the scientific question. Teaching students to create customized statistical tests for custom metrics allows statistics to move beyond the mean and pinpoint specific scientific questions.

Acknowledgments

Without the help of a huge team, this special issue would never have happened. The articles herein are about the equivalent of three regular issues of *The American Statistician*. Thank you to all the authors who submitted papers for this issue. Thank you, authors whose papers were accepted, for enduring our critiques. We hope they made you happier with your finished product. Thank you to a talented, hard-working group of associate editors for handling many papers: Frank Bretz, George Cobb, Doug Hubbard, Ray Hubbard, Michael Lavine, Fan Li, Xihong Lin, Tom Louis, Regina Nuzzo, Jane Pendergast, Annie Qu, Sherri Rose, and Steve Ziliak. Thank you to all who served as reviewers. We definitely couldn't have done this without you. Thank you, TAS Editor Dan Jeske, for your vision and your willingness to let us create this special issue. Special thanks to Janet Wallace, TAS editorial coordinator, for spectacular work and tons of patience. We also are grateful to ASA Journals Manager Eric Sampson for his leadership, and to our partners, the team at Taylor and Francis, for their commitment to ASA's publishing efforts. Thank you to all who read and commented on the draft of this editorial. You made it so much better! Regina Nuzzo provided extraordinarily helpful substantive and editorial comments. And thanks most especially to the ASA Board of Directors, for generously and enthusiastically supporting the "*p*-values project" since its inception in 2014. Thank you for your leadership of our profession and our association.

Gratefully,

Ronald L. Wasserstein
American Statistical Association, Alexandria, VA
ron@amstat.org

Allen L. Schirm
Mathematica Policy Research (retired), Washington, DC
allenschirm@gmail.com

Nicole A. Lazar
Department of Statistics, University of Georgia, Athens, GA
nlazar@stat.uga.edu

References

References to articles in this special issue

Amrhein, V., Trafimow, D., and Greenland, S. (2019), "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication," *The American Statistician*, 73. [2,3,4,5,6,7,8,9]

Anderson, A. (2019), "Assessing Statistical Results: Magnitude, Precision and Model Uncertainty," *The American Statistician*, 73. [3]

Benjamin, D., and Berger, J. (2019), "Three Recommendations for Improving the Use of *p*-Values," *The American Statistician*, 73. [5,7,9]

Betensky, R. (2019), "The *p*-Value Requires Context, Not a Threshold," *The American Statistician*, 73. [4,6]

Billheimer, D. (2019), "Predictive Inference and Scientific Reproducibility," *The American Statistician*, 73. [5]

Blume, J., Greevy, R., Welty, V., Smith, J., and DuPont, W. (2019), "An Introduction to Second Generation *p*-Value," *The American Statistician*, 73. [4]

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J. (2019), "The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making," *The American Statistician*, 73. [5]

Calin-Jageman, R., and Cumming, G. (2019), "The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known," *The American Statistician*, 73. [3,5,9,10]

Campbell, H., and Gustafson, P. (2019), "The World of Research Has Gone Berserk: Modeling the Consequences of Requiring 'Greater Statistical Stringency' for Scientific Publication," *The American Statistician*, 73. [8]

Colquhoun, D. (2019), "The False Positive Risk: A Proposal Concerning What to Do About *p*-Value," *The American Statistician*, 73. [4,7,14]

Fraser, D. (2019), "The *p*-Value Function and Statistical Inference," *The American Statistician*, 73. [6]

Fricker, R., Burke, K., Han, X., and Woodall, W (2019), "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their *p*-Value Ban," *The American Statistician*, 73. [7,9]

Gannon, M., Pereira, C., and Polpo, A. (2019), "Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels," *The American Statistician*, 73. [5]

Goodman, S. (2019), "Why is Getting Rid of *p*-Values So Hard? Musings on Science and Statistics," *The American Statistician*, 73. [7,8,10]

Goodman, W., Spruill, S., and Komaroff, E. (2019), "A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting Its Use," *The American Statistician*, 73. [5]

Greenland, S. (2019), "Valid *p*-Values Behave Exactly as They Should: Some Misleading Criticisms of *p*-Values and Their Resolution With *s*-Values," *The American Statistician*, 73. [3,4,5,6,7]

Heck, P., and Krueger, J. (2019), "Putting the *p*-Value in Its Place," *The American Statistician*, 73. [4]

Hubbard, D., and Carriquiry, A. (2019), "Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness and Relevance," *The American Statistician*, 73. [5]

Hubbard, R. (2019), "Will the ASA's Efforts to Improve Statistical Practice Be Successful? Some Evidence to the Contrary," *The American Statistician*, 73. [8]

Hubbard, R., Haig, B. D., and Parsa, R. A. (2019), "The Limited Role of Formal Statistical Inference in Scientific Inference," *The American Statistician*, 73. [2,7]

Hurlbert, S., Levine, R., and Utts, J. (2019), "Coup de Grâce for a Tough Old Bull: 'Statistically Significant' Expires," *The American Statistician*, 73. [8,9]

Ioannidis, J. (2019), "What Have We (Not) Learnt From Millions of Scientific Papers With *p*-Values?," *The American Statistician*, 73. [6]

Johnson, V. (2019), "Evidence From Marginally Significant *t* Statistics," *The American Statistician*, 73. [7]

Kennedy-Shaffer, L. (2019), "Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize *p*-Values and Significance Testing," *The American Statistician*, 73. [7]

Kmetz, J. (2019), "Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values," *The American Statistician*, 73. [8]

Lavine, M. (2019), "Frequentist, Bayes, or Other?," *The American Statistician*, 73. [6]

Locascio, J. (2019), "The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration," *The American Statistician*, 73. [4,5,8,9]

Manski, C. (2019), "Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," *The American Statistician*, 73. [5]

Manski, C., and Tetenov, A. (2019), "Trial Size for Near Optimal Choice between Surveillance and Aggressive Treatment: Reconsidering MSLT-II," *The American Statistician*, 73. [5]

Matthews, R. (2019), "Moving Toward the Post $p < 0.05$ Era Via the Analysis of Credibility," *The American Statistician*, 73. [4,9]

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer, J. (2019), "Content Audit for *P*-Value Principles in Introductory Statistics," *The American Statistician*, 73. [8]

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J. (2019), "Abandon Statistical Significance," *The American Statistician*, 73. [4,6,7]

McShane, B., Tackett, J., Böckenholt, U., and Gelman, A. (2019), "Large Scale Replication Projects in Contemporary Psychological Research," *The American Statistician*, 73. [9]

O'Hagan, A. (2019), "Expert Knowledge Elicitation: Subjective But Scientific," *The American Statistician*, 73. [5,6]

Pogrow, S. (2019), "How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings," *The American Statistician*, 73. [4]

Rose, S., and McGuire, T. (2019), "Limitations of p-Values and R-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment," *The American Statistician*, 73. [7]

Rougier, J. (2019), "p-Values, Bayes Factors, and Sufficiency," *The American Statistician*, 73. [6]

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C. (2019), "Inference and Decision-Making for 21st Century Drug Development and Approval," *The American Statistician*, 73. [10]

Steel, A., Liermann, M., and Guttorp, P. (2019), "Beyond Calculations: A Course in Statistical Thinking," *The American Statistician*, 73. [8]

Trafimow, D. (2019), "Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < .05$ Universe," *The American Statistician*, 73. [7]

Tong, C. (2019), "Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science," *The American Statistician*, 73. [2,3,4,6,8,9]

van Dongen, N., Wagenmakers, E. J., van Doorn, J., Gronau, Q., van Ravenzwaaij, D., Hoekstra, R., Haucke, M., Lakens, D., Hennig, C., Morey, R., Homer, S., Gelman, A., and Sprenger, J. (2019), "Multiple Perspectives on Inference for Two Simple Statistical Scenarios," *The American Statistician*, 73. [6]

Ziliak, S. (2019), "How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little 'P' is Not Enough," *The American Statistician*, 73. [2,3]

Other articles or books referenced

Boring, E. G. (1919), "Mathematical vs. Scientific Significance," *Psychological Bulletin*, 16, 335–338. [2]

Cumming, G. (2014), "The New Statistics: Why and How," *Psychological Science*, 25, 7–29. [3]

Davidian, M., and Louis, T. (2012), "Why Statistics?" *Science*, 336, 12. [2]

Edgeworth, F. Y. (1885), "Methods of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume, 181–217. [2]

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd. [2]

Gelman, A. (2015), "Statistics and Research Integrity," *European Science Editing*, 41, 13–14. [8]

——— (2016), "The Problems With p-Values Are Not Just With p-Values," *The American Statistician*, supplemental materials to *ASA Statement on p-Values and Statistical Significance*, 70, 1–2. [3]

Gelman, A., and Hennig, C. (2017), "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society*, Series A, 180, 967–1033. [5]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [2]

Ghose, T. (2013), "'Just a Theory': 7 Misused Science Words," *Scientific American* (online), available at *https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/*. [2]

Goodman, S. (2018), "How Sure Are You of Your Result? Put a Number on It," *Nature*, 564. [5]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: Sage. [1]

Junger, S. (1997), *The Perfect Storm: A True Story of Men Against the Sea*, New York: W.W. Norton. [10]

Locascio, J. (2017), "Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39, 239–246. [8]

Mayo, D. (2018), "Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars," Cambridge, UK: University Printing House. [1]

McShane, B., and Gal, D. (2016), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62, 1707–1718. [9]

——— (2017), "Statistical Significance and the Dichotomization of Evidence." *Journal of the American Statistical Association*, 112, 885–895. [9]

Mogil, J. S., and Macleod, M. R. (2017), "No Publication Without Confirmation," *Nature*, 542, 409–411, available at *https://www.nature.com/news/no-publication-without-confirmation-1.21509*. [8]

Rosenthal, R. (1979), "File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin* 86, 638–641. [2]

Wasserstein, R., and Lazar, N. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [1]

Wellek, S. (2017), "A Critical Evaluation of the Current p-Value Controversy" (with discussion), *Biometrical Journal*, 59, 854–900. [4,9]

Ziliak, S., and McCloskey, D. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [1,16]