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: 280104

View related articles

View Crossmark data

Citing articles: 822 View citing articles

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 $1/(1 + \text{LR})$, 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]

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: 27679

View related articles

View Crossmark data

Citing articles: 46 View citing articles

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

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

Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics

Steven N. Goodman

To cite this article: Steven N. Goodman (2019) Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics, The American Statistician, 73:sup1, 26-30, DOI: 10.1080/00031305.2018.1558111

To link to this article: https://doi.org/10.1080/00031305.2018.1558111

© 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: 12881

View related articles

View Crossmark data

Citing articles: 19 View citing articles

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, 26–30: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1558111

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics

Steven N. Goodman

Departments of Medicine and Health Research and Policy, Stanford University, Stanford, CA

ABSTRACT
The current concerns about reproducibility have focused attention on proper use of statistics across the sciences. This gives statisticians an extraordinary opportunity to change what are widely regarded as statistical practices detrimental to the cause of good science. However, how that should be done is enormously complex, made more difficult by the balkanization of research methods and statistical traditions across scientific subdisciplines. Working within those sciences while also allying with science reform movements—operating simultaneously on the micro and macro levels—are the key to making lasting change in applied science.

ARTICLE HISTORY
Received April 2018
Revised November 2018

KEYWORDS
P-values; Reproducible research; Statistical inference, Scientific inference

My first statistical graduate advisor and mentor, Richard Royall at Johns Hopkins, warned that if I took on inference as a professional path, I would be lonely. He was not wrong, but the times have dramatically changed; I cannot recall a time in my professional career when so many statisticians and scientists have discussed issues in statistical inference with such energy and urgency. This is being produced by the largest metacontroversy that science may have yet faced—the "reproducibility crisis" —with one of its root causes being seen as misunderstanding and misuse of statistical methods. If we are interested in change, we cannot afford to waste a crisis.

To start to understand the trajectory going forward, it can be useful to see what its slope has been up to now. Consider this quote:

> …scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level. Then the statistician can grandly draw obvious conclusions about data from any scientific field by proclaiming significance or non-significance. Such nonsense is taught usually by professors who have had minimal contact with the applications of statistical methods to scientific problems. As a result, the number of scientific papers which use statistical methods for window dressing is increasing. It appears that the P-value next to a contingency table is beginning to mean what the "Seal of Good Housekeeping" means …(Cutler et al. 1966).

Those are the words of Marvin Zelen in 1965, summarizing an NIH symposium on hypothesis testing in clinical trials, featuring talks from Sam Greenhouse, Jerry Cornfield, and Marvin Schneiderman. It is fair to say that the situation today is little different, and a major impetus of the ASA *P* value statement. If we are contemplating change, it is critical that we understand why things have improved so little over the last half century, and even since RA Fisher's warnings decades earlier. The answer is to be found more in the sociology and organization of science than in its technical or inferential foundations. I will focus mainly on the application of statistics within biomedical science.

Disciplinary Siloing of Methods

Almost all scientific methods are used, taught and communicated within disciplinary communities. It is within these communities that establish what will count for legitimate knowledge claims; cardiologist write for cardiologists, oncologists for oncologists, psychologists for psychologists, economists for economists. Each of these disciplines has somewhat different modes of experimentation and epistemic justification for recommendations. Most medical specialties look to randomized clinical trials on which to base clinical guidelines, whereas most surgical journals are filled with case series that rarely appear in the top journals of medical specialties.

Different disciplines have different historical memories about achievements and failures, like cultures that pass down archetypal stories that transmit values and customs to the next generation.

A clinical trial that shook the core of cardiology was the Cardiac Arrhythmia Suppression Trial (CAST), completed in 1991 (Echt et al. 1991). The therapies tested in the CAST trial were known to stop the arrhythmias thought responsible for

CONTACT Steven N. Goodman *steve.goodman@stanford.edu*
This essay was adapted from a keynote address given at the ASA Symposium on Statistical Inference, 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.

sudden cardiac death, which was then killing more than one person every minute, or about 1200 per day. CAST revealed that these drugs, the most widely prescribed drugs in the United States, were almost quadrupling the sudden death rate, killing more Americans in the preceding decade than had died in most wars. This taught the cardiology community the danger of using surrogate endpoints, and the unreliability of mechanistic knowledge more effectively than could a thousand lectures, papers or causal diagrams. Yet few physicians in other disciplines know of it and have learned that lesson. These other areas have their own stories, yet even the most dramatic, like the Duke "omics" story (Micheel et al. 2012) are little known outside their own domains.

This siloing includes statistical methods. A methods paper in a cancer journal will usually be little read in another specialties; if it is in pediatrics, it will be little noticed in adult medicine, if it is in a surgical journal, not noted in medicine, and if in epidemiology, not read by most clinical investigators. In the mid-2000's many sports medicine researchers adopted a statistical method called "magnitude-based inference" (Batterham and Hopkins 2006). Subsequent papers showed the approach to be unfounded (Butson 2018; Sainani 2018), but it is still used, and the fact that a subfield of medicine could adopt its own brand of statistical inference in isolation from the rest of the biomedical literature is telling. So, methods education must be both disseminated across fields and targeted within them.

Methods traditions move across scientific disciplines slowly. In psychology, a one-and-done tradition of theory confirmation by a single small randomized trial, while weakening, is still dominant. Multiple articles in the psychology literature over the past decade have raised awareness of the pernicious consequences of single low-power studies, many decades after that became appreciated in clinical medicine. Worse, replication studies in psychology are sometimes taken as an affront to the original researcher (Yong 2012), and the movement to replicate studies has only recently gotten traction.

In the life sciences, the movement has been yet slower, perhaps because the involvement of statisticians in the research is less, and training in statistical reasoning, inadequate. Two of the premier life-science journals, Science and Nature—only added a statistical review component relatively recently (Van Noorden 2014), in contrast to top clinical research journals for which this has been standard since the 1980s. In a poll of life scientists conducted by Nature on ways to improve research reproducibility, the number one remedy named was better training in statistics. (Baker 2016) The leaders of both the NIH and NSF have taken steps and have stated that they are committed to improving research reproducibility in laboratory science.

Gigerenzer and colleagues have written that among disciplines, one finds flavors of statistics, methodologies, standards of proof and acceptable design that are profoundly different (Gigerenzer et al. 1989). Unlike other scientific theories, statistical inference is taught without names or seeming controversy, with the merger of Fisherian and Neyman-Pearsonian approaches called the "silent hybrid solution." (Gigerenzer et al. 1989) Bayesian ideas are rarely mentioned outside of the treatment of Bayesian statistics. It is hard to teach non-frequentist approaches to inference when students are unaware of non-frequentist definitions of probability.

Students who readily accept and understand foundational disputes in physics, biology and economics are treated in statistics courses as though, in Jack Nicholson's immortal cinematic words, they "can't handle the truth." (*https://bit.ly/1sB5MXg*).

Why It Is Hard to Eliminate *P*-Values?

This brings us to the question of why eliminating P-value is so hard. The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It is 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 does not matter if the P-value does not mean what people think it means; it becomes valuable because of what it buys.

Ted Porter, in his book "Trust in Numbers" has written that statistics are a refuge of objectivity for disciplines that are fighting other battles, usually against some form of claimed expertise of its members. Statistics in medicine pushed back against claims of physicians that they could predict what worked and in whom using physiologic or biologic reasoning:

> …the use of statistics tests has become obligatory in…scientific research. …they work mainly as social technologies, not as guides to private thinking.
> The advances of statistics in medicine must be understood as responses to problems of trust, which have been most acute in the context of regulatory and disciplinary confrontations. This, and not any inherently statistical character of clinical medicine, explain why inferential statistics entered medicine through therapeutics. (Porter 1995, p. 209)

P-values are part of a rule-based structure that serves as a bulwark against claims of expertise untethered from empirical support. It can be changed, but we must respect the reason why the statistical procedures are there in the first place. This partly explains why there is so much resistance to Bayesian approaches, which are often viewed as a back-door way to reintroduce the subjectivity that conventional statistical methods were introduced to counter.

The use of *P*-values is a social phenomenon upon which many social rewards and penalties rest. I am faced by this dilemma in my own teaching; when I have biomedical research students for only a day, the first priority is to teach them to be able to read the literature as it is, not as I would like it to be. When I have them for a week, I can include "alternative" material, mainly at the end. I need students for a full term or more, in part to un-teach what they have been taught so I can introduce a different perspective. That's the consequence and power of social norms, which are difficult to change.

A question asked both by Benjamini (2016) and many others in their response to the *P*-value statement is whether *P*-values are really at the heart of the problem. *P*-values carry within themselves conceptual and theoretical baggage that has led to their inappropriate reification and misuse (Goodman 1999a). So it is tempting to think that if we get rid of *P*-values and null hypothesis significance testing, we get rid of that baggage. But that has been the argument for 90 years and thus far is not working.

So what is it that we really want? The ASA statement says it; we want good scientific practice. We want to measure not just the signal properly but its uncertainty, the twin goals of statistics. We want to make knowledge claims that match the strength of the evidence. Will we get that by getting rid of *P*—values? Will eliminating *P*—values improve experimental design? Would it improve measurement? Would it help align the scientific question with those analyses? Will it eliminate bright line thinking? If we were able to get rid of *P*-values, are we sure that unintended consequences wouldn't make things worse? In my idealized world, the answer is yes, and many statisticians believe that. But in the real world, I am less sure.

The problems of *P*-values are in large part from the thoughtless implementation of null hypothesis significance testing (NHST) that even Pearson in his later years said he never envisioned, i.e., divorced from design and from consequences. (Pearson 1962)

The Bayes factor alternative (Goodman 1999b; Kass and Raftery 1995) is attractive but may be the bitcoin equivalent; people are not sure what it means, have little clue where it will be accepted, and it has variations in value. Using Bayes factors or posterior probabilities are akin to forcing people to use the metric system; it makes sense and it fits into a coherent universal system of measurement, but many feel lost when they cannot use familiar measures whose meaning they have internalized, like pounds, Fahrenheit and yards. Bayes factors can also be used in an inappropriate bright-line way if their conceptual foundation as adjusting prior odds are not understood. Bayes factors and methods involve asking for much more than a substitution of one index for another, but rather a new conceptual framework for dealing with scientific uncertainty, which traditional methods do not require probing, and standard courses do not teach how to assess. This is not an argument against some form of Bayes Factor as a replacement for the *P*-value, which I believe would be an improvement (Goodman, 1999b), but an explanation of why its acceptance will be difficult.

The Five "W" Questions

Twenty four of the 31 people who contributed to the ASA *P* value statement development wrote commentaries (Wasserstein and Lazar 2016). That shows how surprising it was that a consensus statement was written at all, akin to a Supreme Court opinion with nine dissents. Of these 24, 21 were generally laudatory, although most were still quite critical of various details. 17 made suggestions for change, 5 being specific (e.g., using a Bayes factor bound in lieu of a *p*-value), and 12 aspirational, w/o a specific strategy (e.g., teach scientists to better understand uncertainty). That showed us that statisticians are passionately interested in the role of statistics in the actual practice of science, but that there is more agreement on what we should *not* do than what we should, a reason why the ASA statement had more "don'ts" then "do's."

In my commentary (Goodman, 2016), I posed the question of what the next steps of statisticians on the road to inferential and research reform are, breaking it down into the 5 "W" questions: who, what, where, when and why. As these are not completely separable, I will address some in combination.

There are a variety of both established and emerging international organizations concerned with improving research conduct, such as the Cochrane and Campbell Collaborations, the Equator Network and the World Congress of Research Integrity. Quite a few centers and initiatives have been established within the last decade to promote reproducible research, such as the Center for Open Science (COS), Berkeley Initiative for Transparency in the Social Sciences (BITSS), and the Meta-Research Innovation Center at Stanford (METRICS). Working with and through any of these entities offer the opportunity to influence the methods and conduct of scientific research locally or worldwide.

Institutions

Some institutional and regulatory doors have already been opened for this, but statisticians must take advantage. A 2018 report released by the National Academy of Sciences entitled "Fostering Integrity in Research" expanded the definition of research integrity beyond the traditional categories of falsification, fraud, and plagiarism to a variety of misleading inferential practices called "detrimental research practices." It suggests that institutions be responsible to monitor and improve both (National Academies of Sciences, 2017) Many if not all of these practices involve deviation from sound statistical design, conduct, analysis and interpretation. Putting proper statistical practice within the frame of scientific integrity, with institutional structures to oversee it may provide an avenue for statisticians to affect the way institutions teach, monitor and promote good science. It means putting structures in place—both training and monitoring—where statistical input can be introduced outside of individual collaborations or standard courses. New teaching models that put inferential issues front and center can also teach and inspire the next generation of both statisticians and scientists to make change.

We also need to create pathways to promotion at our institutions to value and reward faculty who are effective in those activities in addition to than traditional methodological research, coupled with promotion criteria for scientists not based on bibliometrics, which drives the perceived need for significant results (Moher 2018). With the proper professional incentives and rewards, some statisticians might have more time to spend on improving statistical practice, and their collaborators might feel less pressure to game the tools of inference (Wang 2018).

Journals

Applied journals are a very good place for different approaches to be demonstrated, as they can shape the norms of the discipline. Published examples using non-NHST inferential approaches are enormously powerful in demonstrating that papers using these methods will be published in that discipline, as well as illustrating how those methods are to be implemented. Recent publications in JAMA of Bayesian-analyzed clinical trials (Laptook 2017; Goligher 2018) coupled with editorials (Lewis and Angus 2018; Quintana, Viele, and Lewis 2017) serve that role well. Another avenue for influence is for statisticians to serve as statistical editors, of which there are far too few.

Publishing in disciplinary journals also reaches the many teachers and practitioners of statistical methods within disciplines who are not professional statisticians. Writing commentaries that focus on methodology and joining researchers in that field to do their own meta-research, pointing out the prevalence and consequences of suboptimal statistical practice can have great impact, especially if it highlights errors made through practices condemned in the ASA statement.

Funders

Other venue to exert positive influence is with the research funders. It may be that we could do with a little less support for developing new methods and more for tools to train the next generation properly, to produce user friendly code to implement alternatives to *P*-values (*https://osf.io/7dzmk/*), to develop new curricula and hands-on learning tools, and new ways to disseminate and role model proper inferential practice. This is happening to some extent, but the market for such tools is large and constantly evolving.

The most powerful leverage funders have is to require certain levels of methodological rigor in the research they fund. But the transformation of methodological recommendations into implementable policies is difficult. I will present two examples of methods-related policies—one very clear, the other less clear and more problematic in practice. The first is the requirement that any clinical trial submitted to a top medical journal be registered at inception to be considered for publication many years later (DeAngelis et al. 2004). It spurred a dramatic increase in clinical trial registration well before it was required by law in 2007. It was a rule first imposed by the journals, and for the most part, it worked. It was operationally clear, there was a strong technical consensus behind it, the target activity was unambiguous, and the enforcement agent and consequence was spelled out. With the information being public, once this was required by law compliance started to be monitored by others, including the media. (Piller 2015) This preregistration had the effect of increasing the proportion of nonsignificant results in clinical trial publications, and perhaps indirectly decreasing the perceived importance of significant results as a criterion for submission or acceptance (Kaplan, 2015). What the rule and law was less clear about—the completeness and accuracy of the registration—resulted in problems that we are now confronting. (Zarin, 2017).

Another set of methods policies has had only partial success. These are the methodology standards of the Patient-Centered Outcomes Research Institute, PCORI (*https://bit.ly/2xaPcEA*). PCORI was established by Congress with just two boards: the governing board and the methodology committee. This is a surprising arrangement and an opportunity, that a committee of methodologists (currently including a former ASA president) should sit at the highest level of a major research funding agency to advise on how to ensure the best quality science and statistics.

The PCORI methods standards aspire to be minimal, in that they are rules that most would agree if they were violated would seriously weaken or invalidate a given study. They cover how to formulate research questions, the operational meaning of patient centeredness, data integrity, missing data, treatment heterogeneity, data registries, data networks, causal inference, Bayesian trials, diagnostic tests, systematic reviews, data management, and complex interventions.

These are the kinds of standards that one might imagine could prevent the worst problems in scientific practice and inference. But the PCORI effort to translate the standards into policies, assuring that they are met in all research they fund has been only partly successful (Mayo-Wilson, 2017). The reason is that to enforce these standards, a staff member must check a proposal against a long list of multi-part standards that have many grey areas in practice. Even methodologically sophisticated staff find it difficult to judge when the requirements of a standard have been adequately met. If 4% of the data are missing, which cannot be known at the outset, is multiple imputation required? Sensitivity analyses? The consequences of violation are not clear either; should funding be affected by issues like this? This is made yet more complicated when adherence can only be assessed once the project is completed and the data analyzed.

So, when we say that funding agencies have the power to improve the quality of research, whether in technical or inferential dimensions, we have to specify how that power is to be exercised. Often, we find that the entity charged with enforcing change does not have adequate personnel, processes, or enforcement capacity, and that the many gray areas make "violations"—a bright-line designation—hard to assess. The difference between a statistical guideline and a research policy is typically quite wide.

A regulator with a particularly strong influence on statistical practice is the FDA. In addition to issuing guidances on the use of Bayesian methods in device regulation (FDA, 2010), and adaptive trials (including Bayesian approaches) for drugs and biologics, (FDA 2018) the 21st Century Cures Act has directed them to develop further guidances on novel adaptive trial approaches. They are currently funding both internal and external researchers to help advance this effort, and we should expect to see outlines of new FDA research paradigms in the next several years.

The Future

Progress in changing a dominant paradigm is difficult, and lack of movement by any of the key players—funders, journals, institutions, regulators, or payors—constrains movement by the others. The converse is that movement in each of these realms affects the others, and all of these institutions are currently changing. 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 ball is rolling downhill, the current generation is inspired and impatient to carry this forward, and I trust that by the meetings in 2027, no less 2067, more progress will have been made than we have since 1967. The crisis is upon us, the inference iron is hot, and it is time to get this done.

References

Baker, M. (2016), "1,500 Scientists Lift the Lid on Reproducibility," *Nature*, 533, 452–454. [27]

Batterham A.M., Hopkins W.G. (2006), "Making Meaningful Inferences About Magnitudes," *International Journal of Sports Physiological Performance*, 1, 50–57. [27]

Benjamini, Y. (2016), "It's Not the *P*-values' Fault," *The American Statistician*, ASA Statement supplementary materials, available at *utas_a_1154108_sm5354.pdf* [27]

Butson, M. (2018), "Will the Numbers Really Love You Back: Re-examining Magnitude-Based Inference," [Internet] 2018, available at *https://osf.io/yvj5r/*. [27]

Camerer, C.F., Dreber, A., Holzmeister, F. Ho T-H., Huber J., Johannesson M., Kirchler M., Nave G., Nosek B.A., and Pfeiffer T. (2018), "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644.

Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman, M. A. (1966), "The Role of Hypothesis Testing in Clinical Trials," *J Chron Disease*, 19, 857–882. [26]

DeAngelis, C.D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J., Schroeder, T. V., Sox, H. C., and Van Der Weyden, M.B. (2004), "Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors," *JAMA*, 292, 1363–1364. [29]

Echt, D.S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A., Friedman, L., Greene, H. L., Huther, M. L., Richardson, D. W., and The Cast Investigators (1991), "Mortality and Morbidity in Patients Receiving Encainide, Flecaininde, or Placebo. The Cardiac Arrhythmia Suppression Trial," *The New England Journal of Medicine*, 234, 12 781–788. [26]

Food and Drug Administration, Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials Document (2010), available at *https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf* [29]

Food and Drug Administration (2018), "Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry — Draft Guidance," available at *https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf* [29]

Gigerenzer, G, Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Kruger, L. (1989), *The Empire of Chance*, Cambridge, UK: Cambridge University Press. [27]

Goligher, E.C., Tomlinson, G., Hajage, D. et al. (2018), "Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome and Posterior Probability of Mortality Benefit in a Post Hoc Bayesian Analysis of a Randomized Clinical Trial," *JAMA*, 320, 2251–2259. [28]

Goodman, S. N. (1999a), "Towards Evidence-based Medical Statistics, I: The P-value Fallacy," *Annals of Internal Medicine*, 130 995–1004. [27]

——— (1999b), "Towards Evidence-based Medical Statistics, II: The Bayes Factor," *Annals of Internal Medicine*, 130 1005–1013. [28]

——— (2016) "The Next Questions: Who, What, When, Where and Why?" *The American Statistician*, 73(S1), this issue. [28]

——— (2018), "How Confident Are You in a Research Finding? Say It with a Number!" *Nature*, 564, 7.

Institute of Medicine (2012), Committee, on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, 'Evolution of Translational Omics'.

Ioannidis, J. P., Fanelli, D., Dunne, D. D. and Goodman, S. N. (2015), Meta-research: Evaluation and Improvement of Research Methods and Practices," *PLoS Biology*, 13 (10), e1002264.

Kaplan, R. M., and Irvin, V. L. (2015), "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time," PLoS One, 10, e0132382. [29]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *JASA*, 90, 773–795. [28]

Laptook, A. R., Shankaran, S., Tyson, J. E., Munoz, B. , Bell, E. F., Goldberg R. N., Parikh, N. A., Ambalavanan, N., Pedroza, C., Pappas, A., Das, A., Chaudhary, A.S., Ehrenkranz, R. A., Hensman, A. M., Van Meurs, K. P., Chalak, L. F., Khan, A. M., Hamrick, S. E. G., Sokol, G. M., Walsh, M. C., Poindexter, B. B., Faix, R. G., Watterberg, K. L., Frantz, I. D., Guillet, R., Devaskar, U., Truog, W. E., Chock, V. Y., Wyckoff, M. H., McGowan, E. C., Carlton, D. P., Harmon, H. M., Brumbaugh, J. E., Cotton, C. M., Sánchez, P. J., Hibbs, A. M., and Higgins, R. D. (2017), "Effect of Therapeutic Hypothermia Initiated After 6 Hours of Age on Death or Disability Among Newborns With Hypoxic-Ischemic Encephalopathy: A Randomized Clinical Trial," *JAMA*, 318 (16), 1550–1560. [28]

Lewis, R. J., and Angus, D. C. (2018), "Time for Clinicians to Embrace Their Inner Bayesian?: Reanalysis of Results of a Clinical Trial of Extracorporeal Membrane Oxygenation," *JAMA*, 320(21), 2208–2210. [28]

Mayo-Wilson, E., Vander Ley, K., Dickersin, K., Helfand, M. (2017), "Patient-Centered Outcomes Research Institute (PCORI) Methodology Standards to Improve the Design and Reporting of Research." [Abstract] International Congress on Peer Review and Publication. *https://peerreviewcongress.org/prc17-0326*. [29]

Micheel C. M. , Nass S. J., Omenn G. S. (2012), *Evolution of Translational Omics : Lessons Learned and the Path Forward*, National Academies Press; Washington, DC. [27]

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., and Goodman, S. N. (2018), "Assessing Scientists for Hiring, Promotion, and Tenure," *PLoS Biol.* 16:e2004089. [28]

National Academies of Sciences, Engineering, and Medicine (2017), "Fostering Integrity in Research," Washington, DC. [28]

PCORI Methodology Committee (2018), "PCORI Methodology Report." available at *https://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-report*.

Pearson, E.S. (1962), "Some Thoughts on Statistical Inference," *Annals of Mathematical Statistics*, 33, 394–403. [28]

Piller, C. (2015), 'Failure to Report: A STAT Investigation of Clinical Trials', *Stat News*, available at *https://www.statnews.com/2015/12/13/clinical-trials-investigation/* [29]

Porter, T. (1995), *Trust In Numbers: The Pursuit Of Objectivity in Science and Public Life*. Princeton: Princeton University Press. [27]

Quintana, M., Viele, K., and Lewis, R. J. (2017), "Bayesian Analysis: Using Prior Information to Interpret the Results of Clinical Trials," *JAMA*, 318 (16), 1605–1606. [28]

Sainani K. L. (2018), "The Problem with Magnitude-based Inference," *Medicine and Science in Sports and Exercise*, 50 (10), 2166–2176. [27]

Van Noorden, R. (2014), "Science Joins Push to Screen Statistics in Papers," *Nature*, available at *doi:10.1038/nature.2014.15509* [27]

Wang, M.Q., Yan, A. F., and Katz, R. V. (2018), "Researcher Requests for Inappropriate Analysis and Reporting: A U.S. Survey of Consulting Biostatisticians," *Annals of Internal Medicine*, 169 (8), 554–558. [28]

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. [28]

Yong, E. (2012), "A Failed Replication Draws a Scathing Personal Attack from a Psychology Professor," *Discover Magazine* , available at *https://bit.ly/2qc5bkz*. [27]

Zarin, D. A., Tse, T., Williams, R. J., and Rajakannan, T. (2017), "Update on Trial Registration 11 Years after the ICMJE Policy Was Established," *The New England Journal of Medicine*, 376 (4), 383–391. [29]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

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

Raymond Hubbard

To cite this article: Raymond Hubbard (2019) Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary, The American Statistician, 73:sup1, 31-35, DOI: 10.1080/00031305.2018.1497540

To link to this article: https://doi.org/10.1080/00031305.2018.1497540

© 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: 6543

View related articles

View Crossmark data

Citing articles: 8 View citing articles

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, 31–35: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1497540

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

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

Raymond Hubbard

Department of Marketing, College of Business and Public Administration, Drake University, Des Moines, IA

ABSTRACT

Recent efforts by the American Statistical Association to improve statistical practice, especially in countering the misuse and abuse of null hypothesis significance testing (NHST) and *p*-values, are to be welcomed. But will they be successful? The present study offers compelling evidence that this will be an extraordinarily difficult task. Dramatic citation-count data on 25 articles and books severely critical of NHST's negative impact on good science, underlining that this issue was/is well known, did nothing to stem its usage over the period 1960–2007. On the contrary, employment of NHST increased during this time. To be successful in this endeavor, as well as restoring the relevance of the statistics profession to the scientific community in the 21st century, the ASA must be prepared to dispense detailed advice. This includes specifying those situations, if they can be identified, in which the *p*-value plays a clearly valuable role in data analysis and interpretation. The ASA might also consider a statement that recommends abandoning the use of *p*-values.

ARTICLE HISTORY
Received February 2018
Revised May 2018

KEY WORDS
Citation analysis; Null hypothesis significance testing; *p*-Values; Statistical practice; Teaching statistics

"Much has been written about the misunderstandings and misinterpretations of *p*-values. The cumulative impact of such criticisms in statistical practice and on empirical research has been nil" (Berry 2017, p. 896).

1. Introduction

The American Statistical Association recently launched a number of unprecedented initiatives aimed at improving statistical practice. These include the ASA statement on *p*-values (Wasserstein and Lazar 2016), the October 11–13, 2017 Symposium on Statistical Inference in Bethesda, MD, which in turn spawned this TAS Special Issue, "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$," an online, permanently open access issue of *The American Statistician*. For these efforts, the ASA is to be warmly congratulated. But will they be successful?

Leaving aside the tricky definition as to what constitutes "successful," and the time horizon this might involve, it is clear that the ASA faces a daunting uphill battle. As Berry's (2017) introductory quotation spells out, previous attempts to tackle this problem, especially the rampant misuse and abuse of null hypothesis significance testing (NHST) and its ubiquitous *p*-value, have been totally ineffective.

The present contribution provides empirical support for this claim. More specifically, it is shown how a series of highly cited articles and books exposing the fundamental weaknesses associated with NHST, and its detrimental effects on knowledge development, nevertheless were unable to prevent its inexorable rise in the social and management sciences over the period 1960–2007. This does not bode well for the ASA's initiatives.

2. Data on the Spread of NHST, 1960–2007

Data on the spread of NHST in the social and management sciences were obtained from Hubbard (2016, pp. 16–30). He content-analyzed a randomly selected issue of leading journals in the social (geography, political science, psychology, and sociology) and management (accounting, economics, finance, management, and marketing) sciences for every year from 1960 through 2007 to estimate the incidence of empirical research using NHST in these areas. Leading journals were targeted since they would be expected to feature best research practices. For each social science discipline, these journals are given in parentheses: Geography (*Annals of the Association of American Geographers, Economic Geography, Professional Geographer*); Political Science (*American Journal of Political Science, American Political Science Review, Public Administration Review*); Psychology (*Journal of Applied Psychology, Journal of Comparative Psychology, Journal of Consulting and Clinical Psychology, Journal of Educational Psychology, Journal of Experimental Psychology: General, Psychological Bulletin, Psychological Review*); and Sociology (*American Journal of Sociology, American Sociological Review, Social Forces*).

This examination resulted in a total of 5541 empirical articles employing NHST across the four disciplines. The percentage of empirical articles employing NHST in these four disciplines, on a decade-by-decade basis (the final "decade" being 2000–2007), is given in column 10 of Table 1.

The above analysis was repeated for the management disciplines. Being younger disciplines, some journals did not span the entire 1960–2007 time period. In these cases, their inaugural publication dates are supplied: Accounting (*The Accounting Review, Journal of Accounting and Economics*, 1979, *Journal of Accounting Research*, 1963); Economics (*American Economic*

CONTACT Raymond Hubbard drabbuhyar@aol.com Department of Marketing, College of Business and Public Administration, Drake University, Des Moines, IA 50311.

© 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.

Table 1. Citation counts of selected articles and books critical of NHST, 1960–2017, and the percentage of empirical social and management science research employing it, 1960–2007.

		Citation counts								
Years	Article/Book	1960–1969	1970–1979	1980–1989	1990–1999	2000–2009	2010–2017	1960–2017	Social Sciences	Management Sciences
1960–1969	Rozeboom (1960)	28	55	71	104	213	283	754	56%	52 %
	Bakan (1966)	27	138	215	219	192	301	1092		
	Meehl (1967)	9	95	85	134	220	291	834		
	Lykken (1968)	8	137	200	215	226	376	1162		
	Total	72								
1970–1979	Morrison and Henkel (1970)	—	121	124	167	193	176	781	72%	80%
	Greenwald (1975)	—	54	194	215	193	272	928		
	Carver (1978)	—	7	208	267	334	342	1158		
	Meehl (1978)	—	9	225	475	517	568	1794		
	Total		616							
1980–1989	Leamer (1983)	—	—	204	427	903	876	2410	84%	89%
	Oakes (1986)	—	—	18	216	332	243	809		
	Berger and Sellke (1987)	—	—	38	230	328	440	1036		
	Gigerenzer and Murray (1987)	—	—	14	282	254	166	716		
	Gigerenzer et al. (1989)	—	—	5	244	503	430	1182		
	Rosnow and Rosenthal (1989)	—	—	2	224	344	393	963		
	Total			1603						
1990–1999	Cohen (1990)	—	—	—	469	786	890	2145	92%	92%
	Cohen (1994)	—	—	—	543	1590	2070	4203		
	Kirk (1996)	—	—	—	69	620	802	1491		
	Schmidt (1996)	—	—	—	186	576	447	1209		
	Harlow, Mulaik, and Steiger (1997)	—	—	—	39	374	287	700		
	Wilkinson et al. (1999)	—	—	—	12	1410	1590	3012		
	Total				4737					
2000–2009	Nickerson (2000)	—	—	—	—	321	582	903	92%	93%
	Gigerenzer (2004)	—	—	—	—	82	446	528		
	Kline (2004)	—	—	—	—	298	734	1032		
	Ziliak and McCloskey (2008)	—	—	—	—	75	681	756		
	Total					10,884				
2010–2017	Wasserstein and Lazar (2016)	—	—	—	—	—	762	762		
	Total						14,448	32,360		

NOTE: The figures listed in the last two columns show the percentage of empirical work using NHST. They are from Hubbard (2016, p. 233).

Review, Economic Journal, Journal of Political Economy, Quarterly Journal of Economics, Review of Economics and Statistics); Finance (*Journal of Finance, Journal of Financial Economics,* 1974, *Journal of Financial and Quantitative Analysis,* 1966, *Journal of Money, Credit and Banking,* 1969); Management (*Academy of Management Journal, Administrative Science Quarterly, Human Relations, Journal of Management,* 1975, *Journal of Management Studies,* 1964, *Organizational Behavior and Human Decision Processes,* 1966, *Strategic Management Journal,* 1980); and Marketing (*Journal of Consumer Research,* 1974, *Journal of Marketing, Journal of Marketing Research,* 1964).

All told, some 7762 empirical articles in the management sciences relied on NHST. Column 11 of Table 1 shows this as a percentage on a decade-by-decade basis.

3. Citation Analysis

Table 1 presents the Google Scholar citations earned by 25 works (18 articles, 7 books) from the dates of their initial publication through December 10, 2017. It reveals that these ostensibly influential publications highly critical of the uses and abuses of NHST were nevertheless unable to reduce its prevalence. On the contrary, every decade from 1960 to 2007 saw an *increase,* or no diminution, in its adoption.

3.1. Decade-by-Decade Citation Analysis

During the 1960s, when approximately 52%–56% of empirical studies in the management and social sciences employed NHST, four of the earliest and most damning indictments of how its use retards good science appeared in the literature. The articles by Rozeboom (1960), Bakan (1966), Meehl (1967), and Lykken (1968) ignited a debate that continues to this day. Between them they garnered 72 citations in this decade.

The beginning of the 1970s saw the publication of Morrison and Henkel's (1970) anthology, *The Significance Test Controversy*, which included the four articles mentioned above, draw further attention to this issue. Later that decade, Greenwald (1975), Carver (1978), and Meehl (1978) expressed concerns over the damaging effects of NHST. These four efforts attracted 191 citations for 1970–1979. The cumulative citations earned by the eight works listed in Table 1 for 1960–1979 now stood at a respectable 688. Yet 1970–1979 revealed a substantial increase over 1960–1969 in the proportion of empirical research using NHST in both the social (from 56% to 72%) and management (from 52% to 80%) sciences (see Table 1).

Six further prominent critiques of NHST appeared in the 1980s, most of them in the latter half. During this time they gathered 281 citations, with Leamer (1983), at 204, responsible for the lion's share of these. For 1980–1989, a total of 1603 citations accrued to the 14 contributions in Table 1 decrying NHST. Meanwhile, the cumulative citations of these reached 2291. Despite this, the percentage of empirical research based on NHST forged ahead to 84% in the social, and 89% in the management, sciences.

The 1990s witnessed the arrival of two juggernauts by Cohen (1990, 1994) on the damage to scientific progress caused by NHST. At 543, his 1994 article gained the most citations for this decade (or any before it), while his 1990 publication, at 469, took third place. Sandwiched between these, on 475, was Meehl (1978). A total of 4737 citations of research challenging the ubiquity of NHST occurred during 1990–1999. Cumulatively, this total rose to 7028. By now, almost predictably, the incidence of NHST featured in empirical research moved ahead in both the social, 92%, and management, 92%, sciences.

Four additional, noteworthy entrants lamenting the deleterious consequences of NHST for advancing good science debuted in 2000–2009. Consisting of two articles, Nickerson (2000) and Gigerenzer (2004), and two books, Kline (2004) and Ziliak and McCloskey (2008), these acquired some 776 citations during that time. Elsewhere for 2000–2009, the numbers increased dramatically. Weighing in with 1590, Cohen (1994) again procured the most citations for this decade, while Wilkinson et al. (1999), on 1410, claimed the second spot. Leamer (1983), at 903, captured third place. All told, the 24 contributions in Table 1 gained 10,884 citations for 2000–2009; and 17,912 cumulatively. At 92% and 93%, respectively, the proportion of empirical research relying on NHST (which data cover only the period 2000–2007) in the social and management sciences may (or may not) have plateaued.

The largest number of citations attracted by the 25 studies in Table 1 is for the (partial) decade 2010–2017. Over this period, these works were cited 14,448 times. Cohen (1994) and Cohen (1990) took the first and third places with 2070 and 890, in turn. Wilkinson et al. (1999), on 1590, are runners-up. A recent newcomer to this debate warrants special attention. This is Wasserstein and Lazar's (2016) "The ASA's Statement on *p*-Values: Context, Process, and Purpose." In less than 2 years this statement has gone viral, from a citations-generated perspective, with 762 of them. Incredibly, the cumulative total number of citations won by the 25 articles and books in Table 1 for the period 1960–2017 is 32,360.

In closing this section, it is instructive to note that over these same decades very few articles were published attempting to defend the practice of NHST. A notable exception is Hagen (1997). Another is Wainer (1999), although he could only muster "One cheer for null hypothesis significance testing." And while virtually all of the 25 works listed in Table 1 are severely critical of NHST and its baneful consequences for scientific progress, some, for example, Nickerson (2000, p. 241), remark "that when applied with good judgment it [NHST] can be an effective aid to the interpretation of experimental data." The only problem is that, in a 60-page journal article, Nickerson never provides us with examples showing how.

3.2. Further Insights from Table 1

Other patterns in the data in Table 1 deserve notice. The first is that some 14 of the 24 (58%) entries show monotonic increases in their citation rates over the decades (Wasserstein and Lazar, 2016, were excluded because there is no prior decade for comparison). This is especially impressive for the earlier-published articles, such as Rozeboom (1960), Lykken (1968), Carver (1978), and Meehl (1978).

Second, for 17 of the 24 (71%) entries, the period 2010–2017 is the one recording their maximum number of citations. This latest surge in numbers, which is evident also in the years preceding 2010–2017, possibly could reflect greater concern

with bad research practices. On the other hand, it is far more likely that this growth in citations is attributable to the profusion of journals, the development of more extensive and sophisticated citation-tracking systems, and manipulations to inflate them, over recent decades.

Reiterating from Section 3.1, across the period 1960–2017 the 25 works listed in Table 1 have amassed a total of 32,360 citations, or an average of 1294 apiece. These figures are staggering when it is acknowledged that most research goes uncited (Hubbard 2016, p. 238). The five most-cited pieces in Table 1 are Cohen (1994), 4203; Wilkinson et al. (1999), 3012; Leamer (1983), 2410; Cohen (1990), 2145; and Meehl (1978), 1794. Amazingly, the late Jacob Cohen authored 40% of these! Not surprisingly, given that they were among the earliest adopters—and subsequent critics—of NHST (Hubbard 2016, p. 21), 80% of this group are psychologists. These percentages are mirrored in Table 1 as a whole, with 19 of the 25 (76.0%) entries written by scholars trained in psychology, the remainder by economists (Leamer 1983; Ziliak and McCloskey 2008), sociologists (Morrison and Henkel 1970), statisticians (Berger and Sellke 1987; Wasserstein and Lazar 2016), and an interdisciplinary mix (Gigerenzer et al. 1989).

4. Discussion

Before dispensing advice to the scientific community on *p*-values and related matters, it appears that the statistics profession must first get aspects of its own house in order. This is because it is not just practitioners who can misinterpret *p*-values, but also statisticians (Hubbard 2016, pp. 207–208; McShane and Gal 2017). Indeed, Gelman and Carlin (2017, p. 900) sympathize with those who may question the merits of recommendations by statisticians to improve practice given "the mess we have helped to create." The proliferation of *p*-values is central to this mess. Here is Berry's (2017, p. 896) take: "We have saddled ourselves with perversions of logic—*p*-values … [which] are fundamentally un-understandable … We created a monster … The only reasonable route forward is to kill it." This sentiment is echoed, in more understated fashion, by Matthews (2017, p. 40). Or consider Briggs's (2017, p. 897) view that "There are no good reasons nor good ways to use *p*-values. They should be retired forthwith." It is in this context that I applaud the courage of David Trafimow to actually ban the use of *p*-values in *Basic and Applied Social Psychology*, a journal he edits (Trafimow and Marks 2015). This is a policy that Cumming (2014), whose article "The New Statistics: Why and How" has already gained an impressive 1180 Google Scholar citations, would surely approve.

The ASA statement on *p*-values (Wasserstein and Lazar 2016), of course, had to be of a general nature. Subsequent publications on the topic of the appropriate and inappropriate uses/interpretations of *p*-values, whether from the ASA or elsewhere, must be *specific*; the more specific the better. They must amount to a list of Do's and Don'ts concerning *p*-values. A good place to start would be for the ASA to articulate those circumstances, if they exist, in which use of NHST clearly is beneficial. At the same time this will serve to illustrate that its rank and file usage is little more than scientist window dressing.

5. Conclusions

This article has furnished overwhelming evidence supporting Berry's (2017) introductory quotation regarding the abject failure to rein in the use and abuse of NHST and *p*-values. With 32,360 citations between them—the kind of publicity only dreamed of in academic circles—25 publications severely critical of NHST have not been able to arrest, never mind reverse, its growing popularity in the social and management sciences. In fact, I would not be at all surprised to learn that its percentage usage in empirical work in these areas has actually inched ahead of the 92% and 93% highs for 2000–2007 in more recent years.

If the ASA's late intervention in this decades-long saga to improve statistical practice is to have any hope of success, its advice must be both specific and possibly radical. By specific I mean identifying those, if any, situations where *p*-values make a genuine contribution. As amply demonstrated, general statements have been spectacularly unsuccessful. By radical, I mean that the ASA might contemplate issuing a statement banning the use of *p*-values. Such actions will be difficult, but necessary, if the statistics profession is to be relevant in both the classroom, and in scientific method, in the 21st century.

References

Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 77, 423–427. [32,33]

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *p* Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [32,34]

Berry, D. (2017), "A *p*-Value to Die For," *Journal of the American Statistical Association*, 112, 895–897. [31,34]

Briggs, W. M. (2017), "The Substitute for *p*-Values," *Journal of the American Statistical Association*, 112, 897–898. [34]

Carver, R. P. (1978), "The Case Against Statistical Significance Testing," *Harvard Educational Review*, 48, 378–399. [32,33]

Cohen, J. (1990), "Things I Have Learned (So Far)," *American Psychologist*, 45, 1304–1312. [32,33,34]

——— (1994), "The Earth is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [32,33,34]

Cumming, G. (2014), "The New Statistics: Why and How," *Psychological Science*, 25, 7–29. [34]

Gelman, A., and Carlin, J. (2017), "Some Natural Solutions to the *p*-Value Problem—and Why They Won't Work," *Journal of the American Statistical Association*, 112, 899–901. [34]

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [32,33]

Gigerenzer, G., and Murray, D. J. (1987), *Cognition as Intuitive Statistics*, Hillsdale, NJ: Lawrence Erlbaum. [32]

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Krüger, L. (1989), *The Empire of Chance: How Probability Changed Science and Everyday Life*, Cambridge, UK: Cambridge University Press. [32,34]

Greenwald, A. G. (1975), "Consequences of Prejudice Against the Null Hypothesis," *Psychological Bulletin*, 82, 1–20. [32,33]

Hagen, R. L. (1997), "In Praise of the Null Hypothesis Statistical Test," *American Psychologist*, 52, 15–24. [33]

Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds.) (1997), *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum. [32]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: SAGE Publications, Inc. [31,34]

Kirk, R. (1996), "Practical Significance: A Concept Whose Time Has Come," *Educational and Psychological Measurement*, 56, 746–759. [32]

Kline, R. B. (2004), *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, Washington, DC: American Psychological Association. [32,33]

Leamer, E. E. (1983), "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, 31–43. [32,33,34]

Lykken, D. T. (1968), "Statistical Significance in Psychological Research," *Psychological Bulletin*, 70, 151–159. [32,33]

Matthews, R. (2017), "The ASA's *p*-Value Statement One Year On," *Significance*, April, 38–40. [34]

McShane, B. B., and Gal, D. (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112, 885–895. [34]

Meehl, P. E. (1967), "Theory-Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, 34, 103–115. [32,33]

——— (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," *Journal of Consulting and Clinical Psychology*, 46, 806–834. [32,33,34]

Morrison, D. E., and Henkel, R. E. (eds.) (1970), *The Significance Test Controversy: A Reader*, Chicago, IL: Aldine. [32,33,34]

Nickerson, R. S. (2000), "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, 5, 241–301. [32,33]

Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioural Sciences*, Chichester, UK: Wiley. [32]

Rosnow, R. L., and Rosenthal, R. (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44, 1276–1284. [32]

Rozeboom, W. W. (1960), "The Fallacy of the Null-Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428. [32,33]

Schmidt, F. L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers," *Psychological Methods*, 1, 115–129. [32]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [34]

Wainer, H. (1999), "One Cheer for Null Hypothesis Significance Testing," *Psychological Methods*, 4, 212–213. [33]

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. [31,33,34]

Wilkinson, L., and American Psychological Association Task Force on Statistical Inference (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist*, 54, 594–604. [33,34]

Ziliak, S. T., and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [32,33,34]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values

John L. Kmetz

To cite this article: John L. Kmetz (2019) Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values, The American Statistician, 73:sup1, 36-45, DOI: 10.1080/00031305.2018.1518271

To link to this article: https://doi.org/10.1080/00031305.2018.1518271

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 9853

View related articles

View Crossmark data

Citing articles: 7 View citing articles

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, 36–45: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518271

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values

John L. Kmetz*

Department of Business Administration, University of Delaware, Newark, DE

Abstract

p-Values and Null Hypothesis Significance Testing (NHST), combined with a large number of institutional factors, jointly define the Generally Accepted Soft Social Science Publishing Process (GASSSPP) that is now dominant in the social sciences and is increasingly used elsewhere. The case against NHST and the GASSSPP has been abundantly articulated over past decades, and yet it continues to spread, supported by a large number of self-reinforcing institutional processes. In this article, the author presents a number of steps that may be taken to counter the spread of this corruption that directly address the institutional forces, both as individuals and through collaborative efforts. While individual efforts are indispensable to this undertaking, the author argues that these alone cannot succeed unless the institutional forces are also addressed. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2018
Revised August 2018

KEYWORDS
Corrupt research; Generally Accepted Soft-Social-Science Publishing Process (GASSSPP); Peer review; *p*-Values; Reform; Replication.

1. Introduction

Nearly 60 years ago two reports from the Ford Foundation (Gordon and Howell 1959) and Carnegie Foundation (Pierson 1959) were published, both of which were highly critical of the unscientific nature of most business-school research and strongly recommended the adoption of more rigorous and scientific research methods in business education. Business schools responded by hiring many new faculty who came from the social sciences, and with them brought the research methods of the social sciences (Bass 1965; Webber et al. 1970). As a social science itself, the study of management and organization seemed a natural fit for the capabilities of these faculty. Unfortunately, these faculty were mostly trained during the era of rapid growth of empirical research using statistical methods, and chief among these was the use of Null Hypothesis Significance Testing (NHST) which is based on the demonstration of statistical significance; the reliance on NHST also expanded greatly in psychology between 1940 and 1955 (Hubbard, Parsa, and Luthy 1997; Hubbard and Ryan 2000) and has since become a "cult" (Ziliak and McCloskey 2008) and the basis for a huge body of corrupt research in business, management, and other social sciences (Hubbard 2016), which, in turn, inspired the title of this paper.

In this article, NHST is but the core issue in a complex of problems of research in the social sciences.[1] In the following pages, I will summarize what I refer to as the "mythology" of *p*-values, which have become the principal criterion for evaluating the outcomes of NHST, and summarize the issues involved with it. In Section 2, I develop a model of forces surrounding NHST and *p*-values that I refer to as the Generally Accepted Soft Social Science Publishing Process, or GASSSPP. Some find this term cumbersome, but it is an accurate portrayal of the corrupt research that constitutes so much of what we now see published (Hubbard 2016; Ioannidis 2005, 2017). It has its roots in the "soft" social sciences (Meehl 1978); is "generally accepted" in the same sense as the generally accepted rules of financial accounting, but like these, lacks a scientific basis for application; and it defines a self-reinforcing system or process by which these criteria are expected and enforced, primarily within the domain of academic publishing. In Section 3, I will suggest a number of steps that might be taken to remediate these problems, a number of which are already underway, but in other cases will be more challenging and take a longer time to accomplish.

The principal focus of this article is on the GASSSPP model detailed in Section 2. The are many articles, books, and book chapters enumerating the problems and issues of *p*-values, their misinterpretation, and the consequences.[2] What these works make clear is that there is an enormous problem with the dominant social-science methodology (refs. 56, 57, and 58 in

CONTACT John L. Kmetz *kmetz@udel.edu; johnkmetz@yahoo.com* Department of Business Administration, University of Delaware, Newark, DE 19711
Supplementary materials for this article are available online. Please go to *www.tanfonline.com/r/TAS*.
*J. L. K.: Retired.

[1] The literature on misinterpretation and misuse of *p*-values is quite large, and such misuse is but a part of the problem with corrupt research. In the interest of providing full scholarly support for these arguments without excessive demand for journal page space, I have provided only a select set of references at the end of this article, while providing full references in the supplement material. Wherever I refer to more literature than that immediately cited, I cite these as "refs" and provide the number(s) for the publication as enumerated in the supplement material, which provides the full citation(s) for the work(s).

[2] Refs. 7, 14, 24, 33, 37, 50, 51, 52, 59, 60, 61, 62, 63, 74, 80, 91, 95, 97, 102, 111, 117, 132, 142, 147, 148, 152 (supplementary material).

© 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.

supplementary material), sufficient to make it impractical to review all relevant literature in the space of an article.[3]

A p-value is simply the probability that in a research study the data would be at least as extreme as those observed, if the null hypothesis were true (i.e., the assumption that there is no true effect or true difference), referred to as statistical significance. But over decades of confusion, in part a result of academic debate between Neyman and Pearson (1933) and Fisher (1932), a mythology about p-values has become widely accepted. All of the following myths are untrue (Kmetz 2011, 2017): (1) p tells us the odds that our rejection of the null hypothesis is due to chance; (2) statistical significance establishes existence of a statistical effect; (3) $p < 0.05$ proves we have support for an hypothesis; (4) $p < 0.05$ is a "significant" outcome, $p < 0.01$ is "very significant," and $p < 0.001$ is "highly significant;" (5) p is the appropriate metric for those interested in theory development, and effect sizes matter only when practical application is the issue; (6) the p level indicates the likelihood that an outcome would not replicate if the study were repeated; (7) the p level predicts the number of statistical outcomes that would be significant by chance; (8) a null hypothesis is a scientific hypothesis; (9) rejecting a null hypothesis means the alternative is correct; (10) p is the same as α (Hubbard 2016); (11) in addition to these myths, NHST increasingly treats reliability as a substitute for validity (ref. 120 in supplementary material).

These mistakes have become so entrenched that a great many statistics and methodology texts make incorrect statements and claims regarding the meaning of p (refs. 52, 75, and 111 in supplementary material). It is arguable that a very large proportion of research practitioners do not know the correct interpretation of p, having been incorrectly taught in the age of statistical software that makes the mechanics of NHST quite simple but requires limited insight into interpretation (Vickers 2010).

One might think that in the face of demonstrated flaws the scientific community would immediately respond by taking action to correct them. In the case of the GASSSPP, this has not only not been the case, but the widespread adoption of these methods has spread beyond economics and the social sciences into biomedicine, genetics, and other fields where data require statistical analysis (refs. 64, 111, and 142 in supplementary material). In the specific case of the social sciences, the general reaction of the profession to having such problems called to researchers' attention has been to ignore the bad news.

It is my conviction that the long-standing resistance to methods change is only partially based on lack of dissemination or understanding of the information that justifies it. It is the force of a large number of interacting institutional factors which makes lasting change nearly impossible to achieve, and if we are to retain our reputations as credible scientists (Bedeian, Taylor, and Miller 2010), we must address these directly. Most of these forces are part of a self-reinforcing organizational structure (Kmetz 2012) and cannot be effectively countered by individual scientists acting alone. Sustained, concerted effort is required to effect change.

[3] Refs. 7, 8, 10, 11, 12, 15, 18, 22, 23, 26, 28, 29, 32, 34, 35, 38, 40, 41, 42, 51, 53, 54, 60, 61, 63, 65, 70, 71, 73, 79, 80, 84, 87, 88, 90, 91, 92, 94, 96, 100, 101, 102, 114, 115, 116, 120, 121, 123, 136, 138, 139, 140, 145, 152, among others—this list is not exhaustive (supplementary material).

In the remainder of this article, I will present a model of the institutional forces that I consider to be most significant to the GASSSPP and the continuing grip that it holds on our research and then propose steps that I believe can be taken to confront these forces by individuals and concerned groups of scientists. These proposals are based on two earlier articles (Kmetz 2011, 2012), revised, expanded, and significantly updated. Those actions which can be supported by organizations like the American Statistical Association (ASA) will be included as well.

2. Institutional Forces Comprising the GASSSPP

The problem of GASSSPP is not merely a function of adherence to outdated methods and criteria by an obstinate professorate. The institutions surrounding the research community are as much responsible for its persistence as individual researchers, and in the opinion of the author, even more so. These institutional forces are portrayed in Figure 1 and create and sustain what is effectively a self-reinforcing organization. It is this combination of forces, inclusive of and surrounding NHST, that fully defines the GASSSPP. As this figure shows, there are many individual forces at work in this structure; these may be grouped into four force clusters, each interacting with each other through the principal feedback loops shown. These clusters are (1) a Community of Misinterpretation in the center of the model, consisting of four myths about the meaning of p (specifically, that p means a statistical effect has been found; that it indicates the rarity of this finding; that it indicates the likelihood of replication of the finding; and that it indicates the probability of the hypothesis given the data, when it is actually the probability of the data given the hypothesis). (2) The Quality Delusion, based on journal and Social Science Citation Index (SSCI) and similar journal rankings; peer review, which is central to scientific legitimacy but gives the appearance of approval of the mythology of p (and thereby fails to correct it); and university and institutional accreditation practices which implicitly approve of the forces at work in the Quality Delusion. The Quality Delusion and the Community of Misinterpretation clusters contribute to the Illusion of Scientific Validity; this relationship is indicated by the heavy arrow linking them. (3) "Groupthink" properties (the label chosen for its descriptive value) consisting of four influences, which are dysfunctional uniqueness, that is, the expectation that studies should all be novel and not previously published, an apparent carryover from the writing of theses and dissertations; the lack of replication studies, also partly a product of dysfunctional uniqueness and the p mythology; and editorial practices, in particular the oft-noted editorial bias toward positive findings. These are reinforced by the "file drawer" problem wherein studies lacking statistically significant outcomes are not submitted to journals (Rosenthal 1979). (4) Finally, there is the Academic Reward System, consisting of both personal and professional recognition and career advancement, which surrounds and embraces all of the other factors. Many of these are mutually reinforcing, as shown by the two-way links between personal identification with the GASSSPP and professional recognition. Obtaining both the capabilities to do research and a position to pursue one's interests involves an

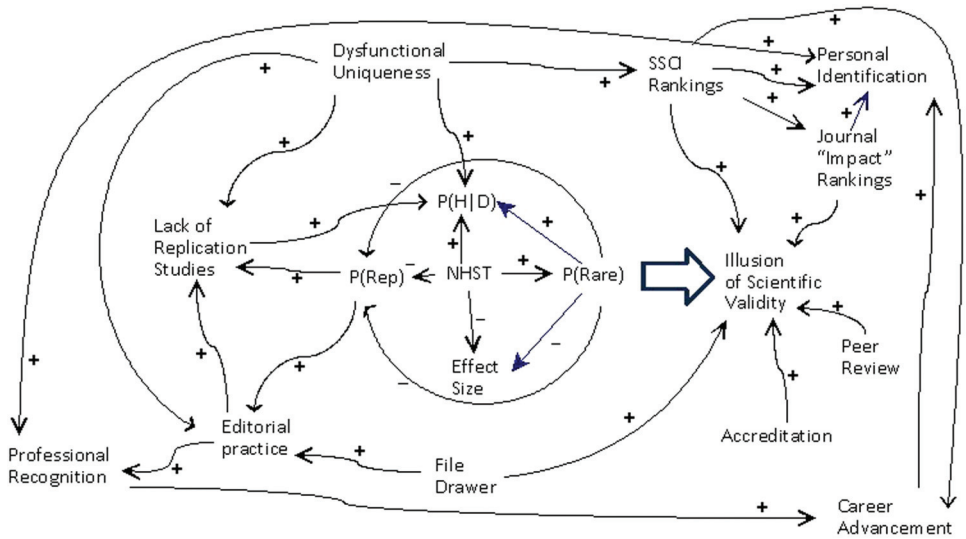


Figure 1. The Self-Reinforcing Structure of the Generally Accepted Soft Social Science Publishing Process (GASSSPP). This figure illustrates the author's conception of the primary institutional forces and their feedbacks which account for the persistence of Null Hypothesis Significance Testing (NHST) in the social sciences; both positive (+) and negative (−) forces contribute to reinforcement of NHST, which is in turn central to the GASSSPP. Source: Adapted from Kmetz (2012).

extensive socialization process, and thus those individuals who choose this path are likely a self-selected group who identify with the field.

Since our interest is in correction of corrupt research these forces will not be discussed in great detail here. I have attempted to be parsimonious in identifying forces in the model, in that nearly every individual force shown in it has been at least the subject of implicit contribution to the GASSSPP, and most could be expanded into a larger constellation of components.

When observing the large number of reinforcing and interacting forces at work in this model, those of us who have attempted to speak out against NHST and the GASSSPP should not be surprised at their persistence. A reversal of acceptance of NHST could call the competence of many researchers, reviewers, and editors into question; it could damage the reputations of institutions which have partly built their accreditation on publication of seemingly important findings; it would call into question the system of journal impact rankings that depend largely on citation and rejection rates, both perceived as indices of quality; and all of this could be disastrous to individual careers. If nothing else, wisdom in the face of these forces suggests that a strategy of confrontation is unlikely to succeed.

In addition, once published, research frequently takes on a "mantra of unassailability"—that which has survived the vetting of the GASSSPP is "truth" and cannot be questioned. Moreover, future work must accept this result and its conditions and base further investigation of the fundamental question on prior reported results. One cannot compare studies, cumulate studies, or estimate parameters from such a fragmented body of work; even in cases like the reified concept of "absorptive capacity," where Lane, Koka, and Pathak (2006) found 289 studies published between 1991 and 2002, mostly focused on R&D environments, they concluded (p. 858) that "the cross-citations between the articles in this body of literature show little evidence of an accumulated body of knowledge" (Kmetz 2011).

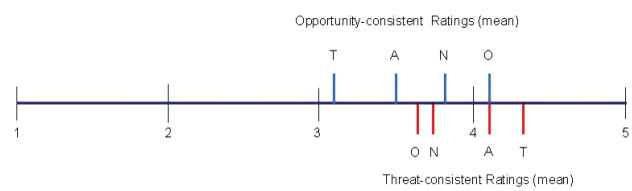


Figure 2. Plots of means in Jackson and Dutton (1988) study of discerning threats and opportunities in decision making. Letters indicate nature of information condition (manipulated script in text) embedded in stimulus booklet given to subjects (opportunity frame above line, threat frame below). Condition means are located to scale. T = Threat-distinctive text, O = Opportunity-distinctive text, A= Ambiguous text, and N = Neutral text. Source: Kmetz (2011).

As another example, Jackson and Dutton (1988) published a study in *Administrative Science Quarterly* which contends that subjects (employed MBA-program alumni) use entirely different information processing procedures for decision making when evaluating cues suggesting threats versus cues suggesting opportunities. Based on p-value differences in F tests, Jackson and Dutton concluded

> The results suggest, however, that threat and opportunity inferences cannot be accurately predicted from such a simple model of information processing. Instead, they indicate that managers are more sensitive to information that suggest the presence of a threat than they are to information that suggests the presence of an opportunity…(p. 384).

I used the authors' original study data to examine effect sizes by plotting the means for each of four cue types (threat, opportunity, ambiguous, and neutral) to scale, shown in Figure 2. As one can see, neutral ratings are nearly the same, opportunity and threat means are highest and lowest under their respective decision frames, and the order of the two cue sets is a mirror image of the other under each decision frame. This kind of symmetry suggests that managers in fact use a highly consistent, predictable form of information processing when making decisions.

Jackson and Dutton are both excellent scholars who did what everyone in the GASSSPP community considered correct practice. Having been published in what is arguably the most prestigious journal in management, their findings have been cited 984 times since 1990 (and four times in 2018 as of the date of this article); I have found no replications of the study, and the authors informed me that they are unaware of any (personal communications with Jackson and Dutton 2018). The number of times this study has been considered a definite, conclusive finding on decision making is unknown, but given the "voting" procedure common to most social-science research, it has probably gone largely, if not entirely, unquestioned. My analysis was nothing more than a reexamination of their effect sizes irrespective of p-values, but once published by reputable scholars following established procedures with no replications, findings become incontrovertible.

3. What Can Scientists Do?

3.1. Changes to the Environment of the GASSSPP Model

Since I first spoke out on this subject years ago the development of the internet and the social media it enables have dramatically altered the environment of science. In this section, I will discuss how both evolving groups and individuals can interact to promote movement beyond "business as usual" (Sohn 2000).

It is conceptually quite simple to correct the interpretational issues inherent to the GASSSPP mythology and move science away from reliance on statistical difference toward "statistical sameness," a status enjoyed in the early days of statistical methods prior to Fisher (Hubbard 2016, pp. 258–261). That is, if researchers focused their assessment of outcomes on effect sizes; used confidence intervals as the principal tool to assess research outcomes; and relegated p levels to the status of providing limited field-appropriate information about research outcomes, much of current GASSSPP misinterpretation would immediately change, of necessity. Indeed, some have questioned whether the null hypothesis is even necessary (refs. 12 and 45 in supplementary material), implying that p-values could become unneeded in much research; indeed, in 2015 *Basic and Applied Social Psychology* banned the use of p-values in its publications.

But as Hubbard (2016) and many others have noted, the reality of change is much more difficult to achieve than to conceptually envisage, and that is especially so because of the self-reinforcing structure that surrounds and supports the GASSSPP. Hubbard shows (2016, pp. 229–234) that there has been little reason to expect that publication of arguments against the GASSSPP through existing journals and books has had any impact on researcher practice; indeed, the only change seems to have been the opposite, that is, increased acceptance of the GASSSP over time. Particularly in the last decade, however, with the growth of the Internet and social media, scholars have greater access to academic journals and alternatives to them, as will be shown below. This also allows those with interests in reform a ready search capability and provides both the impetus and the medium for change, as well as increased public scrutiny of the problem external to the academic literature (refs. 2, 3, 125, and 126 in supplementary material). Early signs are encouraging.

Additionally, in 2016 the American Statistical Association (ASA) published its *Statement on Statistical Significance and P-Values* (Wasserstein and Lazar 2016). This "statement's six principles, many of which address misconceptions and misuse of the p-value," provides the support of an expert organization for researchers and reformers, it clarifies the interpretation of p-values, and it is supported by a brief bibliography of references on its correct interpretation. This bibliography is an excellent example of a general reference set on p-values which could be tailored and expanded to the needs of different researcher clusters (Ioannidis 2017).

3.1.1. Use the Internet and Social Media to Full Advantage

The internet has begun to emerge as a force to ease and improve the research process, and it can be a major resource to assist in overcoming the GASSSPP, a benefit which is already becoming evident. In addition, several internet-based projects backed by private foundations and organizations have been formed to promote the advancement and openness of science, among other objectives, and have been noticed by more general readership as well as professionals (Time's up 2017). Indeed, these are among a number of efforts and projects underway globally to speed all manner of scientific publication, and may be very helpful in ending the hegemony of the GASSSPP, given that lengthy review and revision processes also slow the publication of articles.

Thus, the internet leverages these changes in several ways. It would seem almost obvious today, but social media can direct readers to websites that explain and illustrate the correct interpretation of a p-value, can give tutorials on the interpretation of statistics (for example, see Geoff Cumming's *YouTube* series on "Significance Roulette"), and remain available indefinitely. There are other advantages. First, it is now possible to post all work done for a study, from initial design through the final publication. Second, review by both known colleagues and anonymous reviewers can be performed. Third, the growth of e-journal sites such as *PLOS ONE* (*http//journals.plos.org/plosone/*), with its objectives of making research freely available while maintaining standards of rigor and review, can hasten the reform of print journals; this is already underway with many journals issuing both hardcopy and electronic versions of content. Finally, internet sites can act as a repository for original study data and possibly replications (see Section 3.2). While some journals have attempted to enforce such requirements, few have fully achieved it; a new electronic journal or website is free to require such cooperation or at least enforce reporting requirements that would indicate when an author did not comply.

An underused, in my view, application of the internet is to serve as a training tool for doctoral candidates and emerging researchers. While this has been done privately by many for quite a while, there is no reason not to have doctoral students openly publish their research (other than dissertations, of course) while in their programs. This can have two major benefits to the profession: first, it socializes doctoral students to the world of science and prepares them to receive both accolades and criticism; second, it enables replication of existing research, which can be an excellent training tool with respect to both having the original study to serve as a benchmark, and to reinforce

healthy skepticism when many of those studies, particularly those done using *p*-values as criteria, do not replicate (Open Science Collaboration 2015).

Other, existing internet-based organizations like the Social Science Research Network (SSRN; *https://ssrn.com/en/*) can also be used to expedite making information available, without having to go through the gatekeeping process of peer review. There has long been need for a place to publish articles that resulted in negative outcomes—sound studies that did not find what was expected, have generally been rejected for publication for that reason, and end up in the "file drawer" (refs. 113 and 129 in supplementary material). A bias against negative findings is a problem in all of the sciences (refs. 65 and 85 in supplementary material), and SSRN provides a repository where such studies can be made public. The creation of PsychFileDrawer (*www.psychfiledrawer.org/*) to encourage replication and discussion of widely cited articles in experimental psychology, and more recently The Replication Network (*https://replicationnetwork.com/*) for economics, are examples of these budding efforts.

The past five years suggest there may be dramatically improved chances for similar initiatives. In the past few years, two sites, Figshare, and Dryad (*http//www.datadryad.org*), have been formed to store raw data and allow free access to datasets and articles in a wide array of scientific disciplines. During the past four years, the Center for Open Science has grown from a project concerned with the reproducibility of psychological research into an organization covering ten broad disciplines, oriented primarily toward academics who share the philosophy of freely accessible research, and to serve as a single source for all such projects, including not only articles but methodology, software, and networking for researchers with similar values. Like Nelson, Simmons, and Simonshon (2018), a few years ago I hardly dared to hope such resources would exist, and it is greatly encouraging to see them.

I think it is also wise to include a note of caution. All of these groups are voluntary, and for the most part depend on donations and in-kind support, primarily in the form of individual efforts for their survival. They typically are not supported by a permanent organization nor included in the "permanent" parent structure of one; such organizations can wither and die, and there is no guarantee this will not happen. I would, therefore, recommend that reform organizations should strongly consider maintaining linkages with established organizations. In business and management, the Academy of Management has long offered consortiums to assist member universities and faculties in the development of future researchers. Given the extent of membership overlap between many of these organizations, the Academy might well offer such consortiums with the intention of reforming and improving research practices across organizations. Even when recognized institutions have supported reform, there has been no assurance of success—one only needs to observe the limited effect of the American Psychological Association's Task Force on Statistical Inference (Wilkinson 1999), or the efforts of the American Assembly of Collegiate Schools of Business International (refs. 4, 5, and 6 in supplementary material) to improve the impact and relevance of their members' research to see this. The Academy of Management has attempted to allow those who contend that the NHST is ineffective to influence its members (see, e.g., Starbuck 2008), but little effect has been evident thus far. While I am optimistic that this time we have a different environment and more momentum for change, I do not take these initiatives for granted.

3.1.2. Form Ties to the Practitioner Community

Within the social sciences, there is considerable variability in the extent to which researchers are linked to practitioners in their fields. But as observed by Hubbard (2016) and numerous other scholars over past decades (refs. 4, 49, 72, and 141 in supplementary material), ties to the business community, management practitioners, and to the practicing social science community similar to the professional relationships found in medicine, engineering, law, and other professional disciplines are largely lacking. Many GASSSPP scholars take the position that research collaboration between academics and practitioners is neither necessary nor desirable, and this general sentiment has increased over time (Hubbard 2016, pp. 241–246). The problem seems to be spreading—for example, recent research indicates almost no relationship between education and research on education (Makel et al. 2016); personnel and human resource management (HRM) have long decried the absence of connections between research production and research consumption (Rynes, Giluk, and Brown 2007); one study found an almost total disregard of management research on the part of best-selling business book authors, nearly half of whom were themselves academics (Kmetz 2016).

The different "business models" for research provided by the internet opens an excellent opportunity for scholars, their universities, and administrators alike. Any or all of the relatively new reform websites should solicit ideas for novel but ambitious projects and researchers, in addition to obviously needed replication studies. Given the proliferation of unrelated articles generated in large part by the dysfunctional uniqueness expectation of the GASSSPP, there exists the potential to solicit projects that could answer fundamental questions.

Collaboration of this kind may seem unlikely, and will take time and persistence to develop, but one of the underexploited paths to such collaboration has been illustrated by the McKinsey consulting firm and the London School of Economics, who collaborated to study the impact of management methods between the U.S., U.K., France, and Germany (refs. 25 and 89 in supplementary material). The second, most comprehensive, stage of their study involved 731 companies, none of whom were McKinsey clients, several McKinsey researchers, and a group from the London School of Economics. The title of the McKinsey study, "Management Matters," basically expresses what was found. Interestingly, while some minor attention was given to statistical significance and *p*-values in the academic version of the article, the final McKinsey report relied entirely on graphical and tabular demonstrations of effect sizes on the dependent variables. The study itself is an excellent illustration of the payoff of such collaborative efforts, but has garnered little attention in the academic world.

3.1.3. Sponsor and Participate in Conferences Requiring Improved Science

Conferences and research forums such as the October, 2017, ASA Symposium on *p*-values are a hallmark of academic

research, and our disciplines have seen remarkable global growth in the number of such conferences. Following the 2015 lead of *Basic and Applied Social Psychology*, preparation of a short "procedure and style guide" which makes it clear that standard GASSSPP research is ineligible for consideration would be appropriate and might well be an effort ASA should undertake following the publication of its *Statement*. Specialized or irregular conferences can now be organized virtually for any reasons researchers find helpful.

3.2. Changes Internal to the GASSSPP Model

I will next turn attention to the factors comprising the model itself, beginning with discussion of the "Groupthink" factors on the left side. These are more subject to short-term change, and there is already some evidence of this.

3.2.1. Perform Replication Studies

In speaking directly to the GASSSPP forces, much of the recent growth in the groups mentioned above has come from increasing recognition of the need for replication studies in social science research (refs. 86, 103, and 104 in supplementary material) and other science fields (Baker 2016), which many have called for over decades (e.g., refs. 27 and 81 in supplementary material), and for a brief time it appeared that some limited progress had been made with respect to this need. The formation of groups such as the Reproducibility Project at the Open Science Foundation is intended to encourage replication studies and serve as a repository for them, among other objectives. Within the past two years an additional step in this direction has been taken for economics by The Replication Network. Progress has been erratic (Hubbard 2016), but the recent surge of attention to this problem gives hope that replication studies may finally find their place in the published scientific literature (Nelson, Simmons, and Simonshon 2018).

Replication studies done by students may be an excellent training tool for student researchers, as noted—they would have the original study to serve as a comparison benchmark, and learn healthy skepticism when many of those studies do not replicate. When published datasets are available the projects can serve as practice problems, and both datasets and replications can be published into repositories maintained online by the groups, a feature several groups have begun. In addition, it should be a requirement that all replication studies will be made available on these sites, along with source code for the analytical program used; this requirement must be strictly enforced, given the past failure of similar efforts on the part of several journals and the reluctance of some to require publication of code and datasets (Young 2017).

Given the current promotion and tenure practices at most universities and similar practices at other institutions, it is unlikely that promotion will be gained doing replications alone. To overcome researcher reluctance to perform replication studies, these groups can announce that frequently cited articles will be replicated by the groups; these announcements should take the form of regularly published updates on the group website. This practice can provide a needed alternative to the many quantitative indices of research quality which factor heavily into promotion and tenure decisions. Obviously, if inappropriate decision rules were applied, this information will be published with comments on the article. This should be done with a right for authors of articles published prior to group formation to decline to have their studies replicated; however, all studies published after the specified date would be eligible for replication.

An additional benefit of replication studies may be reduction of the dysfunctional uniqueness problem mentioned earlier. By definition, replication studies are not unique, nor are they expected to be. As replications become more acceptable as evidence of improved science, there may well be at least two benefits to research in general. One is that whenever a study is published, it is more likely that it will replicated if it is considered noteworthy, and that is an incentive to examine its initial outcomes closely before publishing them; the second is that it will be possible to cumulate the results of future studies more effectively, a property of science that is distinctly not compatible with dysfunctional uniqueness.

Most of the reform websites feature a weblog or rely on *Twitter* accounts to promote and store commentary on not only research, but to generally open lines of communication to interested parties. Part of the success of replication work will depend on regular monitoring and commentary on studies published on these media, especially since exact replications or reproductions of those studies may not be possible (Goodman, Fanelli, and Ioannidis 2016). These sites create centralized locations to store both the replication studies and reviews, as well as original study data. What has also begun to change with the reform websites is the ability to provide training materials other than methods textbooks and work with actual raw data and exploration of replicability. In addition, publication of the R statistics package enables students everywhere to obtain and apply software at no cost. Those individual weblogs that exist now tend to be on the periphery of the profession (or are perceived so), and are frequently more difficult to find, problems that the centralized reform sites and commercial media help to overcome.

3.2.2. Challenge Editorial Practices

Nearly every academic article begins with a review of relevant literature, especially in the social sciences and management. Challenge and questioning of prior work are practices that are fundamental to real science, but are notable for their absence in the GASSSPP. The use of p levels as the primary criterion of merit in research exacerbates this problem, and given that the "voting" methods used in reviews to summarize previous literature on a research question yields incorrect and misleading results (refs. 35, 60, 61, 90, and 93 in supplementary material), authors have every reason to challenge reviewers claims of "methodological or statistical weaknesses" that are frequently cited as reason for rejection of submissions. However, there is also reason to question and challenge that which has been published.

An important part of the challenge process can be through peer review, which nearly every published author is asked to do. Peer review will be discussed in more detail below, but reviewer contentions of errors in research designs, outcomes, and conclusions drawn from them can now be partially countered by the ASA statement on p-values and perhaps a standard reference

set such as that accompanying the *Statement*, detailing problems with the GASSSPP mythology; this can enable authors and reviewers to exert gentle but anonymous pressure for constructive change on other editors and authors alike, to their benefit and that of science.

If reform groups were to recognize those journals and editors who applied tools and methods consistent with their standards, powerful motivations would be unleashed to emulate them on the part of both editors and contributors. Praise and recognition are far more powerful than negativity and attack, and other editors (and reviewers) would be likely to seek such external recognition for their work (Review and prosper 2017). Editing is a difficult and often thankless task, and for an editor to be given recognition for an exceptional job is an important source of personal job satisfaction as well as professional prestige.

3.2.3. Revise Accreditation Standards

The cluster of forces on the right side of Figure 1 (to the right of the heavy arrow) are all so interdependent that several reform efforts will be needed to move them, and this will take some time to achieve. Accreditation is considered to bestow scientific and educational legitimacy to research institutions. While one or more bodies will generally accredit an institution, the most significant accreditation is awarded by specific accrediting bodies and procedures using peer review, such as engineering, nursing, and the like. In the realm of business schools, AACSB is considered the most prestigious U.S. accreditor; it has called for increased "impact" of research as part of its accreditation standards, as noted above. If reform groups were to make explicit statements of support for such demonstration of impact and on-going contact with professional colleagues and publicize their benefits to both academics and practitioners, these could reinforce the efforts of accreditors to recognize research with more validity and impact. Permanent, continuous liaison with AACSB, the APA Commission on Accreditation, and other recognized accrediting bodies should be established to support efforts to improve the quality and demonstrate the impact of scholarly research.

Similar expansion of scope can apply to accreditation standards. These are typically slow to change since they are subject to multiple levels of articulation and review, before eventual adoption. Again, to use AACSB and APA as examples, if these organizations were to provide a service similar to that of the Marketing Science Institute by soliciting research needs, suggesting specific (and often long-term and risky) projects, and adding recognition of such projects to their definition of university missions, new avenues to accreditation would be opened for the universities, and more diverse, but still mission-consistent paths for faculty promotion and tenure. Similarly, if these were to endorse the ASA's *Statement on p-Values* and include the expectation that institutional research should recognize and adhere to the six principles in the *Statement* in reporting results, much of the onus for change would be lifted from their clients.

Although not accrediting agencies in their own right, several large U.S. government organizations have so much implicit influence over research standards that relationships should also be established between them and reform organizations. For example, accrediting organizations must comply with regulations from the Department of Education; in addition, Education, Defense, and Health and Human Services all provide significant funding for the social sciences; the National Institutes of Health and Mental Health all fund biomedical and pharmaceutical projects; the Environmental Protection Agency funds environmental science, and so on; relationships to speak for improved statistical science in government-funded projects should be promoted through these agencies by reform organizations.

3.2.4. Improve Peer Review

In Figure 1, peer review is shown as contributing directly to the illusion of scientific validity, but of course it contributes to many other forces in the model and is fundamental to NHST and the GASSSPP. An unfortunate property of the GASSSPP is that the "peer review" process is not really peer review per se, but rather a process of "serial editorship" that requires authors to respond to different reviewer expectations and places authors in the role of being supplicants before superiors who have published before (Starbuck 2003). There is long-standing evidence of reviewer bias against null outcomes (refs. 13, 30, and 48 in supplementary material), and recognition that peer review is really a means of reinforcing professional norms (Bedeian 2003, 2004); however, peer review does not assure quality in complex multivariate problems like social and biological sciences (refs. 68, 69, and 144 in supplementary material). Nevertheless, it is considered the *sine qua non* of scientific quality.

These properties of peer review make it difficult to change, but they also create various points of leverage to improve its outcomes. Since peer review is almost always anonymous, the identities and reputations of reviewers are protected. Peer reviewers thus have the freedom to point out the incorrect interpretation of p-values when they arise, and by referring to the ASA *Statement* and perhaps supplemental reference material appropriate to their field (Ioannidis 2017), both authors and editors would be given incentive to change away from existing GASSSPP practices.

Editors are under great stress from their reviewers and frequently are unwilling to challenge their reviews even when shown to be incorrect. I strongly suspect that many editors would welcome having the support of a multidisciplinary group, and to the extent that those of us opposed to the GASSSPP have potential supporters among journal editors, there is an opportunity to provide such support. Reinhart (2015) tells the story of Kenneth Rothman, associate editor of the *American Journal of Public Health,* who took a strong stand against reporting p-values and the discussion of statistical significance in the mid-1980s. Both types of reporting fell dramatically while he was editor, although many subsequent contributors resumed their submissions of p-values as had been the case prior to Rothman's editorship. Rothman later founded the journal *Epidemiology*, again with strong standards, and after about 10 years was successful in establishing the use of confidence intervals and the nearly total exclusion of significance tests and p. Such editors need to be encouraged, recognized, and rewarded.

In my field, a relatively new organization called Responsible Research in Business and Management (RRBM; *https://www.facebook.com/RRBMnetwork/*) has formed and, as the name

suggests, is motivated to improve the quality and contribution of research in the field. As stated on their homepage

> The core vision of RRBM is that "business can be a means for a better world if it is informed by responsible research.…If nothing is done, business research will lose its legitimacy at best; at worst, it will waste money, talent, and opportunity".

Given that the ASA *Statement* has taken steps to "usher in a post-$p < 0.05$ era," the reviewers who have been relegated to methodological obscurity for their opposition to p now have several umbrella organizations for support; for example, while its primary focus is on biomedical research, PubPeer (*https://pubpeer.com/static/about*) has created a forum for anonymous review of articles. This group would also create an opportunity to set limited standards and establish "best practices" for peer reviewers; these could include a field-specific standard reference set as suggested above to provide authors and reviewers with a well-established case against reliance on p-values.

We might even promote the practice of publishing detailed peer-reviewed research designs in journals, with a guarantee that results will be published (along with an account of what differed from the original design) on the website, and a brief summary of the outcomes in the journal. This would reduce demand for space in journals and have the additional benefit of freeing peer reviewers from their often thankless present jobs and instead allow them to review research designs for their potential to provide useful knowledge. This approach to review would enable researchers and reviewers to break with present GASSSPP traditions regardless of whether they applied them in the past.

3.2.5. Rankings and Citations

A major currency in academic reputation is one's citation count, and in addition, journals compete to achieve high rankings through citation counts, such as the Social Science Citation Index (SSCI). This has both positive and negative implications for the ability to reform our research, but concerted efforts on the part of reformers can emphasize the positive effects. As noted, the primary reaction of the profession to those of us who demonstrate the misinterpretation of p-values is to ignore such information. This not only legitimates ignoring such work by other authors, but is seen as drawing into question the validity of the basic arguments against these misinterpretations. Thus, reformers can bolster their position by frequently (and regularly) citing those who support alternatives. The 2016 *Statement* by ASA provides a very important form of institutional support for those who submit articles using alternative forms of analysis and interpretation. Whenever appropriate, article sections discussing methods should cite a substantial number of authors and articles opposing the GASSSPP. To the credit of the profession, exchanges and disagreements over appropriate methods and interpretations of results are already well-established.

Some may find such an argument to be little more than "gaming the system," but in the author's view it is not. In view of the explosion of research publications in total, and the proportion of them mistakenly interpreting p-values, such support is necessary for correct alternatives to have any chance of advancement; it is worth pointing out that by many accounts (refs. 52, 64, 65, and 152 in supplementary material) the profession has failed to progress over past years. This serves the long-term interests of no one. Visibility of reformers position requires citation, plain and simple.

3.2.6. Career Advancement

Perhaps the most difficult and longest-term change to effect will be to standards for career advancement. As AACSB (2008, p. 28) noted 10 years ago

> The predominant model for faculty support found in business schools today focuses primarily on systems that reward excellence in scholarship and teaching with tenure and other forms of security and compensation. Promotions, especially to full professor, tend to be based on academic contributions and reputation largely determined by success in publishing in the most respected peer-reviewed discipline-based journals.

It is evident that there is enormous pressure to publish, but whether this is beneficial to science is a very open question. Among the problems attributable to excess pressure are the lack of replications (Baker 2016), plus faulty and incomplete citation of sources (Wright and Armstrong 2008) and a tendency to cite highly ranked journal articles on the assumption that ranking connotes quality even when that is questionable (refs. 9, 105, 122, 124, and 131 in supplementary material).

These and other issues are properly the subject of book-length discussions, and what will be suggested here will surely be controversial to some. It has been said that "quantity times quality is a constant" in the world of academic research. Whether true or not, one way to evaluate applications for promotion and tenure in the academic world is to allow only three publications to be used in the application (Wachtel 1980), and further, ensure that these are totally blind-reviewed. While there are cases where widely cited articles may be difficult to review, the experience of Peters and Ceci (1982) suggests otherwise—they found that of 12 articles resubmitted to the journals that had originally published them, under changed (fictitious) author names and unknown institutional affiliations, and within 18–32 months of their original publication, only 3 of the 12 articles were detected. Eight of the nine undetected articles were rejected for publication, with reviewers citing "serious methodological flaws." Limited-publication procedures would allow applicants to select their work for quality evaluation, and these quality ratings could be the basis for future applicants; these could be modified to suit varying requirements in different disciplines and in different institutions. For example, where researchers are responsible for generating their own funding, success at writing grant proposals could be added to research evaluations or partially substituted for them. Other criteria, which most university faculty on both sides of the issue would likely prefer to remain internal to the application process, could continue to be evaluated by existing criteria.

Other modifications in promotion procedures are possible. Coupled with the recent interest in research reproducibility, an investigator could be evaluated positively for taking a role in a lengthy replication project, and the value of such research work made explicit through changes in promotion criteria reflecting

that. Similarly, taking a role as a team member engaged in a risky, multiyear, multi-institutional effort could become a criterion in research quality. At present, the publish-or-perish regime that has become the norm seems impossible to sustain. It forces many potentially excellent scientists to sell out to a system that rewards short-term individual work rather than more sustained and interesting projects; whether this will be tolerated in the face of rising costs for education and competition from scientific efforts with better-established credentials is highly uncertain and risky in its own right.

4. Conclusion

Statistics is an indispensable tool to understand complex systems where many variables can influence diverse outcomes. But it has been heavily corrupted by NHST and the reinforcing properties of the GASSSPP through which research results are published. This article has been an attempt to recommend several concrete steps to correct this corrupt research and change the forces that maintain it. The proposals made here are relatively embryonic and may undoubtedly be improved by the contributions of others in the field. Most of these proposals address the institutional factors that surround individual researchers, and these are admittedly more difficult to change because they are part of a self-reinforcing web. I have focused on several of these that might provide some leverage in the near term, as well as those that might change in the longer term with patience and persistence. In all cases, I have tried to keep these as simple and nonthreatening as possible—simple because power comes from simplicity, not complexity; and nonthreatening because if we have learned anything in the past 50 years, it is that frontal attack on the GASSSPP will not work. These efforts will take time to become manifest, but I am optimistic that with patience, we may yet achieve the "butterfly effect" where small initial events may culminate in major impact. As I noted, at least one journal editor has succeeded in nearly eliminating *p*-values from its publications, a number of reformer internet sites have been formed and have influence and a level of visibility unimaginable a decade ago, and we have held a symposium specifically formed to address the need for such change. Perhaps this is where the butterfly begins to emerge from its chrysalis.

Supplementary Materials

The online supplementary materials contain additional references for the article.

References

(2017), "Review and prosper," *The Economist*, 423, 99. [42]

(2017), "Time's up," *The Economist*, 422, 69–71. [39]

AACSB International (2008), "Final Report of the AACSB International Impact of Research Task Force," St. Louis, MO: AACSB International. [43]

Baker, M. (2016), "Is There a Reproducibility Crisis?" *Nature*, 533, 452–454. [41,43]

Bass, B. M. (1965), "The Psychologist and the Business School," *The Industrial Psychologist*, 2, 14–18. [36]

Bedeian, A. G. (2003), "The Manuscript Review Process: The Proper Role of Authors, Referees, and Editors," *Journal of Management Inquiry*, 12, 331–338. [42]

——— (2004), "Peer Review and the Social Construction of Knowledge in the Management Discipline," *Academy of Management Learning & Education*, 3, 198–216. [42]

Bedeian, A. G., Taylor, S. G., and Miller, A. N. (2010), "Management Science on the Credibility Bubble: Cardinal Sins and Various Misdemeanors," *Academy of Management Learning & Education*, 9, 715–725. [37]

Fisher, R. A. (1932), *Statistical Methods for Research Workers*, Oxford, UK: Oliver & Boyd. [37]

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016), "What Does Research Reproducibility Mean?" *Science Translational Medicine*, 341, 1–6. [41]

Gordon, R. A. and Howell, J. E. (1959), *Higher Education for Business*, New York: Columbia University Press. [36]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Los Angeles: Sage. [36,37,39,40,41]

Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997), "The Spread of Statistical Significance Testing in Psychology: The Case of the Journal of Applied Psychology," *Theory & Psychology*, 7, 545–554. [36]

Hubbard, R. and Ryan, P. A. (2000), "The Historical Growth of Statistical Significance Testing in Psychology—and its Future Prospects," *Educational and Psychological Measurement*, 60, 661–681. [36]

Ioannidis, J. P. A. (2005), "Why Most Published Research Findings Are False." *PLOS Medicine*, 2, e124. [36]

——— (2017), "What Have We (not) Learnt From Millions of Scientific Papers With *p*-Values?," Presentation to the American Statistical Association, Bethesda MD. [36,39,42]

Jackson, S. E. and Dutton, J. E. (1988), "Discerning Threats and Opportunities," *Administrative Science Quarterly*, 33, 370–387. [38]

Kmetz, J. L. (2011), "Fifty Lost Years: Why International Business Scholars Must not Emulate the US Social-Science Research Model," *World Journal of Management*, 3, 172–200. [37,38]

——— (2012), "Self-Reinforcement and Negative Entropy: The Black Hole Of Business-School Research," in *28th European Group for Organization Studies (EGOS) Colloquium, Self-reinforcing Processes in Organizations, Networks and Professions*, Helsinki, Finland, available at *https://sites.edul.edu/mjs*. [37,38]

——— (2016), "Neither 'Food Chain' nor 'Translation Problem'? The Disregard of Academic Research in Best-Selling Business Books," *International Journal of Business Administration*, 7, 101–122. [40]

——— (2017), *Management Junk Science*, available at *https://sites.udel.edu/mjs*. [37]

Lane, P. J., Koka, B. R., and Pathak, S. (2006), "The Reification of Absorptive Capacity: A Critical Review and Rejuvenation of the Construct," *Academy of Management Review*, 31, 833–863. [38]

Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonson, B., and Coyne, M. (2016), "Replication of Special Education Research: Necessary but far too Rare," *Remedial and Special Education*, 37, 205–121. [40]

Meehl, P. E. (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," *Journal of Consulting and Clinical Psychology*, 46, 806–834. [36]

Nelson, L. D., Simmons, J., and Simonshon, U. (2018), "Psychology's Renaissance," *Annual Review of Psychology* 69, 511–534. [40,41]

Neyman, J, and Pearson, E. S. (1933), "On the Problem of the Most Efficient Test of Statistical Hypotheses," *Philosophical Transactions of the Royal Society*, Series A, 231, 289–337. [37]

Open Science Collaboration. (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, 4716. 40

Peters, D. and Ceci, S. J. (1982), "Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again," The *Behavioral and Brain Sciences*, 5, 187–195. [43]

Pierson, F. C. (1959), *The education of American Businessmen: A Study of University-College Programs in Business Administration*, New York: McGraw-Hill. [36]

Reinhart, A. (2015), *Statistics Done Wrong: The Woefully Complete Guide*, San Francisco, CA: No Starch Press. [42]

Rynes, S. L., Giluk, T. L., and Brown, K. G. (2007), "The Very Separate Worlds of Academic and Practitioner Periodicals in Human Resource Management: Implications for Evidence-Based Management," *Academy of Management Journal*, 50, 987–1008. [40]

Sohn, D. (2000), "Significance Testing and the Science," *American Psychologist*, 55, 964–965. [39]

Starbuck, W. H. (2003), "Turning Lemons Into Lemonade: Where is the Value in Peer Reviews?," *Journal of Management Inquiry*, 12, 344–351. [42]

——— (2008), "Croquet with the Queen of Hearts," Presentation to the National Academy of Management, Anaheim, CA. [40]

Vickers, A. (2010), *What is a p-Value Anyway?: 34 Stories to Help You Actually Understand Statistics*, Boston, MA: Addision-Wesley. [37]

Wachtel, P. L. (1980), "Investigation and its Discontents," *American Psychologist*, 35, 399–408. [43]

Wasserstein, R. L. and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," 70, 129–133. [39]

Webber, R. A., Ference, T. P., Fox, W. M., Miles, R. E., and Porter, L. W. (1970), "Behavioral Science and Management: Why the Troubled Marriage?" in eds. T. J. Atchison, and J. V. Ghorpade, *Academy of Management Proceedings: Annual Meeting*, 377–395. [36]

Wilkinson, L., and the Task Force on Statistical Inference (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54, 594–604. [40]

Wright, M. and Armstrong, J. S. (2008), "The Ombudsman: Verification of Citations: Fawlty Towers of Knowledge?" *Interfaces*, 38, 125–139. [43]

Young, C. (2017), "'See for Yourself': The Pleasures and Sorrows of Transparency in Social Science Research," Paper presented to the American Statistical Association Symposium on Statistical Inference, October 17, Bethesda, MD. [41]

Ziliak, S. T. and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [36]

The American Statistician

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

Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness, and Relevance

Douglas W. Hubbard & Alicia L. Carriquiry

To cite this article: Douglas W. Hubbard & Alicia L. Carriquiry (2019) Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness, and Relevance, The American Statistician, 73:sup1, 46-55, DOI: 10.1080/00031305.2018.1543138

To link to this article: https://doi.org/10.1080/00031305.2018.1543138

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 9750

View related articles

View Crossmark data

Citing articles: 4 View citing articles

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, 46–55: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543138

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness, and Relevance

Douglas W. Hubbard[a] and Alicia L. Carriquiry[b]

[a]Hubbard Decision Research, Glen Ellyn, IL; [b]Department of Statistics, Iowa State University, Ames, IA

ABSTRACT

Efforts to address a reproducibility crisis have generated several valid proposals for improving the quality of scientific research. We argue there is also need to address the separate but related issues of *relevance* and *responsiveness*. To address relevance, researchers must produce what decision makers actually need to inform investments and public policy—that is, the probability that a claim is true or the probability distribution of an effect size given the data. The term responsiveness refers to the irregularity and delay in which issues about the quality of research are brought to light. Instead of relying on the good fortune that some motivated researchers will periodically conduct efforts to reveal potential shortcomings of published research, we could establish a continuous quality-control process for scientific research itself. Quality metrics could be designed through the application of this statistical process control for the research enterprise. We argue that one quality control metric—the probability that a research hypothesis is true—is required to address at least relevance and may also be part of the solution for improving responsiveness and reproducibility. This article proposes a "straw man" solution which could be the basis of implementing these improvements. As part of this solution, we propose one way to "bootstrap" priors. The processes required for improving reproducibility and relevance can also be part of a comprehensive statistical quality control for science itself by making continuously monitored metrics about the scientific performance of a field of research.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
p-Values; Process control; Relevance; Replication; Reproducibility

1. Introduction

The failure of researchers to reproduce statistical results in many different disciplines has led to what we know as the *reproducibility crisis*.[1] The reasons for the difficulty to reproduce results include sample deficiencies and low power, incorrect application and interpretation of statistical tests and misalignment of data with research questions (Stodden 2015; Hubbard 2016). The reluctance of researchers to share data and code contributes to the reproducibility crisis as well, since it is often difficult to reproduce the exact statistical analysis without access to the software that was used to produce the original results (e.g., Stodden 2013). This is not a new problem; as early as 1935, Fisher mentioned the need for replication for scientific validity, an idea repeated in Popper (1959). Sterling (1959), Rosenthal (1979), and Sackett (1979) discussed the effect of publication bias on reproducibility in various scientific disciplines over 50 years ago. Dickersin (1990) revisited the question of publication bias in the context of medical research. Buckheit and Donoho (1995) pioneered the concept of open-access software and datasets to address the issue of reproducibility, using as an example a suite of programs developed by them to analyze images. Casadevall and Fang (2010) present a good discussion about reproducible science.

In recent years, the pace at which articles in the peer-reviewed literature and in the popular literature have been published has increased tremendously. Even after warnings as early as Fisher's, decades pass as evidence mounts with no fundamental change in how statistical inference methods are applied.

The *Manifesto for Reproducible Science* (Munafó et al. 2017) thoroughly describes the problem with replication in certain areas of science and it brings together multiple valid suggestions from many authors for making improvements. Among other suggestions, the manifesto promotes research registries, better statistical training and transparency of data to reduce problems like publication bias, p-hacking, and HARKING. Publication bias occurs when the chance that the work is published depends not only on its quality but also on the outcome of the research. Therefore, projects that have a statistically significant result have a larger probability of being submitted and published (e.g., Dickersin 1990). The term p-hacking is used to describe the situation where a researcher tests all possible hypotheses or uses the subset of the data to obtain statistically significant results (Goodman 1999). HARKing stands for "hypothesis after results are known" and is related to the idea of p-hacking. A good reference is Kerr (1998). Munafó et al. (2017) do not

CONTACT Alicia L. Carriquiry alicia@iastate.edu Department of Statistics, Iowa State University, Ames, IA 50011.

[1] We recognize the debate about the use of "reproduction" versus "replication" in this context. We will use the former term unless we refer to the work of earlier authors who used the latter. In that case, we will use their terminology to avoid confusion.

© 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.

propose any significant change to the statistical inference methods themselves and they assume the continued reliance of existing methods like p-values and null hypothesis significance testing (NHST). In fact, in our view, many researchers interested in the problems of reproducibility, appear to see the problem as primarily one of the incorrect use of p-values and NHST, not the reliance on these methods themselves, correctly used or not.

The incorrect use of NHST surely has something to do with reproducibility. Perhaps one of the most common misunderstandings, for example, is that a p-value is an estimate of the probability that the null hypothesis is true. Instead, the p-value is actually defined as the probability of observing a statistic as or more extreme than the one we obtained, when the null hypothesis is true. Used correctly, p-values can be useful as part of the measure of the discrepancy of the data and the null hypothesis, recent negative press notwithstanding. This said, the correct use of NHST alone does not address reproducibility because it does not address publication bias. It also does not address two other issues which need to be resolved together with reproducibility: *relevance* and *responsiveness*.

The issue of relevance refers to answering the right question: given the data, what is the probability that a hypothesis is true? The most important and difficult decisions by either corporate or government policy makers are virtually always made under some state of uncertainty. A systematic approach to making decisions under uncertainty is formally described in the fields of decision analysis and utility theory and these methods also have practical applications. In these fields, a fundamental requirement of making decisions under uncertainty is assigning probabilities to possible outcomes. The p-value and NHST alone do not and, in fact, cannot be sufficient for computing the required probabilities on outcomes.

The issue of *responsiveness* refers to how quickly systematic problems in research are detected and corrected. The relatively recent publication by Ioannidis (2005) provocatively titled "Why Most Published Research Findings are False" would apply to research that goes back decades. Even when problems with one or more studies are found, corrections can take years. Additional reproducibility research on psychology studies was conducted seven years after the publication dates of the original studies (Open Science Collaborative 2015). Reproducibility will likely continue to be a problem if the detection and subsequent corrective actions have such delays. Delays on this scale in the detection and correction of problems would surely be untenable in manufacturing and other areas of business and government.

The solution to the problem of responsiveness requires continuous "statistical process control" for science itself. One component of this approach is to compute a baseline "prior" probability for a field. The baseline prior for a field of study (e.g., childhood nutrition, clinical psychology, etc.) is the probability that the hypothesis being tested in a randomly selected proposed study is true. We will show the surprising finding that such a prior can be derived from knowing only the power, significance level and past reproduction study results from a set of studies in a particular field. This, together with controls prescribed by others (e.g., The Manifesto for Reproducible Science) may be a solution for improving the overall quality and usefulness of scientific research.

Before we introduce the approach to carry out quality control of scientific research that we propose, we note that not everyone believes that there is a reproducibility crisis in scientific research. In a recent publication, Jager and Leek (2014) dispute the notion that the reproducibility crisis in medical studies is as dire as it is in other scientific areas. The authors reviewed every manuscript published in the major medical journals between 2000 and 2010, and found that among the 5200 papers that included p-values, the false discovery rate was estimated to be no larger than $14 \pm 1\%$. Jager and Leek then conclude that the medical literature continue to provide an adequate record of real scientific progress.

2. An Argument for Priors

2.1. Decision Making Under Uncertainty

Assigning probabilities to outcomes from alternative courses of action is fundamental in the context of making decisions in virtually every field where decisions under uncertainty is analyzed quantitatively. This includes game theory and decision theory, operations research, quantitative portfolio management, and actuarial science. We would argue that even for decision makers who may not explicitly state probabilities of outcomes, this is at least an intuitive consideration for many of them. When a company invests in developing a new product, when an insurance company computes a premium, or when an entrepreneur starts a business, there is at least an implicit belief that some outcomes are more likely than others. Of course, this has been acknowledged by many previous authoritative sources so, instead of repeating their entire arguments here, we will select a few of their most salient statements on this point. For one, an early and perhaps most influential work in decision theory (Von Neumann and Morgenstern 1944) consider this fundamental to making decisions under uncertainty when they said the following:

> We have assumed only one thing, and for this there is good empirical evidence, namely that imagined events can be combined with probabilities. And, therefore, the same must be assumed for the utilities attached to them, whatever they may be.

Von Neuman and Morgenstern repeated throughout their work similar points about the necessity of assigning probabilities to all possible outcomes. Others who find priors an unavoidable necessity for decisions include Savage (1954), Wald (1950), Jaynes (1958), and—more recently—Howard (1977, 1984, 2015). If we accept the positions of individuals like these on the need for assigning probabilities to outcomes, it should also be clear that to do that we must state prior probabilities. As Jaynes put it as early as 50 years ago, the argument was sufficiently convincing to conclude "priors could no longer be ignored."

Priors can be anything from a highly individual notion about "the states of nature" to a naïve, standardized prior. We argue below that we can think of priors at different levels, and that at each level, the prior would reflect a different amount of information. (Each level is effectively a posterior probability relative to the previous level and a prior to the next.) Of course, the users of the study or future researchers do not have to accept

the stated prior. On the contrary, we are proposing that we limit restrictions on users of research results, as well as on future researchers, by providing many options for the analysis of the data. First and foremost among these, is the option to have access to all of the research data as recommended in the *Manifesto*. But even if this access to data is provided—and especially if it is not—researchers should provide users and future researchers multiple methods of analyses of these data. As the previously quoted sources would surely agree, researchers should also provide a recommended prior and the resulting posterior probabilities for any hypothesis in the study. We address the question of constructing these priors in subsequent sections. We also recommend providing Bayes factors (BFs) especially if, for any reason, the access to data and priors is not provided. And none of this needs to displace conventional test results such as p-values. Providing this much analytical detail is possible (but perhaps not trivial) when we fully exploit the era of "post-print" (i.e., digital, online) media for publishing scientific research. In such an environment, the practical data limitations of printed and bound issues can be ignored and inferences can be dynamically updated with replication studies and other data.

2.2. The Probability That a Study's Results Are True

So now, we proceed as if the reader has reviewed the arguments of those like Savage and Von Neuman and has ultimately accepted them. In the Bayesian context, prior distributions are used to summarize knowledge about "the state of nature" (often represented in the form of a parameter such as a mean or the probability of occurrence of some event) before any data are collected. Once data become available from an experiment or by some other means, prior information is combined with the information provided by the data using Bayes' rule. Bayes' rule establishes that the posterior odds in favor of a hypothesis equals the product of the prior odds and the likelihood ratio. The rule is prescriptive, and its application results in a posterior distribution that reflects all of the information about the parameter that is available to the investigator. The posterior distribution is simply the distribution of likely values of the parameter given the data, and can be thought of as the updated prior distribution after the data have been observed.

A rich literature argues that the Bayesian approach is rational and coherent, but even after the strong cases made by these early authors on the topic, there is one persistent objection to the use of priors in scientific inference. This objection is based on the idea that objectivity is preferable to the lack of it, that significance tests are "objective" and that the use of priors is merely "subjective." The claimed objectivity of p-values and significance tests was asserted and reasserted to researchers in their earliest statistics courses while challenges to this claim, it seems, are not given the same exposure. We agree with Jaynes (2003) on the confusion in the objectivity versus subjectivity distinction when he observed that "These words are abused so much in probability theory that we try to clarify our use of them."

The word "objectivity" itself bears some analysis. There are two specific meanings of this word that are often conflated when applied to tests of significance. In one sense, objective is used as having reality independent of the mind (as in "objective reality"). Used this way, an arbitrary standard of human creation like $p < 0.05$ is not "objective." In another sense, objective is used to mean "impartial"—that is, unbiased and equal treatment by a judge, arbitrator, or referee. Taken further, impartiality means that different people would have gotten the same answer when the same inputs are provided. Even Fisher (1973), certainly not known for being an advocate of priors in statistical inference, appeared to use it in this latter sense when he said

> . . . the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to, and verifiable by, other rational minds.

It is in this latter sense, the same sense used by Fisher, that the word "objective" applies to significance tests. But in this latter more relevant sense, objectivity is not an exclusive feature of significance levels. Priors are just as objective in this sense.

If either a prior or significance level are stated as an explicit, impartial assumption, then a mathematically correct calculation subsequent to it is objective in this latter sense. For example, the statement "if we assume $\Pr(X) = 0.05$, $\Pr(Y) = 0.5$, and $\Pr(X|Y) = 0.5$, then $\Pr(Y|X) = 0.05$" is necessarily true regardless of how we determined $\Pr(X)$, $\Pr(Y)$, and $\Pr(X|Y)$, subjectively or not. Furthermore, if an independent reviewer or other entity (i.e., not the researcher) is the source of the assumption $\Pr(X) = 0.05$, then the entire statement is as impartial as a significance level used as a matter of convention for a field of research. Used in this sense, mathematically correct statements containing assumed but impartial priors are just as objective as a statement containing an assumed and impartial significance level. In fact, the researcher using a significance test should and often does state caveats like "using a significance level of 0.05" or "assuming $\alpha < 0.05$" before reporting whether they reject of fail to reject the null hypothesis.

3. Do Decisions Require the "Rejection" or "Acceptance" of a Hypothesis?

Since the beginning of statistical tests, there has been an implied need to use arbitrary thresholds as a the basis for acceptance or rejection of some hypothesis, as if evidence for a claim could be best summarized by sorting claims into one of the two buckets of "accept" or "reject." In the case of the significance test that compares the p-value with some α-level, some go further by defining multiple Type I error levels such as $\alpha = 0.05, \alpha = 0.01, \alpha = 0.001$. Researchers may often indicate when p-values meet these significance levels with one, two, or three asterisks, respectively.

Such discretization is not needed and is bad for decision making. Discrete, universally applied thresholds simply do not appear in various quantitative fields related to decision making under uncertainty. One such field, decision analysis, is the practical application of decision theory for decision making under uncertainty (Von Neumann and Morgenstern 1944; Howard and Matheson 1977, 1984; Howard and Abbas 2015). At no point in decision analysis does the "count of asterisks" become a variable in optimizing a decision. Other practical (and highly

overlapping) analytical fields, such as actuarial science, management science, and operations research also do not depend on such distinctions. In these approaches, optimal decision making is modeled as a function of consequences of outcomes of alternatives and the probability distributions over those outcomes. If the possible rewards for a given action are large enough compared to losses, the action may be desirable even if the probability of the reward is low. If, for example, the methods of decision theory were the basis of health policy decisions, the adoption of a new, very expensive disease treatment with serious side effects may require a higher probability of a large benefit than is required for the adoption of a treatment which is less expensive, with fewer side effects, for a fatal condition with no other treatment.

One possible advantage of moving away from discretized, arbitrary thresholds is that apparently lukewarm findings may provide some incentive for replication. After all, the need to reproduce has probably seemed like a low priority once a researcher determines an effect to be "significant." The phrase "statistically significant" may give the typical user of the research the idea that they can be very confident in the results and that further research is not justified. But if $\Pr(H|\text{data}) = 0.43$, then maybe it will seem more clear that the research is not yet over. As we are about to show, posterior probabilities of that magnitude may be the norm for much of the published research in the social sciences. And, of course, whether or not a finding is statistically significant does not by itself confirm whether it is likely to be true, whether an effect is clinically important, or even whether it is worthy of further study. At no point do counts of asterisks or thresholds like "high" or "significant" come into consideration.

4. Levels for Constructing Priors

We propose three possible *levels* in which we generate priors.

Level 0 *Arbitrary, uninformed, universal prior*: In this level, we would use a single prior, set as an arbitrary convention for any hypothesis in an entire field, in much the same way that $P < 0.05$ has been used as the threshold de jure or de facto for statistical significance. Of course, even this conventional prior can be used in a perfectly objective way when, as previously argued, it is stated as an explicit caveat such as "If we assume $P(H) = 0.05$, then $P(H|\text{data}) = 0.45$." However, we consider this to be an extremely rudimentary approach which provides no advantage over conventional significance levels. We propose that more advanced levels should be pursued.

Level 1 *Informed universal prior:* If replication studies in a given field become available, then the prior for any single hypothesis in an entire field can be informed. To do so, we treat the previous replications themselves as a sample of the population of potential research in a field. This could be used to produce a kind of "implied prior" consistent with observed replication rates. We will show a simple example of how this can be done in Section 5.

Level 2 *Discriminated prior:* To derive a *discriminated prior* for specific research studies, we need additional, study-level information. For example, it has been shown that the sponsor of research has bearing on the replication success (e.g., Rochon et al. 1994; Friedberg et al. 1999). This and other factors can be considered to differentiate priors among different studies. The Level 2 prior can also be informed by the judgment of experts. We will review some potential methods.

As just mentioned, Level 0 prior would be a trivial but not recommended solution. Using an arbitrary threshold defined by convention is essentially what the current use of a significance levels does. Level 0 would not be an improvement, although it would not be worse, either (both are equally "objective"). So we will start with how a Level 1 solution to produce priors could be estimated.

Like Level 0, Level 1 produces a single prior for an entire field of research (i.e., the prior does not discriminate among different proposed research studies). Unlike Level 0, Level 1 is an informed prior based on previous replication studies in a given research field. The challenge is to infer this from previous replication studies which, themselves, relied on conventional significance tests. The prescribed approach asks "what is the prior that would be most consistent with observed replication rates?" If an initial Level 1 prior is estimated for a given field and if priors and posteriors will be used from then on (instead of conventional significance tests), the following method may no longer need to be employed.

5. Constructing a Level 1 Prior

We begin by framing a population for which inferences will be made by using a sample from the population. Consider a set of all potential research studies in a field. We leave it to researchers to define what constitutes their fields as they see fit. A "field" could be defined by existing professional bodies and institutions or may be all the articles submitted to a particular publication.[2]

We can think of a field as a set of research studies proposed by some body of researchers. Only a subset of these proposed studies have been conducted to date. That is, studies conducted so far are a sample of the population of studies potentially proposed by the researchers of a field. Furthermore, unless we have reason to believe that there is some order to which studies are conducted first, we propose to treat the studies conducted to date as a random sample of the population. Note that this is not an assumption about whether topics of research will change over time. Of course, that will change but that, too, can be part of computing the prior for a field. It is only a proposed starting point unless and until future observations show otherwise. We do not wish to suggest that converting these ideas into practice is without difficulty.

Also, we treat reproduction studies as a random sample of studies that were published. Here, by replication studies we refer to studies that address the same or a sufficiently similar issue as the original study. If we can make an estimate of the distribution of "true hypothesis" in the population of potential studies, then

[2] Like any definition of a population, what constitutes a field will inevitably be somewhat arbitrary. (However, researchers refer to their "fields" usually without need for further clarification, especially when speaking to others they see as peers in their "field.")

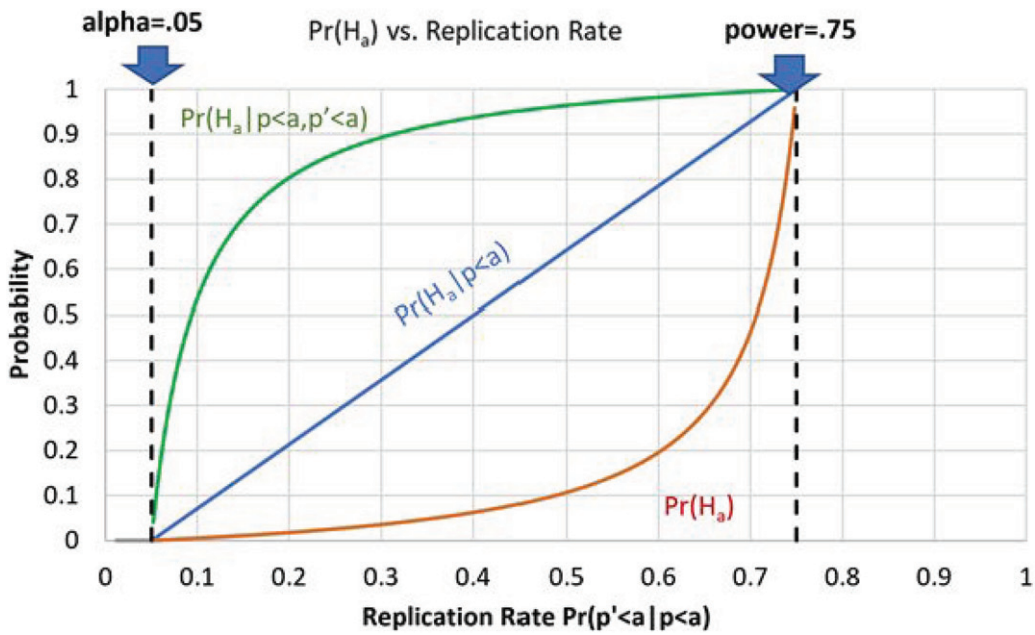


Figure 1. Probability that the research proposition is true for different levels of replication in the field and number of significant results in independent studies. $\Pr(H_a)$, $\Pr(H_a|P \leq \alpha)$, and $\Pr(H_a|P \leq \alpha, P' \leq \alpha)$ are shown in the bottom red curve, the middle blue line, and the top green curve, respectively.

we use that as the prior for a subsequent proposed study (which, again, we treat as being randomly selected from the population of potential studies).

Now consider the simplest situation where a proposition H is either true or false (we will ignore the question of effect size for now). Suppose that in a given field hundreds of such propositions have been researched and published over the years in the peer-reviewed literature, and suppose further that we are interested in investigating the overall quality of this body of research.

We could, for example, proceed as in Open Science Collaboration (2015) which conducted a study that attempted to reproduce the findings of 100 published psychology papers. Only 36 of the 100 studies were successfully reproduced. How would we use the results from those studies—both original and reproduced—to inform our prior about a particular hypothesis H? While the fact that a study is reproduced lends some credence to the original claim of significance, it does not confirm that the proposition was true, since the results from the replication study are based on significance tests, just like the original results. Yet, we are able to make an inference for H (randomly selected from the proposed hypothesis of a "field" of research) being true using the approach we describe below.

We first establish some notation. Consider a study with two simple hypotheses H_0, H_a ("simple" meaning that in each case, the model is completely specified). The probability that a study rejects the null hypothesis H_0 is $\Pr(P \leq \alpha)$, where P denotes the p-value of the test and α is the significance level. If the study is independently reproduced, and if the Type I error is set to the same level α, then the joint probability of observing a significant result in both studies is $\Pr(P \leq \alpha) \times \Pr(P' \leq \alpha) = \Pr(P \leq \alpha, P' \leq \alpha)$, for P' the p-value of in the reproduction study. The Type I error α is defined as the probability of incorrectly concluding H_a or $\Pr(P \leq \alpha|H_0)$, and the power $\pi = 1 - \beta$ is the probability of correctly concluding H_a or $\Pr(P \leq \alpha|H_a)$, for β the Type II error. We use R to denote the reproduction rate, or the probability that a study that rejected the null hypothesis is independently reproduced. If the information is available, R can be estimated as the number of reproduced studies over the number of studies that had earlier concluded H_a.

In the Appendix, we show the derivation for three probabilities that a hypothesis is true (before the initial study, after the initial study, and after the study is reproduced) given results in reproducibility studies across a given field and a given power and significance level:

- The prior probability that a proposed hypothesis is true

$$\Pr(H_a) = \frac{\alpha(R - \alpha)}{\pi^2 - \alpha^2 + R(\alpha - \pi)}. \tag{1}$$

- The posterior probability after the initial study

$$\Pr(H_a|P < \alpha) = \frac{(R - \alpha)}{(\pi - \alpha)}. \tag{2}$$

- The posterior probability after the study is reproduced

$$\Pr(H_a|P \leq \alpha, P' \leq \alpha) = \frac{\pi^2 \Pr(H_a)}{\pi^2 \Pr(H_a) + \alpha^2(1 - \Pr(H_a))}. \tag{3}$$

Figure 1 illustrates the relationship between the replication rate, the implied prior, α, and π. In the figure, we assume that $\alpha = 0.05$ and $\pi = 0.75$, and plot $\Pr(H_a)$, $\Pr(H_a|P \leq \alpha)$, and $\Pr(H_a|P \leq \alpha, P' \leq \alpha)$ in the bottom red curve, the middle blue line, and the top green curve, respectively, for different values of the replication rates in the field.

As an illustration, suppose that in the field of adolescent depression, there was an attempt to reproduce 60 studies which initially had significant results where the power and significance level are consistently 0.8 and 0.05, respectively. Of these, 18 studies were found to have a statistically significant result using the

same power and significance level. This gives us a reproduction rate of 0.3. Using Equation (1), we can compute that the prior probability now for any proposed study in the field is estimated to be

$$\Pr(H_a) = \frac{0.05 \times (0.3 - 0.05)}{0.8^2 - 0.05^2 + 0.3 \times (0.05 - 0.8)} = 0.0303.$$

Then, lacking any other information about a proposed study other than these stated parameters describing the field of research from which the study is drawn, we would say the hypothesis has a 0.0303 probability of being true. If the study then produces a statistically significant result, we could apply equation (2) above and conclude that the tested hypothesis only has a one-in-three chance of being true

$$\Pr(H_a|P < \alpha) = \frac{0.3 - 0.05}{0.8 - 0.05} = 0.333.$$

Does $\Pr(H_a|P \leq \alpha) = 0.333$ seem low? It should be low with such a low rate of reproducible studies. Remember we stated the reproduction rate of studies—which were initially found to have statistically significant results—was only 18 out of 60. If *every* hypothesis being proposed was false, then we would still find that the reproduction rate approaches α after a large number of reproduction studies have been conducted. If every hypothesis were true, then the rate of reproduction would approach the power in the long run. (Of course, in practice it may not be the case that $\alpha < R < \pi$ if the number of studies is not large.) See the Appendix to confirm this result.

Finally, when and if those results are reproduced using the same alpha and power, we can again update the probability to be 0.889 using Equation (3)

$$\Pr(H_a|P < \alpha, P < \alpha) = \frac{0.8^2 \times 0.0303}{0.8^2 \times 0.0303 + 0.05^2 \times (1 - 0.0303)} = 0.889.$$

Most importantly, this approach allows us to simply abandon the use of significance tests altogether once we compute $\Pr(H_a)$ from the previous reproduction studies. We can update $\Pr(H_a)$ with a set of observations by using a BF $\Pr(\text{data}|H_a)/\Pr(\text{data}|H_0)$ and update it in a similar fashion again when this study is reproduced. The proposed equations need only be a way of computing a starting point for priors in a given field of research.

6. "Statistically Significant" May Still Only Mean "Slightly More Likely Than Not to Be False"

Suppose that a replication study was done similar to the Open Science Collaboration project, where 36 out of 100 studies were successfully reproduced, and where power was 0.75 for both the original research and its replication, and α was set at 0.05 for both in all initial studies and replication studies.

Note that we can and should treat the estimate of the replication rate R as itself being estimated from a population. This would be done with a beta distribution on R. We recommend naïve priors for this, such as Beta(1/2,1/2), the Jeffreys prior for R. For the estimate of $\Pr(H_a)$ for a single proposition, we only need the mean of the beta distribution which, with the Jeffreys prior, would be $(a + 1/2)/(a + b + 1)$ where a is the number of successful replications in a field and b is the number of replications which were attempted but were unsuccessful. The prior will make a larger difference in a field when few replication studies have been attempted. Using the observed replications to compute the mean of an estimate of a population proportion of successful replication, this would produce an $R = (36 + 0.5)/(100 + 1) = 0.361$. In this case, the number of replications is large enough that the Jeffreys priors have only a small impact on the estimated replication rate.

This replication rate would produce estimates of $\Pr(H_a) = 0.05$ and $\Pr(H_a|P \leq \alpha) = 0.44$. This is roughly in agreement with findings by Berger and Sellke (1987) where they demonstrated that $\Pr(H_0|P \leq 0.05) = 0.52$ for $n = 50$, which results in $\Pr(H_a|P \leq \alpha) = 0.48$. If these findings hold for other fields then, based on conventional significance tests, we should expect to see that the probability of a proposition being true given a "significant" result is on the order of a coin flip. This alone may leave researchers—as well as news media and other users—with the sense that more needs to be done before making a confident declaration about a new finding.

Note that an important feature of this method is that publication bias, p-hacking or even outright fraud of the original study does not affect these calculations, as long as the reproduction study is conducted independently of the original study. If such factors are common in a field, then naturally the replication rate will be lower. Furthermore, once we establish this prior for a population of research studies, $\Pr(H_a|P \leq \alpha)$ can be updated using the updating property of Bayes' theorem.

7. Level 2 Priors: A Conceptual Overview

Level 2 priors are informed priors that allow for discrimination among different propositions. Recall that Level 0 and (the strongly preferred) Level 1 both only provide priors for a randomly selected proposition out of a population of potential propositions. They do not differentiate among propositions based on their individual characteristics. In Level 1, we do not take into account any knowledge about the proposition that may differentiate it from others in the field. However, in Level 2 priors we utilize characteristics of individual studies to differentiate among them. Level 2 could be much more involved than Level 1 but our goal here is to review the potential benefits of more advanced analytics for statistical inference.

One example of a property that may differentiate an informed prior on a study from other studies in a field is who sponsors the study. Research has shown that the sponsor of the research may have bearing on outcomes. Rochon et al. (1994) and Friedberg et al. (1999) are two early references that show that studies sponsored by drug companies are much more likely to result in favorable results for the new drug than studies funded by not-for-profit organizations. The replication rates would just be estimated separately for different sponsors of research. This could also be done for considerations such as whether the study was properly registered (indicating a lower potential for publication bias) and previous retractions of a research institute. Equivalently, the field-level prior could be adjusted using study-level information to improve the estimation of $\Pr(H_a|P \leq \alpha)$ in a specific study.

Another way to tailor these implied, informed priors to specific propositions is to rely on the judgment of reviewers, who are experts in the field. It may be beneficial to involve reviewers earlier in the research process, so that they can provide input at the time when the study is registered but has not yet been conducted. These reviews could involve subjective adjustments to the previously objectively computed priors. This would allow the prior to be informed on characteristics of the proposed research other than what was explicitly parametrized, such as how well the hypothesis supports or contradicts previously known principles in the field.

There is ample evidence that reviewers could be made to be well *calibrated* in these estimates. The literature on calibration of subjective probabilities defines well-calibrated as a process that produces probabilities that, in subsequent observations, closely matches observed frequencies. That is, when an individual (e.g., a weather forecaster) applies a large number of subjective probabilities to events (rain the following day), it is determined that of all the times when the individual said something had a probability of 0.2, it was observed to occur 20% of the time. Likewise, events occur at a rate of 10% when they say the probability is 0.1, and so on (Murphy and Winkler 1977; Lichteinstein 1981; Ungar et al. 2012). Studies have shown that individuals do not initially perform well in such tasks. They may be extremely overconfident that the observed frequency of an event is much lower than the stated probability and in some contexts may be under-confident. Yet it has also been shown that it is possible to almost eliminate either bias using some combination of training, elicitation techniques, incentives, aggregation, and mathematical transformations of the estimates. These methods are referred to as *calibrating* or *de-biasing* the subjective probabilities (Ferrell and Goey 1980; Fellner and Krüger 2012). A related body of literature addresses the question of elicitation of information from experts. Brownstein et al. (2019) discuss the critical role of expert judgment in the various steps in the scientific process and in decision making, and conclude that establishing reliable protocols for obtaining input from experts is important. O'Hagan (2019) agrees and argues that subjective judgment is inevitable in scientific inference and decision making. He adds that the fact that expert judgment is subjective does not mean that it cannot be elicited in a scientifically valid way, so that the chance of bias is minimized.

At this point, however, we are only recommending that the data about reviewer estimates at least begin to be recorded. How those estimates could subsequently be used to improve the performance of the prior estimation process should be a subject of further study, along with other data analysis about sponsors and other objective characteristics. But at least that field of research would have the data to start evaluating alternative methods. It would be ironic, in fact, if a field like psychology objected to at least evaluating the approach we propose.

8. Quality Control for Science

A feature of both Level 1 and 2 priors is the potential for greatly improved responsiveness. Again, by responsiveness, we mean the ability of researchers in a field to detect and respond to problems in their field in a more timely manner instead of decades of delay. Level 0 creates a naïve prior for an entire field and, like the traditionally recommended significance levels, does not change as replication rates are updated. Levels 1 and 2, however, offer the basis for more frequently evaluated performance of a field. We can purposely collect and record the *metadata* resulting from the process of conducting research. By metadata, we refer to attributes of the publication *process* rather than the content of the publication itself, the type of information about submissions that journals and professional societies currently do not track. This allows for a type of "quality control" for science.

At the level of the study itself, information to be recorded might include effect size, p-value, reputation of investigator, source of funding for the work, degree of "surprise" associated with claim(s), sample size, and many others. At the level of the field itself, we could record the number of retractions in the field, the proportion of studies that get reproduced successfully, and such. Finally we could also keep track of attributes associated with the journal in which a study appeared, and that might include the impact factor for the journal, its acceptance rate, number of volumes, pages per year, etc.

In principle, as more metadata become available, we could think of deploying the type of tools that are used in other fields to classify items into two or more categories. In our case, the "item" is a research claim, and the categories might be "true" and "false." It is standard to train classification algorithms using a large set of items for which we know "ground truth," that is, a large number of research claims for which we "know" whether they are really true or false. These algorithms, generally known as supervised learning algorithms (e.g., Breiman 2001) "learn" which combinations of the various metavariables are typically observed when two items belong in the same category or in a different category. The output is a score, or an empirical probability of membership in each of the two classes.

In the research enterprise, we do not typically get to observe ground truth, as we do not know which proposition is true and which is false. But as described in Section 4, we can estimate the probability that a claim is true, and agree on a convention that says that for the purpose of training a classifier, we will declare that any claim with a probability of occurrence above say 0.75 is "true." Given a large number of "true" and "false" claims and the corresponding metadata, we could compute the empirical class probability for each item in each class, and then construct (and periodically update) a distribution of classification score for "true" and for "false" claims. With the same algorithm, a potential user of a specific research paper would then proceed as follows:

- First, compute the corresponding score, or empirical probability that the claim is true using the classification algorithm.
- Next, compare the empirical probability to the distribution of scores obtained for "true" claims and the distribution of scores obtained from the "false" claims.

The user can then decide whether her score is more typical if the claim is "true" than if the claim is "false."

For illustration, we use the following simple example. In a specific field, investigators attempted to reproduce 200 studies, and in 100 of those studies, the results were the same as

those obtained by the initial researchers. In the other 100, the reproduced study failed to reach the same conclusion. Suppose that for each of the 200 original studies we have information on metadata including the source of funding for the study, the number of retractions associated with the investigator, the sample size and whether the investigator had "registered" the study or not. A logistic regression model with a binary response ($Y = 1$ if results were reproduced, and $Y = 0$ otherwise) and with the meta-variables as predictors in the model can be fit using the 200 studies. If a new paper is published that concludes H_a in some problem, we can evaluate the logistic regression model using the metadata for the new study, and compute an empirical probability that the results will reproduce if the study is repeated. If, from the training studies, we know that when results reproduce, the empirical probabilities vary between 0.6 and 0.99, and when results do not reproduce, the range of empirical probabilities is between 0.0 and 0.45, then this provides information that we can use to predict whether the new study's claim is or is not likely to hold should the study be reproduced.

In the ideal scenario (and in the simple example above), the strength of the score we compute for a proposition would correspond closely with the probability that the proposition is true, computed as described earlier. Here, the classifier plays the role of the expert in that it combines many different pieces of information in some optimal way to produce the best (in some sense) assessment of $\Pr(H_a)$. The effectiveness of the data-science approach we discuss is critically dependent on the information that is used to train the classifier, so at this time we present it more as an idea for the future than as something that can be implemented soon.

9. Conclusions

We propose that the generation of priors for a field of research is both feasible and necessary. We have argued that uninformed priors are as objective as significance levels if objective is taken to mean "impartial" or in the sense that a calculation is objective when it assumptions are explicit. We also know that, unlike significance levels, priors are necessary to compute the probability that a proposition is true given the data and the assignment of probabilities to outcomes is necessary for decisions under uncertainty. This addresses the issue of *relevance*.

We also show how such priors could be computed from reproduction studies even though the reproduction studies themselves rely on conventional significance tests and do not directly compute the probability that a proposition is true. Also, using examples from existing research, we show that the probability that a hypothesis is true after the initial positive finding will often be somewhere in the range of a coin flip (0.5) or lower.

We believe that explicitly publishing such a probability in the abstract of a paper will make users, media, and other researchers more likely to see the finding as tentative instead of final. This may either encourage further reproduction itself or even encourage the original researchers to increase the sensitivity of a study so that the initial finding produces a probability of greater than 0.95 or 0.99.

Priors may be further improved by using information that discriminates among different proposals (as opposed to estimating a single prior for a random proposition out of a population of potential propositions). Using metadata about the performance of studies as described earlier (e.g., reproduction rates, sponsors, reviewer estimates, etc.), would allow for different priors for propositions of different types. Further differentiation may be made by utilizing subjective estimates of reviewers once sufficient data has been collected on estimates from reviewers and a de-biasing/calibration method has been determined which best predicts results of original studies and reproduction studies.

Finally, consistently updating priors for all studies in a field based on reproduction results is one way to address the issue of *responsiveness*. This creates a process of continuous statistical quality control for science similar to what is done in real-time manufacturing environments. Reproductions, retractions, reviews and more can be tracked and used to update priors (Barlow and Irony 1992; Deming 2000). In these methods, often known as statistical process control, when any one of these metrics meets a "control limit" a specific intervention action would be required. Instead of waiting years or even decades to detect a problem, potential issues are monitored with each publication, reproduction, or retraction. If responsiveness is not improved, reproduction can only continue to improve at a snail's pace, if at all.

We are certainly not the first authors to address these types of issues, but we do offer a different perspective from what has been published in the literature. In addition to publications that have already been referenced in this work, we highlight the work of Sellke et al. (2001), Nuzzo (2014), and Fraser (2017). Sellke et al. (2001) were early critics of science's reliance on p-values for decision making, and proposed an approach to calibrate p-values so that those calibrated p-values better correspond to the $\Pr(H_a|\text{data})$. Nuzzo discussed the key role of $\Pr(H_a|\text{data})$ in the decision making process, but she did not elaborate on how to implement a process to get to the desired outcome. Finally, Fraser (2017) demonstrates that p-values can be the bases for a complete inferential framework, and highlights the notion that the fact that p-values have been mis-used and even abused, this does not subtract from their worth as indicators of the distance between the data and a parameter value.

To conclude, we recommend that an entity (e.g., the journal, professional association, university, or other agency) be identified to implement and manage the following:

1. A field should implement the research registries, data sharing, and other standards recommended by the Manifesto.
2. Implementing Level 1 priors can begin in parallel with 1) above. Jeffreys prior for reproduction rates would be updated with reproduction studies from that field. Methods for generalizing this approach to the broader question of effect sizes will need to be developed. More precisely,

   (a) A journal, professional association or other entity should implement, at a minimum, a Level 0 prior or, preferably skip directly to a Level 1 prior if there are adequate reproductions.
   (b) If reproductions in a given field of research are too few, too old, or nonexistent, then that field should imme-

diately organize reproduction studies to bootstrap the informed *implied* prior of Level 1.

(c) Published research should clearly disclose the updated priors and posterior probabilities of a proposition in the abstract or directly under the authors' names. In an online source, these probabilities would be updated dynamically with new reproduction studies.

3. A plan for implementing Level 2 priors should be specified including the following:

(a) Parameters of studies which could be used to discriminate priors among different propositions can be investigated. Where such candidate parameters have sufficient predictive utility, then they can be incorporated as part of a prior estimation and continuous quality control.

(b) A research field should change the peer review process so that reviewers are shown proposed research before it is completed. Those reviewers should be asked to estimate subjective probabilities for proposed research including the probability of a positive result and the probability that reproduction, if attempted, will support the earlier study. As these data are gathered, training, transformations and other calibration methods can be tested. If such a method is found to be well-calibrated, then subjective probabilities of reviewers can be incorporated into the estimate of priors.

(c) Decision criteria for interventions may be established by identifying control limits for each of the tracked metrics. Corrective interventions in a field may include changing the recommended power of studies, research methods, or how they are reviewed.

(d) More advanced data analytic methods can be used (and, of course, monitored for performance) to further improve reproduction, relevance and responsiveness.

In summary, the reproducibility crisis is best handled when managed in concert with relevance and responsiveness. We believe that our recommendations are feasible, reasonable, flexible and ultimately necessary, and that beginning the discussion on how to best implement the approach we propose can be accomplished.

Appendix

Recall from the notation in Section 6 that the Type I error α is defined as the probability of incorrectly concluding H_a, or $\Pr(P \le \alpha | H_0)$. Given the null hypothesis and if the original and the reproduction studies have the same α and are statistically independent, the probability that both studies will exhibit significant results is

$$\Pr(P \le \alpha | H_0)\Pr(P' \le \alpha | H_0)$$

or simply α^2. Similarly, if H_a is true, then the probability that both studies would produce significant results is π^2, assuming that the Type II error β was the same in both studies.

Now, using the law of total probability, we can write

$$\begin{aligned}\Pr(P \le \alpha) &= \Pr(P \le \alpha | H_a)\Pr(H_a) + \Pr(P \le \alpha | H_0)[1 - \Pr(H_a)] \\ &= \pi \Pr(H_a) + \alpha[1 - \Pr(H_a)]. \end{aligned} \tag{4}$$

By the same line of reasoning, we have that

$$\begin{aligned}\Pr(P \le \alpha, P' \le \alpha) &= \Pr(P \le \alpha, P' \le \alpha | H_a)\Pr(H_a) \\ &\quad + \Pr(P \le \alpha, P' \le \alpha | H_0)[1 - \Pr(H_a)] \\ &= \pi^2 \Pr(H_a) + \alpha^2[1 - \Pr(H_a)]. \end{aligned} \tag{5}$$

Using the equality in (4) and letting $R = \Pr(P' \le \alpha | P \le \alpha)$ denote the reproduction rate, we find that

$$R\{\pi \Pr(H_a) + \alpha[1 - \Pr(H_a)]\} = \pi^2 \Pr(H_a) + \alpha^2[1 - \Pr(H_a)],$$

so that after some rearrangement we can solve for $\Pr(H_a)$ to get

$$\Pr(H_a) = \frac{\alpha(R - \alpha)}{\pi^2 - \alpha^2 + R(\alpha - \pi)}. \tag{6}$$

The prior $\Pr(H_a)$ we just obtained in (6) can now be used to estimate the probability that the proposition in the first study was true, given that the investigators rejected H_0 as follows

$$\begin{aligned}\Pr(H_a | P \le \alpha) &= \frac{\Pr(H_a)\Pr(P \le \alpha | H_a)}{\Pr(H_a)\Pr(P \le \alpha | H_a) + [1 - \Pr(H_a)]\Pr(P \le \alpha | H_0)} \\ &= \frac{\pi \Pr(H_a)}{\pi \Pr(H_a) + \alpha[1 - \Pr(H_a)]}. \end{aligned} \tag{7}$$

It follows from all of the above that

$$\Pr(H_a | P \le \alpha, P' \le \alpha) = \frac{\pi^2 \Pr(H_a)}{\pi^2 \Pr(H_a) + \alpha^2[1 - \Pr(H_a)]}. \tag{8}$$

The linearity of $\Pr(H_a | P \le \alpha)$ shown in Figure 1 may be surprising, but when it is simplified it becomes more obvious. Substitute the $\Pr(H_a)$ in the function for $\Pr(H_a | P \le \alpha)$ in (7). Then multiply the numerator and denominator of by $(R - \alpha)/\Pr(Ha)\pi$ and expand the expressions. We would find that a number of terms cancel each other out and we are left with

$$\Pr(H_a | P < \alpha) = \frac{(R - \alpha)}{(\pi - \alpha)}.$$

Funding

Alicia L. Carriquiry gratefully acknowledges partial support from the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

References

Barlow, R. E., and Irony, T. Z. (1992), "Foundations of Statistical Quality Control," in *Current Issues in Statistics: Essays in Honor of D. Basu*, Lecture Notes—Monograph Series (Vol. 17), New York: Springer. [53]

Berger, J. O., and Sellke, T. (1987), "Testing Point-Null Hypotheses: The Irreconcilability of *p*-Values and Evidence," *Journal of the American Medical Association*, 82, 112–122. [51]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [52]

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*. [52]

Buckheit, J. B., and Donoho, D. L. (1995), "WaveLab and Reproducible Research," Technical Report #474, Department of Statistics, Stanford University, Stanford. [46]

Casadevall, A., and Fang, F. C. (2010), "Reproducible Science," *Infection and Immunity*, 78, 4972–4975. [46]

Deming, W. E. (2000), *Out of the Crisis* (1st ed.), Cambridge, MA: MIT Press. [53]

Dickersin, K. (1990), "The Existence of Publication Bias and Risk Factors for Its Occurrence," *Journal of the American Medical Association*, 263, 1385–1389. [46]

Fellner, G., and Krüger, S. (2012), "Judgmental Overconfidence: Three Measures, One Bias?," *Journal of Economic Psychology*, 33, 142–154. [52]

Ferrell, W. R., and Goey, P. J. (1980), "A Model of Calibration of Personal Probabilities," *Organizational Behavior and Human Decision Processes*, 26, 32–56. [52]

Fisher, R. A. (1973), *The Design of Experiments* (9th ed.), New York: MacMillan. [48]

Fraser, D. A. S. (2017), "*p*-Values: The Insight to Model Statistical Inference," *Annual Reviews in Statistics and its Applications*, 4, 1–14. [53]

Friedberg, M., Saffran, B., Stinson, T. J., Nelson, W., and Bennett, C. L. (1999), "Evaluation of Conflict of Interest in Economic Analyses of New Drugs Used in Oncology," *Journal of the American Medical Association*, 282, 1453–1957. [49,51]

Goodman, S. N. (1999), "Toward Evidence-Based Medical Statistics: 1. The *p*-Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004. [46]

Howard, R. A., and Abbas, A. E. (2015), *Foundations of Decision Analysis* (1st ed.), New York: Pearson. [47,48]

Howard, R. A., and Matheson, J. E. (1977), *Readings in Decision Analysis*, Menlo Park, CA: SRI International. [47,48]

——— (1984), *Readings on the Principles and Applications of Decision Analysis* (2 vols.), Menlo Park, CA: Strategic Decisions Group. [47,48]

Hubbard, R. (2016), *Corrupt Research: The Case for Re-conceptualizing Empirical Management and Social Science* (1st ed.), Los Angeles, CA: Sage. [46]

Ioannidis, J. P. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124, DOI: 10.1371/journal.pmed.0020124. [47]

Jager, L. R., and Leek, J. T. (2014), "An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature," *Biostatistics*, 15, 1–12. [47]

Jaynes, E. T. (1958), "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics*, 4, 227–241. [47]

——— (2003), *Probability Theory: The Logic of Science*, London: Cambridge University Press. [48]

Kerr, N. L. (1998), "HARKing: Hypothesizing After the Results are Known," *Personality and Social Psychology Review*, 2, 196–217. [46]

Lichteinstein, S., Fischhoff, B., and Phillips, L. (1981), Calibration of Probabilities: The State of the Art to 1980. *Technical Report PTR-1092-81-6 to Offfice of Naval Research*, Contract No. NOOO14-80-C-0150. [52]

Munafò, M. R., Nosek, B., 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 Behavior*, 1, 0021. [46]

Murphy, A. H., and Winkler, R. L. (1977), "Reliability of subjective Probability Forecasts of Precipitation and Temperature," *Journal of the Royal Statistical Society*, Series C, 26, 41–47. [52]

Nuzzo, R. (2014), "Scientific Method: Statistical Errors," *Nature*, 506, 150–152. [53]

O'Hagan, A. (2019), "Expert Knowledge Elicitation: Subjective But Scientific," *The American Statistician*. [52]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [47,50]

Popper, K. (1959), *The Logic of Scientific Discovery*, London: Routledge. [46]

Rochon, P. A., Gurwitz, J. H., Simms, R. W., Fortin, P. R., Felson, D. T., Minaker, K. L., and Chalmers, T. C. (1994), "A Study of Manufacturer Supported Trials of Non-steroidal Anti-inflammatory Drugs in the Treatment of Arthritis," *Archives of Internal Medicine*, 154, 157–163. [49,51]

Rosenthal, R. (1979), "The File Drawer Problems and Tolerance for Null Results," *Psychological Bulletin*, 86, 638–641. [46]

Sackett, D. L. (1979), "Bias in Analytic Research," *Journal of Chronic Diseases*, 32, 51–63. [46]

Savage, L. J. (1954), *The Foundations of Statistics*, New York: Wiley. [47]

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of *p*-Values for Testing Precise Null Hypothesis," *The American Statistician*, 55, 62–71. [53]

Sterling, T. (1959), "Publication Decisions and Their Possible Effects on Inferences Drawn From Tests of Significance," *Journal of the American Statistical Association*, 54, 30–34. [46]

Stodden, V. (2015), "Reproducing Statistical Results," *Annual Reviews of Statistics and its Application*, 2, 1–19. [46]

Stodden V., Guo, P., and Ma, Z. (2013), "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," *PLoS One*, 8, e6711. [46]

Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., and Swift, S. (2012), "The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions," *AAAI Technical Report FS-12-06*, Association for the Advancement of Artificial Intelligence, University of Pennsylvania. [52]

Von Neumann, J., and Morgenstern, O. (1944), *The Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press. [47,48]

Wald, A. (1950), *Statistical Decision Functions*, New York: Wiley. [47]

The American Statistician

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

The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making

Naomi C. Brownstein, Thomas A. Louis, Anthony O'Hagan & Jane Pendergast

To cite this article: Naomi C. Brownstein, Thomas A. Louis, Anthony O'Hagan & Jane Pendergast (2019) The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making, The American Statistician, 73:sup1, 56-68, DOI: 10.1080/00031305.2018.1529623

To link to this article: https://doi.org/10.1080/00031305.2018.1529623

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 10148

View related articles

View Crossmark data

Citing articles: 22 View citing articles

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, 56–68: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529623

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making

Naomi C. Brownstein[a,b,c], Thomas A. Louis[d], Anthony O'Hagan[e], and Jane Pendergast[f]

[a]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center,Tampa, FL; [b]Department of Oncologic Sciences, University of South Florida, Tampa, FL; [c]Department of Behavioral Sciences and Social Medicine, Florida State University, Tallahassee, FL; [d]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; [e]School of Mathematics and Statistics, The University of Sheffield, Sheffield, UK; [f]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC

ABSTRACT

This article resulted from our participation in the session on the "role of expert opinion and judgment in statistical inference" at the October 2017 ASA Symposium on Statistical Inference. We present a strong, unified statement on roles of expert judgment in statistics with processes for obtaining input, whether from a Bayesian or frequentist perspective. Topics include the role of subjectivity in the cycle of scientific inference and decisions, followed by a clinical trial and a greenhouse gas emissions case study that illustrate the role of judgments and the importance of basing them on objective information and a comprehensive uncertainty assessment. We close with a call for increased proactivity and involvement of statisticians in study conceptualization, design, conduct, analysis, and communication.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Bayesian paradigm; Clinical trials; Collaboration; Elicitation; Scientific method; Subjectivity

1. Introduction

As participants in the October 2017 Symposium on Statistical Inference (SSI), organized and sponsored by the American Statistical Association (ASA), we were challenged to host a session and write a paper inspired by the question, "Do expert opinion and judgment have a role in statistical inference and evidence-based decision-making?" While we work from different perspectives and in different statistical paradigms (both frequentist and Bayesian), there was a resounding "yes!" among us, with ample common ground in our thinking related to this infrequently discussed and often under-appreciated component of statistical and scientific practice.

Expert judgment is a feature of both frequentist and Bayesian inference. Judgments are subjective, and subjectivity is a topic that has generated heated debate in the Statistical community (Gelman and Hennig 2017). The subjectivity in a Bayesian prior distribution has been criticized by frequentists. Implicit claims are that frequentist methods are objective and that they, "let the data speak for themselves." However, frequentist methods also require some components of subjectivity in the choice of a model and of estimators, test statistics, etc. These choices are subjective in the sense that every expert builds knowledge and judgment into their own personal framework of understanding, which is not likely to be identical to that of anyone else. Expert judgment is clearly needed for valid statistical and scientific analyses. Yet, questions regarding how, when, how often, and from whom judgment is helpful, rather than leading to biased, misleading and nonreproducible results, is less clear.

The ASA commissioned this special issue of *The American Statistician* to stimulate "a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical-science—and science itself—into a new age." We consider the roles of expert judgment in scientific practice. This article is a distillation of our common ground, along with our advice, warnings, and suggestions for incorporating expert judgment into science, while maintaining the integrity and scientific rigor of the research. Note that although we were asked in the Symposium to consider expert opinion and judgment, we avoid the word "opinion" because it risks being equated to uninformed rhetoric. We emphasize "judgment" that should be informed, carefully considered and transparent. While we might still disagree on specific details or the relative importance of various points, we present a strong, unified statement related to this key component of statistical practice. Additional literature on expert judgment in statistics may be found in a companion paper (Brownstein 2018).

Our article is organized in the following sections where we share our thoughts on when and how expert judgment has a legitimate and necessary role in scientific inquiry. Section 2 presents the role of judgment in the scientific method and more generally, in four stages of scientific inquiry, inference, and decision-making. In Section 3, the four stages are examined in more detail, focusing on their needs for expert judgment and the qualifications implied by "expert." Two case studies are presented in Section 4, with emphasis on the principled and scientific application of expert judgments. Finally, Section 5 summarizes our key conclusions.

CONTACT Naomi C. Brownstein naomi.brownstein@moffitt.org Department of Behavioral Sciences and Social Medicine, College of Medicine Florida State University, 1115 West Call Street, Tallahassee, FL 32306.

Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/TAS.

© 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.

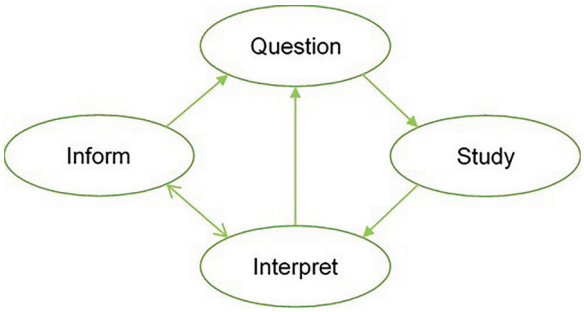


Figure 1. The cycles of inference and decision.

2. The Cycles of Inference and Decision in Science

The practice of science is often described in terms of the scientific method, defined in Oxford University Press (2018d) as involving, "systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses." More generally, one might describe the scientific method as a collection of methodologies and processes of gathering observable, measurable evidence using experimentation and careful observation, and of analyzing the evidence to draw conclusions, inform decisions and raise new questions. The purpose may be to inform about pathways or mechanisms by which results are obtained or to aid in prediction or estimation of quantities of interest. A thorough review of the history of the scientific method is found in Anderson and Hepburn (2016). While we believe that not all scientific inquiry, and certainly not all decision-making, falls within that prototypical guide, it is important to understand how the scientific method fits into our framework. Namely, science is data-driven in the ascertainment of objective information to draw scientific conclusions, yet with expert judgment feeding into the processes.

We base our four-stage graphic (Figure 1) on the depiction of Garland (2016), adding the Inform stage to accommodate decision-making and broadening the definitions of the other stages. It illustrates four stages in the perpetual process of scientific inquiry and evidence-based decision-making: Question, Study, Interpret, and Inform. These stages are formulated in highly general terms, because the practice of scientific inquiry, and hence the remit of statistical analysis, is also very wide.

We first describe activities that comprise each of these stages. Our framework allows us to underscore the varied roles of expert judgment throughout the perpetual cycles of the scientific method.

1. *Question*: Scientific inquiry can be characterized as beginning with one or more questions. A Question might be a formal scientific hypothesis arising either out of observation of real-world phenomena or from the current status of scientific knowledge and debate. On the other hand, a Question could be posed outside the scientific community, such as a request for evidence-based input to inform an impending policy decision. It might also simply express a wish to estimate more accurately certain quantities or parameters in current scientific theories.

2. *Study*: To address the Question scientifically, it is necessary to gather evidence in a rigorous manner. In the Study stage, we include all aspects of study design, including design of observational studies, experiments, questionnaires, systematic literature reviews, and meta-analyses. The Study may also involve sequential design or the design of a number of distinct, possibly concurrent studies. We also include in this stage the conduct of the study, resulting in some form of data.

3. *Interpret*: In the Interpret stage, data resulting from the Study are employed to address the Question. Typically, this may involve descriptive statistics and statistical inference, such as parameter estimation and hypothesis testing. In a Bayesian analysis, the primary statistical result might simply take the form of a posterior distribution. However, the "Interpret" stage should also embed the findings of the analysis into the wider body of science to which it refers, thereby updating that body of knowledge. In doing so, the wider implications of the findings will emerge.

4. *Inform*: The Interpret stage will often suggest new Questions, and a new cycle of scientific investigation will thereby be initiated. First, however, the Interpret stage usually will be followed by the Inform stage. For a formal scientific study, findings should be formally written and communicated in peer-reviewed outlets, such as conferences, journals, and books. In fact, peer-review may lead to revisions in the Interpret stage before formal publication of the findings. Subsequent examination and evaluation of published studies by the scientific community may in turn lead to new interpretations of existing studies and a new Question, leading to a new Study. Where the Question is a request for input to a decision, the Inform stage is when the results of the Study are communicated to facilitate the decision-making process. New Questions may arise based on the output produced in the Inform stage. Alternatively, the original Question may need to be revisited in one or more future Studies, especially when the evidence is not yet adequate to merit a robust conclusion.

3. Science and Subjectivity

In all stages of the scientific method defined in Section 2, expert judgment and knowledge are required. The relevant definitions are first presented. First, knowledge is defined (Oxford University Press 2018b) as, "facts, information, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject." Judgment is defined (Oxford University Press 2018a) as, "the ability to make considered decisions or come to sensible conclusions." By contrast, opinion (Oxford University Press 2018c) is, "a view or judgment formed about something, not necessarily based on fact or knowledge."

We consider that judgment implies using information and knowledge to form an assessment or evaluation. While it is possible that opinions can be similarly well-informed, by definition, opinion does not necessarily imply that external information was incorporated into the evaluation. It may simply be a view, judgment, or appraisal that fits within a belief system or is com-

fortable for some other reason. While "judgment" is normally thought of as being based on observable truth, "opinion" can be simply based on preference. The subjectivity in this article refers to judgment of experts based on knowledge, skill, and experience. We examine the roles of expert judgment in each of the four stages presented in Figure 1, paying particular attention to integration of statistical and *content* expertise. We use content expertise to refer to the discipline expertise in which the Question arises. For more detail on expert scientific judgment, please see Brownstein (2018).

3.1. The Question Stage

When developing the Question, we rely heavily on the expert judgment of the content experts. The Question may arise from identification of a barrier or problem in need of an answer, a quest to understand the "why" behind some phenomenon, event, or process, or simply to better quantify some parameter, effect or disturbance. For example, one might ask "Why are some people able to fend off the negative impacts of an HIV infection while others cannot?" Knowledge of the literature and what other experiments or studies others have done to address this Question or related Questions is critical. Where the Question arises from a request for scientific input to inform a decision, the content experts serve key roles in formulating specific questions, such as "What can we say about the toxicity of this pollutant for fish in European rivers?", or "Which areas in this catchment will be flooded if the catchment experiences a once-in-100-years weekly rainfall?"

While the content experts have primary responsibility to develop the Question, statisticians can elicit clarity on the framing of the Question by asking pertinent questions from their perspective. Inquiry from the statistician serves not only to establish and confirm the statistician's understanding of the Question and its scientific context, but also to translate the question to a statistical framework, which may guide analytic decisions in the Study Stage. Indeed, strong listening and communication skills are critical for both the content and statistical experts! Barry Nussbaum, 2017 President of the ASA, shared his mantra, "It's not what you said. It's not what they heard. It's what they say they heard" (Nussbaum 2017). Communication can be improved by echoing back your understanding of how you interpreted what you heard, and asking others to echo their understanding of the points you have made. Producing a written summary of the Question, evidence-based justification for the importance of the Question, and evidence needed to answer the Question allows the research team members a chance to check on how well they are communicating and where points of disagreement may still exist.

If the Question is seeking information on potential pathways or mechanisms, evidence for or against competing theories is presented, and decisions must be made on the rationale for how the Question will be pursued. The content experts play the main role here, but statistical expertise can be helpful when framing the Question to bring up potential statistical issues with the proposed approach.

Part of defining the Question is determining what evidence, measures and parameters would be useful and adequate to arrive at an answer. Properties of those measurements, including validity, reliability, cost, and distributional properties are considered. This is an area in which both content and statistical experts can contribute. For example, content experts may be focused on what data would be needed and whether primary data collection would be needed or existing secondary data would suffice. If the study needs to collect primary data, there will be a need for discussion of exactly what information will be desired and how to elicit that information. In turn, the statistician will seek to better understand properties of the desired measurements. Moreover, the statistician may highlight unmeasured influences, the impact of missing data, and whether sources of bias, confounding, or variability could be reduced or eliminated by appropriate study design or data collection methods.

In addition, a Study may involve more than one Question. In this case, discussions are needed regarding which Questions are considered primary or secondary, the interrelatedness or independence of the Questions, and whether any of the Questions can be addressed jointly. Statistical and content experts should collaborate in making these decisions.

Thinking through the issues that can (and will!) arise when defining the Question and information needed to address it requires effective collaboration and team effort. Effective collaboration requires mutual respect for team members and personal authenticity; an understanding of the strengths and skill sets of each member; an understanding and buy-in for a common goal; a willingness not only to hear and understand what another has said, but to embrace ideas different from yours; a willingness to ask for evidence and assumptions behind a statement of fact or belief; reliability and consistency in thought and behavior; a willingness to offer alternative ideas or approaches; a willingness to compromise, when appropriate. All of this rely on fluency, both written and oral, in the language used, including technical and discipline-specific terms. Cognizance of acronyms and language shortcuts is the first step to reducing them to enhance communication. We recommend that researchers create a written summary of decisions made in such discussions, and make sure such documents are accessible electronically to all study members for review and editing.

3.2. The Study Stage

Once the Question and pertinent measures are defined, the approach to gathering information (data) must be developed. Can the Question be answered in one study, or will it take a series of planned studies? What studies have been done before? Could any component of prior studies be replicated with modification or improvement in this study? Did previous studies report unanticipated problems that could be avoided in this study?

Much of the work in planning a study involves a collaborative effort among all members of the research team. Those researchers who will be "on the ground" collecting information will have expertise on what is feasible and what is not. The statistician can offer their expertise and judgment on many aspects of study design and implementation, such as strengths and weaknesses of different study designs, questionnaire development, psychometric properties of data collection instruments, issues surrounding sources and control of error, replication,

operational randomization, and blinding. The content experts will share their knowledge and judgments on the target scope of inference, such as whether to estimate the 100-year flood plane for a large geographic area or just one river; for all people with a disease or only those in a local area who meet defined inclusion criteria. Content experts bring to the discussion additional information that can be considered in the study design, perhaps stemming from theorized or understood pathways, mechanisms, or concurrent influences by which observed outcomes can differ.

If working within the Bayesian framework, statisticians help elicit information from the content experts to feed into the prior distribution. Those approaching the problem from a frequentist perspective will also be looking to prior studies and expert judgment when developing a study design. No matter what analysis approach is used, the ultimate goal remains the same: to collect enough high quality information to effectively address the Question. Here, the word "quality" can encompass many aspects, including reduced variability of measures and removal or control of sources of biases, along with proper data collection and maintenance systems that protect the integrity of the data.

During the Study stage, prior to data collection, there should be enough information to create a study protocol, data monitoring plan, and a statistical analysis plan (SAP), upon which everyone agrees (Finfer and Bellomo 2009; Ellenberg, Fleming, and DeMets 2003; Ott 1991). Such SAPs are becoming more common and recognized by funders as an important part of the research process. While not every study will require these formal documents, the idea is that there is a common understanding on how the study will be conducted; how the data will be collected, monitored, and maintained; and the analytic approach used to address the Question. With the goal of transparency of the body of work to be accomplished, good documentation and data provenance are important components of scientific inquiry.

Unfortunately, forms of scientific malpractice, such as repeating the analysis with a slightly altered question in an attempt to achieve statistically significant results—a practice called *HARKing* (hypothesizing after the results are known)—or testing for many associations without prior hypotheses (p-hacking) are commonplace in the scientific literature (Kerr 1998; Head et al. 2015). A well thought-out study design and SAP can help safeguard against urges to reanalyze the data later after obtaining disappointing results. However, it should be noted that the data monitoring process itself involves judgment, as exemplified in Section 4.1.3 and discussed further elsewhere (Pocock 2006).

A related practice of publishing a Registered Report, in which methods and proposed analyses are preregistered and peer-reviewed prior to conducting the research, is gaining momentum. The intent is to eliminate questionable research practices, such as selective reporting of findings and publication bias, by provisional acceptance for publication before data collection begins. Currently, the Center for Open Science lists 121 journals that use the Registered Reports publishing format as an option in regular submissions or as part of special issues (Center for Open Science 2018).

Often, before a study can begin, funding must be obtained. When there are commissioned studies, often researchers need to write a formal proposal to seek funding. A grant proposal provides an opportunity to present not only the justification for the study but also detail on the study process, analysis plan, and how the desired information will address the Question. The grant review process brings in a new external set of experts with judgments of their own on the merits of the proposal. Feedback for a proposal can identify strengths and weaknesses of the proposed work. Reviewers' expert judgments on whether a particular study should be funded are intended to weed out studies without strong support and justification for both the importance of the Question and the development of the Study.

As the study progresses, important discussions and decisions are likely to be made at research team meetings. It is important that statisticians, like other key members of the team, are present at the table to listen, question, offer insight and expertise, and retain knowledge of decisions made. Written summaries of research meetings, documentation of any decisions made that were not anticipated in the Question or (earlier in the) Study stages or are necessary to successfully implement the study protocol, all promote transparency and reproducibility. Additional documentation for the Study stage that should be available for all team members to view may include computer code management systems, clearly documented analytic code, well-considered file structures and naming conventions, and common file space for all members to see such documentation. Subsequently, the team will synthesize information from the Study stage for use in the Interpret and Inform stages.

3.3. The Interpret Stage

The methodology used to analyze the data and the specifications of the assumed model feed into the interpretation of the data. Some analytic methods produce estimated probabilities that relate directly to the Question; others may provide information that helps address the Question, but perhaps more indirectly. The chosen methodology may produce parameter estimates that need to be understood in context of the model, and the alignment of the model to the observed data and prior beliefs needs to be considered.

The expertise of the statistician is needed both to understand the nuances of proper interpretation of the analytic results in context of the executed study, assumptions made, and modeling used and to guard against overinterpretation. For example, a statistician may help protect the team from common misconceptions and malpractice, such as tendencies to extend inferences to populations outside those under study or to interpret association as causation, or failure to consider the impact of unmeasured confounders, mediators or moderators in the interpretation of results. The statistical expert can protect against improper use or interpretation of a p-value, discuss the difference between clinical and statistical significance, and highlight the potential impact of missingness mechanisms and violations of statistical and causal assumptions on the results. If working in a Bayesian framework, the statistician can also discuss the impact of potential expertise bias and over or underprecision in specification of the prior.

Similarly, it is not appropriate for statisticians to focus solely on analytic methods and numeric results, or for content experts to delay involving statisticians until the Interpret stage. By not being involved in other study aspects and discussions, the statistician is poorly positioned to make modeling choices, recognize

possible biases, and interpret findings in the context of the study as conducted, which could differ from the study as planned.

As the analytic results are interpreted in the framework of the Question and the Study protocol, both the content and statistical experts help the study team blend the (properly interpreted) new findings into their existing knowledge and understanding. This process will likely include team discussions of the clinical meaning and impact these findings might have in the population under study, and how the results could be explained or interpreted within the current framework of understanding. The discussion will include the strength of the evidence, based on posterior distributions, point estimates of key parameters, confidence or credibility intervals, etc. and consistency with prior studies, as well as potential weaknesses or caveats of the study posed by both the statistical and content experts.

3.4. The Inform Stage

Once the analytic results have been interpreted within the framework of the study design and measures used to address the Question, it is time to assess what was learned and share that information more broadly. Of course, those who developed the Question and often those who funded the study will need and expect a complete summary of the work done, describing how the results have informed the Question. Indeed, there may be interest in the work outside of academia, such as patients curious about new therapies for their conditions, policy-makers seeking to understand actions that may yield societal benefit, and others simply wondering about current topics and trends in various fields of science.

Scientific reports or publications, where the process, methods, results, and conclusions of a study can be shared broadly, are important tangible outcomes of the Inform stage. In each Inform stage, the team members build on what they knew previously and share what they learned from the findings, whether or not the results obtained were anticipated or exciting. To guard against publication bias, null results, in particular, must be communicated, despite disincentives for doing so (Franco, Malhotra, and Simonovits 2014; Easterbrook et al. 1991).

Strong communication and documentation skills in the Inform stage are paramount to ensure that all components of a study are documented as the study progresses and are presented clearly, completely, and with as much transparency as possible. While the statistician is frequently tasked with managing the sections describing the quantitative methods and results, they should also collaborate throughout the report to provide their input to interpretation and implications of the findings (see Section 3.3). Additional documentation from the Study stage, such as clearly written and readable computer code, can also be included as supplementary material.

The process of creating a scientific manuscript and undergoing the peer-review process for publication is another place in which statistical and content expert judgment enters into both the Interpret and Inform stages. Comments from others on drafts of the manuscript can lead to revised interpretation in light of new information or perspective from that feedback before submission of the manuscript for publication. Once submitted, additional expertise is gathered from the peer reviewers, such as calls for clearer evidence of claims or interpretations made, challenges made to stated justifications of assumptions or interpretations, call for greater transparency or detail in information presented, or citations to related work which may blend into the discussion section. Reviewer comments can greatly impact the quality and clarity of information presented in the final publication.

Recently, journal editors are considering re-examining their review processes, focusing acceptance or rejection on strength of evidence, not *p*-values. Locascio (2017b) argues for a results-blind review system in which the reviewer makes a preliminary decision on the strength of the manuscript without knowing the results. Final decisions would not allow rejection on the basis of *p*-values. Commentary on the feasibility of this approach may be found elsewhere (Marks 2017; Grice 2017; Hyman 2017; Locascio 2017a). We encourage journal reviewers to examine the appropriateness of the methods for the study under consideration, rather than accepting justification of methods simply based on their publication and use elsewhere.

When the Question has arisen to facilitate decision-making, the primary purpose of the Inform stage is to convey the scientific evidence to the decision-maker, after it has been assembled and analyzed in the Study and Interpret stages. Here, too, communication skills are particularly important. Governments are increasingly but not exclusively reliant on evidence for policy- and decision-making (Oliver and de Vocht 2017; HM Treasury 2015; Oliver, Lorenc, and Innvær 2014; LaVange 2014), and there is much current interest in the challenges of communicating uncertainty to decision-makers (Cairney and Oliver 2017; National Academies of Sciences, Engineering, and Medicine et al. 2017; National Research Council et al. 2012a,b). Despite these challenges, it is essential to measure uncertainty and important to try to communicate it as effectively as possible. Understanding the perspectives of decision-makers (i.e., priorities and goals, options under consideration, risks/benefits), what processes they must follow, and time-constraints are helpful in such communications.

As depicted in our graphic of the scientific method (Figure 1), learning and discovery is cyclical. Once we address one Question, new Questions often arise. Researchers may be interested in the work of other authors in similar fields, perhaps using information from published studies to inform their next study. Sometimes the results obtained are not definitive, or not adequate for robust decision-making, and ways to redirect the next investigation of the same or a revised version of the Question are planned. Other times, a replication Study is conducted based on the same Question simply to see if the results remain qualitatively similar, despite inevitable lack of perfect duplicability in all aspects of the study environment (Lindsay and Ehrenberg 1993).

3.5. But Is It Science?

When expert judgments are used in any stage of a scientific inquiry, the outcomes contain subjective elements. The inescapable conclusion is that science itself has a subjective component, aspects of which can be communicated probabilistically, and should be interpreted according to the theory of subjective probability (Anscombe and Aumann 1963). That is, while there may be objective information on which probabilities are

based, there also will always be between-individual (scientist) variation.

There can be heated opposition to the notion of subjective science, with objectivity promoted as fundamental to the scientific method, and subjectivity considered as anathema to a true scientist. According to this viewpoint, subjectivity is unscientific, biased, possibly prejudiced. But, science cannot be totally objective. Scientists propose new or amended theories, choose experimental designs or statistical analysis techniques, interpret data, and so on. Although these judgments are subjective, expert judgments result from a synthesis of related prior knowledge and experiences based on observable evidence and careful reasoning. Indeed, making informed subjective judgments is one of the key features that distinguish a top scientist from a lesser one. Statistics similarly involves subjective judgments, as others have recently argued (Gelman and Hennig 2017).

In the design portion of the Study stage, by definition, information is not yet available from the study being planned. Instead, study design decisions are *preposterior* or *preanalysis* and must be based on (prior) external data and judgment. In this sense, regardless of the subsequent analysis, one could consider the design phase as automatically Bayesian. Designers employ varying degrees of formalism in developing the study design and statistical models. A formal Bayesian approach can be used to either develop a frequentist design (Bayes for frequentist) by, for example, finding a sample size or other design components that ensure

$$P(\text{power} > \text{goal} \mid \text{design \& assumptions}) > \gamma,$$

(see e.g., Shih 1995 for an implementation), or to ensure that Bayesian properties are in an acceptable region (Bayes for Bayes). We recommend practitioners consider additional use of Bayesian approaches, even if only to provide a vehicle for documenting the roles of judgments and as a platform for sensitivity analyses.

All researchers, irrespective of their philosophy or practice, use expert judgment in developing models and interpreting results. 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.

Science also demands transparency. Ideally, it should be standard practice in all scientific investigations to document *all* subjective judgments that are made, together with the evidence and reasoning for those judgments. Although we recognize that this may be impractical in large studies with many investigators, we believe that it might be facilitated with suitable collaborative working tools. We suggest, therefore, that journals might consider requiring that authors provide such documentation routinely, either in an appendix to a submitted paper or in online supplementary material.

4. Case Studies

We present three case studies, each exhibits complexities that require collaborative input from content experts and statisticians. They highlight the roles that expert judgments play and the steps that were taken to ensure that judgments were as objective and scientific as possible.

4.1. Expert Judgment in Randomized Clinical Trials

4.1.1. Background on Bayesian Clinical Trials

Bayesian approaches to clinical trial design, conduct and analysis have been shown to offer substantial improvements over traditional approaches in a variety of contexts. See Abrams, Spiegelhalter, and Myles (2004) and Berry et al. (2010), for a range of examples. The Bayesian approach formalizes using prestudy data and expert judgments in design, conduct, and analysis. Importantly, the formalism provides an effective language for discussing interim and final results, for example, by supporting statements such as, "in the light of accruing data, the probability that treatment A is better than treatment B by at least 5 percentage points is 0.95 …" It provides a natural way to compute and communicate predictive assessments such as futility, that is, in the light of current information, what is the probability that, ultimately, the trial will not be definitive. The Block HF and TOXO studies outlined below illustrate many of these characteristics.

4.1.2. The Block HF Study

The BlockHF study (Curtis et al. 2013) provides an example of the benefits of embedding evaluations in the Bayesian formalism. It was an adaptive randomized trial using Bayesian criteria, with specification of all features dependent on collaboration amongst clinical and statistical experts. The abstract states,

> We enrolled patients who had indications for pacing with atrioventricular block; New York Heart Association (NYHA) class I, II, or III heart failure; and a left ventricular ejection fraction of 50% or less. Patients received a cardiac-resynchronization pacemaker or implantable cardioverter defibrillator (ICD) (the latter if the patient had an indication for defibrillation therapy) and were randomly assigned to standard right ventricular pacing or biventricular pacing. The primary outcome was the time to death from any cause, an urgent care visit for heart failure that required intravenous therapy, or a 15% or more increase in the left ventricular end-systolic volume index.

Two pacing regimens, biventricular (BiV) versus right ventricular (RiV) were compared using a Cox proportional hazards model. Analyses were stratified by two cardiac devices, with information on the two, stratum-specific hazard ratios (HRs) combined for monitoring. The Bayesian monitoring rules addressed patient safety, stopping for futility, and stopping if the treatment comparison was sufficiently convincing (probability that a clinically meaningful difference was sufficiently high), based on combining evidence over the two stratum-specific HRs. From the statistical analysis section,

> An adaptive Bayesian study design allowing up to 1200 patients to undergo randomization was used, featuring sample size re-estimation and two interim analyses with prespecified trial-stopping rules …. An intention-to-treat analysis served as the primary analysis for all outcomes.

Table 1. θ = a weighted average of stratum-specific log hazard ratios, each comparing BiV versus RiV pacing; PP_0 = posterior probability that the study objective has been met, PRR = posterior probability the study objective has not been met.

	Decision boundaries				
	Conclude objective is met and stop study early	Conclude that sample size is sufficient to continue	Determine that sample size is insufficient but elect not to increase sample size	Conclude that sample size must be increased in increments of 175	Stop study for safety
First Interim Analysis	$PP_0 > 0.99$	$0.90 \leq PP_0 \leq 0.99$	$PRR > 0.9$	$PP_0 < .90$ and $PRR \leq 0.9$	P ($\theta > 0$|data, prior) ≥ 0.90
Sample Size Re-estimation Phase	N/A	$0.90 \leq PP_0$	$PRR > 0.9$	$PP_0 < .90$ and $PRR \leq 0.9$	N/A
Second Interim Analysis	$PP_0 > 0.99$	If neither the outcome in column 2 nor the outcome in column 6 occurs, then the study will continue with the current sample size.			P ($\theta > 0$|data, prior) ≥ 0.90

NOTE: See Curtis et al. (2013) for details.

The trial resulted in a "win" for biventricular pacing compared to right ventricular, HR 0.74, 95% credible interval (0.60, 0.90). Table 1 presents the monitoring rules which are based on the following,

- θ = a weighted average of stratum-specific log HRs, each comparing BiV versus RiV pacing.
- $PP_0 = P\{\theta < \log(0.90)|\text{data}\} = \text{pr}\{\text{HR} < 0.90|\text{data}\}$; the posterior probability that the study objective has been met.
- $\text{PRR} = P\{\theta > \log(0.90)|\text{data}\} = \text{pr}\{\text{HR} > 0.90|\text{data}\}$; the posterior probability that the study objective has not been met.
- $P\{\theta > 0|\text{data}\} = \text{pr}\{\text{HR} > 1.00|\text{data}\}$; the posterior probability of a safety concern.

For example, if the probability of meeting the study objective is sufficiently high ($PP_0 > 0.99$), stop the study; if moderately high ($0.90 < PP_0 < 0.99$), continue the study with the current sample size target; if too low ($PP_0 < 0.90$) but there is a reasonable likelihood of success ($\text{PRR} < 0.90$), increase the sample size. Inclusion of such probability-based rules was an important benefit of using the Bayesian formalism. It supported complex decision rules that communicated effectively with the clinical experts on the monitoring board.

4.1.3. The "TOXO" Trial

When accruing data are consistent with a prior distribution, a trial can be stopped earlier than with traditional, likelihood-based monitoring. However, there are also situations wherein the prior and the data diverge and stopping is delayed, as is the case in the following, post hoc, rerunning of the Community Programs for Clinical Research on AIDS randomized trial of prevention of TOXO conducted in the 1990s. The study compared pharmacological prevention at a subtherapeutic dose with placebo, with all treatment groups carefully monitored for indications of disease. The premise of such prevention studies is that a low dose of a pharmaceutical that is typically used to treat overt disease may also prevent or delay onset. However, even a low dose of a pharmaceutical may induce adverse effects, such as toxicities or resistance to treatment, and consequently, watchful waiting (the placebo "intervention") may be better than the potentially preventive treatment.

Jacobson et al. (1994) detail the several decisions and expert judgments needed to design and implement the trial, including choice of drugs, inclusion and exclusion criteria, clinical endpoints, and monitoring and analysis plans. The abstract of the article reporting results (Jacobson et al. 1994) states in part,

> Pyrimethamine, 25 mg thrice weekly, was evaluated as primary prophylaxis for toxoplasmic encephalitis (TE) in a double-blind, randomized clinical trial in patients with human immunodeficiency virus (HIV) disease, absolute CD4 lymphocyte count of < 200/microL [CD_4 lymphocytes fight disease; a low level indicates immunodeficiency] (or prior AIDS-defining opportunistic infection), and the presence of serum IgG to Toxoplasma gondii." …"There was a significantly higher death rate among patients receiving pyrimethamine [compared to control] (relative risk [RR], 2.5; 95% confidence interval [CI], 1.3–4.8; P = .006), even after adjusting for factors predictive of survival. The TE event rate was low in both treatment groups (not significant). Only 1 of 218 patients taking [the control intervention] but 7 of 117 taking aerosolized pentamidine for prophylaxis against Pneumocystis carinii pneumonia developed TE (adjusted RR …, 0.16; 95% CI, 0.01–1.79; P = .14). Thus, for HIV-infected patients receiving trimethoprim-sulfamethoxazole, additional prophylaxis for TE appears unnecessary.

The Data and Safety Monitoring Board monitored the trial at prespecified, calendar-determined dates using the O'Brien and Fleming (1979) boundaries. Early in the trial, these boundaries require substantial evidence (e.g., a small p-value) to stop and make a decision; as the trial approaches the predetermined maximum number of follow-up visits, the criterion is close to that for a fixed sample size trial.

The full, calendar time indexed database was available for the after-the-fact example of how monitoring might have proceeded using a Bayesian approach. This illustrative analysis evaluated the "TOXO or death" endpoint using the Cox (1972) proportional hazards model with adjustment for baseline CD_4 count. The illustrative trial was stopped when the posterior probability of benefit or the posterior probability of harm became sufficiently high. Importantly, prior elicitation occurred while the trial was being conducted, before any outcome information was available to the elicitees.

4.1.4. Model for the "TOXO" Trial

The HR (relative "TOXO or death" event rate between the two treatment groups) was modeled using a log-linear model with covariates treatment group ($z_{1j} = 1$, if participant j received pyrimethamine; $z_{1j} = 0$, if placebo), and CD_4 cell count at study entry (z_{2j}). The CD_4 covariate adjusted for possible differences

in immune status at baseline between the two groups. Specifically, the log hazard-ratio is,

$$\log(\text{hazard ratio}) = \beta_1 z_{1j} + \beta_2 z_{2j}$$

with $\beta_1 < 0$ indicating a benefit for pyrimethamine. A flat prior was used for the CD_4 effect (β_2), and a variety of priors were developed for the pyrimethamine effect (β_1). The choice of the Cox (1972) model and the use of a noninformative prior for β_2 were judgments of the statisticians. Though the Cox model has become the default choice when modeling the time to an event, it is based on important assumptions, namely that the hazard functions for both interventions are proportional and censoring is noninformative. As such, the model should only be adopted after careful consideration by experts, such as in this example. The choice of a noninformative prior distribution for β_2 reflects the clinicians' and statisticians' judgments that no information was available in advance of the trial on the association of CD_4 with TOXO incidence.

4.1.5. Elicitation in the "TOXO" Trial

As described in Carlin et al. (1993) and Chaloner et al. (1993), prior distributions for β_1 were elicited from five content experts—three HIV/AIDS clinicians, one person with AIDS conducting clinical research, and one AIDS epidemiologist. Elicitation occurred while the trial was in progress, only pretrial information was available to the elicitees. Two additional priors were included in the monitoring—an equally weighted mixture of the five elicited priors and a noninformative flat prior (equivalent to using the normalized, partial likelihood as the posterior distribution).

Elicitation targeted potentially observable, clinically meaningful features and then transformed responses to parameters of the Cox model. That is, rather than directly elicit a prior for the hazard ratio, each elicitee was asked to report their best estimate of the probability of TOXO or death in two years under placebo (P_0), then draw a picture of the distribution of the two-year probability under pyrimethamine, conditional on the estimate under placebo ($[P_{\text{pyri}} \mid P_0]$). Then, for each elicitee these conditional distributions were converted to a prior distribution for the log(hazard ratio) using the transformation, $\beta_1 = \log(1 - P_0) - \log(1 - P_{\text{pyri}})$.

Figure 2 displays the elicitation results. At trial initiation, there was little known about the baseline incidence of TOXO, and so the content experts based their distributions on general expertise and analogy with other endpoints and contexts. The range of the five reported two-year incidence probabilities under placebo was wide, but elicitees believed that TOXO had a non-negligible baseline incidence. All elicitees were optimistic regarding pyrimethamine's benefit, with experts C and E the most optimistic, placing all of their probability distribution for incidence under pyrimethamine way below their estimate of incidence under placebo.

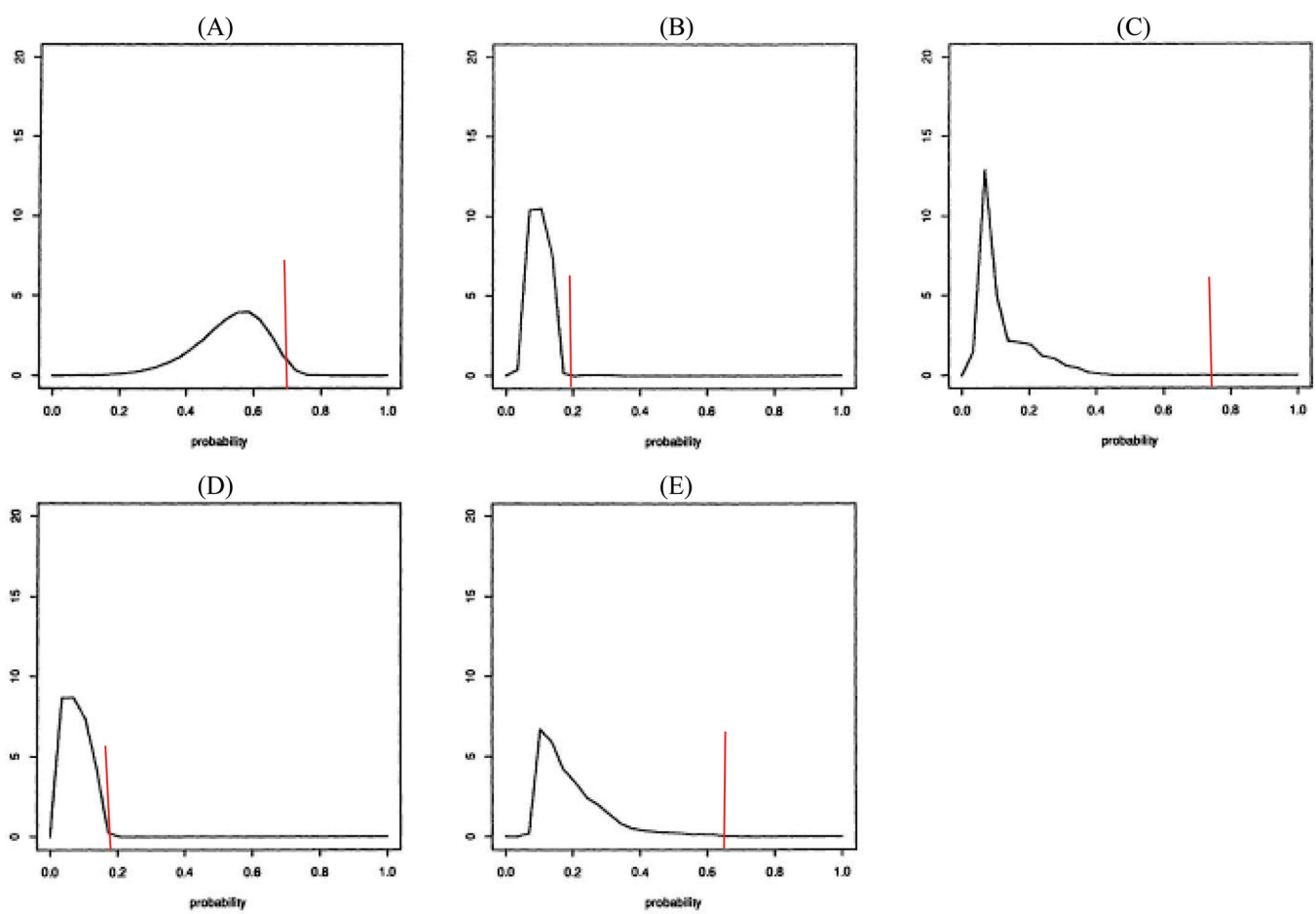


Figure 2. Elicited priors for the five elicitees. The red vertical line at P_0 is the "best guess" two-year incidence of TOXO or death under placebo. The smoothed densities are for two-year TOXO or death incidence under pyrimethamine, conditional on the placebo rate.

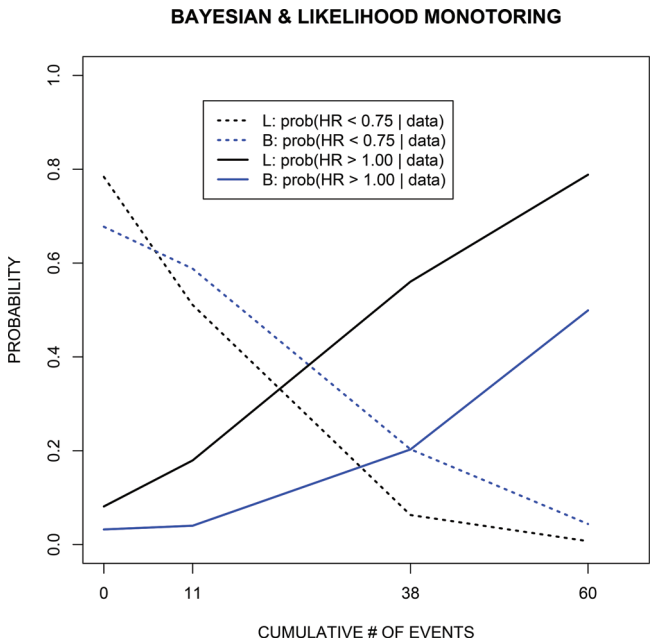


Figure 3. Posterior probability of benefit (hazard ratio $\leq$ 0.75) and harm (hazard ratio $>$ 1.0) for the mixture prior (blue lines) and for monitoring based on the partial likelihood (black lines), which is equivalent to using a flat (improper) prior. Monitoring was at calendar dates (15 Jan 1991, 31 Jul 1991, 31 Dec 1991, 30 Mar 1992) with the X-axis indicating the cumulative number of toxoplasmosis or death events from the combined arms.

While a degree of optimism may be needed to motivate conducting a trial, ethics require that there be sufficient equipoise (uncertainty) regarding which treatment is superior in order to initiate a trial. The elicited priors in this example probably express too strong a belief in the efficacy of pyrimethamine to be used in monitoring an actual trial, but using them in this illustrative monitoring exercise with comparison to likelihood-based monitoring, effectively highlights issues we discuss in Section 4.1.7.

4.1.6. Monitoring Results for the "TOXO" Trial

The actual trial was monitored at calendar dates (15 Jan 1991, 31 Jul 1991, 31 Dec 1991, 30 Mar 1992). At the December 31, 1991 meeting, the monitoring board recommended stopping the trial for futility, because the pyrimethamine group had not shown significantly fewer events, and the low overall event rate made a statistically significant difference in efficacy unlikely to emerge. Additionally, an increase in the number of deaths in the pyrimethamine group relative to the placebo indicated a safety issue.

For the illustrative, after the fact monitoring example, Figure 3 displays posterior probabilities of benefit (HR $\leq$ 0.75, equivalently, $\beta_1 \leq \log(0.75)$ in the Cox model), and harm (HR $>$ 1.0; $\beta_1 > 0$) for an equally weighted mixture of the five elicited prior distributions (denoted by "B"), and for a flat prior that generates partial likelihood-based/traditional monitoring (denoted by "L"). For example, $\beta_1 = \log(0.75)$ indicates that the hazard rate (event rate) in the pyrimethamine group is 75% of that in the control group; $\beta_1 = 0 = \log(1.0)$ indicates equal rates.

As displayed in Figure 2, elicitees believed that TOXO had a nonnegligible incidence and that pyrimethamine would have a substantial prophylactic effect. Thus, each prior and their mixture are to varying degrees far from the accruing information in the trial. Consequently, monitoring based on the partial likelihood, which can be considered "flat prior Bayes," gives an earlier warning of harm compared to monitoring based on the mixture prior. The mixture required considerably more information to overcome the a priori optimism of the elicitees. Also, the posterior probability of harm based on monitoring with any single prior in Figure 2 would lag behind that based on the partial likelihood, with use of prior A, B, or D giving an earlier warning than use of priors C or E.

4.1.7. Discussion of the "TOXO" Trial

Traditional, non-Bayesian monitoring depends on several judgments. For example, trial design including maximum sample size depends on assumptions about baseline event rates, the HR, principal efficacy and safety endpoints, the statistical model, etc. In the Bayesian approach, a subset of these features are given prior distributions, in the TOXO example by independent priors for the slopes on the log HR and on the baseline CD_4 value. These priors, along with other design features determine the monitoring frequency, the shape of monitoring boundaries (flat, increasing, decreasing) and associated values. Boundaries are necessary for all monitored trials, determining them can be either Bayesian or frequentist.

The after-the-fact analysis highlights ethical issues generated by real-time use of expert knowledge. It shows that if the elicited priors had been used in the actual clinical trial monitoring, then due to their optimism it is very likely that trial stopping and other decisions would differ from those produced by likelihood-based (flat prior) monitoring. However, in early phase clinical applications, and in nonclinical applications, use of such optimistic priors can be appropriate and effective.

In the example, stopping would have been delayed, highlighting the question of whether it would be ethical to continue beyond a traditional stopping point due to prior beliefs that pyrimethamine would have a strong, prophylactic effect. Of course, if pyrimethamine had performed well, stopping would have been earlier than that based on traditional methods, also raising an ethical issue for some. And, if the priors had been pessimistic, but the data moderately optimistic, the Bayesian analysis would have tempered enthusiasm (delayed stopping) until convincing, positive, data had accrued.

The TOXO example shows one possible effect of using prior distributions in clinical evaluations designed to be definitive (Phase III). The example is based on the mixture that equally weighted the five prior distributions. The posterior distribution is also a mixture, but the posterior weights give more influence to the priors that are most compatible with the data, so there is some degree of automatic adjustment. Other options include separately monitoring with each prior and then at each monitoring point determine stopping based on the "majority rule" or requiring unanimity. Each choice will produce its own operating characteristic, so extensive simulations are needed to understand properties.

To summarize, in clinical trial monitoring prior information can have two main effects:

1. If prior information conflicts with the available data at the time of monitoring, it may suggest continuing the trial, at least until the next review. This situation can arise when the experts are more optimistic than the emerging data, as in the TOXO trial, or when they are more pessimistic. In either case, continuing the trial long enough to obtain sufficient evidence to convince the content experts may have an added benefit that the results would translate to practice relatively rapidly.

2. If prior information supports the available data at the time of monitoring, it may suggest terminating early, on the grounds that sufficient evidence now exists to make a decision.

An increasing number of trials are utilizing prior information (Abrams, Spiegelhalter, and Myles 2004; Berry et al. 2010), and the advent of such trials emphasizes the importance of prior judgments being made as carefully and as objectively as possible. Judgments should be sought from a sufficient number of content experts to properly represent the range of views. Determining the number and type of elicitees is, itself, an exercise in study design. Of course, prior distributions should be elicited carefully and rigorously, as described, for instance, in O'Hagan (2018).

4.2. Expert Judgment in Environmental Modeling

4.2.1. Background: UK Carbon Flux

This case study describes the estimation of a parameter in a complex environmental modeling problem. As a party to the Kyoto protocol, the United Kingdom is committed to specific target reductions in net emissions of greenhouse gases. Monitoring progress toward these targets is challenging, involving accounting for numerous sources of emissions. One potential mitigating factor is the ability of vegetation to remove carbon dioxide from the atmosphere, thereby acting as a "carbon sink." However, accounting for this is also extremely complex. During the day, through the process of photosynthesis, vegetation absorbs carbon dioxide and releases oxygen, using the carbon to build plant material. Photosynthesis requires sunlight, chlorophyll in green leaves and water and minerals gathered by the plant's roots, so the efficiency of carbon removal depends on factors such as weather, leaf surface area, and soil conditions. Conversely, carbon is released at night, carbon extracted from the atmosphere and converted to biomass will eventually be released as the plant ages and dies, and carbon in leaf litter is released by microbial action in the soil. The Sheffield Dynamic Global Vegetation Model (SDGVM) was built with mathematical descriptions of all these processes to predict net carbon sequestration due to vegetation (Woodward and Lomas 2004). For a given site, the model takes many inputs describing the type of vegetation cover and soil at the site, together with weather data, to estimate Net Biosphere Production (NBP), that is, the net decrease in atmospheric CO_2, from that site over a given time period.

This case study used SDGVM to estimate the total NBP for England and Wales in the year 2000. It is important to recognize that there is inevitably uncertainty about all the model inputs. Uncertainty about inputs induces uncertainty about model outputs, and the study sought to quantify the output uncertainty in the form of a probability distribution for the total NBP. Details are reported in Kennedy et al. (2008) and Harris, O'Hagan, and Quegan (2010).

Before considering the statistical aspects of this case study, we first note that considerable content expertise had already gone into the development of SDGVM. Based on the available scientific knowledge, expert judgments were made in choosing the structure of the model and the equations that describe each of its biological processes. Care went into these choices to ensure that they were reasonable and scientifically defensible. However, in such modeling, it is usually unrealistic to include every process to the greatest level of detail and complexity that is believed to be applicable. First, the more complex and detailed the model, the longer it will take to compute; for practical reasons, it may be necessary to compromise on complexity. Second, more complex models may be less robust and reliable in their predictions, because at the highest level of detail, there is often less scientific consensus about the equations and the parameters within them. In addition, models with a large number of parameters may suffer from overfitting (Hawkins 2004). Judgments about the optimal degree of complexity to achieve a computable, accurate, and reliable representation of the phenomenon being modeled often demand a particularly high level of expertise.

4.2.2. Input Distributions for the SDGVM

The model was run over 707 grid cells covering England and Wales. For each grid cell, given the land cover in that cell, and given input parameters describing the soil composition and properties of the vegetation types, the model was first "spun-up" for 600 years. That is, the model was initialized with a default set of inputs describing the state of the vegetation in terms of ages, heights, leaf density, etc., and the state of the soil in terms of moisture content, age and quantity of organic matter, etc. Then the model was run forward for 600 years using historic climate data at that site from 1400 to 2000, to stabilize the vegetation and soil at states representative of how that site would have been in 2000. The model was then run forward for one more year using weather data from 2000, and the NBP for the year was computed for each grid cell. The NBP for England and Wales in 2000 is the sum of the NBP values across all the grid cells.

Care was taken to quantify the uncertainty in the various inputs as described previously (O'Hagan 2012). Briefly, details are described below.

- *Soil composition*: A publicly available soil map for England and Wales (Bradley et al. 2005) was used to provide estimates of the soil composition in each grid cell. Because the map was created at a higher resolution than the grid cells used for SDGVM, figures were averaged over each grid cell to provide an estimate. The variance of the same figures over a grid cell, divided by the number of map points in the cell, was used to quantify uncertainty around the estimates for a grid cell. However, the variance was increased by a factor to represent (a) additional uncertainty in the map data and (b) spatial correlation within the cell. The decision to use an increased estimate of uncertainty in the model was based on an expert judgment on the part of the statisticians, in consultation with content experts.

- *Land cover*: A map of land cover was also publicly available (Haines-Young et al. 2000), derived from satellite observation. The original analysis reported in Kennedy et al. (2008) did not quantify uncertainty in land cover. However, unlike the soil map, the content experts felt that the uncertainties in land cover were large and there could be biases in the process by which land cover is inferred from the satellite readings. In a subsequent analysis (Cripps, O'Hagan, and Quaife 2013; Harris, O'Hagan, and Quegan 2010), a statistical model was built to quantify uncertainty in land cover maps derived from remote sensing. The analysis of England and Wales NBP in 2000 was then extended to account for the additional uncertainty. The method makes use of the "confusion matrix," which for the Haines-Young et al. (2000) map was given by Fuller et al. (2002). The confusion matrix is a contingency table based on a large survey of actual, ground-truth, land cover, and shows, for each ground-truth vegetation type, the frequency with which it was classified by the Haines-Young et al. (2000) map in each vegetation type. The statistical analysis required probabilistic inversion of the confusion matrix to derive, conditional on the satellite land cover, the probabilities of the various ground-truth cover. The careful expert judgments of statisticians and content experts are delineated in Cripps, O'Hagan, and Quaife (2013).
- *Vegetation properties*: SDGVM classifies land cover into plant functional types (PFT). For England and Wales, we used four PFTs—evergreen needleleaf trees, deciduous broadleaf trees, crops, and grassland. Each PFT is associated with a set of quantities, including maximum age, stem growth rate and leaf lifespan. However, most of these inputs were missing. Some properties had been estimated experimentally, but only for very few individual species within a given PFT. We therefore used expert elicitation to construct a probability distribution for each parameter. Elicitation is an area where it is particularly important to take care to avoid biases that commonly arise in subjective judgments of probabilities (Kynn 2008). Another article arising from the Symposium addresses issues related to elicitation in detail (O'Hagan 2018).
- *Weather*: Weather data, such as temperature, precipitation, and cloud cover, were available for each grid cell for every day in 2000. There are no doubt errors in these data, due not only to errors in the underlying daily measurements, but also to the fact that those measurements have been interpolated to produce the data at the level of grid cells. Nevertheless, it was felt that uncertainty in these inputs was relatively small and could not be quantified without adding extra assumptions. The decision not to trade (potentially unreasonable) assumptions for (potentially increased) precision for weather data was a judgment made jointly by statisticians and content experts.

4.2.3. Propagating the Input Uncertainty in the SDGVM

To propagate input uncertainty through a mechanistic model such as SDGVM, the simplest and most direct approach is by Monte Carlo. A large sample of random draws are made from the probability distributions of the inputs, and the model is run for each sampled input set. The resulting set of outputs is then a random sample from the output uncertainty distribution. However, like many models that are built to describe complex physical processes, the computational load in running the SDGVM model is substantial, and it would not have been feasible to propagate parameter uncertainty through the model in this way at even one of the grid cells. The problem of quantifying uncertainty in complex computer models has acquired the name uncertainty quantification (UQ), and various tools are available to enable uncertainty propagation. The analysis used Gaussian process emulation (O'Hagan 2006; Oakley 2016), which is probably the most popular UQ technique. Even so, it would not have been feasible to build emulators at every one of the 707 grid cells; 33 were chosen by content experts to represent, in their expert judgment, the range of conditions across England and Wales, and GP techniques were adapted to infer the magnitudes of output uncertainty at the other 674 cells. The computational techniques and accompanying statistical theory to manage the analysis are set out in Kennedy et al. (2008) and Gosling and O'Hagan (2006).

4.2.4. Results of the UK Carbon Flux Study

The content experts were certainly interested in knowing how much uncertainty in the total NBP would be induced by uncertainty in the inputs. Their instinctive approach to estimating the total NBP was to run the model just once with all of the inputs set to their expected values; we call their idea the plug-in estimate. Yet, the NBP output from SDGVM is a highly nonlinear function of its inputs. Therefore, the expected value of NBP, when we allow for uncertainty in the inputs, is not equal to the plug-in estimate. Statisticians incorporated the input uncertainty and computed the expected NBP for England and Wales in 2000 as 7.46 Mt C (megatonnes of carbon). By contrast, the plug-in estimate was 9.06 Mt C. Not only is this a substantial difference, but the standard deviation was estimated as 0.55 Mt C. Therefore, the total NBP was probably in the interval (6.36 Mt, 8.56 Mt C) and very likely to be less than the plug-in estimate.

The result was very surprising for the content experts. It seems that the explanation lies in their estimates of vegetation properties. In effect, the experts had estimated values that were approximately optimal for the plants to grow and absorb CO_2. Any deviation from these values led to lower NBP. For the total NBP to be even close to the plug-in estimate required all the parameter values to be close to their estimates, a joint event that had very low probability.

In this case study, the combination of expert judgments from both content experts and statisticians, applied with as much care, rigor, transparency, and objectivity as possible, led to a scientific result that certainly highlighted the role of expert judgment and the statistical quantification of uncertainty, and also prompted new questions regarding the accuracy of current methods for carbon budgeting, with important implications for the science of global climate change. For example, see Cripps, O'Hagan, and Quaife (2013). The cycle of scientific investigation was thereby renewed.

5. Conclusions

Expert scientific judgment involves carefully considered conclusions and decisions based on deep knowledge and skills. The use

of expert judgment is essential in and permeates all phases of scientific and policy studies. Expert knowledge is information; to ignore it or fail to obtain it incurs considerable opportunity costs. Judgments should be as objective as possible and based on data when available. Anything less is unscientific. Yet, deciding what data are relevant always involves degrees of judgment.

Written documentation that logs decisions is critical for informing stakeholders of how and in which study components (design, conduct, analysis, reporting, and decision-making) the principal judgments had impact. By "impact," we mean that such judgments led to decisions that directly affected (or had potential to alter or influence) any aspect of how the study was conducted and/or results interpreted.

Elicitation of expert judgments to produce probability distributions that represent uncertainty about model parameters can be conducted informally, but such judgments are easily affected by unconscious cognitive biases, such as overoptimism or failure to recall all relevant evidence. When such distributions form an important part of a scientific activity, the expert judgments should be elicited scientifically and as objectively as possible, minimizing relevant sources of bias. Doing so requires a carefully designed process, an elicitation protocol, as fully discussed by O'Hagan (2018).

It should be unsurprising that statisticians have essential roles as scientists, ideally serving as leaders or coleaders in all study aspects. Indeed, virtually all aspects of a study have statistical content, though almost no aspects are solely statistical. Consequently, we advise statisticians to more pro-actively promote the needs for statistical principles to permeate all stages of research studies. Furthermore, we advise research authorities, such as journal editors and funding agencies, to recommend or even require thorough collaboration with one or more experts in statistics throughout the duration of all projects.

The 2017 Symposium will not produce a single position document on statistical practice like the "The ASA Statement on *p*-Values" that resulted from the 2015 ASA Board meeting (Wasserstein and Lazar 2016). However, we echo the call in Gibson (2017) for statisticians to better advocate for the importance of their involvement throughout the scientific process. Finally, we applaud stakeholders, such as the National Institutes of Health (Collins and Tabak 2014) and the American Association for the Advancement of Science (McNutt 2014) in the USA, and National Institute for Health and Care Excellence (NICE 2012) and HM Treasury (2015) in the UK for leading calls for increased statistical rigor and understanding of uncertainty. We encourage additional funding agencies, journals, hiring and promotion committees, and others to join in the call for higher scientific standards, statistical and otherwise. Science in the twenty-first century and beyond deserves nothing less.

Funding

The authors gratefully acknowledge support, in part, from the following: NCB: N/A; TAL: NIH-NIAID, U19-AI089680; PMA2020 from the Bill & Melinda Gates Foundation; AO'H: N/A; and JP: NIA grant P30AG028716. Authorship order is alphabetical.

References

Abrams, K. R., Spiegelhalter, D., and Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health Care*, New York: Wiley. [61,65]

Anderson, H., and Hepburn, B. (2016), "Scientific Method," available at *https://plato.stanford.edu/archives/sum2016/entries/scientific-method*. [57]

Anscombe, F. J., and Aumann, R. J. (1963), "A Definition of Subjective Probability," *The Annals of Mathematical Statistics*, 34, 199–205. [60]

Berry, S. M., Carlin, B. P., Lee, J., and Müller, P. (2010), *Bayesian Adaptive Methods for Clinical Trials*, Boca Raton, FL: Chapman&Hall/CRC Press. [61,65]

Bradley, R., Bell, J., Gauld, J., Lilly, A., Jordan, C., Higgins, A., and Milne, R. (2005), "UK Soil Database for Modelling Soil Carbon Fluxes and Land Use for the National Carbon Dioxide Inventory," Report to Defra Project SP0511, Defra, London. [65]

Brownstein, N. C. (2018), "Perspective from the Literature on the Role of Expert Judgment in Scientific and Statistical Research and Practice," arXiv no. 1809.04721. [56,58]

Cairney, P., and Oliver, K. (2017), "Evidence-Based Policymaking Is Not Like Evidence-Based Medicine, So How Far Should You Go to Bridge the Divide Between Evidence and Policy?," *Health Research Policy and Systems*, 15, 35. [60]

Carlin, B. P., Chaloner, K., Church, T., Louis, T. A., and Matts, J. P. (1993), "Bayesian Approaches for Monitoring Clinical Trials, With an Application to Toxoplasmic Encephalitis Prophylaxis," *The Statistician*, 42, 355–367. [63]

Center for Open Science (2018), "Registered Reports," available at *https://cos.io/rr/*. [59]

Chaloner, K., Church, T., Louis, T. A., and Matts, J. P. (1993), "Graphical Elicitation of a Prior Distribution for a Clinical Trial," *The Statistician*, 42, 341–353. [63]

Collins, F. S., and Tabak, L. A. (2014), "NIH Plans to Enhance Reproducibility," *Nature*, 505, 612–613. [67]

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Series B, 34, 187–220. [62,63]

Cripps, E., O'Hagan, A., and Quaife, T. (2013), "Quantifying Uncertainty in Remotely Sensed Land Cover Maps," *Stochastic Environmental Research and Risk Assessment*, 27, 1239–1251. [66]

Curtis, A., Worley, S., Adamson, P., Chung, E., Niazi, I., Sherfesee, L., Shinn, T., and Sutton, M. (2013), "Biventricular Pacing for Atrioventricular Block and Systolic Dysfunction," *New England Journal of Medicine*, 368, 1585–1593. [61,62]

Easterbrook, P. J., Gopalan, R., Berlin, J., and Matthews, D. R. (1991), "Publication Bias in Clinical Research," *The Lancet*, 337, 867–872. [60]

Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. (2003), *Data Monitoring Committees in Clinical Trials: A Practical Perspective*, Chichester: Wiley. [59]

Finfer, S., and Bellomo, R. (2009), "Why Publish Statistical Analysis Plans," *Critical Care and Resuscitation*, 11, 5–6. [59]

Franco, A., Malhotra, N., and Simonovits, G. (2014), "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science*, 345, 1502–1505. [60]

Fuller, R., Smith, G., Sanderson, J., Hill, R., Thomson, A., Cox, R., Brown, N., Clarke, R., Rothery, P., and Gerard, F. (2002), "Countryside Survey 2000 Module 7. Land Cover Map 2000," Final Report, CSLCM/Final. [66]

Garland, T., Jr. (2016), "Scientific Method as an Ongoing Process," available at *https://en.wikipedia.org/w/index.php?title=Scientific_method&oldid=822947033*. [57]

Gelman, A., and Hennig, C. (2017), "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society*, Series A, 180, 967–1033. [56,61]

Gibson, E. W. (2017), "Leadership in Statistics: Increasing Our Value and Visibility," *The American Statistician*, DOI: 10.1080/00031305.2017.1336484. [67]

Gosling, J. P., and O'Hagan, A. (2006), "Understanding the Uncertainty in the Biospheric Carbon Flux for England and Wales," Technical Report 567/06, University of Sheffield, available at *http://tonyohagan.co.uk/academic/pdf/UUCF.pdf*. [66]

Grice, J. W.(2017), "Comment on Locascio's Results Blind Manuscript Evaluation Proposal," *Basic and Applied Social Psychology*, 39, 254–255. [60]

Haines-Young, R., Barr, C., Black, H., Briggs, D., Bunce, R., Clarke, R., Cooper, A., Dawson, F., Firbank, L., Fuller, R., Furse, M. T., Gillespie, M. K., Hill, R., Hornung, M., Howard, D. C., McCann, T., Morecroft, M. D., Petit, S., Sier, A. R. J., Smart, S. M., Smith, G. M., Stott, A. P., Stuart, R. C., and Watkins, J. W. (2000), "Accounting for Nature: Assessing Habitats in the UK Countryside," Natural Environment Research Council and

Centre for Ecology and Hydrology, Department of the Environment, Transport and the Regions, available at *https://trove.nla.gov.au/work/17604300*. [66]

Harris, K., O'Hagan, A., and Quegan, S. (2010), "The Impact of Satellite-Derived Land Cover Uncertainty on Carbon Cycle Calculations," available at *tonyohagan.co.uk/academic/pdf/CTCDPaper_v6.pdf*. [65,66]

Hawkins, D. M.(2004), "The Problem of Overfitting," *Journal of Chemical Information and Computer Sciences*, 44, 1–12. [65]

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015), "The Extent and Consequences of P-Hacking in Science," *PLoS Biology*, 13, e1002106. [59]

HM Treasury (2015), "The Aqua Book: Guidance on Producing Quality Analysis for Government," *HM Government, London, UK, nd https://www. gov. uk/government/publications/theaqua-book-guidance-on-producing-quality-analysis-for-government (accessed July 10, 2017)*. [60,67]

Hyman, M. R.(2017), "Can 'Results Blind Manuscript Evaluation' Assuage 'Publication Bias'?," *Basic and Applied Social Psychology*, 39, 247–251. [60]

Jacobson, M., Besch, C., Child, C., Hafner, R., Matts, J., Muth, K., Wentworth, D., Neaton, J., Abrams, D., Rimland, D., and Perez, G. (1994), "Primary Prophylaxis With Pyrimethamine for Toxoplasmic Encephalitis in Patients With Advanced Human Immunodeficiency Virus Disease: Results of a Randomized Trial," *The Journal of Infectious Diseases*, 169, 384–394. [62]

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I., Gosling, J. P., and Heinemeyer, A. (2008), "Quantifying Uncertainty in the Biospheric Carbon Flux for England and Wales," *Journal of the Royal Statistical Society*, Series A, 171, 109–135. [65,66]

Kerr, N. L. (1998), "Harking: Hypothesizing After the Results Are Known," *Personality and Social Psychology Review*, 2, 196–217, PMID: 15647155. [59]

Kynn, M. (2008), "The 'Heuristics and Biases' Bias in Expert Elicitation," *Journal of the Royal Statistical Society*, Series A, 171, 239–264. [66]

LaVange, L. M. (2014), "The Role of Statistics in Regulatory Decision Making," *Therapeutic Innovation & Regulatory Science*, 48, 10–19. [60]

Lindsay, R. M., and Ehrenberg, A. S. (1993), "The Design of Replicated Studies," *The American Statistician*, 47, 217–228. [60]

Locascio, J. J. (2017a), "Rejoinder to Responses on 'Results-Blind Publishing,'" *Basic and Applied Social Psychology*, 39, 258–261. [60]

Locascio, J. J. (2017b), "Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39, 239–246. [60]

Marks, M. (2017), "Commentary on Locascio," *Basic and Applied Social Psychology*, 39, 252–253. [60]

McNutt, M. (2014), "Journals Unite for Reproducibility," *Science*, 346, 679–679. [67]

National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods, Harris-Kojetin, B. A., Groves, R. M., eds. (2017), *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*, Washington, DC: National Academies Press. [60]

National Research Council, Division on Engineering and Physical Sciences, Board on Mathematical Sciences and Their Applications, and Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification (2012a), *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*, Washington, DC: National Academies Press. [60]

National Research Council, Division of Behavioral and Social Sciences and Education, Division on Engineering and Physical Sciences, Committee on National Statistics, Computer Science and Telecommunications Board, and Panel on Communicating National Science Foundation Science and Engineering Information to Data Users (2012b), *Communicating Science and Engineering Data in the Information Age*, Washington, DC: National Academies Press. [60]

NICE (2012), "The Guidelines Manual: Process and Methods," available at *https://www.nice.org.uk/process/pmg6/*. [67]

Nussbaum, B. (2017), "Introductory Remarks: JSM 2017 Presidential Invited Address," available at *https://ww2.amstat.org/meetings/jsm/2017/webcasts/index.cfm*. [58]

Oakley, J. (2016), "Introduction to Uncertainty Quantification and Gaussian Processes," available at *http://gpss.cc/gpuqss16/slides/oakley.pdf*. [66]

O'Brien, P. C., and Fleming, T. R. (1979), "A Multiple Testing Procedure for Clinical Trials," *Biometrics*, 35, 549–556. [62]

O'Hagan, A. (2006), "Bayesian Analysis of Computer Code Outputs: A Tutorial," *Reliability Engineering & System Safety*, 91, 1290–1300. [66]

——— (2012), "Probabilistic Uncertainty Specification: Overview, Elaboration Techniques and Their Application to a Mechanistic Model of Carbon Flux," *Environmental Modelling & Software*, 36, 35–48. [65]

O'Hagan, A. (2018), "Expert Knowledge Elicitation: Subjective, But Scientific," [submitted]. [65,66,67]

Oliver, K. A., and de Vocht, F. (2017), "Defining 'Evidence' In Public Health: A Survey of Policymakers' Uses and Preferences," *European Journal of Public Health*, 27, 112–117. [60]

Oliver, K., Lorenc, T., and Innvær, S. (2014), "New Directions in Evidence-Based Policy Research: A Critical Analysis of the Literature," *Health Research Policy and Systems*, 12, 34, DOI: 10.1186/1478-4505-12-34. [60]

Ott, M. G. (1991), "Importance of the Study Protocol in Epidemiologic Research," *Journal of Occupational Medicine*, 33, 1236–1239. [59]

Oxford University Press (2018a), "Judgment," available at *https://en.oxforddictionaries.com/definition/judgment*. [57]

——— (2018b), "Knowledge," available at *https://en.oxforddictionaries.com/definition/knowledge*. [57]

——— (2018c), "Opinion," available at *https://en.oxforddictionaries.com/definition/opinion*. [57]

——— (2018d), "Scientific Method," available at *https://en.oxforddictionaries.com/definition/scientific_method*. [57]

Pocock, S. J. (2006), "Current Controversies in Data Monitoring for Clinical Trials," *Clinical Trials*, 3, 513–521. [59]

Shih, J. (1995), "Sample Size Calculation for Complex Clinical Trials With Survival Endpoints," *Controlled Clinical Trials*, 16, 395–407. [61]

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. [67]

Woodward, F. I., and Lomas, M. R. (2004), "Vegetation Dynamics—Simulating Responses to Climatic Change," *Biological Reviews*, 79, 643–670. [65]

The American Statistician

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

Expert Knowledge Elicitation: Subjective but Scientific

Anthony O'Hagan

To cite this article: Anthony O'Hagan (2019) Expert Knowledge Elicitation: Subjective but Scientific, The American Statistician, 73:sup1, 69-81, DOI: 10.1080/00031305.2018.1518265

To link to this article: https://doi.org/10.1080/00031305.2018.1518265

© 2019 The Author(s).

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 11515

View related articles

View Crossmark data

Citing articles: 66 View citing articles

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, 69–81: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518265

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Expert Knowledge Elicitation: Subjective but Scientific

Anthony O'Hagan

University of Sheffield, Western Bank, Sheffield, United Kingdom

ABSTRACT

Expert opinion and judgment enter into the practice of statistical inference and decision-making in numerous ways. Indeed, there is essentially no aspect of scientific investigation in which judgment is not required. Judgment is necessarily subjective, but should be made as carefully, as objectively, and as scientifically as possible.

Elicitation of expert knowledge concerning an uncertain quantity expresses that knowledge in the form of a (subjective) probability distribution for the quantity. Such distributions play an important role in statistical inference (for example as prior distributions in a Bayesian analysis) and in evidence-based decision-making (for example as expressions of uncertainty regarding inputs to a decision model). This article sets out a number of practices through which elicitation can be made as rigorous and scientific as possible.

One such practice is to follow a recognized *protocol* that is designed to address and minimize the cognitive biases that experts are prone to when making probabilistic judgments. We review the leading protocols in the field, and contrast their different approaches to dealing with these biases through the medium of a detailed case study employing the SHELF protocol.

The article ends with discussion of how to elicit a joint probability distribution for multiple uncertain quantities, which is a challenge for all the leading protocols. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2018
Revised August 2018

KEYWORDS
Cognitive bias; Elaboration; Elicitation protocol; Facilitator; Heuristics; Independence in subjective distributions; Rational impartial observer; SHELF; Subjective probability

1. Introduction

This article is a result of a presentation I made to the October 2017 Symposium on Statistical Inference, organized and sponsored by the American Statistical Association, in a session devoted to the role of expert opinion and judgment in statistical inference and decision-making. From the presentations and discussions in that session, a position article by the convenor and speakers in the session presents their view that

> 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. (Brownstein et al. 2018)

One way in which expert opinion and judgment enters into statistical inference and decision-making is through expert knowledge elicitation. Elicitation in this context is the process of expressing expert knowledge in the form of probability distributions for uncertain quantities. It is increasingly used to quantify uncertainty about parameters in models that are employed across the whole spectrum of human endeavour. In the context of the Symposium on Statistical Inference, we can think of the elicited probability distributions as forming prior distributions for Bayesian statistical inference, or as directly informing decision-making under uncertainty.

The literature on elicitation is extensive. with important contributions made in fields as diverse as statistics, psychology, management science, economics, and environmental science. A review of this literature can be found in O'Hagan et al. (2006), while Kurowicka and Cooke (2006), O'Hagan (2012), European Food Safety Authority (2014), and Hanea et al. (2018) provide entry points to some more recent developments. Even works that seek only to elicit estimates of uncertain quantities, rather than to characterize uncertainty in the form of a probability distribution, for instance Burgman (2015), may offer much good advice about eliciting judgements from experts.

Eliciting expert knowledge carefully, and as scientifically as possible, is not simply a matter of sitting down with one or more experts and asking them to tell us what they think. In particular, psychologists have identified numerous ways in which naive questioning can engender cognitive biases in the experts' judgements. To capture expert knowledge as objectively as possible in the form of a probability distribution, the elicitation needs to be structured so as to avoid, or at least to minimize, such biases. Research in the field has led to the development of carefully designed elicitation *protocols*, which set out procedures for the elicitation to follow.

This article begins by addressing these issues of cognitive biases and elicitation protocols. Section 2 introduces the psychological literature on heuristics and biases, highlighting those that are of particular relevance in eliciting probability distributions.

CONTACT Anthony O'Hagan a.ohagan@sheffield.ac.uk University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom.
Supplementary materials for this article are available online. Please go to *www.tanfonline.com/r/TAS*.

© 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.

In Section 3, issues around the use of multiple experts are considered, including further psychological hazards in elicitation from groups of experts. The leading protocols are introduced in Section 4. Section 5 presents a detailed case study, where judgments about the future incidence of long-term medical conditions among elderly people in the UK were elicited from a group of experts using the SHELF protocol. Each step in the protocol is discussed with particular emphasis on how it is designed to address and minimize the biases identified in Sections 2 and 3. At each step, differences between SHELF and the other leading protocols are also discussed. Section 6 considers elicitation of knowledge about two or more uncertain quantities, which is a challenge in elicitation that applies to all protocols. Finally, Section 7 reviews the principal ideas in this article.

2. Biases in Elicitation

Theories of subjective probability, in which probability is characterized as a personal degree of belief, date back at least to Ramsey (1926), with important contributions by de Finetti (1937, 1970) and Savage (1954). The idea of formally representing uncertainty using subjective probability judgments to inform decision-making began to be taken seriously in the 1960s and 1970s. However, at about the same time, psychologists were identifying problems in the ways that people make judgments. The ground-breaking research of Tversky and Kahneman (1974) set in motion the *heuristics and biases* research program, the underlying principle of which is that people's judgments are often made on the basis of heuristics, which are quick, short-cut reasoning processes. These heuristics have served us and our ancestors well in countless everyday situations where we do not have time or need for prolonged thought, but the instinctive application of heuristics can lead to systematic errors of judgment (biases) in more complex, less everyday tasks. The conclusion is that for serious decision-making we need to consciously slow down, taking time to think. Kahneman (2011) expounds in detail this theory of fast and slow thinking.

Of the large number of heuristics that have been identified in the psychology literature, a few are of particular relevance to the judgment of probabilities, and hence to elicitation.

2.1. Anchoring

One of the original heuristics of Tversky and Kahneman (1974) was "anchoring and adjustment": when asked to make a numerical judgment, people start with a readily available value and adjust from that point to assign a judgment. The initial value is called the anchor, because typically the adjustment is insufficient and so the resulting judgment is biased toward the anchor. The canonical case is when two successive judgments are made of related quantities; the first judgment acts as an anchor for the second. An example of this in the case of probability judgments is discussed in Section 2.4.

To illustrate a different form of anchoring in probability judgment, I have conducted an experiment on a sample of participants on courses I have given over the past four years. This is admittedly not a proper random sample from a well-defined population of subjects, but randomization was applied within each course.

Table 1. Average probability judgments in the Muslims exercise.

	2 million first	8 million first
$P(M > 2 \text{ million})$	0.696	0.804
$P(M > 8 \text{ million})$	0.303	0.403

The participants were asked to make judgments about the number of Muslims in England and Wales according to the UK census of 2011. They were given the following information:

> In the 2011 United Kingdom Census, people were asked to specify their religion. Of the 52.0 million people in England and Wales who gave an answer (excluding 4 million non-respondents), 33.2 million stated that they were Christian (which includes Catholic and various Protestant denominations). Other possible answers (18.8 million people in total) were Buddhist, Hindu, Jewish, Muslim, Sikh, "Other religion", "No religion". In this exercise, you are asked to assess probabilities concerning the number, M, of Muslims amongst those 18.8 million.

They were asked to provide two probabilities, that M was more than 8 million and that it was more than 2 million. There were two versions of the exercise: in one version they were asked for $P(M > 8)$ first, and in the other version the first judgment was $P(M > 2)$. They did not see the second question until they had answered the first (and could not go back and revise their first answer). In each course, the allocation of versions to participants was randomized. I gave 10 courses over four years, to participants from widely varying backgrounds. For instance, in some cases, they were predominantly British, in two cases they were all from Nordic countries, and in two more mostly Italian. I have data from approximately 90 participants on each version of the exercise. Their average probabilities are given in Table 1.

The effect of anchoring is seen clearly here. Participants who received the $P(M > 8)$ question first on average give higher probability judgments on *both* questions. Putting the number 8 million in their heads serves as an anchor, suggesting higher values for M than if the anchor is 2 million. Since the data are not from a proper random sample, any formal statistical analysis can only be indicative, but the differences are quite clear. In fact, the same feature, of higher average probabilities on both questions for participants who received the version with $P(M > 8)$ first, was observed separately in every one of the 10 courses, with the sole exception of just one inversion in one course.

2.2. Availability

According to the availability heuristic, also originally identified in Tversky and Kahneman (1974), an event is judged more probable if we can quickly bring to mind instances of it occurring. For instance, I will judge deaths from a particular disease to be more common if I know of people who suffer from, or have died from, this disease. Dramatic events are more memorable, and so

are more easily brought to mind. Thus, aircraft and train crashes make headline news, and will influence a person's judgment of risk with these modes of transport. Car crashes are, in contrast, far more common but are rarely newsworthy, so their probability tends to be underestimated.

More generally, when experts are asked about an uncertain quantity, their judgments will be more influenced by evidence that they can readily bring to mind. This will tend to be recent evidence and evidence that they have personally been involved in deriving. Less salient evidence is likely to be overlooked.

2.3. Range–Frequency

If we divide the possible values of an uncertain quantity into a number of categories, and ask experts to assign a probability to each category, then they will tend to spread the total probability more or less evenly between the categories. Their stated probability judgments are then a compromise between this even spread and their actual beliefs about how probable each category is. This *range–frequency compromise* is related to a more general heuristic of the same name (Parducci 1963), and leads to a bias that assigns less probability to the categories judged most likely, and more probability to the other categories. Furthermore, probability judgments are influenced by the choice of categories. In a seminal experiment, Fischhoff, Slovic, and Lichtenstein (1978) asked subjects to assign probabilities to various reasons that a car might fail to start. One group were given reasons grouped into seven categories. The other group had just four categories, the first three of which were the same as for the first group, but the remaining category was "All other problems." The second group assigned higher probabilities on average to the first three categories than the first group.

To illustrate this bias in the context of elicitation of an uncertain quantity, I carried out another experiment with participants in my most recent training course. They were asked to make judgments about the total number, C, of medals that China will win in the next summer Olympic Games. They were given a table of the number of medals won by China at all previous Summer Olympics, and they were asked to assign probabilities to C being "more than 119," "between 110 and 119," "between 100 and 109," and so on. There were two versions of the exercises which again were randomly assigned to participants. One version had as its final category C being "less than 60." while the other split this into two final categories, "between 50 and 59," and "less than 50." Thus, one version had eight categories, while the other had nine.

The range–frequency compromise suggests that the participants with the first version of the exercise will assign a smaller probability to the single category "less than 60" than the total probability given by participants with the second version to their final two categories combined. This is indeed what I found, with an average probability of 0.077 given to the single category "less than 60," compared with an average total probability of 0.104 given to the split category. This was a small sample, and certainly does not prove that the range–frequency compromise is operating as expected in this example, but it is supportive of that principle and I hope to be able to establish it firmly in future courses.

2.4. Overconfidence

It is often said that experts are typically overconfident in their judgments. The evidence for this is primarily from studies in which subjects were asked to give an interval of values for an uncertain quantity with a specified probability, and where the frequency with which the true values fell within those intervals was less than the specified probability. For instance, in studies where subjects gave 95% probability intervals, fewer than 95%, and perhaps as few as 65% of those intervals were found to contain the corresponding true values. The intervals exhibited overconfidence.

A number of possible explanations can be advanced for this finding.

- Although not a heuristic in its own right, overconfidence may be related to anchoring. Subjects are often asked for an interval after they have first given an estimate, which can serve as an anchor. Then, the interval being too narrow may be due to insufficient adjustment from the anchor.
- For experiments, the true values of quantities must be known to the researchers but not to the subjects. Some studies have asked about "almanac" quantities, that is ones whose values might readily be found in a reference book or online, such as populations of cities, lengths of rivers, heights of buildings, etc. In selecting quantities, the researchers may unconsciously choose ones which are "more interesting." which could equate to them having unexpected values. If so, the low coverage frequency of the subjects' intervals may be quite reasonable, and not evidence of overconfidence.
- When the subjects are presented as experts, they may feel under some pressure to demonstrate their expertise by giving narrower intervals than their knowledge would really justify. Equally, though, experts may give wider intervals when they fear there may be consequences to "getting it wrong," thereby expressing underconfidence.
- It has been suggested that expertise is associated with experts developing their own short-cut heuristics to reach quick answers to questions commonly arising in their field. An instinctive reliance on such specialist heuristics may manifest as overconfidence, because in an elicitation exercise they are likely to be asked about less routine quantities.
- To assess a 95% interval, a subject must judge the probability of being outside that interval as only 5%, but it is difficult to judge small probabilities in this way.

The significance of the last point is that if an expert is asked to give a 95% interval, they simply think in terms of giving a range of values in which they judge the quantity of interest is very likely to lie. They would give the same range if asked for a 99% interval, because "very likely" is an imprecise term. To illustrate this, I conducted another experiment with my most recent course participants. In this exercise, they were given the following information:

> You are a taxi driver and you need to drive a customer from the town of Great Missenden to Terminal 5 of London's Heathrow Airport, to arrive no later than 17:45 on Friday evening. The journey comprises 11 miles on A roads, followed by 11

Table 2. Average bounds in the taxi journey exercise.

	95% Bound first	Plausible bound first
Upper 95% bound	93 mins	67 mins
Upper plausible bound	126 mins	84 mins

> miles on motorways, including 5 miles on the M25. According to Google maps, the journey time in "usual traffic" should be 33 minutes, but you need to arrange to pick the customer up early enough so that there is very little risk of not arriving by 17:45. Let the journey time on Friday be T (minutes).

Participants were then asked to specify an upper 95% bound (a value they were 95% confident that T would be below) and an upper "plausible" bound (that they were "almost sure" T would be below). As in the other exercises, there were two versions in which either the 95% or plausible bound was requested first, and they did not see the second question until they had answered the first. Table 2 shows the average values of their bounds.

Again, this is a single small sample, but it is striking that the participants gave an average of about 90 minutes for their first upper bound, whether they were asked first for a 95% bound or a plausible bound. If they had first given a 95% bound, to give a plausible bound they adjusted their first value upwards, whereas to go from a plausible to a 95% bound they adjusted downwards (adjustments which, because of anchoring, we may expect to have been inadequate!). I hope to be able to replicate this finding in future courses.

3. Multiple Experts

Expert knowledge elicitation is generally conducted by a *facilitator*, somebody who is knowledgeable about the process of elicitation, working with one or more experts. It may equally be a more informal, solo activity, in the sense that a scientist wishes to quantify his or her own knowledge about an uncertain quantity, for the purposes of some scientific endeavour. In that case, the scientist plays the role of both facilitator and expert, and although this may seem an artificial analogy it is useful; the scientist's judgments should be made as carefully and objectively as reasonably possible, according to the principles identified in this article.

When formal expert knowledge elicitation is employed, it is usual to seek judgments from more than one expert. Nevertheless, we generally require the outcome to be a single probability distribution representing the combined knowledge of experts in the field. Resolving the experts' judgments into a single distribution is known as the problem of aggregation. There are two principal approaches.

- *Mathematical aggregation.* In this approach, also known as *pooling*, separate judgments are elicited from the experts and a probability distribution is fitted to each expert's judgments. These are then combined into the aggregate distribution using a mathematical formula (a pooling rule).
- *Behavioral aggregation.* In contrast, the behavioral approach asks the group of experts to discuss their knowledge and opinions, and to make group "consensus" judgments, to which an aggregate distribution is fitted.

Neither approach is without disadvantages. Mathematical aggregation requires a choice of pooling rule, and numerous such rules have been proposed. To limit the choice, we might ask for the rule to have desirable properties, but French (1985) offers two reasonable consistency criteria and reports that no fixed pooling rule can satisfy both.

The behavioral approach has the difficulty of persuading experts with differing opinions to reach "consensus." It is also open to various additional hazards that have been identified by psychologists. For instance, there are clearly problems associated with the personalities of those whose opinions may be sought as experts. More junior experts may defer to a senior person, and thereby fail to contribute their own expertise fully. Similarly, a strong personality may dominate the group discussion and consensus judgments, without necessarily meriting a dominant position by virtue of their knowledge and expertise. Even if one expert does not dominate the group, two or more like-minded experts may do so together. Conversely, the judgments of a quieter and less extrovert expert may be ignored or overlooked.

Another feature of expert groups, which might be called a group heuristic, is a tendency for discussion to be restricted to ideas that will be broadly acceptable to all the group members, a behavior known as *groupthink* (Janis 1972). In the context of behavioral aggregation in elicitation, this term describes a tendency for a consensus view to emerge that is overconfident, because of the very act of seeking consensus.

Mathematical aggregation does not have to face these additional, potentially biasing, psychological effects, a fact which is often presented as a reason for using protocols employing mathematical aggregation. The problems of personalities and groupthink are formally avoided by the experts not meeting, with all judgments being anonymized, but this is achieved at the cost of losing the opportunity for the experts to share and debate their opinions, which is the principal benefit of behavioral aggregation.

A technical question that can be directed to the aggregate distributions derived from all of these methods is, "Whose probability distribution is this?" The point of the question is that the reason for seeking an aggregate distribution is to represent uncertainty about the quantity of interest, and for this purpose it should be interpretable as an expression of a person's subjective beliefs. It is clear that the result of mathematical aggregation is not the belief of any individual. It might be claimed that the result of a behavioral aggregation represents the beliefs of the group, but it is far from clear that the group has any such beliefs when the distribution is more the outcome of compromise than consensus.

4. Elicitation Protocols

In a substantial guidance document on eliciting expert knowledge, EFSA (the European Food Safety Authority 2014)

recommends three protocols for elicitation with multiple experts.

- The Cooke protocol (Cooke 1991) employs mathematical aggregation. Experts separately make judgments about the uncertain quantity or quantities of interest, but they also make judgments about a number of *seed* variables whose true values are known to the facilitator. The seed variables are chosen so that as far as possible they are similar to the quantities of interest (although they inevitably differ in the sense that the true values of the quantities of interest are not known). The view is taken that experts' performance in making judgments about the seed variables will be predictive of how good their judgments will be about the quantities of interest. A special pooling rule (referred to as "the classical model") is used in which the experts are weighted according to their performance on the seed variables.
- The Sheffield protocol in the EFSA guidance is a particular case of the SHELF protocol (Oakley and O'Hagan 2016; Gosling 2018), and employs behavioral aggregation. It is characterized by two rounds of judgments from the experts. In the first round, they make individual judgments privately. These are then revealed and discussed, with a view to sharing and understanding the reasons for differing opinions, before the second round in which the group agrees on "consensus" judgments. The "consensus" is according to the perspective of a rational impartial observer; see the discussion in Section 5.5. The SHELF protocol requires an experienced facilitator to manage the experts, and to address possible sources of bias in group interactions.
- EFSA's third protocol is the classic Delphi method (Rowe and Wright 1999), but adapted to elicit judgments of uncertainty rather than simply estimates. It is known as EFSA Delphi or probabilistic Delphi, and has features of both mathematical and behavioral aggregation. Experts make two or more rounds of judgments, with feedback being given between rounds summarizing the judgments of all the experts. Anonymity is maintained but there is some interaction and sharing of knowledge between experts, although the interaction is strictly limited. At the end, it is necessary to apply a pooling rule to aggregate across the experts' final distributions. The IDEA protocol (Hemming et al. 2018; Hanea et al. 2018) is a version of Delphi with just two rounds, but with more emphasis on a facilitated discussion between experts after the first round. Published examples do not ask for formal probabilistic judgments, but the protocol is flexible enough to accommodate different forms of judgments, including probabilistic judgments as in the above three protocols, and can involve seed variables and weighted pooling.

It is not possible to say which of these leading protocols is best, in the sense of most accurately quantifying the experts' knowledge and beliefs in the form of a probability distribution. The principal reason is what has been seen as the impracticality of conducting experiments in which genuine experts make judgments about uncertain quantities that are within their area of expertise. Even if the true values of these quantities were to become known subsequently, which is rarely the case, an experiment would need to draw on a large number of experts, allocate them randomly to different protocols, and do this on enough different quantities and occasions to constitute adequate replication. Nevertheless, this is a neglected area of research that would benefit from some innovative thinking.

Such experimental evidence as exists is much more limited and can at best be only indicative of good practice. For instance, much of the experimental evidence in psychology addresses a small part of an elicitation (such as the elicitation of single probability) and is based on nonexperts (often psychology students). Cooke and Goossens (2008) and Cooke et al. (2014) analyze applications of the Cooke method and show that the "classical model" weighted pooling performs better than equal-weighted pooling. These analyses perhaps come closest to the goal of studying entire elicitation protocols, but they rely on comparing predictions of some of the seed variables (with the weighting based on judgments of the other seed variables), not on predictions of the quantities of interest.

Protocols represent the judgments of their developers, based on the evidence and their own experience, as to the best way to ensure that the elicited distribution provides a careful, scientific synthesis of the experts' knowledge. And the choice of a protocol for any given elicitation task is another judgment, in this case the judgment of the facilitator in collaboration with the client.

The above brief outlines of the leading protocols will be amplified through the case study in Section 5, where the differences between them are identified for each stage of the elicitation process.

5. Case Study

In July 2014, I facilitated an elicitation workshop as part of a large project managed by the Centre for Workforce Intelligence on behalf of and by commission from the UK Department of Health called "Horizon 2035" (H2035), the purpose of which was to advise government on the likely demand for health services in the year 2035 and improve system-wide workforce planning (CFWI 2015). The elicitation concerned the incidence of Long-Term Conditions (LTCs) in 2035. LTCs cover a variety of health conditions that require long-term care, such as diabetes, heart disease, and respiratory disorders. The term also includes mental disorders, but the elicitation concerned only physical LTCs. This case study focuses on the elicitation of the number of LTCs in 2035 among people aged 85 and above, which is the group with the highest rate of LTCs.

The elicitation was conducted according to the SHELF protocol, and involved a one day face-to-face meeting. In this section, I will describe all the steps of that elicitation, highlighting how the protocol is designed to aid with careful, scientific judgments, including the ways in which SHELF avoids or minimizes the various cognitive biases and challenges identified in Sections 2 and 3. Comparisons will be made throughout with the other leading protocols.

5.1. Preliminaries

Whatever protocol one is using, good elicitation requires extensive preparation. An important early step is the selection and

recruitment of the experts. There is much useful guidance on this in European Food Safety Authority (2014). For the SHELF protocol, some considerations are:

- Aim for around four to eight experts. This is a large enough number generally to cover the range of opinion within the relevant community; more will usually extend the discussion unnecessarily without usefully adding further knowledge and opinion. For Cooke and Delphi protocols, four to eight also works well, although it is possible to accommodate larger numbers.
- Choose experts who will listen to and take account of the opinions of the other experts. People whose opinions are too firmly held will make it difficult to achieve the kind of consensus required. Even if such a person is a very high-ranking expert, it may be better to invite them to submit their judgments in writing, for consideration by the other experts in the elicitation workshop, than to invite them to take part in the workshop. This advice is also relevant for the Delphi protocol, although it is not a concern for Cooke.
- Don't invite an expert who will naturally defer to another invited expert. Pushing that person to express their own opinions is unlikely to be profitable and may cause them discomfort. In the Cooke and Delphi protocols, where experts' judgments are made anonymously or no interaction between experts is allowed, this advice does not apply.

For the H2035 case study, five experts were present—a statistician working in chronic disease modeling, a Director of Public Health, an Allied Health Professions Officer at the National Health Service, a health policy analyst and an epidemiologist.

Another important step before the elicitation workshop is the preparation of an *evidence dossier*. This assembles all the most relevant evidence into a single document in a format that is readily accessible during the workshop. It is reviewed at the start of the SHELF workshop, and during the discussion stage the experts are encouraged to explain their first round judgments by reference to specific evidence in the dossier. The evidence dossier is a key tool in combating the availability heuristic, by ensuring that all the evidence is fresh in the experts' minds at all times. Once experts have been recruited for the elicitation, they are asked to read a first draft of the dossier (which is based on initial research by the client) and to submit any additional evidence that they may have or be aware of, in order for this to be included in the final draft for the benefit of all the experts. The SHELF package of documents (Oakley and O'Hagan 2016) includes advice on preparing the evidence dossier, together with an example dossier.

Reviewing the evidence is obviously also required in other protocols, although there is generally less emphasis on the formal preparation of a dossier.

Finally, experts need to be trained to make the necessary probabilistic judgments. In the H2035 case study, the experts were trained at the start of the workshop, and completed a practice elicitation to familiarize them with the process and the judgments required, and to identify and resolve any initial misunderstandings. Rather than take time during the workshop itself for training, there is now an online e-learning course available at *http://www.tonyohagan.co.uk/shelf/ecourse.html*, sponsored by the U.S. Office of Naval Research. This will familiarize the experts with making the necessary probabilistic judgments in advance, although it will still be necessary to run through a training exercise so that they understand the process of group discussion and judgments.

Training is important for other protocols, too. Indeed, the original impetus for the e-learning course was the failure of the first attempt by EFSA to use the probabilistic Delphi protocol. Experts generally do not meet in Delphi elicitation, and so face-to-face training is not feasible. In that EFSA elicitation, experts were sent detailed written instructions by email, but it was clear from their subsequent judgments that they had either not read or not fully understood the instructions. It is hoped that the e-learning course will lead to experts being better able to make judgments in a Delphi exercise. The specific judgments covered in the course are those required in the SHELF protocol, but could also be used in probabilistic Delphi, and they are readily adapted to judgments required in the Cooke protocol.

5.2. SHELF Templates

An elicitation according to the SHELF protocol follows a number of templates that organize the process of the elicitation though a predefined series of steps. Each template is a document, in which the conduct and outcome of each step are to be recorded. The completed templates form a *record* of the elicitation. A SHELF workshop begins with the completion of a SHELF1 template, which records housekeeping details such as the time and place of the elicitation, the names and areas of expertise of the participants (experts, facilitator, and any other persons present), any conflicts of interest, and the training given. Then, for each uncertain quantity of interest, a SHELF 2 template is completed, which records the elicitation of a probability distribution for that quantity. It is through the prescribed sequence of steps in each template that the SHELF protocol ensures careful and thoughtful consideration of all the required judgments, and avoids or minimizes sources of bias.

Table 3 shows the first part of the SHELF2 record for the H2035 case study. The first three fields repeat information from the SHELF1 record, while the "Quantity" field states the uncertain quantity of interest to be elicited with this template. Note that in the "Definition" step, the quantity (now named as X) is defined more precisely. It is essential that the definition is unambiguous, since otherwise the experts may interpret it differently from each other, and differently from how it was intended. All experts are required to agree on the definition.

The "Anonymity" field is important. Although the participating experts are named in the SHELF1 record, details of discussions and judgments recorded in the SHELF2 record are anonymized, so that a reader will know what was said but not who said it.

Note that the "Start time" was 14:15 because the morning was taken up with preparation, including training, a practice elicitation, and reviewing the evidence dossier. (The dossier is reviewed again for each quantity of interest, identifying data particularly relevant to that quantity.)

The SHELF templates serve the dual role of prescribing a protocol to minimize cognitive biases and documenting the elicitation exercise. The completed SHELF1 and SHELF2 records provide a traceable account of how the final probability dis-

Table 3. H2035 SHELF2 record, preliminaries.

Elicitation title	Horizon 2035: Physical long-term conditions
Workshop	SHELF workshop 1
Date	7th July 2014
Quantity	% change in the demand for care for physical LTCs per 1000 persons aged 85+ by 2035
Anonymity	In this record, the experts are identified by letters A, B, E, F, and G, and the Facilitator by Z
Start time	14:35
Definition	X = % change in the demand for care for physical LTCs per 1000 persons aged 85+ by 2035 relative to current. "Business as now" in terms of technology changes, changes in training, changes in organization. Ruling out radical technology changes and no major policy interventions or restructuring. Essentially, we are assuming that current trends are continuing.

Table 4. H2035 SHELF2 record, individual elicitation.

Individual elicitation	**Method:** Quartile **Judgments:** *Z* asked the experts to write down their median (*M*), to be obtained by consideration of a value for which there was a 50/50 chance of being more or less. If they were offered a prize for guessing whether *X* was above or below their *M* value, they should not have a preference for betting either way. The lower and upper quartiles (Q1 and Q3 respectively) were to be selected on the basis that the experts should judge it to be equally likely for *X* to be below Q1 or between Q1 and *M*, and that they should also judge it equally likely that *X* would be between *M* and Q3 or above Q3. In training, they had been advised that usually they would both be closer to *M* than to *L* or *U* and that the distance from *M* depends on confidence the expert has in their value for *M*. Experts were finally asked to make a coherence check that $L < Q1 < M < Q3 < U$ and a "flip-of-the-coin" test on the true value being inside or outside the range (Q1,Q3).

tributions were reached. Documentation like this represents good scientific practice, and allows any recipients of the elicited distributions to evaluate the extent to which they should adopt them for purposes of inference or decision-making.

Further excerpts from the H2035 SHELF2 record will be shown in subsequent sections of this case study. The full record is available in supplementary material. The SHELF package can be downloaded from *http://tonyohagan.co.uk/shelf*, and contains all the templates together with many documents, PowerPoint presentations and software to guide and assist the facilitator in organizing and conducting an elicitation according to the SHELF protocol.

5.3. Individual Judgments

The next steps comprise the SHELF protocol's individual judgments round. Experts are first asked to write down privately their own *plausible range* for X, in the form of a lower bound L and an upper bound U, such that they would find it very surprising if the true value were found to lie outside the interval $[L, U]$. It is not important whether the expert interprets this interval as being a 95% interval, a 99% interval or an "almost sure" interval. Its function is to establish that some values of X are judged by the expert to be simply not plausible. By beginning with this judgment, the expert is encouraged to think at the outset about the full range of possibility for X, which helps to counter overconfidence.

The facilitator encourages them to challenge their bounds by imagining that somebody reports to them that the true value of X has been determined and it falls just outside $[L, U]$. Their reaction to such a claim should be that it was most likely wrong—either the method of determining the value was flawed or it has been reported incorrectly. If not, if a value outside their range might indeed be plausible, then they should either decrease L or increase U. The facilitator then seeks to determine the lowest value of L given by any of the experts and the highest value of U, pointing out that all judgments can now be confined to $[L_{\min}, U_{\max}]$. Experts are asked again to consider whether their own plausible range is wide enough, and to revise if necessary.

Following the judgments of plausible range, the individual judgments round is completed according to one of three SHELF *methods*. Table 4 shows the relevant field of the H2035 record, where the quartile method was chosen. The following points are illustrated by Table 4.

1. *Median.* The expert is asked to specify their median value M, such that in their judgment X is equally likely to lie above or below M. Notice that the only numbers that have been suggested have come from the expert, rather than being given to the expert in a way that might create anchors. In judging M, the expert has anchors at L and U, and their anchoring effects should tend to cancel out so as to have minimal impact on the expert's judgment of M.
2. *Quartiles.* The expert then specifies their quartiles $Q1$ and $Q3$, such that the expert judges it to be equally likely for X to be below $Q1$, or between $Q1$ and M, and equally likely to be between M and $Q3$, or above $Q3$. Again, any anchoring effect on these judgments of the previously specified values L, U, and M should be minimal.
3. *Challenging the judgments.* Table 4 records some of the ways that experts are challenged to think about their judgments, and to revise them if necessary. The challenges are an important part of helping the experts to make their judgments carefully and thoughtfully.
4. *Training.* The judgments of median and quartiles are not explicitly judgments of probability but simply ask the expert to identify equally likely ranges of possible values of X. Nevertheless, they are not easy for experts who have not encountered these ideas before. Training and careful explanation are essential. For every judgment that experts are asked to make

within the SHELF protocol, the SHELF package includes a PowerPoint slide set to guide the expert in making, challenging, and refining their judgments.

The SHELF tertile method simply replaces the judgment of quartiles with tertiles $T1$ and $T2$. They divide the range of possible values into three equally likely parts—below $T1$, between $T1$ and $T2$, and above $T2$. This method addresses overconfidence and anchoring in the same way as the quartile method.

The third SHELF method is known as the roulette method. After the first step of specifying the plausible range, the expert is then presented with a tableau comprising a series of N "bins," where N is typically about 10, each representing a range of possible values of X so that the bins partition the plausible range into N parts. The expert places counters in the bins to show the relative probability that X will fall in each bin. For instance, if the expert places one counter in the first bin and three counters in the next, he or she is making the judgment that X is three times as likely to lie in the second bin compared to the first. This method has the advantage that the experts find it simple to use, particularly if they are familiar with the concept of a probability density function. However, some bias due to the range–frequency compromise may be expected, as shown for instance in the example in Section 2.3. Moreover, there is a temptation for experts to place their counters to achieve a nice shape, rather than encouraging them to think carefully about probabilities. For these reasons, although some practitioners of the SHELF protocol use the roulette method, the guidance in the SHELF package recommends the tertile or quartile methods.

The other leading protocols ask the experts to make similar judgements. In the Cooke protocol, they are usually asked for their median and 95% range. The use of a 95% range to characterize the expert's uncertainty seems unwise in the light of evidence that experts do not judge such intervals well, and do not distinguish well between 95%, 99%, or "almost certain"; see Section 2.4. SHELF asks for a plausible range, but for the purpose of countering over-confidence and anchoring, using instead the quartiles or tertiles to characterize uncertainty. The EFSA guidance specifies using the SHELF judgments of plausible range, median, and quartiles for the probabilistic Delphi protocol, although others have been used.

5.4. Facilitated Group Discussion

When judgments are elicited only from a single expert, or when an individual is simply using the framework of formal elicitation to express their own knowledge, then there is no need for aggregation; all protocols will use that expert's judgments as the basis for a final fitted probability distribution. In the case of more than one expert, however, once the experts are happy with their individual judgments, which hitherto have been written down privately, then in the SHELF protocol all are revealed and the group discussion phase begins. Other protocols do not use behavioral aggregation. The probabilistic Delphi protocol allows for some limited exchange of information between experts, but then they are aggregated by a pooling rule such as an equally weighted average. The Cooke protocol does not incorporate any interaction between experts, and their individual judgments are aggregated by a weighted pool, the weights being determined from the experts' judgments on the seed variables. For the SHELF protocol, however, the group discussion is an essential part. Ideally, the experts will all be present for face-to-face discussion in what is called a SHELF *workshop*. If that is not feasible, video-conferencing provides a reasonable alternative, although it becomes harder to manage the discussion so that every expert is able to participate fully.

Table 5. Individual judgments in the H2035 elicitation.

Expert	L	$Q1$	M	$Q3$	U
A	0	2.5	3	4	10
B	0	4	6	8	20
E	1	2.25	3	4	6
F	1	1.8	2	3	5
G	−10	−4	0	4	10

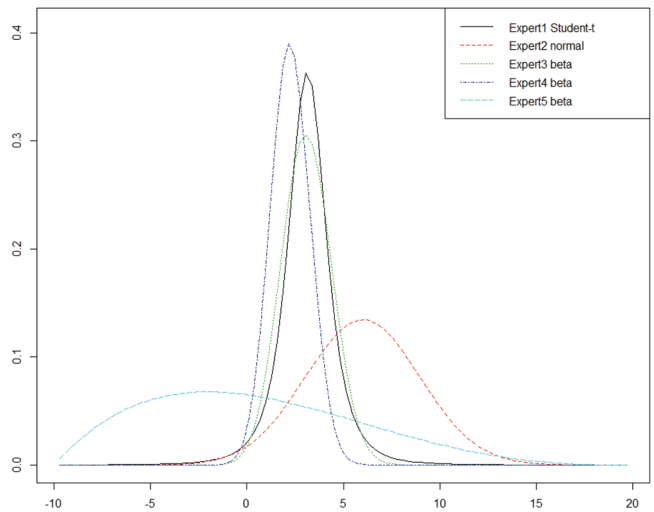


Figure 1. Distributions fitted to the judgments in Table 3.

The judgments of the five experts in the H2035 elicitation workshop are shown in Table 5, while Figure 1 plots probability distributions fitted to those judgments using the SHELF software, which was set to select for each expert the best-fitting distribution from a variety of standard families. Note that in fitting the distributions the experts' plausible bounds are not used, except that $L_{\min}$ and/or $U_{\max}$ may be used if bounded distributions are to be fitted (and this range is also used to display the fitted densities).

In the legend for Figure 1, Expert 1 is A in Table 3, Expert 2 is B, and so on. It is clear that the experts have a wide range of initial opinions, forming a natural basis to start the group discussion stage.

SHELF is based on behavioral aggregation because its developers believe it makes better use of the combined knowledge and expertise of the experts, but Section 3 has identified a number of additional challenges associated with the behavioral approach. For this reason, the SHELF workshop must be led by a facilitator who has expertise in the process of eliciting expert knowledge, and in particular is familiar with SHELF. The expertise of the facilitator in managing the SHELF protocol is as important as the experts' expertise regarding the quantities of interest. The facilitator works with the experts to obtain accurate judgments of their knowledge and manages the group discussion.

Table 6. H2035 SHELF2 record, group discussion.

Group discussion	Z initiated group discussion by asking B and G to explain why they thought X might be, respectively, so high or so low. B felt that diabetes is on the increase and cancer mortality is down so LTCs might be set for a substantial increase. It was also suggested that there could be an increase in social care demands for this age group because fewer families may be able to cope with the care provision. G said that there was far too much uncertainty to rule out very low, that is highly negative, or very high values for X. At this point, the other experts contributed their strong feelings that although X could conceivably be negative, in current conditions a decrease in LTCs in this age group was very unlikely. A also argued that a large increase was equally unlikely, simply because the numbers of LTCs in the 85+ age group was already high and it was hard to see how it could increase much further. There was a discussion around whether the differences in the demands might be taken up by younger age groups. A wide-ranging discussion continued, with A, E, and F being challenged over their distributions being much narrower, representing greater certainty about X, than those of B and G. However, no further material arguments being presented, Z drew the discussion to a close.

Table 7. H2035 SHELF2 record, group elicitation.

Group elicitation	**Method:** Probabilities **Judgments:** Z first asked the experts to consider RIO's probability that X would be less than 2%, then that X would be more than 4%, and finally that X would be negative. After discussion, the following values were agreed. $\Pr(X < 2\%) = 0.6$ $\Pr(X > 4\%) = 0.2$ $\Pr(X < 0\%) = 0.1$ For $\Pr(X < 2\%)$, experts agreed that the true value is more likely to be below than above 2% and that 60% probability seems to convey this. Z pointed out that only one expert had originally given probability more than 0.5 to this, but the experts felt that the discussion had generally led them to give more weight to lower values of X than they had initially thought. Z also questioned their judgment of $\Pr(X < 0\%) = 0.1$, pointing out that G had initially given probability 0.5 to X being negative. G confirmed that he/she would give this a lower probability after the discussion, and that 0.1 is a reasonable judgment for RIO.

The facilitator prompts the experts to explore their areas of disagreement. Experts are asked to explain the reasons for their judgments, with reference wherever possible to the evidence dossier. Long discussions often ensue, and can become heated if not managed carefully by the facilitator. An experienced facilitator will be familiar with the possible sources of bias in behavioral aggregation, and needs to be constantly alert. A checklist for managing the discussion would include the following.

- Allow discussion all the while it seems to be developing ideas. Don't let the experts keep repeating the same arguments.
- Make sure all opinions are heard and properly considered. Keep bringing quieter members into the discussion, and curtail the more voluble experts if necessary.
- Don't allow arguments to be presented aggressively.
- Listen carefully. Try to get a sense of the strengths of competing arguments. Two people giving the same argument doesn't make it twice as valid.
- Highlight and summarize significant points of contention or of agreement before moving on to the next area for discussion. It can be useful to write these on a whiteboard or flipchart. These summaries will be noted in the SHELF2 record, and so it is important for the experts to confirm that they are a fair representation of the discussion.

The facilitator's role may also be found in other elicitation protocols, but it is particularly important in SHELF.

Table 6 shows the H2035 SHELF2 record of the group discussion. This part of a SHELF elicitation can take a substantial amount of time, even running to hours in some cases. As the first bullet point above says, it is important to allow discussion, but only as long as it is yielding additional insight. In the H2035 case study, it was possible to bring the discussion to a close relatively quickly.

5.5. RIO and Group Judgments

The final stage of the SHELF protocol is for the experts to make group judgments leading to a 'consensus', aggregate distribution. It is important to understand, and to make clear to the experts, the nature of this "consensus." Even after discussing and debating, experts will not reach complete agreement (such that they now have the same knowledge and beliefs about an uncertain quantity, represented by the same probability distribution). Their opinions may be modified by the discussion, but they will inevitably leave the workshop with differing beliefs about the quantity of interest. If, during the workshop, they are coerced into an artificial consensus, it is unclear what the resulting distribution will mean, since it would not represent a true convergence of opinion.

In the SHELF method, the experts are asked to judge what a *rational impartial observer*, known informally as RIO, might reasonably believe, having seen their individual judgments and listened to their discussion. Experts are advised that RIO would not completely agree with any one expert, but would see merit in the opinions of all the experts. Some arguments will have been more persuasive in the discussion, so that RIO would give more weight to some opinions than others, and it is the experts themselves who are best able to judge this.

Perhaps surprisingly, by taking the perspective of RIO, in my experience experts have never failed to reach agreement on a distribution that they believe represents a rational impartial view of their combined knowledge. Although RIO is a fictional person, it is an abstraction that allows the SHELF protocol to answer the question, "Whose probability distribution is this?" It is, at least notionally, RIO's distribution. Furthermore, it is this distribution, representing the beliefs of a rational impartial observer, taking into account the combined knowledge, expertise, and reasoning of the group, that in effect is being sought when we conduct an elicitation with multiple experts.

Table 7 shows the group judgments part of the SHELF2 record in the H2035 case study. Whereas only three methods are permitted for the individual judgments, SHELF has a fourth method that is appropriate for the group discussion. The probabilities method requires the facilitator to choose three X values and to ask the experts to agree on probabilities that RIO might assign to X being above or below those values. This method is not used at the individual judgments stage because the chosen values of X will serve as anchors to bias the experts' judgments. At the group stage, there have already been very many X values mentioned in various contexts, and introducing new ones is unlikely to cause bias. The quartile and tertile methods are still available for group judgments, but it is recommended to use different methods in the two judgment stages. The advantage of changing the method, and particularly of switching to the probabilities method, is that the experts have to consider their probabilities in a fresh way. Otherwise, for instance, if the experts are asked to assign RIO's median value they are likely to focus on their individual medians and engage in negotiating a compromise value for RIO, instead of applying the definition of the median.

The three X values are chosen based on the facilitator's feelings of what the experts' RIO judgments are likely to be. The lowest one is placed so that the facilitator thinks the experts will judge there to be a probability of 0.2 to 0.3 below this value, while the highest is placed where they might judge there to be a probability of 0.2 to 0.3 above. The middle one is aiming for a probability of about 0.4 or 0.6 below (but not 0.5, since experts may then find 0.5 an easy, lazy judgment). The objective is for the experts to make a set of judgments that they have to think about and which will span the most probable X values to facilitate fitting a distribution.

Table 7 also shows how the facilitator acts "on behalf of RIO" in challenging the group judgements. The facilitator needs to feel that the final judgments are indeed reasonable for a rational impartial observer. The fact that SHELF has two rounds of judgments is valuable here, because the facilitator can see each expert's initial judgments, as well as hearing all the group discussion. If the group judgments suggest less overall uncertainty than the variability in the individual beliefs, and if this does not seem to be justified by the intervening exchange of views, then groupthink may be at work.

It may be noted that the H2035 SHELF2 fails to record any challenge from the facilitator to the experts' group judgment of $P(X > 4\%)$. This is unfortunate because the probability of 0.2 seems small, when Table 5 shows that only expert F gave less than 0.25 to this event in their individual judgments. In this case, the judgment was questioned and the experts had reasons to retain the probability of 0.2. But without that detail in the record, a reader will rightly wonder whether the final judgments are fully defensible. Even experienced facilitators are not infallible and each elicitation offers learning opportunities. Expertise in this, as in any other area, is hard won!

5.6. The Final Distribution

The group judgments are finally used as a basis for fitting a probability distribution. In the individual elicitation round it is not important how well the fitted distributions represent each expert's initial opinions, because it is primarily their individual judgments that are used to start the group discussion. In contrast, the final fitted distribution is the outcome of the elicitation process, and so must be selected carefully and with full approval of the experts. The SHELF software allows a range of standard distributions to be fitted to the group judgments, including the normal, gamma, beta and lognormal families. In each case, the fit cannot be perfect because the software is trying to match three probabilities with just two adjustable parameters. In discussion with the experts, best-fitting distributions from several families may be considered. The facilitator will give feedback about relevant features of a fitted distribution and how it may fail to match what RIO might think. If necessary, the experts may revise one or more of their group judgments.

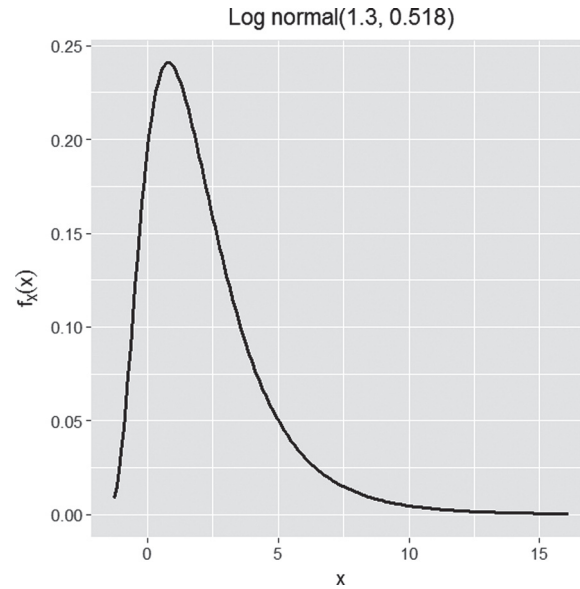


Figure 2. The final fitted distribution in the H2035 elicitation.

Table 8. Original and fitted probabilities in the H2035 case study.

	$P(X < 2\%)$	$P(X > 4\%)$	$P(X < 0\%)$
Original judgments	0.6	0.2	0.1
Fitted distribution	0.57	0.17	0.12

In the H2035 elicitation, a Student-t distribution was fitted first but rejected because it did not capture skewness. The final fitted distribution is shown in Figure 2, and Table 8 compares the three group judgments with the probabilities implied by this distribution.

Figure 2 is a lognormal distribution with origin shifted to -2, and the facilitator pointed out that it implied zero probability for X to be less than -2%. Although this hard constraint might be strictly unrealistic, the experts thought that the fitted distribution was a sufficiently accurate representation of their group judgments, and that the possibility that X might be below -2% could be ignored (from the perspective of RIO). The full H2035 SHELF2 record in the Supplementary Material gives more detail on the process of reaching this final fitted distribution.

6. Multiple Quantities

Before we can conduct an elicitation, no matter which protocol one might favour, it is necessary to decide what quantity or quantities the experts will be asked to provide judgments for. Although this may seem a simple decision—just ask about the quantities that we need their judgments for — in practice, there may be different ways to achieve the study goals, and some may promote careful, scientific judgments better than others.

6.1. Independence and Elaboration

Having more than one uncertain quantity of interest is a common challenge. Even with just two quantities, say X and Y, it is not enough to just elicit a probability distribution for each quantity because the two marginal distributions do not imply the joint distribution of (X, Y). We need to consider how X and Y may be correlated.

To understand what correlation or dependence means in this context, it is important to remember that X and Y are invariably defined to have unique values. They are not repeatable; we cannot imagine plotting a sample of (X, Y) values as a scatter diagram; there will only ever be one point in that diagram, but its location is uncertain. Terms like correlation or independence are defined as properties of an expert's subjective joint distribution for the location of that (X, Y) point. X and Y are independent if (and only if) the expert's beliefs about one quantity would not change if new information emerged about the other. Independence is a subjective judgment. For instance, I am uncertain about both the date of the next general election in the UK and the date of the next total lunar eclipse that will be visible from my home. I could find the date of the next lunar eclipse by a simple internet search, but this information would not change my beliefs about the date of the next general election, nor would an announcement of that latter date by the government change my uncertainty about the next total lunar eclipse, so these two uncertain quantities are independent for me (and probably for all readers of this article). If X and Y are independent, then for a single expert it would be enough to elicit (marginal) distributions for them separately. With multiple experts, in the SHELF protocol we can elicit X and Y separately if they would be judged independent by RIO, so this is another question to ask the experts. For the protocols using mathematical aggregation, there is no individual, real or conceptual, whose judgments are represented by the aggregated distribution, so the decision to assume independence and elicit separately is imposed by the facilitator or client.

If X and Y are not independent, then the nature and magnitude of dependence between them needs to be elicited to construct a joint distribution. This requires more complex judgments from the experts. For instance, if we first elicit median and tertiles for X from an expert we could ask for the median of Y both marginally and then conditionally upon X equalling its median, its lower tertile and its upper tertile. Eliciting conditional probability judgments like this requires the expert to make judgments about one quantity if the true value of the other variable was found to be some specific number, or to lie in some specific range. The cognitive challenge here is appreciably more difficult for experts than judgments about a single quantity. Furthermore, there is essentially no research identifying which judgments about two quantities experts understand most clearly or make most reliably. We do not know what new biases may be induced by the forms of questions. With three or more uncertain quantities, these challenges are multiplied enormously.

Two approaches to eliciting a joint distribution for two or more dependent quantities are used in practice. The first is effectively the same as for eliciting a univariate distribution—elicit as few extra judgments as necessary and then fit a joint distribution from a standard multivariate family. SHELF offers templates for two such families. One is for a set of uncertain proportions that must sum to 1, and a Dirichlet distribution is assumed. In this case, beta distributions are elicited for each proportion separately, no additional judgements are made, and a Dirichlet distribution fitted whose beta marginal distributions come as close as possible to the elicited distributions. The other template fits a Gaussian copula, using the elicited marginal distributions and pairwise "concordance probabilities," that is for each pair of quantities the expert gives their probability that both will be above or both below their separately elicited medians. The concordance probabilities imply correlations in the Gaussian copula, but with three or more quantities the set of such correlations may not correspond to a valid correlation matrix, in which case the facilitator has the challenge of helping the experts to reconcile this noncoherence. Kurowicka and Cooke (2006) describe more complex ways to elicit joint distributions with copulas, but experts are required to make even more difficult judgements, such as partial rank correlation coefficients.

The other principal way to elicit dependence is to avoid it by transforming the quantities. The suggestion above that "there may be different ways to achieve the study goals, and some may promote careful, scientific judgments better than others" refers particularly to this approach. With two quantities, X and Y, if we can identify a one-to-one transformation to $U = f(X, Y)$ and $V = g(X, Y)$, such that the experts judge U and V to be independent, then separate, independent distributions can be elicited for U and V and the joint distribution of (X, Y) deduced by the inverse transformation of variables. This is known as *elaboration*. O'Hagan (2012) discussed elaboration at length, and gives several examples of how elaboration can address different elicitation tasks, not only achieving independence of multiple quantities of interest but also such that the transformed quantities are simpler for the experts to think about individually.

In the H2035 project, the primary quantity of interest was the number of LTCs amongst people aged 85+, but the quantity X defined for elicitation in the case study was not this number but the *percentage change* in the *rate of LTCs per thousand* people. This was a result of elaborating the primary quantity of interest as the product of the elicited X, the rate per thousand in 2012 and the population aged 85+ in 2035. These were considered to be three independent quantities. To derive the distribution for the primary quantity of interest, distributions were needed for the other two quantities. There was quite good data available on the rate of LTCs per thousand in this age group in 2012, with well characterized uncertainty, so that a distribution could be derived without requiring expert elicitation. Similarly, population projections for the UK in 2035 gave an estimate for the 85+ age group, again with a well-understood uncertainty. This was an example of what O'Hagan (2012) calls "elaboration by infor-

mation sources," to simplify the elicitation of a single quantity of interest by expressing it as a function of several independent quantities for which distributions could more easily be obtained.

6.2. Many Quantities

If the number of quantities of interest is very large, methods that are good for a small number of quantities may become infeasible. If the quantities are not independent, methods based on fitting a standard family of distributions are in my opinion unreliable with more than four or five quantities. Even with independent quantities (perhaps after an elaboration step), constraints on resources and the willingness of experts to give their own time will limit the number of quantities that can be the subject of full and careful elicitation, particularly when using the SHELF or Cooke protocols.

One solution to this problem is proposed in the EFSA Guidance (European Food Safety Authority 2014). The process begins with a quick and simple elicitation for each quantity, conducted by the elicitation team using a process of *minimal assessment*, in which each quantity is given a "best estimate" m and an uncertainty measure s such that the quantity is judged 'likely' to be in the range $m - s$ to $m + s$. These terms are vague but adequate for purpose because they are used only to decide which quantities are most important to be elicited carefully. The next step is a simple sensitivity analysis in which the decision or risk model is run with every quantity set to its m, and then varying each one separately out to $m - s$ and $m + s$, to see how much the model output changes. Quantities are prioritized for full elicitation according to their sensitivities. The Guidance says that assuming $N(m, s^2)$ distributions, based on their minimal assessment, is then adequate for quantities for which the output does not vary materially.

An alternative is to apply full elicitation with the SHELF, or possibly Cooke, protocol to a selection of quantities and then to send the same experts a probabilistic Delphi questionnaire. This approach is based on the idea that the experts will first be fully trained to make the necessary probabilistic judgments, thereby mitigating the problem which arises with Delphi of experts not understanding the task properly.

Finally, elaboration can be used to reduce the number of quantities to be elicited if we are prepared to make some assumptions. An example in O'Hagan (2012) is of eliciting judgments about a dose–response relationship. In principle, this means eliciting beliefs about every point on the dose–response curve, an infinite number of quantities. However, it would be natural to assume a standard sigmoid shape, such as a probit function. Then, it is necessary only to elicit judgments about the two parameters of the curve.

In the H2035 project, there was interest not just in the rate of LTCs in people over the age of 85 but in that rate for all age groups. In the SHELF workshop, three distributions were elicited—the change of rate for over 85s, for those between 40 and 65, and for those aged under 18. These were sufficiently separated that the experts felt it reasonable for RIO to judge them independent. Change of rate for intermediate age groups could then be obtained by a quadratic interpolation between these three points, an instance of using elaboration to break correlation and to reduce the number of quantities. All three elicitations in the SHELF workshop were made under conditions of "business as now," but the H2035 project was also interested in how rates would change under a number of alternative scenarios. A few days after the SHELF workshop, the experts were sent a questionnaire to elicit their judgments about various scenarios, as the first step in a probabilistic Delphi elicitation.

7. Conclusions

Subjective expert judgments play a part in all areas of scientific activity, and should be made with the care, rigor, and honesty that science demands. One area where expert judgment is particularly prominent is the elicitation of expert knowledge in the form of probability distributions for uncertain quantities.

There are many aspects of elicitation where care is needed to avoid introducing biases to the experts' judgments. Research in psychology has identified a number of cognitive biases to which expert probabilistic judgments may be subject, but formal elicitation protocols have been developed by practitioners to minimize bias and to encourage experts to make their judgments accurately. The three leading protocols are the Cooke, SHELF, and probabilistic Delphi protocols.

The protocols were contrasted in the context of a case study in which the SHELF method was used. The protocols differ primarily in the following respects.

- In all protocols, the basic format for eliciting from a single expert is to ask for quantiles, to avoid one source of anchoring bias. The Cooke method typically asks for 5th, 50th and 95th percentiles, although in my opinion there is good experimental evidence to suggest that it is unwise to rely on judgments of 5th and 95th percentiles to characterize uncertainty. SHELF usually asks for median and quartiles or median and tertiles, and does so in a structured sequence of judgments designed to minimize other sources of anchoring bias. The probabilistic Delphi protocol, as defined originally in the European Food Safety Authority (2014) guidance, employs the same sequence of judgments as SHELF, although the IDEA variant is less prescriptive and may use the same judgments as in the Cooke protocol.

Although these protocols provide important disciplines for eliciting judgments from a single expert, or even for a single scientist eliciting his or her own judgments, formal expert knowledge elicitation usually involves consulting multiple experts. It is here that the three protocols are most strongly differentiated.

- When multiple experts are consulted, SHELF uses behavioral aggregation, with the experts meeting together and agreeing on a final probability distribution representing what a rational impartial observer (RIO) might believe. The other two methods rely on mathematical aggregation, in which a distribution is elicited from each expert and the distributions combined by a formula. In the Cooke protocol, this involves weighting the experts according to their performance on "seed variables." The distributions resulting from a mathematical aggregation do not represent the belief of any person, real or conceptual, and so their status as subjective probability distributions is unclear.

All the leading protocols demand substantial input of resources, particularly the time of the experts, the facilitator and those organizing the elicitation. Elicitation should not be seen as a cheap option. However, the result is information, elicited from experts in a rigorous and scientific manner. To obtain experimental data of the same information content and quality will almost invariably require much more resources.

The purpose for which the information is sought may not require such high standards, and in that case it would be possible to reduce and simplify the process. Judgments may be sought from fewer experts, possibly just one, and the elicitation may be done without the aid of a trained facilitator. For instance, if the purpose is Bayesian statistical inference, if the elicitation is of prior distributions and if the data will provide strong information, then formal, rigorous elicitation may be needed for only a small number of parameters, and perhaps for none. Similarly, if the elicited distributions will be for inputs to a decision model, the decision may be insensitive to the uncertainty in some of those inputs, in which case it will not be necessary to elicit them so carefully and "minimal assessment" may be adequate.

Acknowledgments

I am grateful to the Department of Health for permission to use the H2035 case study and in particular to Matt Edwards, the Department's Head of Horizon Scanning and International Workforce Information and Analysis, for several helpful comments on this article. It should be noted that some details of the elicitation have been changed in this article, to more clearly illustrate some concepts but without affecting the work's implications within the Horizon 2035 project.

I am also grateful to the reviewers for making many very useful suggestions that have greatly improved this article.

References

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J. (2018), "The Role of Expert Judgement in Statistical Inference and Evidence-Based Decision-Making," *The American Statistician*, this issue, DOI: 10.1080/00031305.2018.1529623. [69]

Burgman, M. A. (2015), *Trusting Judgements: How to Get the Best Out of Experts*, Cambridge, UK: Cambridge University Press. [69]

CFWI (2015), *Horizon 2035 – Future Demand for Skills: Initial results*. National Archives, available at *https://www.gov.uk/government/publications/horizon-2035-future-demand-for-skills-initial-results* [73]

Cooke, R. M. (1991), *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford, UK: Oxford University Press. [73]

Cooke, R. M., and Goossens, L. H. J. (2008), "TU Delft Expert Judgement Database," *Reliability Engineering and System Safety*, 93, 745–756. [73]

Cooke, R. M., Wittmann, M. E., Lodge, D. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., and Mason, D. M. (2014), "Out-of-sample Validation for Structured Expert Judgement of Asian Carp Establishment in Lake Erie," *Integrated Environmental Assessment and Management*, 10, 522–528. [73]

de Finetti, B. (1937), "Foresight: Its Logical Laws, its Subjective Sources," Reprinted (and translated from the original French) in *Studies in Subjective Probability*, eds. H. E. Kyburg, Jr. and H. E. Smokler, New York: Wiley, 1964. [70]

de Finetti, B. (1970), *Theory of Probability: A Critical Introductory Treatment* (translation by A. Machi and A. F. M. Smith, 1974–5), 2 volumes. Wiley. [70]

European Food Safety Authority (2014), "Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment," *EFSA Journal* 2014, 12 (6), 3734. [69,72,74,80]

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978), "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation," *Journal of Experimental Psychology: Human Perception and Performance*, 4, 330–344. [71]

French, S. (1985), "Group Consensus Probability Distributions: A Critical Survey," in *Bayesian Statistics 2*, J. M. Bernardo et al (eds.), Oxford, UK: Oxford University Press, pp. 183–202. [72]

Gosling, J. P. (2018), "SHELF: The Sheffield Elicitation Framework," in *Elicitation: The Science and Art of Structuring Judgement*, eds. L. C. Dias, A. Morton and J. Quigley, Springer, pp. 61–93. [73]

Hanea, A. M., McBride, M. F., Burgman, M. A., and Wintle, B. C. (2018), "Classical Meets Modern in the IDEA Protocol for Structured Expert Judgement," *Journal of Risk Research*, 4, 417–433. [69,73]

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., and Wintle, B. C. (2018), "A Practical Guide to Structured Expert Elicitation using the IDEA Protocol," *Methods in Ecology and Evolution*, 9, 169–1980. [73]

Janis, I. L. (1972), *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*, Boston, MA: Houghton Mifflin. [72]

Kahneman, D. (2011), *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux. [70]

Kurowicka, D., and Cooke, R. M. (2006), *Uncertainty Analysis with High Dimensional Dependence Modelling*, Chichester: Wiley. [69,79]

Oakley J. E., and O'Hagan, A. (2016), "SHELF: the Sheffield Elicitation Framework" (version 3.0), School of Mathematics and Statistics, University of Sheffield, UK, available at *http://tonyohagan.co.uk/shelf*. [73,74]

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006), *Uncertain Judgements: Eliciting Expert Probabilities*, Chichester, UK: Wiley. [69]

O'Hagan, A. (2012), "Probabilistic Uncertainty Specification: Overview, Elaboration Techniques and their Application to a Mechanistic Model of Carbon Flux," *Environmental Modelling and Software*, 36, 35–48. [69,79,80]

Parducci, A. (1963), "The Range-Frequency Compromise in Judgment," *Psychological Monographs*, 77, 1–50. [71]

Ramsey, F. P. (1926), "Truth and Probability," reprinted in *Studies in Subjective Probability*, eds. H. E. Kyburg, Jr. and H. E. Smokler, 2nd edition, New York: R. E. Krieger Publishing Company, pp. 23–52. [70]

Rowe, G., and Wright, G. (1999), "The Delphi Technique as a Forecasting Tool; Issues and Analysis," *International Journal of Forecasting*, 15, 353–375. [73]

Savage, L. J. (1954), *The Foundations of Statistics*. New York: Wiley. [70]

Tversky, A., and Kahneman, D. (1974), "Judgments under Uncertainty: Heuristics and Biases," *Science*, 185, 1124–1131. [70]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing

Lee Kennedy-Shaffer

To cite this article: Lee Kennedy-Shaffer (2019) Before $p<0.05$ to Beyond $p<0.05$: Using History to Contextualize p-Values and Significance Testing, The American Statistician, 73:sup1, 82-90, DOI: 10.1080/00031305.2018.1537891

To link to this article: https://doi.org/10.1080/00031305.2018.1537891

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 17463

View related articles

View Crossmark data

Citing articles: 25 View citing articles

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, 82–90: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1537891

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing

Lee Kennedy-Shaffer

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

ABSTRACT

As statisticians and scientists consider a world beyond $p < 0.05$, it is important to not lose sight of how we got to this point. Although significance testing and p-values are often presented as prescriptive procedures, they came about through a process of refinement and extension to other disciplines. Ronald A. Fisher and his contemporaries formalized these methods in the early twentieth century and Fisher's 1925 *Statistical Methods for Research Workers* brought the techniques to experimentalists in a variety of disciplines. Understanding how these methods arose, spread, and were argued over since then illuminates how $p < 0.05$ came to be a standard for scientific inference, the advantage it offered at the time, and how it was interpreted. This historical perspective can inform the work of statisticians today by encouraging thoughtful consideration of how their work, including proposed alternatives to the p-value, will be perceived and used by scientists. And it can engage students more fully and encourage critical thinking rather than rote applications of formulae. Incorporating history enables students, practitioners, and statisticians to treat the discipline as an ongoing endeavor, crafted by fallible humans, and provides a deeper understanding of the subject and its consequences for science and society.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
Education; Foundational issues; Hypothesis testing; Inference; Probability

1. Introduction

With new journal policies, conferences, and special issues, it is easy to view the debate around p-values and hypothesis testing as a modern invention. For many scientists whose primary connection to statistics is through these methods, the debate may seem like a challenge to the received wisdom of their profession, a rebuke to the way they have been using statistics for decades. For students learning the field, it can seem bewildering, and they might be tempted to replace one decontextualized methodology with another. Indeed, as Gerd Gigerenzer (2004, p. 589) writes, the anonymizing of the roots of the p-value and hypothesis testing has contributed to the idea that "they were given truths" and encouraged the "mindless" use of these procedures, to the point of misuse and abuse. But for those who have studied statistics, and, in particular, studied the progression of statistical theory, the debates are not a sudden attack on a completely accepted paradigm, and the statistics themselves did not arise wholly formed to be prescriptively applied. Rather, the statistics arose through the ongoing process of scientific discovery, with contributions by many along the way.

In order to properly understand the challenges that face statistics and its applications in science, medicine, and policy today, and to meet those challenges in the future, we must consider the history of the discipline and its most prominent methods. It is a history that is too poorly known, even among statisticians, but it is rich in characters, personal grudges, and academic debates. Gigerenzer (2004, pp. 587–588) laments this lack of focus on the history and controversy when he relates the story of a psychological statistics textbook author who removed all mention of Thomas Bayes, Ronald A. Fisher, Jerzy Neyman, and Egon Pearson. In a similar vein, Stephen Ziliak and Deirdre McCloskey (2008, p. 232) argue that the conscious erasure of William S. Gosset from the history contributed to the dominance of Fisher's paradigm and reduced the prominence of competing ideas.

In Section 2, I trace the use of statistical reasoning similar to the modern p-value before 1900, demonstrating that the statistic and the use of thresholds did not arise from Karl Pearson and Fisher alone. In Section 3, I briefly describe the contributions of Pearson, Gosset, and Fisher, both covering the similarities among them and highlighting some of the debates that occurred as early as the 1920s when Fisher's *Statistical Methods for Research Workers* began to put the p-value in the hands of experimenters. In Section 4, I point out some of the challenges that emerged in response to Fisher's paradigm, focusing especially on those arising from Gosset, Neyman, and Egon Pearson, and from the Bayesian paradigm. These sections are far from comprehensive; rather, they seek to provide an overview of the history that can spur thought and encourage further research. In Section 5, I present resources that can be used for that research and as teaching tools. I also discuss how the historical debates relate to modern arguments surrounding the p-value and how that can encourage statisticians to craft a more useful and durable response to this controversy. I further describe the role this history can play in education in formal classroom settings and in research and collaboration settings.

Understanding how p-values and 0.05 came to occupy their prominent role in twentieth century statistics reminds us that

CONTACT Lee Kennedy-Shaffer lee_kennedyshaffer@g.harvard.edu Department of Biostatistics, Harvard T.H. Chan School of Public Health, 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.

these "arbitrary" thresholds came about through work to make mathematical statistics more practical and useful for experimentalists. But these efforts were never without controversy. Learning this history will help statisticians better appreciate the translational challenges of their own work by improving understanding of the fact that, since its inception, the modern field of statistics has grappled with the balance between mathematical rigor and practical use to scientists. Those pushing the boundaries of knowledge in the discipline will surely face this balance in their own work. They will have to consider, like the statisticians of the early twentieth century did, how others will use their theories.

Learning this history will help practitioners understand that no method is sacred and that all methods are products of the era in which they were born and the functions to which they have been applied. As technology, mathematics, and science develop, new methods or adjustments to old methods will be needed as the underlying assumptions no longer apply, whether in a world of early electronic computing devices or a world of big data.

Learning this history will help students access the discipline by learning of the faults, personal and professional, of those who came up with today's commonly used statistics and help them understand statistics as a living discipline rich with ongoing debate and new understandings. Indeed, one can find many parallels between today's debate and the controversies that arose with the development of p-values and significance testing, framing the ASA statement and subsequent discussion as another step in the ongoing evolution of the discipline of statistics.

2. A World Before Fisher

The p-value is generally credited to Karl Pearson's (1900) article in his journal *Biometrika*; Ronald A. Fisher's (1925) *Statistical Methods for Research Workers* then formalized the concept and expanded its reach to experimenters (Hubbard 2016, p. 14). But statistics similar to p-values and probabilistic reasoning akin to hypothesis tests existed well before then. Both Stigler (1986) and David and Edwards (2001) point to John Arbuthnott's (1710) "An Argument for Divine Providence" as perhaps the earliest use of probabilistic reasoning that matches that of a modern null hypothesis test. Using birth data from London, Arbuthnott (1710) notes that births of males exceeded births of females for 82 years in a row. Supposing that the probability of males exceeding females in a year is 50%, and implicitly assuming independence across the years, Arbuthnott calculates the miniscule probability of this 82-year pattern. "From whence it follows," Arbuthnott (1710, p. 189) confidently concludes, "that it is Art, not Chance, that governs." Any modern student who has run a test of proportions would notice the reasoning, see Arbuthnott's calculation of a p-value of 2.07×10^{-25}, and confirm his rejection of the null hypothesis that each year has an independent probability of 50%. The mathematically-inclined physician's goal in this endeavor was to demonstrate the work of "Divine Providence" in the sex distribution (Arbuthnott 1710, p. 186). Many statisticians would recognize the flaw in this reasoning: the lack of a clearly stated alternative hypothesis that would be logically implied by a rejection of the null hypothesis. Gigerenzer (2004, p. 588) decries this "null ritual" used by experimentalists who often fail to properly specify "alternative substantive hypotheses."

In the nineteenth century, French mathematicians used similar methods to analyze a wide variety of data. In celestial mechanics, Pierre-Simon Laplace (1827, p. S.30) found a small value for a statistic closely related to the modern p-value and concluded that it indicated with a high likelihood that the discrepancy in the measurements was thus "not due solely to the anomalies of chance." Stigler (1986, p. 151) notes that Laplace himself appealed to a 0.01 significance level in his work. Stigler (1986, pp. 151–153) further highlights several errors implicit in Laplace's analysis, errors that would be familiar to students and critics of modern hypothesis testing: improper assumptions of independence and improper estimation of variance.

Not long after, Siméon-Denis Poisson used a quantity equal to one minus a modern p-value in describing patterns in the outcomes of French jury trials. Two comparisons he makes are particularly instructive. In one, he finds a p-value of 0.0897, a value not large enough for him to conclude that there has been a change in causes (Poisson 1837, p. 373). Shortly thereafter, a p-value of 0.00468 leads Poisson to believe that in that case there is a "real anomaly in the votes of juries" (Poisson 1837, pp. 376–77). Poisson's conclusions in these two cases, nearly a century before Fisher's work, would comport with a 0.05 (or 0.01) significance threshold, but do not specify a threshold he used. Poisson (1837, p. 375) also refused to make a causal statement from his identified associations, noting that "the calculation cannot teach us" this answer.

Antoine Augustin Cournot formulated the p-value in fairly explicit terms, noting that as a measure of the importance of some discrepancy it combines the size of the effect and the sample size (Cournot 1843, p. 196). Cournot (1843, pp. 196–197) also issues a warning about the narrow-minded use of probabilistic statements, noting that this p-value does not fully capture the importance of the effect size and "does not at all measure the chance of truth or of error pertaining to a given judgment." With a little modernization of language, Cournot could have written principles 2, 5, and 6 of the ASA Statement (Wasserstein and Lazar 2016).

In 1885, Francis Ysidro Edgeworth provided a more formal mathematical underpinning for the significance test and gave a simple example of how to use the standard deviation (he used the "modulus," equal to the standard deviation multiplied by the square root of two) to perform a significance test on a given parameter (Edgeworth 1885, pp. 184–185). Using a threshold of twice the "modulus," Edgeworth (1885) constructed a test that would be equivalent to a modern two-sided $\alpha = 0.005$. Stigler (1986, p. 311) notes that this "was a rather exacting test" and that Edgeworth also considered smaller differences as "worthy of notice, although he admitted the evidence was then weaker."

The existence of these tests of significance and p-value-like quantities long before the twentieth century demonstrate that this method of inference had an alluring rationale for practitioners in a variety of fields. Their errors in interpretations and words of caution, however, presage the controversies that would follow. Throughout the twentieth century, many of the technical probability results needed for modern significance testing arose through the theory of errors, by which astronomers and other physical scientists combined measurements and discarded out-

liers (Gigerenzer, Swijtink, and Daston 1989, pp. 80–84). These developments allowed Pearson, Gosset, and Fisher to make key contributions that formalized, shaped, and popularized the modern form of significance tests.

3. R.A. Fisher: the Experimentalist Statistician

In the early twentieth century, the forerunners of modern statistics began to determine the properties of various useful distributions. Karl Pearson (1900) described the χ^2 distribution and uses of the χ^2 statistic, including its use in tests of independence for proportions. Pearson (1900, pp. 157–158) here denoted by P the "chances of a system of errors with as great or greater frequency than that denoted by χ." In an example involving dice throws, Pearson (1900, pp. 167–168) finds $P = 0.000016$ on a null distribution of equal probability of each face appearing and claims that "it would be reasonable to conclude that dice exhibit bias towards the higher points." The combination of this type of probabilistic reasoning and a distribution with many practical uses made the p-value more approachable and brought it more or less to its modern formulation. W. Palin Elderton built on Pearson's work and produced tables of values for this distribution that would enable investigators to test the goodness of fit. His article, published in *Biometrika* in 1902, devoted roughly half of its space to these tables (Elderton 1902). Ziliak and McCloskey (2008, pp. 199–202) note that Pearson was soon teaching his students, and enforcing as a rule for authors seeking publication in *Biometrika*, that three probable errors, or two standard errors, represented "certain significance."

William Sealy Gosset, the head experimental brewer at Guinness publishing under the pseudonym "Student" (1908, p. 25), found a curve "representing the frequency distribution of values of the means of such samples," that is, samples from a normal or "not strictly normal" distribution, "when these values are measured from the mean of the population in terms of the standard deviation of the sample." This so-called Student's t distribution is now taught in introductory and applied statistics courses, as it forms the basis for a substantial number of inferential procedures. Gosset's initial paper focused as much on illustrating examples of the utility of this curve as on the mathematical justification for its use, and he produced numerous tables to enable others to use it. He calculated statistics akin to the p-value and drew conclusions from extreme values of these. For one drug trial, he regarded a statistic equivalent to $p = 0.0015$ as "such a high probability," it would be in practical matters "considered as a certainty" (Student 1908, p. 21). For Gosset, however, whether an effect existed or not was less important than its impact, and he saw the use of the tests more in determining the "pecuniary advantage" of one decision versus another (Ziliak and McCloskey 2008, pp. 18–19). That is, any conclusion must rest on effect size and the relative loss and gain of any potential decision; this will be a recurring theme in the debate between competing frameworks for testing discussed below.

Ronald A. Fisher, who had corresponded with Pearson and Gosset at various points, was well aware of these advances and thus of the use of significance tests. His work, especially a series of three monographs published in the 1920s and 1930s, would expand the reach of significance tests, promote their use (and the use of statistically rigorous experimental design and analysis more broadly) to researchers, and provide tables that enabled investigators to conduct such tests.

Fisher, employed at the time at Rothamsted Experimental Station, an agricultural research institution, "extended the range of tests of significance" using the theory of maximum likelihood commonly used today and conceived of tests for small sample problems (Box 1978, p. 254). In 1922, he published three key manuscripts which covered the theoretical foundations of maximum likelihood estimation and the concept of the likelihood (Fisher 1922c), the use of Pearson's χ^2 distribution to calculate p-values from contingency tables (Fisher 1922b), and the use of Student's t distribution to conduct significance tests on regression coefficients (Fisher 1922a). In 1925, he published the first edition of *Statistical Methods for Research Workers*, which sought, in his words, "to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data" (Fisher 1925, p. 16). The book discusses in detail the meaning and practical implications of "P", the statistic now known as the p-value, and suggests 0.05 as a useful cutoff:

> The value for which $P = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a negative result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty (Fisher. 1925, p. 47)

This simple paragraph demonstrates the probability-based definition of the p-value that is commonly misunderstood: that it is the probability of a result as or more extreme than the observed result given that the null hypothesis is true (Greenland et al. 2016). Additionally, it makes immediately apparent why 0.05 is convenient: it is roughly equivalent to the probability of being more than two standard deviations away from the mean of a normally distributed random variable. In this way, 0.05 can be seen not as a number Fisher plucked from the sky, but as a value that resulted from the need for ease of calculation at a time before computers rendered tables and approximations largely obsolete. This particular value had the added bonus of corresponding to three "probable errors," a measure of spread of the normal distribution used commonly in early statistics but now largely forgotten (Stigler 1986, p. 230). So a useful rule of thumb could be given to researchers on either of two scales of measuring the spread of the distribution. Later, in applying the statistic to the χ^2 distribution, Fisher (1925, p. 79) remarks that "[w]e shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy."

Statistical Methods was most valuable in the hands of experimentalists due to its explanations of tests and estimation procedures, illustrative examples, and a wealth of user-friendly tables. The tables further entrenched the use of Fisher's preferred p-value cutoffs by displaying the calculated figures so that an

investigator looked up a desired probability level for the distribution and found the quantile of the statistic that corresponded to it. Among the levels presented one almost always found 0.05 and 0.01 (Fisher 1925). As Fisher's biographer and daughter, Box (1978, p. 246), states, "[b]y this means he produced concise and convenient tabulations of the desired quantities" and presented values "that were of immediate interest to the experimenter." It is this accessibility that made the book popular among practicing experimentalists who "had not a hive of staff humming at their desk calculators," but it did not endear him to more rigorous mathematicians (Box 1978, pp. 242–246). And through these presentations, which Fisher (1935; Fisher and Yates 1938) continued and expanded in *The Design of Experiments* and the 1938 compilation of tables with Francis Yates entitled *Statistical Tables for Biological, Agricultural, and Medical Research*, he set the standard for the use of p-values and statistical inference in a variety of forms of research. One may note, however, that Fisher's tables show that he did not think 0.05 was one size fits all; if 0.05 worked in every setting, there would have been only one column in each table.

The history of the tables presented in *Statistical Methods* is interesting in itself and further demonstrates how these values came to be presented; it also foreshadows forthcoming schisms regarding these tests. Hubbard (2004, p. 311) notes that Pearson's *Biometrika* denied Fisher permission to use Elderton's table of χ^2 probabilities in his monograph. When he created his own version, according to Pearson et al. (1990, p. 52), Fisher "gave the values of χ^2 for selected values of P . . . and thus introduced the concept of nominal levels of significance." Because of this change from Elderton's table to Fisher's, for users of the table in *Statistical Methods* and its successors in *Statistical Tables*, it would be easier to compare a calculated χ^2 value to a set threshold of significance rather than find the precise p-value. For tables of the t statistic, as Ziliak and McCloskey (2008, p. 229) note, "Fisher himself *copyrighted again Gosset's tables in his own name*" in *Statistical Methods* (emphasis in original). Through this action, which left Gosset's name out of the book except in the phrase "Student's t", Fisher removed Gosset from the history of his own statistic, hid his contributions, and, more importantly, hid his competing philosophy on how the statistic should be used (Ziliak and McCloskey 2008, pp. 230–232). Reprinters of the table and those who used it in applied research would encounter only Fisher's versions and his interpretations.

Following Fisher, the use of p-values grew among experimentalists. In the United States, they were particularly encouraged by Harold Hotelling of Stanford University, who called some of the tables in *Statistical Methods* "indispensable for the worker with moderate-sized samples" (quoted in Ziliak and McCloskey 2008, p. 234). George Snedecor of Iowa State University played a crucial role as well, continuing to develop the methods and promoting their use in scientific fields (Hubbard 2016, p. 21). Psychologists, sociologists, political scientists, and economists all found the innovations useful (Hubbard 2016, pp. 22–27). Thus, the p-value spread not only across oceans but beyond the natural sciences to the social sciences, echoing its use by Poisson a century earlier.

The use of 0.05 as a cutoff became customary, though not all-encompassing. Fisher's student L. H. C. Tippett (1931, p. 48), wrote in *The Method of Statistics* that the 0.05 threshold was "quite arbitrary" but "in common use." Lancelot Hogben (1957, p. 495), two decades later, wrote that Fisher's claim that the cutoff was in usual practice was "true only of those who rely on the many rule of thumb manuals expounding Fisher's own test prescriptions." For scientists and students today, perhaps the prominence of this admittedly arbitrary cutoff is difficult to comprehend. However, they need only consider a time before computers and compare the calculation of a p-value by hand from one of Fisher's or Gosset's or Pearson's formulae to the ease by which one can determine whether a statistic meets a threshold by reference to one of Fisher's tables. It will immediately become clear how Fisher's standard became the gold standard. Fisher led other tables to adopt his format through his role as secretary of the Tables Committee of the British Association (Box 1978, p. 247), ensuring that future statisticians who sought to reach experimentalists would need to reconcile their methods to this framework. Thus, "$p < 0.05$" could grow to the prominence it holds today.

4. Challenges to Fisher's View

The other piece of history often lost in the presentation of p-values is that statisticians brought many challenges to Fisher's framework as soon as it was presented. As Fisher was writing his manuscripts, Jerzy Neyman and Egon Pearson (1933) were preparing their own framework for hypothesis testing. Rather than focusing on falsifying a null hypothesis, Neyman and Pearson presented two competing hypotheses, a null hypothesis and an alternative hypothesis, and framed testing as a means of choosing between them. The decision then must balance two types of error, one made by incorrectly rejecting the null hypothesis when it is true (Type I Error) and one made by incorrectly accepting the null hypothesis when it is false (Type II Error). More generally, one can consider the class of "admissible alternative hypotheses" of which the null hypothesis is a member (Neyman and Pearson 1933, p. 294); the goal is then to compare the null hypothesis to the alternative that imparts the highest likelihood on the observed data. They propose a class of tests that, for a given limit of Type I Error, minimize the risk of Type II Error, the so-called most powerful tests. The Type I Error risk, often called the significance level and denoted α, is commonly set at 0.05 (or 0.01), as the pair noted in their paper. The Type II Error risk, often denoted β, is equal to one minus what we now call the power of the test.

This procedure has many similarities to Fisher's framework that uses the p-value as a continuous measure of evidence against the null hypothesis; indeed, in many cases, Fisher's choice of test statistic corresponds to a reasonable choice of alternative hypothesis in a Neyman–Pearson most powerful test (Lehmann 1993, pp. 1243, 1246). In those cases, $p < \alpha$ if and only if the most powerful α-level test would reject the null hypothesis. Nonetheless, the two factions debated fiercely the merits of each version. In one sense, the controversy can be regarded as a debate over the role of the statistician and of the test itself: should the test be considered as a step along the way to deeper understanding, a piece of evidence among many to be considered in crafting and supporting a scientific theory? Or should it be considered as a guide to decision-making, a

way to choose the next behavior, whether in a practical or experimental setting? Fisher's writings generally support the former view, taking the test and the p-value as a piece of evidence in the scientific process, one that he wrote "is based on a fact communicable to, and verifiable by, other rational minds" (Fisher 1956, p. 43). For Neyman and Pearson, on the other hand, to accept a hypothesis means "to act as if it were true" and thus the hypotheses and error probabilities should be chosen in light of the consequences of making either decision (Gigerenzer, Swijtink, and Daston 1989, p. 101).

In a practical way, the Neyman–Pearson view also meant considering the reasonability of alternative hypotheses. Berkson (1938, p. 531) provided an application of this question, discussing how someone familiar with the data would only truly reject the null hypothesis "on the willingness to embrace an alternative one." The debate took on a variety of aspects, however, including being somewhat representative of a larger controversy over the role of mathematical rigor in statistics, with Fisher assailing Neyman and Pearson as mathematicians whose work failed to reflect the nuances of scientific inference (Gigerenzer, Swijtink, and Daston 1989, p. 98). It also covered differences in the role assigned to a statistical model of data and decision-making, which in turn relate to fundamental probability questions about defining populations and samples (Lenhard 2006). All of these differences were heightened and perhaps even exaggerated by "the ferocity of the rhetoric" (Lehmann 1993, p. 1242).

While this debate raged in the halls of academic statisticians for decades (and, even today, attempts are made to clearly define the differences or reconcile the two theories), experimentalists began to follow a third way, an "*anonymous* hybrid consisting of the union of the ideas developed by" Fisher and Neyman–Pearson (Hubbard 2004, p. 296, emphasis in original). Often, reporting of results will include a comparison of the p-value to a threshold level (e.g., 0.05) to claim existence of an effect, reporting of the p-value itself, and relative measure of evidence terms such as "highly significant," "marginally significant," and "nearly significant." This leads to what Hubbard (2004, p. 297) calls a "confusion between p's and α's" among applied researchers, as seen in textbooks, journal articles, and even publication manuals. This confusion undermines the rigorous Neyman-Pearson interpretation of limiting error to a prespecified level α. And the role of the value of p as a quantitative piece of ongoing scientific investigation (including using null hypotheses that are not a hypothesis of zero effect) favored by Fisher is lost to the decision-making encouraged by a statement of significance or lack thereof. Neither Fisher nor Neyman and Pearson would approve of this hybrid, though it has been institutionalized by textbooks and curricula, especially in applied settings. Its popularity owes a great deal to its simplicity and the ability of applied researchers to perform this "ritual" of testing in a more mechanized fashion (Gigerenzer 2004).

While this debate was ongoing, a revival of another paradigm of probability gained steam. Based on a crucial theorem by Thomas Bayes that was published in 1763, the "inverse probability" or Bayesian viewpoint embraced the subjectivity of statistical analysis (Weisberg 2014, sec. 10.2). With regard to testing, the Bayesian approach allows a researcher to calculate the probability of a specific hypothesis given the observed data, rather than the converse, which is what the Fisher and Neyman–Pearson approaches do. These views gained considerable traction after Leonard J. (Jimmie) Savage's 1954 publication of *The Foundations of Statistics*, which also replied to anticipated objections to the paradigm. His work builds on that of Bruno de Finetti (1937) and Harold Jeffreys (1939).

Bayesian ideas were present before then, however, as Fisher (1922a, pp. 325–330) included in his article on maximum likelihood a rejection of Bayesian approaches. Fisher (1922c, p. 326) even notes that the works of Laplace and Poisson, discussed above, "introduced into their discussions of this subject ideas of a similar character" to inverse probability. While this article is far too short to cover the debate between the various Bayesian approaches and the frequentist approaches of Fisher, Gosset, and Neyman–Pearson, Savage's book is a useful starting point, and a higher-level summary can be found in Weisberg (2014). Sharon McGrayne (2011) provides a very accessible overview of the Bayesian approach, its history, and the common use of Bayesian methods in practical research even while it was philosophically rejected by statisticians. These debates, too, are ongoing, with Bayesians or frequentists holding more sway in different scientific fields (Gigerenzer, Swijtink, and Daston 1989, pp. 91, 105), and Bayesian approaches are often suggested as alternatives to p-values, as discussed below.

In addition to these broad philosophical challenges, statisticians and scientists objected to Fisher's p-value on practical grounds. Gosset wrote to Fisher and to Karl Pearson of the importance of considering effect sizes and, indeed, arranging experiments "so that the correlation should be as high as possible" (quoted in Ziliak and McCloskey 2008, p. 224). Fisher's own co-author, Francis Yates, wrote in 1951 (p. 33) of his concern that experimenters were regarding "the execution of a test of significance as the ultimate objective." Fisher (1956, p. 42) himself later wrote that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses." His book even included a chapter entitled "Some misapprehensions about tests of significance" (Fisher 1956, p. 75). His writings on the matter, however, are sometimes contradictory and admit several interpretations (Gigerenzer, Swijtink, and Daston 1989, p. 97). Medical statistician Joseph Berkson (1942, p. 326) feared a disconnect between significance testing and "ordinary rational discourse," especially in applying a rule to these tests without regard to "the circumstances in which it is applied" (p. 329). The International Biometric Society's British Regional President warned in 1969 that significance tests might "exercise their own unintentional brand of tyranny over other ways of thinking" (Skellam 1969, p. 474). Psychologist William Rozeboom (1960) wrote of the failings of p-values and significance testing, including the uncritical appeals to 0.05, in 1960. Other psychologists and social scientists soon followed (Bakan 1966; Meehl 1967; Skipper et al. 1967).

These arguments, which began as soon as the paradigm-defining works were published, would all be familiar to those following the modern debates and resonate in the ASA statement (Wasserstein and Lazar 2016). Discussing and teaching the modern debate without acknowledging its historical roots does a disservice not only to those thinkers who engaged in the debate, but to the statistics profession as a whole.

5. History as Context to Inform the Present Debate

In this article, I have endeavored to recount only a small part of the history of *p*-values and significance testing, which itself forms only a small part of the history of probability and statistics. Much more can be found on these subjects. David Salsburg's (2001) *The Lady Tasting Tea* provides a highly accessible treatment. Stephen Stigler's (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900* covers the early history of the *p*-value and how it fit into notions of reasoning about uncertainty; Theodore Porter's book *The Rise of Statistical Thinking, 1820–1900*, also published in 1986, covers the latter end of this pre-Pearson history. H. A. David and A. W. F. Edwards's (2001) *Annotated Readings in the History of Statistics* highlights primary source material relevant to this history. In books published twenty-five years apart, Gigerenzer et al. (1989) and Weisberg (2014) describe the rise of the dominant modern mathematical conception of probability and how that influenced and was influenced by the rise of these statistics and of data-driven sciences. Stephen Ziliak and Deirdre McCloskey (2008) describe the rise of these statistics in the early twentieth century in great detail, focusing on reviving the forgotten role of Gosset through presentation and interpretation of his archival materials; they also describe the spread of the Fisherian paradigm in economics, psychology, and law, and the consequences of that spread. Finally, Donald MacKenzie's (1981) *Statistics in Britain, 1865–1930* discusses Fisher and his immediate predecessors in detail, focusing especially on the effects of the British social context on the work of Francis Galton, Karl Pearson, and Fisher and on how eugenics shaped the statistical work of the three men and the rise of statistics in Britain.

Articles and books exploring the use of statistical inference, especially hypothesis testing, in specific fields can be informative of this history as well: Morrison and Henkel (1970) write of the controversies in the social sciences; Hubbard (2016) discusses the use of statistics in the management sciences as well as the social sciences; Hubbard (2004) and Chambers (2017) describe the controversies in psychology; Kadane, Fienberg, and DeGroot (1986) discuss the use of statistics in the field of law with several case studies; I (Kennedy-Shaffer 2017) cover some of this history with a focus on significance testing at the United States Food and Drug Administration. These various detailed accounts, among others, clarify the lessons that statisticians and practitioners can take from this history and provide ample material for statistics educators to incorporate this history into their formal and informal teaching.

5.1. Lessons for Statisticians and Practitioners

The history of the *p*-value and significance testing is useful for statisticians and scientists who use statistical methods today for a variety of reasons. The history helps clarify today's debates, adding a long-term dimension to modern discussions. In this way, it illuminates the factors that drive the creation of statistical theory and methods and what enables them to catch on in the broader community. Understanding these factors will help statisticians respond to today's debates and consider how proposed solutions to problems that have arisen will play out in the scientific community today and in the future.

First of all, the history clarifies the debates that are occurring today; in particular, many of the objections raised to *p*-values by modern scientists (and in Wasserstein and Lazar (2016) and the accompanying Online Discussion) were raised by contemporaries of Fisher. One particular aspect, the importance of considering effect size rather than simply statistical significance, was the crux of the difference between Fisher's framework and Gosset's (Ziliak and McCloskey 2008). Ziliak (2016) reiterates this connection in an article in the *Online Discussion*, demonstrating the relevance of historical debates to today's discussion. A thread of argument from Fisher's earliest critics (and indeed Cournot and Edgeworth before him) to Rothman (2016) indicates that the de-emphasizing of effect size in favor of the *p*-value is an easy mistake to make and one that needs to be addressed. Similarly, debates have continued over the conflating of Fisher's paradigm with the Neyman-Pearson approach, as discussed above. Lew (2016) describes how these different inferential questions have become hybridized. Discussions of power and the role of statisticians in the design of experiments arise in the commentaries by Berry (2016) and Gelman (2016). While their approaches are quite different, Fisher certainly understood that argument, writing an entire book on how to properly design experiments (Fisher 1935); Gosset, too, participated in this discussion, disagreeing with Fisher on key aspects (Ziliak and McCloskey 2008, p. 218). And the Bayesian-frequentist debate continues today, unresolved after decades of discussion. Among others, Benjamin and Berger (2016) and Chambers (2017, pp. 68–73) promote the potential use of Bayesian hypothesis testing as an alternative to *p*-values and significance testing.

It would be easy to be disheartened by this history. If we have been debating these ideas, raising similar arguments for a century, what hope do we have of solving them now? And, as Goodman (2016) puts it, "what will prevent us from dusting this same statement off 100 years hence, to remind the community yet again of how to do things right?" The history may provide the answer here. In particular, a closer look at how Fisher's ideas spread and how the hybridization of the Fisher and Neyman-Pearson paradigms occurred, processes discussed here only briefly, can inform us of what makes statistical methods catch hold in the broader scientific, policymaking, and public communities. Berry (2016) notes that statisticians should not seek to "excuse ourselves by blaming nonstatisticians for their failure to understand or heed what we tell them." But we can understand why they fail to heed us. Benjamini (2016) notes that the *p*-value was so successful in science because it "offers a first-line defense against being fooled by randomness." That is, it was useful to nonmathematicians in giving them a quantitative basis for addressing uncertainty. Additionally, it has some intuitive meaning, as can be seen by the fact that methods similar to the *p*-value arose repeatedly in various fields even before Fisher. And it had passionate advocates who put the tools into the hands of scientists in a way that was easy to use, like through Fisher and Yates's *Statistical Tables*. Finally, it was responsive to conditions of the time. These approaches addressed questions about variance and experimental design that were frequently raised at the time (Gigerenzer et al. 1989, pp. 73–74). Considering these virtues, Abelson (1997) suggests in a tongue-in-cheek piece that significance tests would be re-invented if they were banned and forgotten.

Table 1. Criteria for a lasting framework for inference beyond $p < 0.05$

Criterion	How $p < 0.05$ Meeting criterion	Does $p < 0.005$ meet criterion?
Provides quantitative basis to address uncertainty	p-value varies with precision of estimator	Same as $p < 0.05$
Rewards increased precision	Significance achieved more easily with higher sample size	Same as $p < 0.05$
Matches intuitive understanding	Statistic akin to p-value developed several times	Stricter threshold easily understood, but requires departure from current intuition
Advocated by statisticians and nonstatisticians	Promoted by Fisher, his students, and scientists in a variety of fields	Supported by some statisticians and practitioners, but value still disputed
Computationally feasible for nonstatisticians	Fisher and Yates made tables user-friendly and accessible to scientists and practitioners	Any threshold feasible with modern software
Responsive to changing conditions	$p < 0.05$ met needs of a time when few tests were conducted while varying thresholds allowed responses to multiple testing	Addresses current preponderance of tests, but viewed by advocates as a stopgap measure

A response that gains traction outside of academic statisticians and that is durable, I argue, must meet these same criteria, summarized in Table 1. And moreover, to remain valuable, it must be able to adjust to changing conditions. For example, as many authors, including Weisberg (2014, sec. 12.3), have discussed, our computational and data-gathering capabilities have changed enormously over the last several years, to say nothing of changes since 1925. We have seen how the lack of computing power at the time rendered Fisher's tables so valuable and thus so influential to practitioners. And the limited computer capabilities of the 1950s may have limited the ability of Bayesian methods to catch on with a wider audience (Weisberg 2014, sec. 8.4). The ease of computation is one cause of the multiplicity issues that are commonly discussed (Ioannidis 2005; Benjamini 2016). However, there is no reason to believe that computing capabilities have plateaued, and so an appropriate response would take into account not only today's conditions, but also those likely to occur in the future. Moreover, as we have seen, statistical methods are not always used with fidelity to the original intents and assumptions, especially decades after their initial formulation. Several of the responses to p-values, as Benjamini (2016) notes, would be susceptible to misuse as well.

Certainly, these are high demands to make of any statistical method, or indeed of any scientific methodology at all. And the sheer variety of alternatives proposed indicate that even the statistical community has not coalesced around one. To take one example, consider the proposal to lower the significance threshold to 0.005 (Benjamin et al. 2018). Table 1 summarizes whether and how $p < 0.005$ addresses the criteria for a lasting framework, not to argue for or against it, but to suggest the utility of this framework in assessing responses beyond $p < 0.05$. This proposal has several advantages: it maintains the ease of use and familiarity that scientists prize and can be viewed as in line with the approaches of Fisher (who often wrote of different thresholds in different settings) or of Neyman and Pearson (if it represents some true cost of a Type I Error and is paired with Type II Error control). It also addresses some of the multiplicity issues that have arisen from changing conditions and the reduced computational burden. This is not even the first time it has been proposed; as discussed, Edgeworth implicitly used this threshold at times, and threshold proposals varied greatly before and even after Fisher. However, it is, as Benjamin et al. (2018) acknowledge, just as arbitrary as current thresholds and just as susceptible to misinterpretation. And the benefits in addressing multiplicity may fade as datasets get bigger and tests are run even more frequently. Little (2016) also notes that lowering the threshold fails to address the longstanding debate between statistical significance and substantive significance. But differing thresholds have worked in other fields and this proposal may have a great deal of value in certain settings. And with tables of significance thresholds no longer necessary thanks to modern computing power, it is quite easy for researchers to use different thresholds at different times. This suggests that no one method and no one response to the controversy will be sufficient.

A multitude of responses, tailored to scientific purposes and fields of study, will be much more likely to be able to address all of these needs. Indeed, one can see this as an extension of arguments made at various points by both Fisher and Neyman-Pearson that different experimenters, working in different contexts, will use different thresholds of significance or set different α and β parameters. As Fisher's work focused on agriculture and biology, perhaps his advice still holds sway there, while other fields face different needs. Beyond just significance thresholds, different scientific questions can be approached with the variety of tools available, from Bayesian approaches to confidence intervals to machine learning, to suit their context. Such an approach, however, relies on a great deal of statistical sophistication among those who use statistical methods. Fortunately, this history can help improve statistics education and guide changes that would enhance that sophistication.

5.2. The Role of History in Statistics Education

The rise in popularity of statistics books aimed at general audiences, including some listed above, demonstrates the desires of many people to learn both the practical uses of the discipline and the way in which it came to be. Statistics educators broadly defined, whether course instructors, statistical collaborators, or writers of articles aimed at nonstatisticians, can benefit from this interest and use history as a teaching tool within this moment of debate in the discipline. The British mathematician John Fauvel (1991, pp. 4–5) presented a variety of reasons for incorporating history into mathematics education, including to "increase motivation for learning," "explain the role of mathematics in society," and "contextualise mathematical studies." A decade later, the Taiwanese educator Po-Hong Liu expounded these ideas. He noted specifically that "[h]istory reveals the humanistic facets of mathematical knowledge" and can challenge students' perceptions "that mathematics is fixed, rather than flexible, relative, and humanistic" (Liu 2003, p. 418).

These reasons all hold for statistics, especially as the discipline faces great change, not just in the use of conventional inferential methods but also with the rise of computing power and big data. As Goodman (2016, p. 1) notes: "that statisticians do not all accept at face value what most scientists are routinely taught as uncontroversial truisms will be a shock to many." To meet Millar's (2016) and Stangl's (2016) challenge of improving statistical education, teachers and collaborators should consider the introduction of this history into their discussions of significance testing. Presenting these controversies requires educators to present other approaches and thus also serves to, as Millar (2016) suggests, "make our students aware that p-values are not the 'only way.' "

The topics covered here can be introduced alongside the presentation of the tables of values of the normal, Student's t, and χ^2 distributions, which still hold a place as early as the Advanced Placement Statistics curriculum (AP 2010). Inviting students to consider how the lack of computers affected the development of statistics may further appreciation for these tables (or, more likely, further appreciation for the computer software that has rendered them obsolete). This in turn will help students appreciate what has changed since 1925 and how methods may need to change to reflect that.

Presenting the debate between Fisher, Gosset, Neyman-Pearson, and the Bayesians, and how that debate has evolved into the current discussion, highlights the human aspect of statisticians and the constantly changing, challenging nature of the field. As discussed above, many of the specific points made in that debate are ongoing points of contention today. In-depth analysis of Fisher's rationale for using the 0.05 standard can highlight how, though arbitrary, it is not without context, and how it responded to the needs of experimentalists at a certain point of history. This understanding will allow students and practitioners to form their own assessment of, for example, the proposal to lower the standard to 0.005. In this way, it becomes harder to dismiss the p-value without providing a substitute that is similarly usable by those who perform statistical analyses today. This teaching will also give students and practitioners the ability to critique the next statistical method that comes along, and to consider alternatives to the p-value in the context of statistical history and the role of statistics in modern science and society.

6. Conclusion

As we consider a world "beyond $p < 0.05$," I invite statisticians and scientists alike to consider the world before $p < 0.05$, a world where statistical analysis was less common and far more difficult an undertaking. It is then easier to see how p-values came to such prominence throughout science, despite the immediate disagreements among statisticians. Statistics is an evolving discipline, but it is in the difficult position of needing to evolve alongside the various disciplines that make use of its tools. In Fisher's teaching and manuscripts, writes Box (1978, p. 242), "he aimed to give workers a chance to familiarize themselves with tools of statistical craft as he had become familiar with them, and to evolve better ways of using them." This approach helped make statistics a fundamental tool in many disciplines, but has led to the challenges discussed in the ASA statement and elsewhere. Presenting this history as context for these discussions provides appropriate recognition of the rich debates that define statistics. It encourages statisticians to consider how their work will be used by practitioners and encourages practitioners to consider whether they are using statistical methodologies as they were intended. Through ongoing discussions and by encouraging this critical thinking, statistics can continue to be a field that helps push forward the boundaries of knowledge.

Acknowledgments

The author thanks the editors and referees for their helpful comments.

Funding

The author's studies are supported by Grant No. 5T32AI007358-28, U.S. National Institute of Allergy and Infectious Diseases, National Institutes of Health 5T32AI007358-28.

References

Abelson, R. P. (1997), "A Retrospective on the Significance Test Ban of 1999 (If There Were No Significance Tests, They Would Be Invented)," in *What If There Were No Significance Tests?*, eds. L. L. Harlow, S. A. Mulaik and J. H. Steiger, Multivariate Applications, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 117–141. [87]

AP (2010), "Statistics Course Description," available at *https://secure-media.collegeboard.org/ap-student/course/ap-statistics-2010-course-exam-description.pdf*. [89]

Arbuthnott, J. (1710), "An Argument for Divine Providence, Taken From the Constant Regularity Observ'd in the Births of Both Sexes," *Philosophical Transactions (1683–1775)*, 27, 186–190. [83]

Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66, 423–437. [86]

Benjamin, D. J. and Berger, J. O. (2016), "Comment: A Simple Alternative to p-values," *The American Statistician*, Online Discussion 70, 1–2. [87]

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., J. Wagenmakers, E., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. and Johnson, V. E. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [88]

Benjamini, Y. (2016), "It's Not the p-values' Fault," *The American Statistician*, Online Discussion, 70, 1–2. [87,88]

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-square Test," *Journal of American Statistical Association*, 33, 526–536. [86]

Berkson, J. (1942), "Tests of Significance Considered as Evidence," *Journal of American Statistical Association*, 37, 325–335. [86]

Berry, D. A. (2016), "p-values Are Not What They're Cracked Up to Be," *The American Statistician*, Online Discussion, 70, 1–2. [87]

Box, J. F. (1978), *R. A. Fisher, the Life of a Scientist*, Wiley Series in Probability and Mathematical Statistics, New York: Wiley. [84,85,89]

Chambers, C. (2017), *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*, Princeton, NJ: Princeton University Press. [87]

Cournot, A. A. (1843), *Exposition de la Théorie des Chances et des Probabilités*, Paris: L. Hachette. [83]

David, H. A. and Edwards, A. W. F. (2001), *Annotated Readings in the History of Statistics*, Springer Series in Statistics: Perspectives in Statistics. New York: Springer. [83,87]

De Finetti, B. (1937), "La Prévision: Ses Lois Logiques, Ses Sources Subjectives," *Annales de l'Institut Henri Poincaré*, 7, 1–68. [86]

Edgeworth, F. Y. (1885), "Methods of Statistics," *Journal of Statistical Society, London*, Jubilee Volume, 181–217. [83]

Elderton, W. P. (1902), "Tables for Testing the Goodness of Fit of Theory to Observation," *Biometrika*, 1, 155–163. [84]

Fauvel, J. (1991), "Using History in Mathematics Education," *For the Learning of Mathematics*, 11, 3–6. [88]

Fisher, R. A. (1922a), "The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients," *Journal of Royal Statistical Society*, 85, 597–612. [84,86]

——— (1922b), "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of *P*," *Journal of Royal Statistical Society*, 85, 87–94. [84]

——— (1922c), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*, Series A, 222, 309–368. [84,86]

——— (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd. [83,84,85]

——— (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd. [85,87]

——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd. [86]

Fisher, R. A. and Yates, F. (1938), *Statistical Tables for Biological, Agricultural and Medical Research*, Edinburgh: Oliver and Boyd. [85]

Gelman, A. (2016), "The Problems With *p*-values Are Not Just with *p*-values," *The American Statistician*, Online Discussion, 70, 1–2. [87]

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [82,83,86]

Gigerenzer, G., Swijtink, Z., and Daston, L. (1989), *The Empire of Chance: How Probability Changed Science and Everyday Life*, New York: Cambridge University Press. [84,86,87]

Goodman, S. N. (2016), "The Next Questions: Who, What, When, Where, and Why?" *The American Statistician*, Online Discussion, 70, 1–2. [87,89]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. and Altman, D. G. (2016), "Statistical Tests, *p*-values, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician*, Online Discussion, 70, 1–12. [84]

Hogben, L. T. (1957), *Statistical Theory: The Relationship of Probability, Credibility, and Error*, London: Allen & Unwin. [85]

Hubbard, R. (2004), "'Alphabet Soup: Blurring the Distinctions Between *p*'s and α's in Psychological Research," *Theory of Psychology*, 14, 295–327. [85,86,87]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: SAGE Publications. [83,85,87]

Ioannidis, J. P. A. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124. [88]

Jeffreys, H. (1939), *The Theory of Probability*, Oxford: Oxford University Press. [86]

Kadane, J. B., Fienberg, S. E. and DeGroot, M. H. (1986), *Statistics and the Law*, Wiley Series in Probability and Mathematical Statistics, New York: Wiley. [87]

Kennedy-Shaffer, L. (2017), "When the Alpha is the Omega: *p*-values, Substantial Evidence, and the 0.05 Standard at FDA," *Food & Drug Law Journal*, 72, 595–635. [87]

Laplace, P. S. (1827), *Traité de Mécanique Céleste, Supplément*, Paris: Duprat. [83]

Lehmann, E. L. (1993), "The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of American Statistical Association*, 88, 1242–1249. [85,86]

Lenhard, J. (2006), "Models and Statistical Inference: The Controversy Between Fisher and Neyman–Pearson," *British Journal for the Philosophy of Science*, 57, 69–91. [86]

Lew, M. J. (2016), "Three Inferential Questions, Two Types of *p*-value," *The American Statistician*, Online Discussion, 70, 1–2. [87]

Little, R. J. (2016), "Comment," *The American Statistician*, Online Discussion, 70, 1–2. [88]

Liu, P.-H. (2003), "Do Teachers Need to Incorporate the History of Mathematics in Their Teaching?" *The Mathematics Teacher*, 96, 416–421. [88]

MacKenzie, D. A. (1981), *Statistics in Britain, 1865–1930; The Social Construction of Scientific Knowledge*. Edinburgh: Edinburgh University Press. [87]

McGrayne, S. B. (2011), *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*, New Haven, CT: Yale University Press. [86]

Meehl, P. (1967), "Theory-testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, 34, 103–115. [86]

Millar, A. M. (2016), "ASA Statement on *p*-values: Some Implications for Education," *The American Statistician*, Online Discussion, 70, 1–2. [89]

Morrison, D. E., and Henkel, R. E. (1970), *The Significance Test Controversy: A Reader*, Methodological Perspectives. Chicago, IL: Aldine Publishing. [87]

Neyman, J. and Pearson, E. S. (1933), 'The Testing of Statistical Hypotheses in Relation to Probabilities a Priori," *Mathematical Proceedings of Cambridge Philosophical Society*, 29, 492–510. [85]

Pearson, E. S., Plackett, R. L. and Barnard, G. A. (1990), *'Student': A Statistical Biography of William Sealy Gosset*, Oxford: Clarendon Press. [85]

Pearson, K. (1900), "X. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it can be Reasonably Supposed to Have Arisen from Random Sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science.* 50, 157–175. [83,84]

Poisson, S.-D. (1837), *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile: Précédées des Règles Générales du Calcul des Probabilités*, Paris:Bachelier. [83]

Porter, T. M. (1986), *The Rise of Statistical Thinking, 1820–1900*, Princeton, NJ: Princeton University Press.

Rothman, K. J. (2016), "Disengaging From Statistical Significance," *The American Statistician*, Online Discussion, 70, 1. [87]

Rozeboom, W. W. (1960), "The Fallacy of the Null-Hypothesis Significance Test," *Psychological Bulletin.* 57, 416–428. [86]

Salsburg, D. (2001), *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, New York: W.H. Freeman and Co. [87]

Savage, L. J. (1954), *The Foundations of Statistics*, New York: Wiley. [86]

Skellam, J. G. (1969), "Models, Inference, and Strategy," *Biometrics* 25, 457–475. [86]

Skipper, J. K., Guenther, A. L., and Nass, G. (1967), "The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science," *American Sociology*, 2, 16–18. [86]

Stangl, D. (2016), "Comment," *The American Statistician*, Online Discussion, 70, 1. [89]

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: Belknap Press of Harvard University Press. [83,84,87]

Student (1908), "The Probable Error of a Mean," *Biometrika*, 6, 1–25. [84]

Tippett, L. H. C. (1931), *The Methods of Statistics: An Introduction Mainly for Workers in the Biological Sciences*, London: Williams and Norgate. [85]

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. [83,86,87]

Weisberg, H. I. (2014), *Willful Ignorance: The Mismeasure of Uncertainty*, Hoboken, NJ: Wiley. [86,87,88]

Yates, F. (1951), "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics," *Journal of the American Statistical Association*, 46, 19–34. [86]

Ziliak, S. T. (2016), "The Significance of the ASA Statement on Statistical Significance and *p*-values," *The American Statistician*, Online Discussion, 70, 1–2. [87]

Ziliak, S. T. and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [82,84,85,86,87]

The American Statistician

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

The Limited Role of Formal Statistical Inference in Scientific Inference

Raymond Hubbard, Brian D. Haig & Rahul A. Parsa

To cite this article: Raymond Hubbard, Brian D. Haig & Rahul A. Parsa (2019) The Limited Role of Formal Statistical Inference in Scientific Inference, The American Statistician, 73:sup1, 91-98, DOI: 10.1080/00031305.2018.1464947

To link to this article: https://doi.org/10.1080/00031305.2018.1464947

© 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: 8944

View related articles

View Crossmark data

Citing articles: 14 View citing articles

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, 91–98: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1464947

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The Limited Role of Formal Statistical Inference in Scientific Inference

Raymond Hubbard[a], Brian D. Haig[b], and Rahul A. Parsa[c]

[a]College of Business and Public Administration, Drake University, Des Moines, IA; [b]Department of Psychology, University of Canterbury, Christchurch, New Zealand; [c]Debbie and Jerry Ivy School of Business, Iowa State University, Ames, IA

ABSTRACT

Such is the grip of formal methods of statistical inference—that is, frequentist methods for generalizing from sample to population in enumerative studies—in the drawing of scientific inferences that the two are routinely deemed equivalent in the social, management, and biomedical sciences. This, despite the fact that legitimate employment of said methods is difficult to implement on practical grounds alone. But supposing the adoption of these procedures were simple does not get us far; crucially, methods of formal statistical inference are ill-suited to the analysis of much scientific data. Even findings from the claimed gold standard for examination by the latter, randomized controlled trials, can be problematic.

Scientific inference is a far broader concept than statistical inference. Its authority derives from the accumulation, over an extensive period of time, of both theoretical and empirical knowledge that has won the (provisional) acceptance of the scholarly community. 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 manifold problems surrounding the implementation and usefulness of formal methods of statistical inference in advancing science do not speak well of much teaching in methods/statistics classes. Serious reflection on statistics' role in producing viable knowledge is needed. Commendably, the American Statistical Association is committed to addressing this challenge, as further witnessed in this special online, open access issue of *The American Statistician*.

ARTICLE HISTORY
Received December 2017
Revised February 2018

KEYWORDS
Analytic studies;
Enumerative studies;
Observational studies;
Randomized controlled trials;
Significant difference;
Significant sameness;
Scientific inference;
Statistical inference

"Much of causal inference is beyond statistics" (Shadish and Cook 1999, p. 298).

"Statistical inference …is fundamentally incompatible with 'most' science" (Gunter and Tong 2016–2017, p. 1).

1. Introduction

Employment of formal methods of statistical inference, that is, frequentist methods for generalizing from sample to population in enumerative studies, is ubiquitous in empirical work in the social, management, and biomedical sciences. Indeed, statistical inference and scientific inference generally are taken to be one and the same in these areas. They are not. The present paper argues that "objective" methods of statistical inference play a limited role in scientific inference.

Their prevalence notwithstanding, strictly pragmatic considerations make employment of such statistical methods challenging. That is, the seemingly prosaic steps listed routinely in textbooks for the legitimate application of these tools in fact pose numerous and compounding obstacles hampering their use. More critically, formal statistical inference is geared to the conduct of enumerative studies (Deming 1975) whose object is the estimation of fixed population parameters from random samples. Yet such investigations are the exceptions, not the rule. Most scientific research involves analytic or predictive studies which do not lend themselves to analysis by formal inferential approaches (Deming 1975; Hahn and Meeker 1993). Even the procedure touted as ideal for showcasing the merits of formal statistical inference, the randomized controlled trial, has issues meeting the assumptions justifying its adoption, and can yield troublesome results (Cartwright 2007).

The much broader concept of scientific inference (versus statistical inference) is then introduced. In our telling, this revolves around the search for *significant sameness* (Ehrenberg 1975; Hubbard 2016; Hubbard and Lindsay 2013a, b; Nelder 1986), that is, the discovery of replicable and empirically generalizable findings. These discoveries, and their causes, are fueled in the main by observational (nonrandomized) studies.

Concluding thoughts reiterate the fact that, while occasionally important, overall the part played by formal statistical inference in scientific inference is relatively minor. This, in turn, raises questions concerning the teaching of this topic, and the texts informing it, which unwittingly exaggerate its contribution.

2. How Did We Get to this State of Affairs?

In many fields of inquiry, like the social sciences, management sciences, and areas of biomedicine, the notions of statistical inference and scientific inference are viewed as all but synonymous. This has led to distorted, naïve impressions about how

CONTACT Raymond Hubbard drabbuhyar@aol.com Thomas F. Sheehan Distinguished Professor of Marketing, Emeritus, College of Business and Public Administration, Drake University, Des Moines, IA 50311.

© 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.

science makes headway which have endured to the present day. Foremost among these is the widespread belief that scientific progress arises from the application of formal methods of statistical inference to random samples of data, something Guttman (1985, p. 3) characterizes as "illogical," and Gigerenzer (2004, p. 587) as "mindless." This, then, begs the question: How did matters get to be this way?

Irony abounds in answering this question, which draws heavily from Hubbard (2016, pp. 258–261). This is because the emergent social sciences, and the field of statistics, were concerned primarily with the discovery of significant sameness in their datasets, the very concept we are now trying to rehabilitate!

Early statisticians and social scientists alike subscribed to the French positivism of August Compte, which philosophy posited that all phenomena, both natural and social, are governed by universal laws. Compte maintained that the researcher's job was to detect the laws influencing societal behaviors.

During the 19th century social statisticians uncovered many laws, manifested as empirical regularities, in social data. Chief among them was the Belgian astronomer, Adolphe Quetelet, generally hailed as the father of both sociology (along with Compte) and statistics itself. Quetelet applied probability theory in detecting these empirical regularities. More specifically, he made use of the astronomer's law of error, better known today as the Gaussian or normal curve. Quetelet fit this distribution to biological data, including human birth and death rates, height, weight, and shoe size, and compared them by sex, age, place of residence, and profession. Others found evidence of this approximate normal curve regularity in social data, such as the incidence of crime, drunkenness, illness, insanity, marriage, suicide, and even in musical and poetic talent.

Quetelet demonstrated that although individual behaviors may be unpredictable, they are lawlike in the aggregate, conforming to a normal distribution. It was through this that he viewed the statistical regularity as the cardinal means of understanding social science. In addition, Quetelet regarded Poisson's "law of large numbers" to be the essential axiom of his "social physics." Succinctly put, statistical regularities were the means of discerning the universal laws of society.

Quetelet's belief in the universality of the error law had enormous repercussions on the development of 19th century science and statistics. As an example, the renowned physicist James Maxwell was intrigued by Quetelet's work. In particular, during the 1860s he wanted to know whether Quetelet's normal law might be adapted from its social setting involving large numbers of people to studying the behavior of gas molecules. Maxwell's success in this effort was a landmark in exposing the probabilistic nature of the physical world, previously thought to be deterministic. Likewise, brilliant scientists like Francis Edgeworth, Francis Galton, Gustav Fechner, and Karl Pearson were descendants of Quetelet, who all understood the crucial role played by what we call the pursuit of significant sameness.

Yet the notion of significant sameness lost currency, from which it has never recovered, when Ronald Fisher's ideas began dominating the statistics and social science disciplines in the 20th century. Whereas his predecessors favored the idea of large-scale descriptive studies, Fisher ushered in the age of the significant difference paradigm, with its focus on small sample theory, inferential methods, and statistical significance tests. Another question that arises, then, is this: How was the significant difference paradigm able to displace the emphasis on significant sameness?

In addressing this question, it must be acknowledged that the distinguished individuals mentioned above all had strong mathematical backgrounds. However, the rank and file of members of the Royal Statistical Society of London (of which Quetelet was a founding member in 1834) during the 19th, and well into the 20th, centuries, were not. Most were economists whose charge was the accumulation and sorting of facts generated by population censuses and other government reports. And, of course, the same applied to other fledgling social scientists.

But Fisher's methods seemed to offer something of a "cookbook" approach to statistical/scientific inference, rote and accessible to all. Under Fisher's authoritative sway, researchers came to believe that $p \leq 0.05$ results from null hypothesis significance tests (NHSTs) are the ultimate objective of a study. And it is difficult to argue with them given the force of his recommendations. For example, "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance" (Fisher 1966, p. 13). Or consider his endorsement of the p-value's epistemic role in certifying knowledge claims: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this [$p < 0.05$] level of significance" (Fisher 1926, p. 504). And his proclamation that: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher 1966, p. 16). [It is true that Fisher later changed his mind regarding the insistence on a fixed 0.05 level, yet the sacredness of this value remains cemented in the research community.] Lastly, he assured that: "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences" (Fisher 1970, p. 2). In short, Fisher promoted the use of formal statistical inference, particularly significance testing, as an integral component of scientific method. Small wonder that researchers in embryonic disciplines looking for scientific respectability followed suit.

It is easy to appreciate the reasoning behind Fisher's initial impact on data analysis. What is harder to comprehend is that despite withering criticism of its usefulness in the scientific enterprise, NHST (mostly in the form of the incompatible Fisher–Neyman-Pearson hybrid, e.g., Gigerenzer 1993; Goodman 1993; Hubbard and Bayarri 2003) continues to feature prominently in methods texts. Admissible employment of NHSTs requires that the data are a random sample from a clearly defined population, the latter belonging to the Neyman-Pearson component of the hybrid. [Fisher's views on the nature of populations, probability, and random sampling were confusing. See Hubbard (2016, pp. 182–183) for additional discussion.] In a nuts-and-bolts fashion, section 3 assesses the feasibility of meeting these two demands.

3. Practical Difficulties in Implementing Formal Statistical Inference

When presented in statistics and research methods texts, obtaining a random sample from a target population for purposes of statistical inference seems straightforward enough. The following steps from Iacobucci and Churchill's (2010, p. 283) highly

regarded textbook, now in its 11th edition, are typical of the genre: 1. define the target population, 2. identify the sampling frame, 3. select a random sampling procedure, 4. determine the sample size, 5. select the sample elements, and 6. collect the data from the designated elements. In fact, each of these steps can be fraught with difficulties which impede the use of formal methods of statistical inference in scientific work. We comment briefly on these six steps.

3.1. Define the Target Population

In an important but under-appreciated (only 82 Google Scholar citations as of September 5, 2017) article titled "Assumptions for Statistical Inference," Hahn and Meeker (1993, p. 4) acknowledge that this step sometimes is omitted. Indeed they say that textbooks commonly evade the issue by stating "Assume a random sample …from the population of interest" (Hahn and Meeker 1993, p. 1). Or as Gelman (2016) more recently conceded, statistics courses and textbooks (including his own) mostly treat the data as given.

Shaver and Norton's (1980a, b) examination of empirical papers in notable education journals led them to conclude that few gave an explicit definition or description of the target population. Likewise, Bottai (2014, p. 236) admitted that ill-defined populations are common in medical studies. Finally, Gigerenzer and Marewski (2015, p. 422) wrote that social scientists typically "do not draw random samples from a population or define a population in the first place."

3.2. Identify the Sampling Frame

The sampling frame is a list of the population members from which the sample will be drawn. But such lists may not exist. For example, in his book *Practical Sampling*, Henry (1990, p. 85) noted that getting lists of the general U.S. population is difficult, if not impossible. This helps to explain the relative popularity of *cluster* samples, where sampling frames are required only for the randomly selected clusters (subgroups) employed in the study.

But assessing the external validity (generalizability) of an investigation's results demands the sampling of settings, treatments, and observations as well as people. And while locating comprehensive lists for the latter is challenging enough, it is incredibly bothersome for the other three categories (Shadish, Cook, and Campbell 2002, p. 344).

Even if sampling frames are available they will rarely (ever?) exhibit a one-to-one correspondence with the target populations. Some are obsolete, dated, otherwise varying in quality, or, like the phone book, only rough approximations of a population. Since discrepancies between frames and populations may affect a study's conclusions, Hahn and Meeker (1993, p. 5) urge researchers to examine them. This is seldom done in practice.

3.3. Select a Random Sampling Procedure

In the social and other sciences, the generalizability of findings is conveyed almost exclusively in the language of formal methods of statistical inference. That is, matters of external validity center on the notion of *statistical* generalization (Hubbard and Lindsay 2013a) or the so-called *representative* model of generalization (Cook and Campbell 1979).

The centrality of representativeness was highlighted by survey researchers who wished to generalize from sample *to* population with known probability of error (Mook 1983). And before long, "the inference from a sample to population grew to be considered the most crucial part of research" (Gigerenzer and Marewski 2015, p. 425). It is reflected in methods texts which invariably state that samples must be representative of the relevant population. This has led to an insistence on random (representative) sampling, computing sample averages, and conducting tests of statistical significance.

We have already seen that selecting random samples from target populations is difficult to achieve. The importance of this limitation is reinforced in Section 4.

3.4. Determine the Sample Size

In applying formulae to calculate the sample size necessary to meet desired levels of precision and confidence regarding population parameters, textbook presentations on this topic (e.g., Iacobucci and Churchill 2010, pp. 313–317) sometimes invoke questionable assumptions. For example, it is posited that we know in advance the population variance, σ^2. Or if we don't know σ^2 (usually the case), we can estimate it using the sample variance, s^2, by doing either a pilot study, or possibly dividing the range of the data (if we know it) by 6, to get an estimate of the sample standard deviation, s.

Such practices are uncommon. Researchers rarely calculate required sample sizes (see, e.g., Bezeau and Graves 2001), employing instead whatever data are, or can be made, available. This is further borne out, indirectly, by the fact that much of the empirical social science literature is noticeably underpowered (Cohen 1988; and as summarized by Hubbard 2016, pp. 53–56), and that this state of affairs has been prevalent for some 60 years (Hubbard 2016, pp. 229–230).

3.5. Choose the Sample Elements

In order to draw a *simple* random sample we must be able to number the population elements from 1 to N (population size). But it is commonplace that this supposedly innocuous task is beyond our reach. The same stricture holds if we want to adopt *systematic* sampling.

Matters of selection get more complicated still if we wish to employ, say, a *stratified, cluster*, or *area* random sampling design. Here, in addition to being able to list the population members from 1 to N, we need additional information about each member in terms of the variable(s) used to assign them to a particular stratum, cluster, or area. Yet these kinds of real world details typically are ignored, or downplayed, in the textbooks.

3.6. Collect Data from Designated Elements

Again, an ostensibly simple-sounding task can be forbidding. The impact of unlisted numbers, not-at-homes, and refusals to cooperate, for example, contribute to the major problem of *nonresponse bias*. People are inundated with requests for

information and are increasingly unlikely to comply (Blair and Zinkhan 2006).

Practical considerations alone make application of formal statistical inference problematic. Yet even if this were not the case, the more salient issue is that this methodology offers little assistance in fostering scientific progress.

4. Statistical Inference is Ill-Suited for Promoting Scientific Inference

Substantiating the above claim necessitates a distinction between two different kinds of studies, enumerative and analytic. The representative, or statistical inference, model plays a decisive role in what Deming (1975, p. 147) labels *enumerative* studies. Here the goal is to estimate some fixed population parameter based on choosing a random sample from its constituents. As an example: "What is the Grade Point Average of high school juniors in Baltimore, MD?" Canvassing this class ranking supplies the "right" answer.

But Deming's (1975, p. 147) *analytic* studies are concerned with outcomes in the future, so the aim is to offer predictions about some process or cause system. As an example: "Will increasing the length of the school day by one hour raise SAT scores among U.S. high school seniors?" In such cases, representativeness often is impractical and possibly harmful to knowledge development.

To appreciate this, consider Hahn and Meeker's (1993) expansion of Deming's (1975) views. They underline that use of formal statistical inference mandates adherence to two basic requirements. First, the population must be well defined, finite, and unchanging. Second, a random sampling procedure is imperative in the selection of elements. In methods texts this sounds as uncomplicated as following the steps in a basic recipe. In practice, Hahn and Meeker (1993, p. 4) relate that for analytic (and even enumerative) studies, things are not nearly so simple because:

> We define an analytic study to be a study in which one is *not* dealing with a finite, identifiable, unchanging collection of units, and, thus is concerned with a process, rather than a population. … [I]n our experience, the great majority of applications encountered in practice, especially in industrial and in medical and other scientific applications, involve analytic, rather than enumerative studies. Moreover, it is inherently more complex to draw conclusions from analytic than from enumerative studies; analytic studies require the critical (and often unverifiable) added assumption that the process about which one wishes to make inferences is statistically identical to that from which the sample was selected.

In analytic studies random sampling seldom is an attainable option since "one is no longer dealing with 'an aggregate of identifiable units,' [hence] there is no relevant frame from which one can take a random sample" (Hahn and Meeker 1993, p. 5).

Other statisticians, such as Berk and Freedman (2003, p. 2), Chatfield (2002, p. 4), Gunter and Tong (2016–2017, p. 2), and Kass (2011, p. 8) share Hahn and Meeker's (1993, p. 10) position that authentic random samples are the exception rather than the rule. So do methodologists like Gigerenzer (2004, p. 599) and Grice et al. (2017, p. 236), who remark that experimental subjects are hardly ever sampled randomly from adequately defined populations. Or consider Shadish, Cook, and Campbell (2002, p. 91), who claimed that "we have not put much emphasis on random sampling for external validity, primarily because it is so rarely feasible in experiments." Despite this restriction, virtually all data in practice are analyzed as if random samples drawn from clearly specified populations are the norm.

In addition, there is the largely unrecognized admission that the formal statistical inference model is *misleading* in the context of analytic/predictive studies. This is because appraisals of external validity must involve characteristics included in the study along with others which are not (Hubbard and Lindsay 2013a; Shadish, Cook, and Campbell 2002). As Leviton (2001) noted, to be confident in making broad generalizations necessitates sampling from a "super-population" composed of every circumstance imaginable which may impact the result. It is obvious that no single study has this capability and so offers scant evidence of the external validity of an outcome (Leviton 2001, p. 5197).

And paradoxically, seeking representativeness can actually reduce the precision of results in regards to their application or generalizability to particular situations. That is, when making a prediction we must anticipate the different factors (subpopulations) that may affect a result and deal with them directly. This idea is developed further in the next section which addresses randomized controlled trials.

5. Randomized Controlled Trials (RCTs)

5.1. Advantages of RCTs

A major area where the superiority of formal methods of statistical inference are taken mostly for granted is the medical field's adoption of randomized controlled trials (RCTs). Popularized by Fisher (1935), use of randomization is often viewed as a prerequisite for drawing *causal* inferences in scientific experiments. As explained in most methods texts, the random assignment of subjects to the treatment (receiving, say, the new drug) and control (placebo) groups helps ensure their *equivalence* before the experiment commences. After some appropriate passage of time, differences in the mean incidence of illness in the two groups is subjected to a test of statistical significance. If a $p < 0.05$ result is found in favor of the treatment group, the new drug is regarded as "successful." And because the two groups were deemed equivalent before the trial began, improved results for the treatment group must have been caused by the new drug. Otherwise expressed, RCTs are thought to exhibit high levels of *internal validity*.

Indeed, RCTs are touted as the gold standard of experimental research (Jadad and Enkin 2008; Meldrum 2000). A principal reason for this acclaim is because they employ a *deductive* methodology—a causal inference may be implied if the assumptions of the method are met (Cartwright 2007, p. 11). Echoing Cartwright, the gold standard of evidence in the pursuit of evidence-based medicine is believed to come from the execution of large (phase III) RCTs (Williams 2010, p. 107).

But do RCTs constitute the gold standard for causal inference? Recent work disputes this moniker, including the articles by Cartwright (2007) and Williams (2010). In the first place, as discussed throughout this article, there is much information to indicate that meeting the assumptions required to employ formal methods of statistical inference is difficult. As one example

alone, RCTs rarely involve a random sample of patients targeted for the treatment (Williams 2010, p. 109). Second, even if these assumptions could be met, it does not eliminate the potentially serious drawbacks associated with the conduct of RCTs, as explored below.

5.2. Limitations of RCTs

The highest standard of evidence sought in evidence-based medicine is judged to come from large (phase III) RCTs. This is because, in the face of *heterogeneous* patient populations, large samples are necessary to obtain $p < 0.05$ results in comparisons of treatment and control group averages. Pharmaceutical companies view such findings as obligatory in order to win Food and Drug Administration (FDA) approval for the new medicine (Williams 2010, p. 109). The problem is that physicians do not treat the *average* patient, they treat real individuals (Demidenko 2016, p. 34; Kent and Hayward 2007, p. 68; Ruberg, Chen, and Wang 2010, p. 574). As such, averages may have only weak predictive validity for the latter (Williams 2010, p. 106).

Successful ($p < 0.05$) RCT outcomes suggest that patients react in like manner to the effects of the new drug. But this is not the case. It is just as, or more than, likely that reactions depend on the attributes of the patients themselves, implying that only some subset(s) of them may benefit (Williams 2010, p. 108). Here, Williams is reiterating Cartwright's (2007, p. 15) point that, when employing Suppes's (1970) probabilistic theory of causality, the only logical deduction we are entitled to make from a successful RCT is that the treatment causes the outcome in at least *some* members of that population. Disturbingly, Bottai (2014, p. 235), Cartwright (2007, p. 16), and Kent and Hayward (2007, p. 62) warned, this same treatment could have no beneficial outcomes for some subpopulations, and might even be harmful to still others. So much for the value of $p < 0.05$ results.

On the other hand, RCTs with $p > 0.05$ findings, while failing to pass muster with the FDA, can nevertheless be of great help to many patients (Kent and Hayward 2007, p. 62). For instance, the FDA initially denied approval for the drug temozolomide as a treatment for glioblastoma brain tumors because of overall $p > 0.05$ results. This meant delayed access to a new medicine which subsequently became the treatment of choice for patients with brain cancer (Williams 2010, p. 113).

As shown above, because of the focus on the average patient from a heterogeneous population, RCTs likely include patients who may benefit, not benefit, or possibly be harmed by the new drug. This suggests that trial investigators examine their data to see whether these subgroups exist (Kent and Hayward 2007, p. 62; Ruberg, Chen, and Wang 2010, p. 574; Williams 2010, p. 108). Yet this rarely happens. For example, in a review of 108 trials, Kent and Hayward (2007, p. 67) found that only one followed this approach.

The discussion in this section supports the belief that the high levels of internal validity said to be characteristic of large (phase III) RCTs is exaggerated. Moreover, claims regarding the external validity of such RCT results are difficult to justify (Cartwright 2007, p. 19), even though they are routinely overgeneralized (Williams 2010, p. 118). In short, RCT findings can be deficient in terms of *both* their internal and external validities. Or as Cartwright (2007, p. 18) summed it up: "The RCT, with its vaunted rigor, takes us only a very small part of the way we need to go for practical knowledge."

6. Scientific Inference

Unlike its statistical inference counterpart, the concept of scientific inference defies reduction to a series of allegedly neat and tidy methodological steps whose dutiful observance renders the output "science." Believing otherwise is wishful thinking. Scientific inference transcends *by far* the purview of statistical inference. This is why it is inherently more difficult to describe, let alone codify.

Scientific inferences in any given area are not made in a vacuum. Rather, they are offered in light of the totality of empirical and theoretical subject-matter knowledge that has accumulated over a long period of time. In progressive sciences this background knowledge, whose cogency rests on conditional acceptance by members of the field, plays a defining role in evaluating the plausibility of new findings. So scientific inferences are made within a dynamic context of what we believe we know, and hope to know, about our world and beyond. Importantly, these scientific conjectures and judgments do not often involve applications of formal statistical inference. They arise instead from researchers employing rules-of-thumb in their daily, and mostly exploratory (Gunter and Tong 2016–2017), work.

While no single approach to knowledge acquisition is likely to be optimal, a particularly fruitful candidate would be one centered on the search for *significant sameness* among phenomena. This idea is outlined in what follows.

6.1. Looking for Significant Sameness (Replicability)

The academy's historical, and ongoing, fixation with revealing "significant differences" in studies actively impedes cumulative science. In direct opposition, pursuit of "significant sameness" nurtures the latter.

Advocates of significant sameness tend to follow a *realist* philosophy of science (Haig 2014; Haig and Evers 2016; Hubbard 2016). Realism is based on how researchers actually behave, and is concerned with the lengthy process of theory *development*.

Under the auspices of realism, the significant sameness model postulates a facts-before-theory approach. In a nutshell, the argument is as follows: Data from individual studies rarely speak for themselves—they are ephemeral, affected by idiosyncratic boundary conditions, and usually irreproducible. As such, we are unlikely to build sound theories around individual data sets, especially those replete with $p < 0.05$ results. Better to construct theories *after* the discovery of what are variously described as repeatable facts, empirical regularities, or phenomena: their relative stubbornness deserves an explanation. Indeed, it should be emphasized that more Nobel prizes have been awarded for such discoveries than for advancing theories (Haig 2005, p. 384).

Because of widespread unfamiliarity with the issue, it is important to clarify further the distinction between data and phenomena/stubborn facts (Woodward 1989). Theories are mostly concerned with the explanation of phenomena, not data (Haig 2014, p. 34). But phenomena, typically, are not observable, and they are more durable than data. It is the role of data to provide evidence with respect to phenomena (Haig 2014, p. 35).

Accordingly, the progression is data → phenomena → theory, and it underpins the significant sameness model.

The quest for significant sameness endorses the pivotal role of replication research in cultivating healthy science. This is because successful replications are the primary means available for identifying phenomena. So we must ask: How do we gauge whether a replication is successful? We can do so by employing the criterion of overlapping confidence intervals (or Bayesian equivalents) around point values of interest, but always in a heuristic fashion and not as some cut-and-dried rule. In this endeavor, statistical analysis plays a vital role.

Additionally, replication research is crucial in placing bounds on the application of stubborn facts, that is, exploring their empirical generalizability. Moreover, empirical generalization focuses on projecting quantitative results *across* multiple, relatively homogeneous, (sub)populations. Of note, failure to generalize results can further enrich theory building by explaining why this limitation occurred. Counter to popular misconceptions, replication research in no way legislates "brute empiricism." Rather, its practice is responsible for eliciting facts, determining their empirical range, and assisting in the initial and subsequent refinement of causal explanations for them. This requires the conduct of many studies over a long period of time. Science in action is a process, not a product.

Scientific inference is better viewed as being grounded in *abductive* (explanatory) reasoning. Abduction—sometimes termed inference to the best explanation (Harman 1965)—takes as its locus the studying of facts and proposing a theory about the causal mechanisms generating them. Thus, abduction is a method of scientific inference geared toward the development of feasible and best explanations for the stubborn facts we possess. Like detective work, this approach mirrors the behavior of practicing scientists. And it is not beholden to methods of formal statistical inference.

Of course, the significant sameness model is not without its own limitations. For example, it might be objected that this model could be potentially expensive to apply in practice. There is merit to this argument as things stand today. If, however, the editorial-reviewer biases favoring the publication of "novel," $p < 0.05$, results, and against the publication of replication research—which biases institutionalize bad science—could be successfully confronted, then substantial resources would be freed to promote significant sameness.

As well, there are shortcomings associated with replication studies which could cause some editors, reviewers, and authors to question their usefulness (Hubbard 2016, pp. 156–157). For instance, let's say that Smith attempts to replicate Jones's initial work and gets different results. This situation could degenerate into the pair blaming one another for the conflicting outcomes. Jones could accuse Smith of failing to conduct an accurate replication—a charge understandable given that limited journal space often precludes a full disclosure of the methods used. Smith could respond that Jones made mistakes. And on and on. Yet while all of this is possible, there is no escaping the fact that "Replicability is almost universally accepted as the most important criterion of genuine scientific knowledge" (Rosenthal and Rosnow 1984, p. 9).

6.2. The Indispensable Role of Observational Studies

It is necessary at this juncture to revisit sampling issues. Since the 1960s and 1970s, when various user-friendly software packages became widely available, use of formal methods of statistical inference have monopolized empirical work in large areas of academe. At this same time, the impression is easily gained that nonrandomized (observational) studies came to be seen as lacking the exactitude of their randomized counterparts, and therefore constituted inferior research. This, despite the fact that virtually all scientific advances over the centuries have accrued—and continue to do so—without the "benefit" of randomization.

Scientific knowledge is gleaned chiefly from nonrandom samples. For instance, the process of discovering the scope and limits of empirical generalizations, mentioned earlier, relies on the use of purposive or judgment samples (Hubbard and Lindsay 2013a, p. 1384), where researcher choices are central (Rosenbaum 2001, p. 223). Yet there is no reason for concern. Nonrandom samples, allied with a replication strategy, can yield robust findings. Which is to say that, over time, via replication research, the point estimates and confidence intervals of additional (new) results about the phenomena at hand may be compared with the increasingly sturdy benchmarks established by their numerous predecessors to check for consistency. Consistency gained in this manner leads to deserved support for the veracity of the findings.

Beyond this, it should not be forgotten that entire disciplines such as archaeology, dendrochronology, and paleontology have been forged on the basis of *convenience* or *accidental*, never mind purposive, samples (Hubbard 2016, p. 180). And striking breakthroughs in other fields have arisen from the employment of such samples. For example, Freedman (1999) tells of how the use of convenience samples was key to establishing the causal links between contaminated water and the spread of cholera, and cigarette smoking and lung cancer. Contrary to the entrenched beliefs of many in the scientific community, "most of what we know about causation …is derived from observational studies" (Freedman 1999, p. 255).

In the context of RCTs, Williams (2010) concurs. A major criticism of the generalizability of results from a phase II trial (which does not feature randomization or a control group) is that, owing to possible selection bias, they may be due to characteristics in the patient sample, rather than the treatment. He adds that because phase III samples may not be representative of the universe of patients targeted for the treatment, coupled with the vagaries in individual patient attributes, the claimed advantages of phase III over phase II trials mostly evaporate. Along with Bottai (2014, p. 236), Williams (2010, p. 109) concludes that neither phase II nor phase III trials eliminate the worry about results not generalizing beyond the patients involved in that specific trial. Consequently, he favors the use of small trials, because when it comes to predicting the efficacy of a treatment on a particular patient, phase III RCTs are often "only marginally better than simply guessing" (Williams 2010, p. 111).

Consistent with the description of scientific inference sketched above, Shadish, Cook, and Campbell (2002, pp. 353–354) list five principles of generalized causal inferences that scientists employ every day in their work with mostly purposive samples. First is the idea of *surface similarity*, as when scientists

generalize by evaluating the likenesses among phenomena. Second is the *ruling out of irrelevancies*, meaning scientists generalize by deciding on those aspects of persons, settings, treatments, and results which are immaterial because they do not alter a generalization. Third is *making discriminations*, that is, finding limits to a generalization. Fourth is *interpolation* and *extrapolation*, where scientists generalize within the range of sampled persons, settings, treatments, and outcomes, together with those beyond them. Fifth is *causal explanation*, in which scientists generalize by developing and testing explanatory theories. It must be underscored that none of these five principles espouse the virtues of formal statistical inference. But this is not problematic. As Shadish, Cook, and Campbell (2002, p. 356) note, even though "these purposive sampling methods are not backed by a statistical logic that justifies formal generalizations …they are more practical than formal probability sampling …and are by far the most widely used sampling methods for generalized causal inference problems."

7. Conclusions

It is quite extraordinary that a model of inference that resists prescribed implementation, and anyway plays a quite limited part in furthering scientific progress, nevertheless has dominated the pages of journals in the social, management, and biomedical sciences for decades. Clearly, some serious re-thinking about statistics' role in scientific advance is needed. Fortunately, help looks to be on the way.

Recently the American Statistical Association (ASA) drew attention to problems associated with statistical significance and *p*-values (Wasserstein 2016; Wasserstein and Lazar 2016) in the context of science's "reproducibility crisis" (Peng 2015). This is a welcome and bold initiative offering as it does an unprecedented opportunity to publicly debate the value of statistical methods in the creation of robust scientific knowledge. It was followed by the ASA's Symposium on Statistical Inference in October 2017, which in turn led to this TAS Special Issue, "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$," an online, permanently open access issue of *The American Statistician*.

It seems certain that these initiatives will require major educational reforms in statistics classes and textbooks (Gelman 2016). Berry (2016, p. 1, p. 4) maintains that since the collective credibility of statisticians in the science community presently is jeopardized, such reforms are vital. They must reflect the circumscribed role of formal statistical inference in scientific inference.

References

Berk, R. A., and Freedman, D. (2003), "Statistical Assumptions as Empirical Commitments," in *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, (2nd ed.), eds. T. G. Blomberg and S. Cohen, Hawthorne, NY: Aldine de Gruyter, pp. 235–254. [94]

Berry, D. A. (2016), "*P*-values Are Not What They're Cracked Up to Be," online supplemental comment to Wasserstein (2016). [97]

Bezeau, S., and Graves, R. (2001), "Statistical Power and Effect Sizes of Clinical Neuropsychology Research," *Journal of Clinical and Experimental Neuropsychology*, 23, 399–406. [93]

Blair, E., and Zinkhan, G. M. (2006), "Nonresponse and Generalizability in Academic Research," *Journal of the Academy of Marketing Science*, 34, 4–7. [94]

Bottai, M. (2014), "Inferences and Conjectures About Average and Conditional Treatment Effects in Randomized Trials and Observational Studies," *Journal of Internal Medicine*, 276, 229–237. [93,95,96]

Cartwright, N. (2007), "Are RCTs the Gold Standard?" *BioSocieties*, 2, 11–20. [91,94,95]

Chatfield, C. (2002), "Confessions of a Pragmatic Statistician," *The Statistician*, 51, 1–20. [95]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum. [93]

Cook, T. D., and Campbell, D. T. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston, MA: Houghton Mifflin. [93]

Demidenko, E. (2016), "The *p*-Value You Can't Buy," *The American Statistician*, 70, 33–38. [95]

Deming, W. E. (1975), "On Probability As a Basis For Action," *The American Statistician*, 29, 146–152. [91,94]

Ehrenberg, A. S. C. (1975), *Data Reduction* (Rev. reprint), London, UK: Charles Griffin. [91]

Fisher, R. A. (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. [92]

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd. [94]

——— (1966), *The Design of Experiments* (8th ed.), Edinburgh: Oliver and Boyd. [92]

——— (1970), *Statistical Methods for Research Workers* (14th ed.), New York: Hafner. [92]

Freedman, D. (1999), "From Association to Causation: Some Remarks on the History of Statistics," *Statistical Science*, 14, 243–258. [96]

Gelman, A. (2016), "The Problems with *P*-Values are Not Just With *P*-Values," online supplemental comment to Wasserstein (2016). [93,97]

Gigerenzer, G. (1993), "The Superego, the Ego, and the Id in Statistical Reasoning," in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, eds. G. Keren and C. A. Lewis, Hillsdale, NJ: Lawrence Erlbaum, pp. 311–339. [92]

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [92,94]

Gigerenzer, G., and Marewski, J. N. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440. [93]

Goodman, S. N. (1993), "*p* Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496. [92]

Grice, J., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E., and Vest, A. (2017), "Four Bad Habits of Modern Psychologists," *Behavioral Sciences*, 7, 1–21. [94]

Gunter, B., and Tong, C. (2016–2017), "A Response to the ASA Statement on Statistical Significance and *P*-values," *NV–ASA Newsletter*, 14, 1–3. [91,94,95]

Guttman, L. (1985), "The Illogic of Statistical Inference for Cumulative Science," *Applied Stochastic Models and Data Analysis*, 1, 3–10. [92]

Hahn, G. J., and Meeker, W. Q. (1993), "Assumptions for Statistical Inference," *The American Statistician*, 47, 1–11. [91,93,94]

Haig, B. D. (2005), "An Abductive Theory of Scientific Method," *Psychological Methods*, 10, 371–388. [95]

——— (2014), *Investigating the Psychological World: Scientific Method in the Behavioral Sciences*, Cambridge, MA: MIT Press. [95]

Haig, B. D., and Evers, C. W. (2016), *Realist Inquiry in Social Science*, Thousand Oaks, CA: Sage. [95]

Harman, G. H. (1965), "Inference to the Best Explanation," *Philosophical Review*, 74, 88–95. [96]

Henry, G. T. (1990), *Practical Sampling*, Newbury Park, CA: Sage. [93]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: Sage. [91,92,93,95,96]

Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence (*p*'s) Versus Errors (α's) in Classical Statistical Testing," (with comments), *The American Statistician*, 57, 171–182. [92]

Hubbard, R., and Lindsay, R. M. (2013a), "From Significant Difference to Significant Sameness: Proposing a Paradigm Shift in Business Research," *Journal of Business Research*, 66, 1377–1388. [91,93,94,96]

Hubbard, R., and Lindsay, R. M. (2013b), "The Significant Difference Paradigm Promotes Bad Science," *Journal of Business Research*, 66, 1393–1397. [91]

Iacobucci, D., and Churchill, G. A., Jr. (2010), *Marketing Research: Methodological Foundations* (10th ed.), Mason, OH: South-Western. [92,93]

Jadad, A. R., and Enkin, M. W. (2008), *Randomized Controlled Trials: Questions, Answers and Musings* (2nd ed.), Malden, MA: Blackwell. [94]

Kass, R. E. (2011), "Statistical Inference: The Big Picture," *Statistical Science*, 26, 1–9. [94]

Kent, D., and Hayward, R. (2007), "When Averages Hide Individual Differences in Clinical Trials," *American Scientist*, 95, 60–68. [95]

Leviton, L. C. (2001), "External Validity," in *International Encyclopedia of the Social and Behaviour Sciences*, eds. N. J. Smelser and P. B. Baltes, Oxford, UK: Elsevier, pp. 5195–5200. [94]

Meldrum, M. L. (2000), "A Brief History of the Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard," *Hematology/Oncology Clinics of North America*, 14, 745–760. [94]

Mook, D. G. (1983), "In Defense of External Invalidity," *American Psychologist*, 38, 379–387. [93]

Nelder, J. A. (1986), "Statistics, Science and Technology," (with comments), *Journal of the Royal Statistical Society*, Series A, 149, 109–121. [91]

Peng, R. (2015), "The Reproducibility Crisis in Science: A Statistical Counterattack," *Significance*, 12, 30–32. [97]

Rosenbaum, P. R. (2001), "Replicating Effects and Biases," *The American Statistician*, 55, 223–227. [96]

Rosenthal, R., and Rosnow, R. L. (1984), *Essentials of Behavioral Research: Methods and Data Analysis*, New York: McGraw–Hill. [96]

Ruberg, S. J., Chen, L., and Wang, Y. (2010), "The Mean Does Not Mean as Much Anymore: Finding Sub-Groups for Tailored Therapeutics," *Clinical Trials*, 7, 574–583. [95]

Shadish, W. R., and Cook, T. D. (1999), "Comment—Design Rules: More Steps Toward a Complete Theory of Quasi-Experimentation," *Statistical Science*, 14, 294–300. [91]

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, MA: Houghton Mifflin. [93,94,96]

Shaver, J. P., and Norton, R. S. (1980a), "Populations, Samples, Randomness, and Replication in Two Social Studies Journals," *Theory and Research in Social Education*, 8, 1–10. [93]

Shaver, J. P., and Norton, R. S. (1980b), "Randomness and Replication in Ten Years of the *American Educational Research Journal*," *Educational Researcher*, 9, 9–15. [93]

Suppes, P. (1970), *Probabilistic Theory of Causality*, Atlantic Highlands, NJ: Humanities Press. [95]

Wasserstein, R. L. (2016), "ASA Statement on Statistical Significance and *P*-Values," *The American Statistician*, 70, 131–133. [97]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–131. [97]

Williams, B. A. (2010), "Perils of Evidence-Based Medicine," *Perspectives in Biology and Medicine*, 53, 106–120. [94,95,96]

Woodward, J. (1989), "Data and Phenomena," *Synthese*, 79, 393–472. [95]

The American Statistician

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

Large-Scale Replication Projects in Contemporary Psychological Research

Blakeley B. McShane, Jennifer L. Tackett, Ulf Böckenholt & Andrew Gelman

To cite this article: Blakeley B. McShane, Jennifer L. Tackett, Ulf Böckenholt & Andrew Gelman (2019) Large-Scale Replication Projects in Contemporary Psychological Research, The American Statistician, 73:sup1, 99-105, DOI: 10.1080/00031305.2018.1505655

To link to this article: https://doi.org/10.1080/00031305.2018.1505655

© 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: 6969

View related articles

View Crossmark data

Citing articles: 24 View citing articles

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, 99–105: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1505655

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Large-Scale Replication Projects in Contemporary Psychological Research

Blakeley B. McShane[a], Jennifer L. Tackett[b], Ulf Böckenholt[a], and Andrew Gelman[c]

[a]Department of Marketing, Kellogg School of Management, Northwestern University, Evanston, IL; [b]Department of Psychology, Northwestern University, Evanston, IL; [c]Department of Statistics and Department of Political Science, Columbia University, New York, NY

ABSTRACT

Replication is complicated in psychological research because studies of a given psychological phenomenon can never be direct or exact replications of one another, and thus effect sizes vary from one study of the phenomenon to the next—an issue of clear importance for replication. Current large-scale replication projects represent an important step forward for assessing replicability, but provide only limited information because they have thus far been designed in a manner such that heterogeneity either cannot be assessed or is intended to be eliminated. Consequently, the nontrivial degree of heterogeneity found in these projects represents a lower bound on the true degree of heterogeneity. We recommend enriching large-scale replication projects going forward by embracing heterogeneity. We argue this is the key for assessing replicability: if effect sizes are sufficiently heterogeneous—even if the sign of the effect is consistent—the phenomenon in question does not seem particularly replicable and the theory underlying it seems poorly constructed and in need of enrichment. Uncovering why and revising theory in light of it will lead to improved theory that explains heterogeneity and increases replicability. Given this, large-scale replication projects can play an important role not only in assessing replicability but also in advancing theory.

ARTICLE HISTORY
Received October 2017
Revised June 2018

KEYWORDS
Between-study variation; Heterogeneity; Hierarchical; Meta-Analysis; Multilevel; Null hypothesis significance testing; *p*-value; Psychology; Replication

1. Introduction

The validity of research in the biomedical and social sciences is under intense scrutiny at present, with published findings failing to replicate at an alarming rate. This problem appears particularly acute in psychology, where the failure to replicate several prominent findings (for example, Bargh, Chen, and Burrows (1996); Bargh et al. (2001); Carney, Cuddy, and Yap (2010); Bem (2011)) has attracted the attention of both academics and the popular press.

One positive development that has emerged from this unfortunate situation is an increased interest in planning and conducting replications, as evidenced by *inter alia* large-scale replication projects such as the Many Labs project (Klein et al. 2014), the Open Science Collaboration (OSC) project (Open Science Collaboration 2015), and Registered Replication Reports (RRRs; Simons, Holcombe, and Spellman 2014) in which one or more psychological phenomena is investigated across multiple laboratories. The projects involve heroic coordination efforts and require tremendous resources but offer the promise of allowing for the assessment of the replicability of individual phenomena and, perhaps taken together, psychological research as a whole.

Nonetheless, replication is complicated in psychological research because studies of a given psychological phenomenon can never be direct or exact replications of one another (Rosenthal 1991; Tsang and Kwan 1999; Brandt et al. 2014; Stroebe and Strack 2014; Fabrigar and Wegener 2016). Instead, studies differ at minimum in their method factors, that is, anything—known or unknown—pertaining to the implementation of the study that is not directly related to the theory under investigation. Method factors can include seemingly major factors such as the operationalization of the dependent measure(s), the operationalization of the experimental manipulation(s), and unaccounted for moderators but also seemingly minor factors such as the social context, the subject pool, and the time of day (for a comprehensive list, see Brown et al. 2014). Differences in method factors result in heterogeneity, that is, effect sizes that vary from one study of a given phenomenon to the next—an issue of clear importance for replication.

While current large-scale replication projects represent an important step forward for assessing replicability, they provide only limited information. Specifically, and consistent with the fact that heterogeneity has heretofore been underappreciated in psychology, these projects have thus far been designed in a manner such that heterogeneity either cannot be assessed or is intended to be eliminated. Consequently, the nontrivial degree of heterogeneity found in these projects represents a lower bound on the true degree of heterogeneity. Given this, we recommend enriching large-scale replication projects going forward by embracing heterogeneity through the systematic variation of method factors. We argue this is the key for assessing replicability: if effect sizes are sufficiently heterogeneous from one study to the next—even if, for example, the sign of the effect is consistent—the phenomenon in question does not seem particularly replicable and the theory underlying it seems poorly constructed and in need of enrichment. In particular, what are believed to be method factors either interact with theoretical moderators or are, in fact, unaccounted for theoretical moder-

CONTACT Blakeley B. McShane b-mcshane@kellogg.northwestern.edu Marketing Department, Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208.
We thank the Defense Advanced Research Projects Administration (agreement number D17AC00001) for partial support of Andrew Gelman's work.
© 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.

ators. Uncovering this and revising theory in light of it will lead to improved theory that explains heterogeneity and increases replicability. Given this, large-scale replication projects can play an important role not only in assessing replicability but also in advancing theory—a fact that has been overlooked to date.

In the remainder of this article, we briefly review current large-scale replication projects. We then discuss assessments of heterogeneity based on them, offer recommendations for enriching them, and discuss challenges with large-scale replication in domains of psychological research where they have yet to take place. We conclude with a brief discussion.

2. Large-Scale Replication Projects

2.1. Many Phenomena, One Study: The Open Science Collaboration Project

The OSC project is a large-scale replication project consisting of replications of 100 social and cognitive psychology phenomena published in three journals (*Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*) and conducted by 270 authors. Fundamental to the approach was that each phenomenon was replicated only once thus allowing for broad coverage across a host of phenomena.

The OSC authors examined five indicators of replicability, and results were, broadly speaking, consistent across them:

> Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects.

In sum, the results seem grim.

The major strength of the OSC project is its broad coverage of studies across social and cognitive psychology. However, this comes with a major limitation: because each phenomenon was replicated only once, the OSC project replication setting is the canonical and classical one of a single "original" study and a single "replication" study of each phenomenon.

Because there is only a single replication study of each phenomenon, heterogeneity cannot be assessed and accounted for. While this is a limitation in and of itself, it also means many of the indicators of replicability discussed above necessarily ignore heterogeneity and are thus miscalibrated.

For example, consider the second—that only 36% of replication studies had statistically significant results whereas 97% of the original studies did—which has gained traction in the popular press and reflects the classic definition of replication employed in practice (i.e., a subsequent study is considered to successfully replicate a prior study if either both fail to attain statistical significance or both attain statistical significance and are directionally consistent). When heterogeneity exists but is ignored, the Type I error of the single study significance tests on which these 36% and 97% figures are based is inflated above the nominal size α and standard sample size and power formulas overstate power and thus understate Type II error (McShane and Böckenholt 2017).

2.2. One Phenomenon, Many Studies: The Many Labs Project and Registered Replication Reports

Overcoming the major limitation of the OSC project are the Many Labs project and RRRs which feature multiple replication studies of one phenomenon (RRRs) or a small number of phenomena (Many Labs) across multiple independent laboratories thus allowing for deeper examination of these phenomena. In particular, the Many Labs project investigates 16 classic and contemporary psychological research phenomena across 36 independent samples totaling 6344 subjects; each of the 36 laboratories involved in the project used identical materials and administered them through a web browser to ensure procedural consistency across laboratories. Similarly, RRRs consist of "a set of studies from a variety of laboratories that all followed an identical, vetted protocol designed to reproduce the original method and finding as closely as possible" (Simons, Holcombe, and Spellman 2014) and have thus far investigated phenomena that include the facial feedback hypothesis, ego depletion, the effect of time pressure on cooperative decisions, and the link between commitment to and betrayal of a romantic relationship (Wagenmakers et al. 2016; Hagger et al. 2016; Bouwmeester et al. 2017; Cheung et al. 2016). Thus, the Many Labs and RRR approach allow data to be integrated via meta-analysis—with the data from each laboratory treated as an independent replication study—to provide more definitive and informative inferences and conclusions.

The primary indicator of replicability used by the Many Labs and RRR authors is whether a meta-analysis that pools the data across the multiple replication studies of a given phenomenon matches the statistical significance and direction of the original study. Empirically, evaluations of replicability diverge between the Many Labs and RRR authors, with the Many Labs authors replicating 14 of the 16 phenomena they examined and RRR authors generally failing to replicate phenomena thus far. Rather than reflecting any differences in, for example, the replication approach taken, we believe this divergence more likely reflects the choice of the phenomena examined by the Many Labs and RRR authors.

The major strength of the Many Labs and RRR approach is that pooling data across multiple replication studies allows for more powerful evaluations of replicability—ones that allow heterogeneity to, at least to some degree, be assessed and accounted for. However, this comes with a major limitation: due to the vast resources required to conduct replications across multiple laboratories, the approach is necessarily more focused on one (or a small number of) phenomena and thus does not provide broad coverage across psychology or domains of it. In sum, the strengths and limitations of the Many Labs and RRR approach are opposite those of the OSC project. Yet, like the OSC project, this approach has thus far defined replication in terms that focus on the sign and statistical significance of results.

3. Assessing Heterogeneity in Large-Scale Replication Projects

In this section, we discuss assessments of heterogeneity—that is, variation in effect sizes from one study of a given phenomenon

to the next resulting from differences in method factors—based on large-scale replication projects, in particular those following the one phenomenon, many studies approach (i.e., because those following the many phenomena, one study approach do not allow for such assessments). While by no means the only or necessarily the best measure, we here discuss heterogeneity in terms of a relative measure, namely the I^2 statistic (i.e., the proportion of the variation in observed effect sizes due to heterogeneity rather than sampling error); we do so because it is easily interpretable and comparable across phenomena. However, as we discuss in the next section, the absolute level of heterogeneity is also relevant for assessing replicability. To place I^2 in context, Pigott (2012) defines low, medium, and high heterogeneity in psychological research as I^2 of 25%, 50%, and 75% respectively (see also Higgins and Thompson 2002).

Heterogeneity has generally been regarded as important across sets of studies that consist of general (i.e., systematic or conceptual) replications, and a host of recent data convincingly demonstrate that it is indeed rife and large across them. Consider, for example, the sets of studies included in the comprehensive meta-analyses of the sort published in *Psychological Bulletin*, the premier outlet for meta-analyses in psychology. van Erp et al. (2017) examined heterogeneity estimates from 705 meta-analyses published there between 1990 and 2013 and found a median I^2 of 71%. Similarly, Stanley, Carter, and Doucouliagos (2017) surveyed 200 recent meta-analyses published there and found a median I^2 of 74%.

While the finding that heterogeneity is so large in these meta-analyses may perhaps not be so surprising given that the studies included in them tend to feature tremendous variation in their method factors, more surprising—as well as more interesting and more important—is that heterogeneity can also be substantial even across sets of studies that consist entirely of close replications (i.e., studies that use identical or very similar materials). In particular, heterogeneity persists, and to a reasonable degree, even in large-scale replication projects such as the Many Labs project and RRRs where rigid, vetted protocols with identical study materials are followed across laboratories in a deliberate attempt to eliminate differences in method factors and thus heterogeneity.

Specifically, random effects meta-analyses conducted by the Many Labs authors yielded nonzero estimates of heterogeneity for all 14 of the phenomena they found to be nonnull and the average I^2 across these phenomena was 40% (see Table 3 of Klein et al. 2014). In addition, among the 6344 Many Labs subjects were 1000 recruited via Amazon's Mechanical Turk. The study materials were administered to these 1000 subjects over seven unique days, beginning on August 29, 2013 and ending on September 11, 2013 (i.e., seven consecutive days excluding Fridays, weekends, and the Labor Day holiday). Restricting attention to only these subjects and treating each unique day as a separate study/laboratory yields seven extremely close replications of each phenomenon. Again, however, despite the extreme degree of closeness, heterogeneity is nontrivial: random effects meta-analyses yield nonzero estimates of heterogeneity for nine of the 14 nonnull phenomena, and across these, the average I^2 was 21% (McShane, Böckenholt, and Hansen 2016). Similarly, a moderate amount of heterogeneity has been found in RRRs (see, e.g., Eerland et al. 2016 and Hagger et al. 2016). In sum, despite protocols explicitly designed to eliminate differences in method factors and thus heterogeneity, heterogeneity was nonetheless nontrivial in large-scale replication projects following the one phenomenon, many studies approach.

This is astounding not only from a substantive perspective but also from a purely statistical perspective. In particular, when heterogeneity is in fact nonzero but there are a relatively small number of studies/laboratories (as is unfortunately the case in the Many Labs project and RRRs), standard estimators of heterogeneity are biased downward and estimates of zero heterogeneity result implausibly often (see, e.g., Chung et al. 2013b, and Chung, Rabe-Hesketh, and Choi 2013a). For this and related reasons, the Type I error of the significance test of zero heterogeneity is often inflated above the nominal size α and power is low (Huedo-Medina et al. 2006; Ioannidis, Patsopoulos, and Evangelou 2007).

Based on this poorly calibrated significance test, many have wrongly concluded that large-scale replication projects do not provide evidence for heterogeneity. For instance, the Many Labs authors write "tests of heterogeneity suggested that most of the variation in effects is attributable to measurement error" (by which we believe they mean sampling error as their analyses do not account for measurement error). In contrast, we note that, even if one believes that homogeneity is plausible and that significance testing is a reasonable thing to do, one should be very skeptical of this particular significance test unless the number of studies is indeed very large (i.e., to allow for sufficient power). Further, laying aside any skepticism of significance testing, we note that the privileged role given to the null hypothesis of homogeneity by this particular significance test does not seem reasonable in psychological research because studies of a given psychological phenomenon can never be direct or exact replications of one another; instead, rather than assuming homogeneity and only rejecting it given sufficient evidence to the contrary, it seems much more reasonable to assume some degree of heterogeneity and only conclude in favor of homogeneity given sufficient evidence for it.

Given that large-scale replication projects have typically shared their data even at the most granular level, there is great opportunity to move beyond poorly-performing estimators and significance tests to better assess heterogeneity. For example, consider again the Many Labs project. Current assessments of heterogeneity are based on the standard two-parameter (one intercept, one variance component) so-called random effects meta-analytic model fit to a single summary statistic computed from the individual-level data from each of the 36 laboratories and fit to each of the 16 phenomena examined by the project entirely separately. Instead, it would be far superior to base assessments of heterogeneity on a hierarchical (or multilevel) model fit to the totality of the individual-level observations across all laboratories and phenomena jointly. The partial pooling, particularly of the variance components, possible with such a model would yield much more reliable and precise estimates.

In sum, in large-scale replication projects such as the Many Labs project and RRRs, we should—for substantive reasons (i.e., protocols designed to eliminate heterogeneity) and statistical reasons (i.e., estimators and significance tests that perform poorly in a manner that falsely suggests homogeneity)—expect to observe little to no heterogeneity. The very fact we observe a

nontrivial degree of it is compelling evidence that heterogeneity is not only the norm but also cannot be avoided in psychological research—even if every effort is taken to eliminate it.

4. Enriching Large-Scale Replication Projects

In this section, we make four recommendations pertaining to the design of large-scale replication projects, the statistical analysis of the data resulting from them, and assessments of the replicability of the phenomena investigated by them. We then provide an illustration based on the anchoring effect (Jacowitz and Kahneman 1995) of how these recommendations can improve theory in a manner that explains heterogeneity and increases replicability.

Because heterogeneity cannot be avoided in psychological research and because assessing and accounting for heterogeneity is not possible when analyzing single studies in isolation but only when analyzing multiple studies jointly, our *first recommendation* is that future large-scale replication projects follow the one phenomenon, many studies approach of the Many Labs project and RRRs rather than the many phenomena, one study approach of the OSC project. This offers substantive as well as statistical benefits: beyond allowing for the "mere" assessment of heterogeneity, it can suggest that method factors either interact with theoretical moderators or are, in fact, unaccounted for theoretical moderators (as discussed in greater depth below); it also provides better calibration of Type I and Type II error and yields more efficient estimates of overall average effects (McShane and Böckenholt 2014, 2017).

Before proceeding to our second recommendation, it is important to understand that the degree of heterogeneity across set of studies depends critically on the variation in the method factors across the set. For example, it is unsurprising that studies from the Many Labs project and RRRs are less heterogeneous than studies included in meta-analyses published in *Psychological Bulletin* as the former used identical materials while the latter did not. Similarly, it is unsurprising that the Mechanical Turk subsample of the Many Labs sample is less heterogeneous than the full sample as, for example, differences in the subject population across days on Mechanical Turk are likely to be smaller than those across the 36 Many Labs. Put differently, the Mechanical Turk subsample features less variation in method factors than the full Many Labs sample which, in turn, features less variation in method factors than the published meta-analyses. In sum, the less (more) varied a set of studies are in their method factors, the smaller (larger) heterogeneity will be.

Given this, the Many Labs and RRR approach—due to the rigidness of the protocol designed to eliminate differences in method factors and thus heterogeneity—provides at best limited information about—indeed, a lower bound on—the true degree of heterogeneity. Specifically, because all laboratories use identical materials, any information on heterogeneity provided by this approach relates almost exclusively to differences in the subject populations across the laboratories.

Instead, systematically varying various method factors would provide much more information about the phenomenon in question. For example, when a set of studies features great variation in their method factors and an assessment of heterogeneity is relatively low, this suggests that the phenomenon in question is highly stable with regards to variation in these factors; in other words, these method factors are unlikely to interact with theoretical moderators and the underlying theory may be relatively tight (i.e., in terms of accounting for all relevant theoretical moderators). A large assessment suggests just the opposite, namely that the phenomenon in question is not stable with regards to variation in these factors and the underlying theory may be in need of improvement. On the other hand, when a set of studies features little variation in their method factors and an assessment of heterogeneity is relatively low, this is not particularly informative or probative (although in this setting, a large assessment suggests even more greatly that the phenomenon in question is not stable with regards to variation in these factors and the underlying theory may be in need of improvement).

Consequently, our *second recommendation* is that future large-scale replication projects systematically vary method factors across the laboratories involved in the project (see also Ehrenberg and Bound 1993; Lindsay and Ehrenberg 1993; Baribault et al. 2018); ideally, in doing so, they would also take an explicit stance about which ones are likely to be major and which are likely to be minor.

Our *third recommendation* is that researchers analyze the data resulting from these projects using a hierarchical model fit to the individual-level observations, and specifically that all theoretical moderators should be modeled via covariates while all other potential moderators—that is, method factors—should induce variation (i.e., heterogeneity). In classical terms, this would amount to treating theoretical moderators as fixed and method factors as random (though we would not limit ourselves or others to classical models).

Such a modeling approach requires researchers to take an explicit stance on which moderators pertain to theory and which are heterogeneity-inducing method factors (for examples, see Chernev, Böckenholt, and Goodman 2015 and McShane and Böckenholt 2018). While one might argue this is difficult or perhaps even unrealistic, we believe it is necessary and beneficial. Carefully constructed theory involves the delineation of relevant constructs, interactions, boundary conditions, and related matters. The modeling approach we suggest necessitates this thus resulting in clearer and likely improved theory. In addition, the assessments of heterogeneity resulting from this approach have the potential to enrich theory.

Finally, our *fourth recommendation* is that assessments of replicability should not depend solely on estimates of effects, or worse, significance tests based on them. While we do not claim to be able to offer a universal definition of replicability (indeed, we doubt that one single definition would do across domains of research), we believe heterogeneity is an important consideration in assessing replicability. For example, if an effect is estimated to be large but highly heterogeneous, it may be quite replicable according to the classic definition based on sign and statistical significance. However, in our view, sufficiently high heterogeneity does not necessarily indicate replicability of a phenomenon—even if one can be quite sure of its sign. Instead, while recognizing that what constitutes "sufficiently high" heterogeneity will vary across domains of research and is a subject matter rather than statistical issue, large variation in effect sizes

to us indicates an underlying theory that is poorly constructed and in need of enrichment. In particular, what are believed to be method factors either interact with theoretical moderators or are, in fact, unaccounted for theoretical moderators. Uncovering this and revising theory in light of it will lead to improved theory that explains heterogeneity and increases replicability.

As an example, consider the anchoring effect, the tendency for individuals to rely too heavily on an initial piece of information offered (known as the "anchor") when making decisions. While there is an extremely large literature on this phenomenon (see, e.g., Chapters 6-8 of Gilovich, Griffin, and Kahneman 2002), one popular paradigm for it, used *inter alia* in the Many Labs project, is to present subjects with a number that is clearly too small or too large and then to ask them to estimate the distance between San Francisco and New York City; subjects anchor on the number presented and thus tend to provide smaller (larger) estimates when presented with the small (large) number.

Suppose a future large-scale replication project follows our recommendations and employs the one phenomenon, many studies approach; systematically varies method factors; analyzes the resulting data using a hierarchical model fit to the individual-level observations with the small versus large anchor specified as the only theoretical moderator; and finds a very large average effect size and high heterogeneity such that the sign of the effect is virtually always in the hypothesized direction. We would not stop here deeming the replication a success.

Instead, we would ask what variation in method factors drove this high heterogeneity. Suppose the researchers examine their data and find that it results from the fact that the studies were conducted across, say, the United States and Europe. Given the operationalization of the dependent measure employed (i.e., estimate of the distance between San Francisco and New York City), it might not be surprising if European subjects relied more heavily on the anchor than American subjects. This may or may not be interesting for theory. On one hand, it could be deemed a mere nuisance interaction between method factors and theoretical moderators—important to know about and perhaps even to account for in the analysis but not of theoretical interest. On the other hand, it could be deemed extremely theoretically interesting: assuming that American subjects are more knowledgeable about the distance between San Francisco and New York City than European subjects, perhaps this result suggests that expertise attenuates or even eliminates the anchoring effect. This could be examined by reformulating the hierarchical model to account for expertise and by designing future studies to explicitly test this.

Of course, if method factors were systematically varied as we recommend, there would be multiple operationalizations of the dependent measure employed (e.g., estimate of the distance between Rome and Berlin). Consequently, one could examine also whether the reverse holds under this operationalization (i.e., whether American subjects rely more heavily on the anchor than European subjects).

We recognize that conducting multiple replications as opposed to a single replication and systematically varying various method factors is inherently quite challenging in terms of coordination efforts, resources required, and related matters. It also requires more complex designs for replication projects and more elaborate statistical models. Nonetheless, we believe the benefits for assessing replicability (in particular, the degree to which method factors moderate the effect in question) and advancing theory are sufficiently valuable to warrant this investment.

5. Large-Scale Replication in Other Domains of Psychological Research

Because current large-scale replications projects have been prospective in nature, they have thus far been restricted to phenomena for which data can be collected quickly and cheaply. For example, one criterion the OSC project authors used to choose the three journals from which they selected phenomena to replicate was that they "represent psychology sub-disciplines that have a high frequency of studies that are feasible to conduct at relatively low cost" (Open Science Collaboration 2015). Similar criteria were applied in the Many Labs project and RRRs. Consequently, these projects have been constrained to certain types of research (e.g., self-report or behavioral data collected from convenience samples) and thus certain domains (e.g., cognitive and social psychology).

Given this, one might ask how to conduct large-scale replication projects to assess replicability and understand the role of method factors in other domains of psychological research (e.g., clinical psychology, developmental psychology, behavioral genetics, behavioral neuroscience, health psychology, counseling psychology, and community psychology) where data collection is slow and costly and individual datasets are typically much richer (see also Tackett et al. 2017). Since researchers in these domains often have access to large amounts of shared (or shareable) archival data, we argue that a retrospective approach to replication is valuable. In particular, researchers in these areas could pool data across laboratories and model it using either a traditional meta-analytic model fit to summary statistics or a hierarchical model fit to the individual-level observations. They could then assess how effects replicate and vary across laboratories and investigate whether method factors either interact with theoretical moderators or are, in fact, unaccounted for theoretical moderators.

Further, a prospective approach to replication (that is respectful of the pace and cost of data collection in these domains) is also possible and valuable. For example, a researcher planning efforts to investigate some new phenomenon could ensure that variables that speak to previously investigated phenomena are built into the data collection protocol; in doing so, the researcher could even consciously vary method factors. Then, the new data could be pooled with previously collected data from other laboratories and analyzed via meta-analytic or hierarchical models allowing replicability to be assessed in light of this variation in method factors.

We recognize that barriers to large-scale replication in these domains are likely to be quite different from (and larger than) those encountered in current large-scale replication projects. For example, a move toward widespread data sharing and thus likely also from single laboratory work to many laboratories work will require shifts in culture, incentives, and infrastructure (e.g., standards for authorship, publication, tenure, and funding). It will also require researchers to better understand

statistical methods appropriate for analyzing pooled data (i.e., hierarchical models) and more complex effects (e.g., curve or function estimates as opposed to point estimates); laboratory-specific and other moderators most relevant to include in such analyses; additional method factors that drive heterogeneity (e.g., drop out mechanisms in longitudinal studies); ethical and legal implications of sharing sensitive data; and how to harmonize measurements across laboratories (e.g., if they use different measures of depression). Finally, as discussed in the prior section, assessments of replicability should not depend solely on estimates of effects or significance tests based on them; particularly in the rich datasets common in these domains where there are multiple effects of interest all of which are nonzero but which vary tremendously across laboratories, heterogeneity simply must be a consideration in such assessments.

6. Discussion

We have discussed large-scale replication projects as a positive development that has emerged from recent difficulties with replication in psychological research. While we believe current large-scale replication projects represent an important step forward, we believe they can be made even more informative.

Specifically, replication is complicated in psychological research because studies of a given psychological phenomenon can never be direct or exact replications of one another. As a consequence, heterogeneity cannot be avoided. Rather than trying in vain to eliminate it, we instead argue that large-scale replication projects going forward should embrace it through the systematic variation of method factors.

There is already growing appreciation for the important role that method factors—and the heterogeneity that results from varying them—play in traditional work outside of large-scale replication projects. For instance, researchers are increasingly considering how method factors might moderate effects—if only to attempt to explain recent replication failures. More systematically, Simons, Shoda, and Lindsay (2017) make the excellent suggestion that all primary research papers include a "Constraints on Generality" statement that identifies key method factors, thereby "help[ing] other researchers...when conducting a direct replication" and "encourag[ing] follow-up studies that test the boundary conditions of the original finding." Responding to this suggestion, Kennedy and Gelman (2018) propose that a Constraints on Generality statement provides the necessary information to allow for inference, via hierarchical modeling and poststratification (Park, Gelman, and Bafumi 2004), regarding how quantities of interest might vary along with the key method factors identified in the statement. Thus, our recommendations can be seen as part of an increasing call for researchers to think systematically about method factors and heterogeneity.

If our recommendations prove fruitful, perhaps a move to a many phenomena, many studies approach to large-scale replication may be warranted. This could potentially yield even deeper benefits for theory. Specifically, if seemingly distinct phenomena are subject to the same heterogeneity-inducing method factors, this might point to commonalities at a higher theoretical level.

Our recommendations will not solve recent difficulties with replication in psychological research. Indeed, they are not meant to. Rather, given that heroic coordination efforts and tremendous resources are being invested in large-scale replication projects, we would like their payoff to be as large as possible. Large-scale replication projects that systematically vary method factors will provide much more information on the degree of heterogeneity and thereby can play an important role not only in assessing replicability but also in advancing theory.

References

Bargh, J. A., Chen, M., and Burrows, L. (1996), "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action," *Journal of Personality and Social Psychology*, 71, 230–244. [99]

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., and Trotschel, R. (2001), "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals," *Journal of Personality and Social Psychology*, 81, 1014–1027. [99]

Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., White, C., De Boeck, P., and Vandekerckhove, J. (2018), "Meta-Studies for Robust Tests of Theory," *Proceedings of the National Academy of Sciences*, 115, 2607–2612. [102]

Bem, D. J. (2011), "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect," *Journal of Personality and Social Psychology*, 100, 407–425. [99]

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G., Cornelissen, G., Døssing, F. S., and Espín, A. M. (2017), "Registered Replication Report: Rand, Greene, and Nowak (2012)," *Perspectives on Psychological Science*, 12, 527–542. [100]

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., and Van't Veer, A. (2014), "The Replication Recipe: What makes for a Convincing Replication?" *Journal of Experimental Social Psychology*, 50, 217–224. [99]

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Power, L. P., and Jacobs, W. J. (2014), "A Duty to Describe: Better the Devil You Know than the Devil You Don't," *Perspectives on Psychological Science*, 9, 626–640. [99]

Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010), "Power Posing Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance," *Psychological Science*, 21, 1363–1368. [99]

Chernev, A., Böckenholt, U., and Goodman, J. (2015), "Choice Overload: A Conceptual Review and Meta-Analysis," *Journal of Consumer Psychology*, 25, 333–358. [102]

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R., Aykutoğlu, B., Bahník, Š., Bowen, J., Bredow, C., Bromberg, C., and Caprariello, P. (2016), "Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002)," *Perspectives on Psychological Science*, 11, 750–764. [100]

Chung, Y., Rabe-Hesketh, S., and Choi, I.-H. (2013a), "Avoiding Zero Between-Study Variance Estimates in Random-Effects Meta-Analysis," *Statistics in Medicine*, 32, 4071–4089. [101]

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013b), "A Non-Degenerate Estimator for Hierarchical Variance Parameters via Penalized Likelihood Estimation," *Psychometrika*, 78, 685–709. [101]

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J., Aucoin, P., Berger, S., Birt, A., Capezza, N., and Carlucci, M. (2016), "Registered Replication Report: Hart & Albarracín (2011)," *Perspectives on Psychological Science*, 11, 158–171. [101]

Ehrenberg, A. S. and Bound, J. A. (1993), "Predictability and Prediction," *Journal of the Royal Statistical Society*, Series A, 156, 167–206. [102]

Fabrigar, L. R. and Wegener, D. T. (2016), "Conceptualizing and Evaluating the Replication of Research Results," *Journal of Experimental Social Psychology*, 66, 68–80. [99]

Gilovich, T., Griffin, D., and Kahneman, D. (2002), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge, UK: Cambridge University Press. [103]

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., and Bruyneel, S. (2016), "A Multilab Preregistered Replication of the Ego-Depletion Effect," *Perspectives on Psychological Science*, 11, 546–573. [100,101]

Higgins, J. P. and Thompson, S. G. (2002), "Quantifying Heterogeneity in a Meta-Analysis," *Statistics in Medicine*, 21, 1539–1558. [101]

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., and Botella, J. (2006), "Assessing Heterogeneity in Meta-Analysis: Q Statistic or I^2 Index?" *Psychological Methods*, 11, 193–206. [101]

Ioannidis, J. P., Patsopoulos, N. A., and Evangelou, E. (2007), "Uncertainty in Heterogeneity Estimates in Meta-Analyses," *BMJ*, 335, 914–916. [101]

Jacowitz, K. E. and Kahneman, D. (1995), "Measures of Anchoring in Estimation Tasks," *Personality and Social Psychology Bulletin*, 21, 1161–1166. [102]

Kennedy, L. and Gelman, A. (2018), "How the Recognition of Constraints on Generality can Provide a Direction for Generalizing Research Findings," Columbia University Technical Report. [104]

Klein, R. A., Ratliff, K., Nosek, B. A., Vianello, M., Pilati, R., Devos, T., Galliani, E. M., Brandt, M., van 't Veer, A., Rutchick, A. M., Schmidt, K., Bahnik, S., Vranka, M., IJzerman, H., Hasselman, F., Joy-Gaba, J., Chandler, J. J., Vaughn, L. A., Brumbaugh, C., van swol, L., Wichman, A., Packard, G., Brooks, B., Cemalcilar, Z., Storbeck, J., Bocian, K., Levitan, C., Bernstein, M. J., Krueger, L. E., Eisner, M., Davis, W. E., Nier, J. A., Nelson, A. J., Steiner, T. G., Mallett, R., Thompson, D., Huntsinger, J. R., Morris, W., Skorinko, J., and Kappes, H. (2014), "Investigating Variation in replicability: A "Many Labs" Replication Project," *Social Psychology*, 45, 142–152. [99,101]

Lindsay, R. M. and Ehrenberg, A. S. (1993), "The Design of Replicated Studies," *The American Statistician*, 47, 217–228. [102]

McShane, B. B. and Böckenholt, U. (2014), "You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic," *Perspectives on Psychological Science*, 9, 612–625. [102]

——— (2017), "Single Paper Meta-analysis: Benefits for Study Summary, Theory-testing, and Replicability," *Journal of Consumer Research*, 43, 1048–1063. [100,102]

——— (2018), "Multilevel Multivariate Meta-analysis with Application to Choice Overload," *Psychometrika*, 83, 255–271. [102]

McShane, B. B., Böckenholt, U., and Hansen, K. T. (2016), "Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes," *Perspectives on Psychological Science*, 11, 730–749. [101]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [99,103]

Park, D. K., Gelman, A., and Bafumi, J. (2004), "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls," *Political Analysis*, 12, 375–385. [104]

Pigott, T. (2012), *Advances in Meta-Analysis*, New York: Springer. [101]

Rosenthal, R. (1991), *Replication Research in the Social Sciences*, Newbury Park, CA: Sage, chap. Replication in Behavioral Sciences, pp. 1–30. [99]

Simons, D. J., Holcombe, A. O., and Spellman, B. A. (2014), "An Introduction to Registered Replication Reports at Perspectives on Psychological Science," *Perspectives on Psychological Science*, 9, 552–555. [99,100]

Simons, D. J., Shoda, Y., and Lindsay, D. S. (2017), "Constraints on Generality (COG): A Proposed Addition to all Empirical Papers," *Perspectives on Psychological Science*, 12, 1123–1128. [104]

Stanley, T., Carter, E. C., and Doucouliagos, H. (2017), "What Meta-Analyses Reveal about the Replicability of Psychological Research." Tech. Rep., Deakin Laboratory for the Meta-Analysis of Research. [101]

Stroebe, W. and Strack, F. (2014), "The alleged crisis and the Illusion of Exact Replication," *Perspectives on Psychological Science*, 9, 59–71. [99]

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., and Shrout, P. E. (2017), "It's Time to Broaden the Replicability Conversation: Thoughts for and from Clinical Psychological Science," *Perspectives on Psychological Science*, 12, 742–756. [103]

Tsang, E. W. and Kwan, K.-M. (1999), "Replication and Theory Development in Organizational Science: A Critical Realist Perspective," *Academy of Management Review*, 24, 759–780. [99]

van Erp, S., Verhagen, J., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2017), "Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in Psychological Bulletin From 1990–2013," *Journal of Open Psychology Data*, 5. [101]

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., Albohn, D., Allard, E., Benning, S., and Blouin-Hudon, E.-M. (2016), "Registered Replication Report: Strack, Martin, & Stepper (1988)," *Perspectives on Psychological Science*, 11, 917–928. [100]

The American Statistician

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

Valid *P*-Values Behave Exactly as They Should: Some Misleading Criticisms of *P*-Values and Their Resolution With *S*-Values

Sander Greenland

To cite this article: Sander Greenland (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:sup1, 106-114, DOI: 10.1080/00031305.2018.1529625

To link to this article: https://doi.org/10.1080/00031305.2018.1529625

© 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: 26374

View related articles

View Crossmark data

Citing articles: 82 View citing articles

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, 106–114: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529625

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Valid *P*-Values Behave Exactly as They Should: Some Misleading Criticisms of *P*-Values and Their Resolution With *S*-Values

Sander Greenland

Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA

Abstract

The present note explores sources of misplaced criticisms of *P*-values, such as conflicting definitions of "significance levels" and "*P*-values" in authoritative sources, and the consequent misinterpretation of *P*-values as error probabilities. It then discusses several properties of *P*-values that have been presented as fatal flaws: That *P*-values exhibit extreme variation across samples (and thus are "unreliable"), confound effect size with sample size, are sensitive to sample size, and depend on investigator sampling intentions. These properties are often criticized from a likelihood or Bayesian framework, yet they are exactly the properties *P*-values *should* exhibit when they are constructed and interpreted correctly within their originating framework. Other common criticisms are that *P*-values force users to focus on irrelevant hypotheses and overstate evidence against those hypotheses. These problems are not however properties of *P*-values but are faults of *researchers* who focus on null hypotheses and overstate evidence based on misperceptions that $p = 0.05$ represents enough evidence to reject hypotheses. Those problems are easily seen without use of Bayesian concepts by translating the observed *P*-value p into the Shannon information (*S*-value or surprisal) $-\log_2(p)$.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Compatibility; Dichotomania; Evidence; Information; Logworth; Nullism, *P*-values; S-values; Significance testing; Surprisal

1. Introduction

There are many reasons to be critical of traditional testing (which tethered *P*-values to decision rules); nonetheless, they provide no basis for blaming *P*-values for behaving as they were designed to do, or for their misuse and misinterpretation (Benjamini 2016). As has been argued before (e.g., Senn 2001, 2002), the problems are instead a failure of textbooks and tutorials to describe correctly the inferential meaning of *P*-values, and a failure to describe test hypotheses appropriate for practical needs. There is also an egregious tendency to blame users for these problems, even though the statistics literature up to the highest levels displays descriptions and definitions founded on jargon that is inconsistent across sources and which violate ordinary language meanings. These problems require far more expertise to sort out than ordinary users could be reasonably expected to have, and set the stage for confusions and misinterpretations such as those discussed by Gigerenzer (2004), Hoekstra et al. (2006), Hurlbert and Lombardi (2009), Greenland et al. (2016), Amrhein et al. (2017), McShane et al. (2017, 2018), Wasserstein and Lazar (2016), and many other sources.

The present note describes how authoritative definitions of "significance levels" and "*P*-values" vary, leading to misinterpretations of *P*-values as error probabilities. It then discusses several criticisms of *P*-values that are instead problems of poor teaching and terminology, and several properties of *P*-values that are often presented as fatal flaws but reflect instead how valid *P*-values should behave. The underlying view is that, while *P*-values are certainly limited in scope and difficult to understand properly, many strident criticisms judge these statistics according to inappropriate criteria or in comparison to methods that are subject to parallel criticisms. Especially pernicious are criticisms that overlook how demands for conclusive inferences can undermine any method, including confidence intervals and Bayesian statistics.

Concepts will be illustrated with a cohort study which compared adverse event rates among infants receiving ibuprofen (Advil™, Motrin™), alone or in addition to other drugs, to rates among those receiving only acetaminophen (paracetamol, Tylenol™) (Walsh et al. 2018). A highlighted result for renal (kidney) adverse events was an adjusted estimated rate ratio (*RR*) of 1.84 with 95% confidence limits of 0.66, 5.19, which correspond to an event-rate increase of 84% for ibuprofen compared to acetaminophen alone, an interval ranging from a 34% decrease to a 419% increase, and a *P*-value of 0.25 for testing the null hypothesis of no association (rate ratio of 1, a 0% difference). Theoretical descriptions will focus on testing of a hypothesis **H** about a coefficient β embedded in a regression model **A** (which encodes the set of background assumptions used for the test), such as **H**: $\beta = b$ where b is a fixed, hypothesized value for β; in multiplicative rate models $\beta = \ln(RR)$. The tested value b is usually *but need not be* zero. The discussion will also consider the more general case of tests of fit of the embedding model **A**.[1] For simplicity, however, I will assume the dataset is large enough so that all the usual continuous

[1] This case is more general in that a test of $\beta = b$ given **A** can be treated as a test of the fit of the model in which $\beta = b$ holds relative to the model in which β is unconstrained but the embedding model **A** holds.

CONTACT Sander Greenland lesdomes@ucla.edu Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA
© 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.

large-sample approximations hold. The most technical asides will be in footnotes. As will be described, some of the problems in understanding P-values as evidence measures can be reduced by converting the observed P-value p to an information measure such as the S-value $s = \log_2(1/p) = -\log_2(p)$.

2. The Confusing and Unacknowledged Variation in Basic Definitions and Descriptions

Some problems can be traced to variation in definitions and terminology across authoritative sources with no mention of the variation by those sources. At least two definitions of "the observed P-value" are in wide use. In the usual inferential (Fisherian) definition, a P-value is the tail probability p under **H** that a test statistic (such as an absolute Z-score or χ^2) would be as large or larger than what was observed, given the embedding model **A** (Cox and Donnelly 2011, sec. 8.4). Fisher (1925) called p the "significance level" or "value of P," and "significance level" is common in British sources thereafter (e.g., Cox and Hinkley 1974). In the decision-theoretical (Neyman–Pearsonian) definition, the observed p is often defined as the smallest α level (testing cutoff) that would allow rejection in an α-level decision rule (Neyman–Pearson hypothesis test) which rejects **H** when $p \leq \alpha$ (e.g., Lehmann 1986). While the two definitions appear superficially distinct, they are mathematically equivalent and thus represent logically a single definition stated in two different ways.

Unfortunately, some authors use "significance level" to refer to α rather than p (Lehmann 1986, p. 70). This second usage of "significance level" contradicts the original usage and leads to confusion of p and α, often in very subtle ways (as discussed in the next section). In the face of such inconsistent usage among authoritative texts, it should hardly be surprising when basic textbooks as well as researchers confuse p and α. The observed p is a sample feature relating the observed data to the hypothesis **H** and model **A** used to compute p. Specifically, $100p$ is the percentile location of the observed test statistic in a distribution computed from **H** and **A** (Perezgonzalez 2015), and in this special sense p can be considered as describing something about the data. In contrast, α is a fixed known number (like 0.05) that tells us nothing about the data.

Despite the common definition of "P-value" as a probability, many decision-theoretical authors instead focus only on repeated-sampling properties, and thus define a "P-value" not as the observed value p, but instead as the random variable P whose value (realization) in a given sample is the observed p (e.g., Kuffner and Walker 2017; Murdoch et al. 2008). Thus, we have two logically distinct definitions of "P-value." But, as with the conflicting terminology, this conflict in definitions is rarely noted—so it should be hardly be surprising when basic texts and researchers confuse the random P with the observed p, which is similar to confusing the name of a variable with an unspecified value for it (e.g., confusing the variable X called "weight in kg" with the unspecified value "x kg" that one might observe upon weighing someone).

The distinction between the two definitions is important, not the least because frequentists further define validity of the P-value in terms of the random P: The random variable P is said to be *valid for testing* **H** given **A** (or properly calibrated, or a U-value) if it is uniformly distributed when **H** and **A** are correct; in that case, for every α the rule "reject **H** when $p \leq \alpha$" will falsely reject **H** with frequency α (Bayarri and Berger 2004).[2] Of course, issues of power must enter into choice of the test statistic from which p is computed, but those issues are outside the present scope.

Confusion of the observed P-value p with the random variable P may be a major contributor to some of the fallacies described below and earlier (e.g., Greenland et al. 2016). Worse, however, is that the tail-probability definition of the observed p is often equated to or replaced by wholly incorrect descriptions such as "p is the probability that chance alone produced the association," which reflects confusion beyond mere terminology (since the probability of chance alone is none other than the probability that **H** and **A** are correct; see item no. 2 in Greenland et al. 2016). Setting such outright errors aside, more subtle problems arise when the observed p is described as measuring evidence against **H**, because it is inversely related to that evidence: *smaller* values of p correspond to *more* evidence against **H** in the data, given the model **A**.

In an attempt to forestall misinterpretations, p can be described as a measure of the degree of statistical compatibility between **H** and the data (given the model **A**) bounded by $0 =$ complete incompatibility (data impossible under **H** and **A**) and $1 =$ no incompatibility apparent from the test (Greenland et al. 2016). Similarly, in a test of fit of **A**, the resulting p can be interpreted as a measure of the compatibility between **A** and the data.[3] The scaling of p as a measure is poor, however, in that the difference between (say) 0.01 and 0.10 is quite a bit larger geometrically than the difference between 0.90 and 0.99. For example, using a test statistic that is normal with mean zero and standard deviation (SD) of 1 under **H** and **A,** a p of 0.01 vs. 0.10 corresponds to about a 1 SD difference in the statistic, whereas a p of 0.90 vs. 0.99 corresponds to about a 0.1 SD difference.

One solution to both the directional and scaling problems is to reverse the direction and rescale P-values by taking their negative base-2 logs, which results in the S-value $s = -\log_2(p)$. Larger values of s do correspond to more evidence against **H**. As discussed below this leads to using the S-value as a measure of evidence against **H** given **A** (or against **A** when p is from a test of fit of **A**).

3. A *P*-value is Not an Error Probability (Except in a Useless Hypothetical Sense)

Describing or defining an observed P-value p as a minimum α-level for rejection along with references to both p and α as "significance levels" seem to have led to p being misinterpreted as an error probability for actual decision problems. The latter interpretation is just a mistake (Goodman 1999; Sellke et al. 2001; Hubbard and Bayarri 2003; see also Greenland et al. 2016, item 9): The theory of α-level hypothesis tests requires an α

[2] In parallel, P is *conservatively valid* if the rule "reject **H** when $p \leq \alpha$" will reject **H** with frequency no greater than α when **H** and **A** are correct; conservative validity is usually the best one can do with discrete data.

[3] One could also say "p measures the compatibility of **A** with the data" when the subject is **A**, or "p measures the compatibility of the data with **A**" when the subject is the data.

that is a constant (such as 0.05) specified by the analyst *before* seeing the data as the desired upper bound for the error rate of a rejection rule over an entire sequence of datasets generated from the model (Neyman 1977; Lehmann 1986). Specifically, α is determined independently of the data, based on the cost of false rejections: Higher α would be used by those for whom false positives are of minor consequence, and lower α would be used by those for whom false positives are of major consequence (Lakens et al. 2018). Since decision consequences could vary for different stakeholders in the same setting, different readers of the same study report may well have different losses from false positives and false negatives, and thus different α-levels, in turn leading to conflicting decisions based on seeing the same data and thus the same observed *P*-value p.

Nonetheless, no such α is needed for the definition or presentation of *P*-values, nor does comparison of p to α make the number p an error rate. We can imagine a decision rule that has an error rate α equal to the observed p, but α and p are different quantities conceptually since p varies across samples and thus cannot be prespecified. In the example, the rate ratio of 1.84 for renal adverse events corresponds to an estimated ibuprofen coefficient of $\hat{\beta} = \ln(1.84) = 0.610$ in an exponential-rate model, with standard error $\sigma = \ln(5.19/0.66)/2(1.96) = 0.526$ and test statistic for **H**: $\beta =$ of $\hat{\beta}/\hat{\sigma} = 1.159$, yielding a *P*-value p of 0.25. We can imagine a rule defined by "reject $\beta = 0$ if its *P*-value is less than $\alpha = 0.25$," but no one stated that rule *before* seeing the data let alone derived it from error costs. To describe the observed p as a 25% Type-I error rate is thus at best a statement referring to someone who had a prespecified α of 0.25 which the *P*-value just happened by chance to equal. There is surely no such person in the example, and thus the statement is a completely unnecessary distraction: An accurate description of the test result is simply that p for $\beta = 0$ was 0.25. Readers who wanted to base a decision on that *P*-value alone could immediately see whether 0.25 was above or below their own cutoff α.

A small related point is the common confusion of the α-level α with the actual probability that the test rejects **H** when **H** is true (the Type-I error rate of the test, or test size). The actual Type-I error rate of a test of the hypothesis **H** given the assumptions **A** is often unknown to the investigator, because it may deviate from α due to defects in **A** or discreteness of the data.[4] In contrast, α is defined as the maximum tolerable Type-I error rate, which is set by the investigator and thus is known; p is then compared to this α to make decisions, on the assumption that the corresponding random *P* is valid (which makes α equal to the Type-I error rate).

4. Misleading Criticisms of *P*-values

4.1. *P*-Values Do Not Force Users to Focus on Null Hypotheses—But Nullistic Jargon Does

A common criticism of statistical testing is that it forces users to focus on irrelevant null hypotheses. There is no question that many null hypotheses are indeed scientifically irrelevant (Cohen 1994). This irrelevancy problem is not, however, a fault of *P*-values, but is instead a product of traditional training and an academic environment that makes users focus on such hypotheses. While there are ongoing calls to abolish misleading jargon involving "significance" (Amrhein et al. 2018; McShane et al. 2018), almost no attempt has been made to correct Fisher's mistake of using "null hypothesis" for any tested hypothesis **H**, ignoring that in ordinary English "null" is a synonym for zero or nothing.[5] This tradition has led users to think and certain experts to claim that statistical science is only about testing "null hypotheses," where "null" means "no association" or "no effect" rather than any and all hypotheses of importance or concern (Greenland 2004, 2017).

Breaking from Fisher's misleading terminology, Neyman (1977, pp. 104–106) instead called **H** the targeted or *tested* hypothesis. But Fisher's jargon has prevailed with the back-rationalization that **H** is the hypothesis to be "nullified" by the test, and was sustained by attempts to distinguish hypotheses like $\beta = 0$ as "nil hypotheses" (Cohen 1994). Unsurprisingly, then, the ill effects of Fisher's usage continue to be seen in the justly maligned cult of null-hypothesis significance testing (NHST) (Ziliak and McCloskey 2008) in which *P*-values are computed only for hypotheses of no effect, when they should also be given for alternatives of relevance (e.g., the hypotheses used to compute power in funding applications).

A more technically involved problem is excessive focus on point hypotheses, which will be discussed below under sensitivity to sample size.

4.2. *P*-Values Are "Unreliable"—Exactly as They Should Be

It is often noted that *P*-values vary dramatically from sample to sample even under ideal experimental replications (Goodman 1992; Senn 2001,2002; Gelman and Stern 2006; Murdoch et al. 2008; Boos and Stefanski 2011). As a consequence, some researchers criticize *P*-values for being "unreplicable" or "unreliable" or "noisy," or for failing to converge to some constant upon replication or increase in sample size, as if a *P*-value is estimating a parameter or "statistical significance" is a state of nature.

Parameter measurement is not, however, done by the *P*-value, but instead by the estimator $\hat{\beta}$: This estimator and its standard error extract information on the systematic component of data variation represented by β in the model **A**. The observed p is computed from the remaining variation, as captured by the absolute *Z*-score $|\hat{\beta} - b|/\hat{\sigma}$ for testing $\beta = b$; this test statistic can be viewed as a summary standardized residual for comparing the submodel of **A** in which **H** is correct ($\beta = b$) to the embedding model **A** which has no restriction on β. Upon recognizing that a *P*-value is a function of what would be purely random error under **H** and **A**, it should be seen that *by definition* it does not measure any parameter or population quantity—quite the opposite: the *P*-value is a measure of random error in the estimate, *given* **H** and **A**, rescaled (standardized) to a uniform distribution. Similarly, a *P*-value for a test of fit of the embedding model **A** may be based on a residual sum of squares which is a rescaled measure of the noise or random

[4] When the test is a conservatively valid test of the entire model (assumption set) **A**, discreteness becomes the only source of discrepancy.

[5] See, e.g., "null" in Oxford 2017: adj. 2. Having or associated with the value zero; noun 1. Zero. Merriam-Webster 2017: adj. 6. of, being, or relating to zero; 7. zero.

error left after extracting what the model **A** says is the signal or systematic variation (the fitted regression equation or table of fitted values, which is the signal remaining after the noise is filtered out).

Thus, if **H** and **A** are correct, the random P will (if valid) bounce around uniformly across the unit interval, producing $p \leq \alpha$ about $100\alpha\%$ of the time. This is exactly what P should do, because in that case it should be pure uniform noise (random error) which is supposed to vary unpredictably from study to study (Senn 2001, 2002). If across replications of data collection and analysis P does *not* look like uniform noise, we are warned that at least one of **H** or **A** is incorrect, i.e., we are using a wrong hypothesis or wrong model for inference (signal extraction) or decision. That is exactly what statistical tests are for. If instead P is uniform, we can only say that this particular test is insensitive to whatever violations of **H** or **A** are present (in decision-theoretical terms: for these violations, the test based on P has power equal to its Type-I error rate, meaning it conveys no information about those violations). Thus, P-values provide diagnostic or warning mechanisms for hypothesis or model problems, and like all such mechanisms are fallible.

4.3. P-Values Confound Effect Size with Sample Size—Exactly as They Should

The concept of "statistical compatibility" has been a major obstacle to proper understanding and interpretation of P-values, as the concept involves comparing the magnitude of discrepancies between observations and expectations against the estimated random variability in those discrepancies. It is in fact often lamented that P-values confound effect size with sample size, and they are even banned on account of this (Lang et al. 1998). That is ironic because this "confounding" reflects how P-values are doing their job correctly: The distance of (say) an estimate from a model prediction should constitute evidence against the model; nonetheless, *how much* evidence that distance corresponds to should depend on the precision of the estimate.

In particular, the information against **H**: $\beta = b$ given the model **A** is a function of both the absolute deviation (distance from estimate to hypothesized value) $|\hat{\beta} - b|$ and the standard error $\hat{\sigma}$ where $\hat{\beta}$ and $\hat{\sigma}$ correspond to the imputed signal and noise level based on the auxiliary (decoding) assumptions in **A**. These quantities are combined in the test statistic $|\hat{\beta} - b|/\hat{\sigma}$ for **H** given **A**. Because the precision $\hat{\sigma}^{-2}$ of $\hat{\beta}$ is proportional to the sample size n, the test statistic depends directly on n and on the distance $|\hat{\beta} - b|$ from the estimate $\hat{\beta}$ to the hypothesized value b, and so p is inversely related to both n and $|\hat{\beta} - b|$, exactly as a consistent measure of evidence against $\beta = b$ should be. In the example, for **H**: $\beta = 0$ the distance is $|\hat{\beta} - 0| = 0.610$, again with a p of 0.25; if, however, the precision of the study were increased fourfold (e.g., by quadrupling the size n of the cohort), $\hat{\sigma}$ would be halved and the same distance of 0.610 would give a p of 0.02, reflecting the fact that evidence against **H** represented by a given deviation *should* increase with n.

4.4. P-Values Do not Overstate Evidence Against Hypotheses—People Do

Among fair criticisms of P-values are that they are too easily confused with posterior probabilities, and that they are distortive evidence measures that need logarithmic transformation to gauge properly (e.g., Bayarri and Berger 1999; Boos and Stefanski 2011; Greenland 2017, 2018). Thus, consider again the Shannon-information or S-value transform of the observed P-value, $s = -\log_2(p)$, which is a measure of the information against **H** encoded in the test statistic (the refutational information supplied by the test given the model **A**).[6] The negative log transform of a probability is also known as the *self-information* or *surprisal* for observing an event that has probability p (Shannon 1948; MacKay 2003; Fraundorf 2017). With base-2 logs, the units for measuring this information are *bits* (binary digits); the first integer larger than s is the number of binary digits (indicator variables) needed to encode p.[7]

There are other formal definitions of statistical information, but the S-value s is a simple cognitive device for appreciating the information conveyed by p without reference to contextual details outside of those used to specify **H** and **A**. Notably, the information measure s refers to the observed tail probability p (rather than the random P) and thus represents a relation between the observed data and the model formed by combining **H** with **A**. A useful consequence of its log scaling is that it makes the information additive across independent tests, a fact used to create meta-analytic P-values (Cox and Hinkley 1974, p. 80). To provide an intuitive interpretation of the information conveyed by s, let k be the nearest integer to s. We may then say that p conveys roughly the same information or evidence against the tested hypothesis **H** given **A** as seeing all heads in k independent tosses of a coin conveys against the hypothesis that the tosses are "fair" (each independent with chance of heads $= 1/2$) versus loaded for heads; k indicator variables all equal to 1 would be needed to represent this event.

As an illustration, the chance of seeing all heads in 4 fair tosses is $1/2^4 = 0.0625$. Thus, under the model **A**, observing a P-value of 0.05 conveys only $s = -\log_2(0.05) = 4.3$ bits of information against **H**: $\beta = b$, which is hardly more surprising than seeing all heads in 4 fair tosses. In the ibuprofen example, the P-value of 0.25 is no more surprising if $\beta = 0$ than seeing 2 heads in 2 fair tosses, since $s = -\log_2(0.25) = 2$. For contrast, $\beta = \ln(2)$ corresponds to a doubling of the adverse-event rate with ibuprofen; the P-value from the statistic $|\hat{\beta} - \ln(2)|/\hat{\sigma} = 0.158$ is 0.87, for which $s = -\log_2(0.87) = 0.19$, showing that there is even less information against a doubling of the adverse-event rate with ibuprofen than against no difference ($\beta = 0$).

By this measure, and contrary to certain commentaries (e.g., Goodman 1999; Sellke et al. 2001; Hubbard and Lindsay 2008), P-values do *not* overstate evidence against hypotheses or models: The observed p is just a hypothetical probability or percentile computed from **H** and **A**. Any overstatement of the evidence conveyed by p is from those who (based on the entrenchment of $\alpha = 0.05$ in automated decision rules) mistakenly think a p of 0.05 represents just enough evidence to reject the tested hypothesis or model. Furthermore, under this interpretation any changes in the P-value p and thus the S-value s as a result

[6] Good 1956, pp. 1132; 1983, 146 suggested using this measure with centering around its mean $E\{-\log_2(P)\}$, the *Shannon entropy* of P (which for valid P is itself is maximized when **H** and **A** are correct).

[7] In base-10 logs the units are called *Hartleys*; when **H**: $\beta = 0$ this measure is sometimes called the *logworth* of β. The first integer larger than $-\log_{10}(p)$ is the number of decimal digits needed to encode p.

of changing the embedding model or tested hypothesis (e.g., from changing from a single to a multiple comparison) correctly reflects the differences in the information supplied by the data against the different models or hypotheses. Yet, the *S*-value also corroborates the commentaries by revealing that a p of 0.05 maps into only 4.3 bits of refutational information, showing that interpreting a p of 0.05 as "borderline evidence" is nothing more than a bad cultural habit. This is an underlying cognitive issue which can be seen without use of machinery such as prior spikes or Bayes factors (which even some Bayesian sympathizers find objectionable, e.g., Casella and Berger 1987ab; Gelman 2013; Greenland and Poole 2013).

4.5. P-values are Sensitive to Sample Size—Exactly as They Should Be

Using the *S*-value $s = -\log_2(p)$ to measure information in a test statistic, *P*-values obey the refutational version of the *P*-postulate (less accurately termed "the α postulate"): Equal *P*-values correspond to equal refutational information against the tested hypothesis **H** or the tested model **A** (Royall 1986). This property is often criticized, e.g., because two studies may provide the same *P*-values for the same test hypothesis and yet may show very different observed associations (Greenland et al. 2016, point 17). This criticism points to the necessity of considering additional background (contextual) information about what violations of **H** are of practical importance. A related criticism is that the embedding model **A** consists of all auxiliary assumptions used for inference, including explicit assumptions such as homogeneity of effects and random sampling or treatment randomization, and less-often stated assumptions such as no database errors or selective analysis reporting. As such, **A** is never perfectly correct; hence, for large enough samples p will become very small and thus s will become very large even if **H** is correct.

It follows that the test may now indicate that at least one or both of **H** and **A** are strictly false even if both are good enough for practical purposes; thus the criticism could be restated as "In large samples, *P*-values become too sensitive to small deviations from **H** or **A**." Bayesians extend this criticism by placing a point mass of prior probability on **H** and spreading the remainder over a restricted family of alternatives (usually the family of all models with β a known constant in the same embedding model **A**). A consequence of this artifice is that data supplying considerable information against **H** according to some criterion α for p or $-\log(\alpha)$ for s may still increase the posterior probability of **H** given **A** (the Jeffreys–Lindley paradox; see, e.g., Royall 1997; Senn 2001; Spanos 2013).

A crude fix for this large-sample sensitivity of *P*-values to unimportant discrepancies is to lower the α-level for rejecting **H** as the sample-size increases (Royall 1986). This fix is, however, only a demand for more information to reject **H** or **A** as the sample size increases, which ignores actual error costs on which α should be based. Furthermore, these costs and demands are irrelevant to measuring the statistical information given by the test. In fact, any valid and efficient test *should* detect model imperfections when given enough data information, even if those are of no practical consequence. Instead the defect lies in the traditional focus on point test hypotheses like $\beta = b$ (Casella and Berger 1987ab), and a failure to pay due attention to the size of the observed discrepancy $|\hat{\beta} - b|$ in practical terms.

Accounting for practical importance instead requires specification of tolerances for imperfections, such as a maximum tolerance for the actual discrepancy $|\beta - b|$. One way to do so is to replace point targets such as **H**: $\beta = b$ with interval targets such as **H**: $|\beta - b| \leq c$, where $[-c, c]$ represents an interval deemed practically equivalent to no discrepancy ($\beta = b$). It may even make more sense contextually to reverse the role of the test and alternative hypothesis, so that the tested **H** becomes $|\beta - b| \geq c$, as in equivalence testing (Berger and Hsu 1996; Senn 2008; Wellek 2010).[8] Other examples of role reversal include risk-limiting audits (Lindeman and Stark 2012). In all these cases, the observed *P*-value now measures data compatibility with the composite hypothesis **H** given **A**, and the *S*-value measures information against that **H** given **A**. Upon recognizing these hypotheses as more contextually relevant than $\beta = b$, sensitivity to sample size ceases to be a valid objection to *P*-values; it is rather an objection to point test hypotheses, reminding us that all our models have imperfections that become noticeable with enough data. It should be reassuring that *P*-values and *S*-values conform to this sound intuition.

An opposite objection is that *P*-values or *S*-values are *insensitive* insofar as they ignore features of the data that may be taken as evidence against the model (such as implausible estimates). This is true, but reflects nothing more than the limited information capacity of any single number such as a point estimate or confidence limit. A one-dimensional summary of multidimensional information is simply not a sufficient statistic for inference; hence, further information (such as residual plots and *P*-values for alternative hypotheses or models) will also be needed to make sensible inferences. A small p only warns that something may be wrong with the hypothesis **H** or embedding model **A**, not what is wrong or that they are unsafe to use. The kind of model updating that would follow from an initial check depends on focused prior information, diagnostics, and suspicions about model violations; for example, poor fit of the embedding model **A** is often addressed by relaxing (expanding) that model to address detected imperfections (e.g., by adding higher-order terms) (Box 1980).

Conversely, a large p only means the test did not provide much information against the tested hypothesis or model, and is *not* a "safety signal." The *S*-value reveals that no parameter value inside a 95% confidence interval has more than 4.3 bits of information against it, supporting recommendations to view the interval interior as a region of hypotheses highly compatible with the data, rather than overconfidently viewing its exterior as a region ruled out by the data (Poole 1987ab; Greenland et al. 2016). In the ibuprofen example, this region of high compatibility for the rate ratios $\exp(\beta)$ extended from 0.66 to 5.19, revealing that the study (even if otherwise perfect) was

[8] There are intricate logical and technical issues in extending *P*-values to composite **H**, e.g., see Berger and Boos 1994; Berger and Hsu 1996; Bayarri and Berger 1999, 2000, 2004. Some extensions may be rejected on the grounds that they lead to logical incoherencies, but Schervish (1996) also rejects the coherent extension based on maximizing p over **H**; the negative log of that supremum does, however, provide a lower bound on the information against **H**.

in fact practically uninformative about possible adverse effects of ibuprofen on renal outcomes, since anything from a 1/3 rate decrease ($RR = 1/3$) to a fivefold rate increase ($RR = 5$) has less than 4.3 bits of information against it. Yet, the study abstract concluded that rates of renal adverse events among infants "were not different between the ibuprofen (+/–acetaminophen) and acetaminophen-only groups," displaying a common type of nullistic cognitive blindness (McShane and Gal 2017; Greenland 2017) to the practical uninformativeness of the quoted results. In contrast, an accurate report would have said "Our study lacked sufficient information to reach any useful inference about adverse renal events comparing ibuprofen to acetaminophen alone; much more data would be needed to address ibuprofen safety concerns"—albeit under current journal publication criteria such an honest conclusion would make publication difficult. One could further remark that there is almost 10 bits of information against a 2/3 rate decrease ($RR = 1/3$) and a 10-fold rate increase ($RR = 10$) for each have $p \approx 2^{-10}$, but that would only distract from the inadequacy of the study for safety assurances.

4.6. P-Values Depend on Investigator Sampling Intentions—Exactly as They Should

Nowhere do statistical principles seem to clash more persistently than in the role of intentions or analysis plans (protocols) in inference (see, e.g., Bayarri and Berger 2004, sec. 5). Consider the claim that "if we stick to the short-run perspective when measuring evidence, identical data produce identical evidence regardless of the experimenters' intentions" (Goodman 1999, p. 1000). Without possibly questionable qualifications, this claim is false. Consider one experiment (which is a very short run) that reports $p = 0.004$ for $\mathbf{H}$: $RR = 1$ with a point estimate for RR of $\widehat{RR} = \exp(\hat{\beta}) = 1.8$ and a likelihood ratio of over 50 for comparing $RR = 2$ to $RR = 1$. If the experimenter intended to follow best practices and reported doing so, we might well regard those results as providing some credible evidence ($s = -\log_2(0.004) = 8$ bits) against $\mathbf{H}$. But suppose we discover that the experimenter's intention was to produce data that gave $p < 0.005$, if necessary by postrandomization reallocation and exclusions; we should then take the resulting data as providing little or no evidence regarding $\mathbf{H}$ or RR in general.

One could say the experimenter's intentions do matter in this example because in the first case they leave the sample space restricted only by the initial valid study design, whereas in the second case they further restrict the sample space to samples with $p < 0.005$. The only way we might reasonably discount these intentions in measuring the data evidence is if we knew *with certainty* that the intentions produced no action that changed the sample, e.g., if we knew the initial treatment allocation happened to give $p = 0.004$ and so led to no manipulation by the ill-motivated experimenter. If, however, we had no such information, we would have to rely on the probability that manipulation would prove unnecessary, which would be 0.005 if $\mathbf{H}$ were correct. Seeing $p = 0.004$ (or a correspondingly small Bayes factor for $\mathbf{H}$) would then only provide us with evidence against the combined hypothesis that "$\mathbf{H}$ is correct *and* there was no experimental misconduct." Thus, intentions matter even though their effects must be mediated through experimenter actions and reporting.

Discovering deceptive intentions of the sort just described should lead us to alter our embedding model in a way that changes the likelihood function for RR from its form under honest intentions (e.g., the probability of $\widehat{RR} = 1.8$ when $RR = 1$ would be much higher with deceptive than with honest intent); thus accounting for such intentions need not be controversial. The ongoing controversy instead concerns intentions which alter p but leave the likelihood function unchanged, as in optional-stopping problems. This is a vast topic, but briefly: The dependence of p on samples not in fact observed (counterfactual data) leads some to reject it as an evidence measure in favor of pure likelihood or Bayesian statistics (Berger and Wolpert 1988; Edwards 1992; Royall 1997; Goodman 1999). For many statisticians, however, the full sampling distribution (including counterfactual datasets) encodes important information about the data-generating mechanism (e.g., Cox and Hinkley 1974; LeCam 1988; Lane 1988; Ritov et al. 2014, p. 635)[9] leaving P-values as basic tools for checking assumptions even if the final results might be reported as posterior probabilities (Box 1980).

The dependence of P-values on counterfactual data is thus seen as a means of allowing for assumption uncertainties without specifying all alternatives (as in tests of fit that allow for unforeseen types of model violations). From this perspective, the insensitivity of likelihood ratios to counterfactual data reflects an information loss which can be extreme in high-dimensional problems (where likelihood-based procedures may provide no consistent inference yet valid P-values can be constructed from the full sampling distribution; see Robins and Wasserman 2000, sec. 5; Ritov et al. 2014).

5. Conclusions

P-values are often criticized for having properties they *should* possess under correct interpretations, and for encouraging misuse and misinterpretation of data. Unquestionably, P-values have proven problematic for correct teaching, description, and usage, as evidenced by the many common misinterpretations (Greenland et al. 2016). These problems are aggravated by the distortive scaling and unwarranted dichotomization of P-values, the misleading and inconsistent jargon surrounding their definitions and description, and the misplaced criticisms which blame P-values for ill-conceived traditions surrounding their use and interpretation (Hurlbert and Lombardi 2009). These problems can be reduced by

1. focusing on definitions of P-values that make no reference to α levels or decisions, in which the observed p is a tail probability and its random counterpart P is a uniform variate under the hypothesis and model used for its construction;

[9] In terms used in this literature, P-values can violate the likelihood principle (which says proportional likelihood functions should yield the same inferences), but dissenters from the principle do not in practical terms regard the likelihood function as a sufficient statistic outside of some narrowly specified circumstances.

2. computing P-values for contextually important alternatives to null (nil or "no effect") hypotheses, such as minimal important differences;
3. rescaling p to the Shannon information (S-value) $s = -\log_2(p)$ to provide a better scale for measuring the amount of information the test supplies against the hypothesis;

and most of all

4. avoiding dichotomization of p or comparison of p to α-levels (see, e.g., Poole 1987a; Hoekstra et al. 2006; Hurlbert and Lombardi 2009; Greenland et al. 2016, 2017; Amrhein et al. 2017, 2018; McShane et al. 2018, 2017; Wasserstein and Lazar 2016).

It may be argued that exceptions to (4) arise, for example when the central purpose of an analysis is to reach a decision based solely on comparing p to an α, especially when that decision rule has been given explicit justification including both false-acceptance (Type-II) error costs under relevant alternatives and false-rejection (Type-I) error costs (Lakens et al. 2018), as in quality-control applications. Such thoroughly justified applications are, however, uncommon in observational research settings.

It may then seem ironic that many writers who have campaigned vigorously against rigid α-level hypothesis testing promote in its place a version of it in the form of the 95% confidence interval. As with the 5% α-level from which it is derived, the conventional 95% confidence level is divorced from any consideration of error costs, and the resulting interval is typically nothing more than the set of all values b for β for which the test of $\beta = b$ yields $p > 0.05$. It should thus be no surprise that numerous examples (like the ibuprofen study) show that confidence intervals have not provided the hoped-for cure for testing abuse, but have instead perpetuated the dichotomania and excessive certainty that plague research reports. And there is no basis for expecting Bayesian tests or posterior intervals to be treated more wisely.

At best, a 95% confidence interval roughly indicates an entire region of high compatibility between the data and possible parameter values within a given model, as judged for example by having less than 4.3 bits of information against the values inside the interval (Using S-values, a simpler and arguably better range for claiming "high compatibility" would be a 5-bit interval ($\approx$ 97% "confidence".). Thus it would be less misleading to refer to these intervals as compatibility intervals rather than confidence intervals, where "compatibility" only means that the data supply limited information (not even 5 coin-tosses worth) against the parameter values in the interval under the model used to construct the interval. Parallel cautions would apply to posterior probability intervals, with the rephrasing that "compatibility" only means the data supply limited information against the parameter values in the interval under the model *and prior distribution* used to construct the interval.

It should be emphasized that this limitation represents a paucity of information about the parameter in the data and model, rather than decisive evidence in favor of values in the interval or a refutation of values outside the interval. In particular, values not far outside a 95% interval also have limited information against them, and may easily fall inside an interval produced from another, equally plausible model. Such caution is especially important when the model underlying the interval-estimation method (including so-called robust methods) may be incorrect in ways not accounted for by the model (such as unmodeled or mismodeled measurement errors). In such cases (which are the norm in health and social sciences), a 95% coverage or posterior probability description does not adequately incorporate model uncertainty, and thus the interval so described becomes an *overconfidence* interval (A more complete discussion of these issues is given in Greenland 2018.).

Acknowledgements

I thank Valentin Amrhein, Sameera Daniels, Michael Fay, Lawrence McCandless, Mohammad Mansournia, Keith O'Rourke, Philip Stark, David Trafimow, and an anonymous reviewer for helpful comments on earlier drafts of this paper. I also owe a special debt of gratitude to Allen Schirm for exceptionally detailed comments, corrections and suggestions on the original and revised versions.

References

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The Earth is Flat ($p > 0.05$): Significance Thresholds and the Crisis of Unreplicable Research," *Peer J*, 5, e3544. [106,112]

Amrhein, V., Trafimow, D., and Greenland, S. (2018), "Inferential Statistics are Descriptive Statistics," *The American Statistician*, this issue. [108,112]

Bayarri, M. J., and Berger, J. O. (1999), "Quantifying Surprise in the Data and Model Verification," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J.O. Berger, A.P. Dawid, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 53–82. [109,110]

Bayarri, M. J., and Berger, J. O. (2000), "Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142. [110]

Bayarri, M. J., and Berger, J. O. (2004), "The Interplay of Bayesian and Frequentist Analysis," *Statistical Science*, 19, 58–80. [107,110,111]

Benjamini, Y. (2016), "It's Not the P-values' Fault," *The American Statistician*, Online Supplement to ASA Statement on P-values. 70, online supplement 1, available at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/xxxx*. [106]

Berger, J. O., and Sellke, T. M. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–139.

Berger, J. O., and Wolpert, R. L. (1988), "The Likelihood Principle" (with discussion) (2nd ed.), IMS Lecture Notes-Monograph Series, 6, 1–199. [111]

Berger, R. L., and Boos, D. D. (1994), "P Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016. [110]

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trials, Intersection-Union Tests, and Equivalence Confidence Sets," *Statistical Science*, 11, 283–319. [110]

Boos, D. D., and Stefanski, L. A. (2011), "P-Value Precision and Reproducibility," *The American Statistician*, 65, 213–221. [108,109]

Box, G. E. P. (1980), "Sampling and Bayes Inference in Scientific Modeling and Robustness," *Journal of the Royal Statistical Society,* Series A, 143, 383–430. [110,111]

Casella, G., and Berger, R. L. (1987), "Reconciling Bayesian and Frequentist Evidence in the 1-sided Testing Problem" (with discussion), *Journal of the American Statistical Association*, 82, 106–135.

Casella, G., and Berger, R. L. (1987), "Comment," *Statistical Science*, 2, 344–417. [110]

Cohen, J. (1994), "The Earth is Round ($p < 0.05$)," *American Psychology*, 47, 997–1003. [108]

Cox, D. R., and Donnelly, C. A. (2011), *Principle of Applied Statistics*, Cambridge, UK: Cambridge University Press. [107]

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Chapman and Hall. [107,109,111]

Edwards, A. W. F. (1992), *Likelihood* (2nd ed.), Baltimore, MD: Johns Hopkins University Press. [111]

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh, UK: Oliver and Boyd. [107]

Fraundorf, P. (2017), "Examples of Surprisal," available at *http://www.umsl.edu/~fraundorfp/egsurpri.html*. [109]

Gelman, A. (2013), "*P* Values and Statistical Practice," *Epidemiology*, 24, 69–72. [110]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [108]

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [106]

Good, I. J. (1956), "The Surprise Index for the Multivariate Normal Distribution," *The Annals of Mathematical Statistics*, 27, 1130–1135. [109]

Good, I. J. (1983), "Some Logic and History of Hypothesis Testing," in *Philosophical Foundations of Economics*, ed. J. C. Pitt, Dordrecht: D. Reidel, pp. 149–174. Reprinted as Ch. 14 in Good, I.J. (1983), *Good Thinking*, Minneapolis, MN: University of Minnesota Press, pp. 129–148. [109]

Goodman, S. N. (1992), "A Comment on Replication, p-values and Evidence," *Statistics in Medicine*, 11, 875–879. [108]

Goodman, S. N. (1999), "Towards Evidence-Based Medical Statistics, I: The P-value Fallacy," *Annals of Internal Medicine*, 130, 995–1004. [107,109,111]

Greenland, S. (2004), "The Need for Critical Appraisal of Expert Witnesses in Epidemiology and Statistics," *Wake Forest Law Review*, 39, 291–310. [108]

Greenland, S. (2017), "The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. [108,109,111,112]

——— (2018), "The Unconditional Information in P-values, and Its Refutational Interpretation via S-values," manuscript. [109,112]

Greenland, S., and Poole, C. (2013), "Living with Statistics in Observational Research," *Epidemiology (Cambridge, Mass.)*, 24, 73–78. [110]

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.C., Poole, C., Goodman, S.N., and Altman, D.G. (2016), "Statistical Tests, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician*, 70, online supplement 1, available at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf*; reprinted in the *European Journal of Epidemiology*, 31, 337–350. [106,107,110,111,112]

Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006), "Probability as Certainty: Dichotomous Thinking and the Misuse of p-values," *Psychonomic Bulletin & Review*, 13, 1033–1037. [106,112]

Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing," *The American Statistician*, 57, 171–177. [107]

Hubbard, R., and Lindsay, R. M. (2008), "Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing," *Theory & Psychology*, 18, 69–88. [109]

Hurlbert, S. H., and Lombardi, C. M. (2009), "Final Collapse of the Neyman–Pearson Decision Theoretic Framework and Rise of the neoFisherian," *Annales Zoologici Fennici*, 46, 311–349. [106,111,112]

Kuffner, T. A., & Walker, S. G. (2017), "Why Are p-values Controversial?" *The American Statistician, in Press*, 1. [107]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C. L., de Xivry, J.-J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., & Zwaan, R. A.) (2018), "Justify Your Alpha: A Response to 'Redefine Statistical Significance," *Nature Human Behaviour*, 2, 168–171. [108,112]

Lane, D. (1988), "Discussion of Berger and Wolpert," *IMS Lecture Notes-Monograph*, 6, 175–181. [111]

Lang, J. M., Rothman, K. J., and Cann, C. I. (1998), "That Confounded P-value," *Epidemiology (Cambridge, Mass.)*, 9, 7—8. [109]

LeCam, L. (1988), "Discussion of Berger and Wolpert," *IMS Lecture Notes-Monograph*, 6, 182–185. [111]

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, New York: Wiley. [107,108]

Lindeman, M., & Stark, P. B. (2012), "A Gentle Introduction to Risk-limiting Audits," *IEEE Security & Privacy*, 10, 42–49. [110]

MacKay, D. J. C. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge, Cambridge University Press, sec. 2.4, available at *http://www.inference.org.uk/mackay/itila/book.html* [109]

McShane, B. B., and Gal, D. (2017), "Statistical Significance and the Dichotomization of Evidence" (with discussion), *Journal of the American Statistical Association*, 112, 885–908. [106,111]

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2018), "Abandon Statistical Significance," *The American Statistician*, this issue. [106,108,112]

Merriam-Webster Dictionary (2017), "Null," available at *https://www.merriam-webster.com/dictionary/null*. [112]

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008), "P-Values are Random Variables," *The American Statistician*, 62, 242–245. [107,108]

Neyman, J. (1977), "Frequentist Probability and Frequentist Statistics," *Synthese*, 36, 97–131. [108]

Oxford Living Dictionary (2017), "Null," available at *https://en.oxforddictionaries.com/definition/null*.

Perezgonzalez, J. D. (2015), "P-values as Percentiles. Commentary on: 'Null Hypothesis Significance Tests. A Mix-up of two Different Theories: the Basis for Widespread Confusion and Numerous Misinterpretations,'" *Frontiers in Psychology*, 6, 341. [107]

Poole, C. (1987a), "Beyond the Confidence Interval," *American Journal of Public Health*, 77, 195–199. [110]

——— (1987b), "Confidence Intervals Exclude Nothing," *American Journal of Public Health*, 77, 492–493. [112]

Ritov, Y., Bickel, P. J., Gamst, A. C., and Kleijn, B. J. K. (2014), "The Bayesian Analysis of Complex, High-Dimensional Models: Can It Be CODA?" *Statistical Science*, 29, 619–639. [111]

Robins, J. M., and Wasserman, L. (2000), "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts," *Journal of the American Statistical Association*, 95, 1340–1346. [111]

Royall, R. R. (1986), "The Effect of Sample Size on the Meaning of Significance Tests," *The American Statistician*, 40, 313–315. [110]

——— (1997), *Statistical Inference: A Likelihood Paradigm*, New York: Chapman and Hall. [110,111]

Schervish, M. J. (1996), "P-values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206. [110]

Sellke, T. M., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71. [107,109]

Senn, S. J. (2001), "Two Cheers for P-Values," *Journal of Epidemiology and Biostatistics*, 6, 193–204. [106,108,109,110]

——— (2002), "Letter to the Editor re: Goodman 1992," *Statistics in Medicine*, 21, 2437–2444. [106,108,109]

——— (2008), *Statistical Issues in Drug Development* (2nd ed.), New York: Wiley. [110]

Shannon, C.E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423, 623–656. [109]

Spanos, A. (2013), "Who Should Be Afraid of the Jeffreys–Lindley Paradox?" *Philosophy of Science*, 80, 73–93. [110]

Walsh, P., Rothenberg, S. J., and Bang, H. (2018), "Safety of Ibuprofen in Infants Younger than Six Months: A Retrospective Cohort Study," *PLoS One*, 13, e0199493, available at *https://doi.org/10.1371/journal.pone.0199493* [106]

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. [106,112]

Wellek, S. (2010), *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd ed.), New York: Chapman & Hall. [110]

Ziliak, S. T., and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives,* Ann Arbor, MI: University of Michigan Press. [108]

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: 13791

View related articles

View Crossmark data

Citing articles: 49 View citing articles

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]

The American Statistician

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

Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty

Andrew A. Anderson

To cite this article: Andrew A. Anderson (2019) Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty, The American Statistician, 73:sup1, 118-121, DOI: 10.1080/00031305.2018.1537889

To link to this article: https://doi.org/10.1080/00031305.2018.1537889

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 13766

View related articles

View Crossmark data

Citing articles: 13 View citing articles

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, 118–121: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1537889

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty

Andrew A. Anderson

Office of the Comptroller of the Currency, U.S. Department of the Treasury, Washington, DC

ABSTRACT
Evaluating the importance and the strength of empirical evidence requires asking three questions: First, what are the practical implications of the findings? Second, how precise are the estimates? Confidence intervals provide an intuitive way to communicate precision. Although nontechnical audiences often misinterpret confidence intervals (CIs), I argue that the result is less dangerous than the misunderstandings that arise from hypothesis tests. Third, is the model correctly specified? The validity of point estimates and CIs depends on the soundness of the underlying model.

ARTICLE HISTORY
Received January 2018
Revised September 2018

KEYWORDS
Inference; Robustness; Sampling error

1. Introduction

The American Statistical Association statement on p-values (Wasserstein and Lazar 2016), calls for moving beyond "bright line rules" when evaluating empirical results. How? To evaluate a finding requires asking three questions:

1. What are the practical implications of the estimate?
2. Is the estimate precise?
3. Is the model correctly specified?

The first question addresses importance, the second and third pertain to the strength of the evidence. This article discusses each question in nontechnical writing, meant to be accessible for anyone who presents or uses statistical results (e.g., journalists, lawyers, policymakers, and students).

There is no simple rule for evaluating research that encompasses the three questions posed above. Worse, simple rules, such as whether a null hypothesis is rejected, give a false sense of certainty. Proof is only possible within theoretical frameworks such as logic and mathematics. In contrast, empirical research generates evidence. Accumulated evidence from multiple studies may eventually satisfy a "burden of proof," but individual studies are rarely definitive. The inherent uncertainty of statistical evidence means that evaluating empirical work requires judgment, informed by quantitative analysis and subject matter knowledge.

Reviewers of research who do not possess subject-specific expertise must seek it out. Academics can be found via their publications and are usually pleased to discuss their work. Some government agencies provide access to subject matter experts via outreach offices (e.g., the Department of Agriculture and the National Laboratories within the Department of Energy). Experts may also be identified and contacted through outreach groups run by professional associations. Throughout this article, I point out the roles that subject matter expertise plays in the research process, from the crucial first step of model specification to interpreting the magnitudes and precision of estimates.

2. Magnitude

A new fertilizer increases the average height of corn plants by 7 cm. Does it matter? What does it imply for corn yield and profit? Pose these questions to a statistician and you may be left hungry for answers. Judging the importance requires knowledge of corn farming. It requires knowing whether the finding conforms to expectations based on theory and prior evidence, or if the finding is surprising, revealing fertile ground for future research.

Do not mistake statistical significance for practical importance. In informal discussions and the media, the term "significant" is often used without clarifying whether it is meant in a statistical or practical sense. Statistical significance depends on an estimate's magnitude as well as its precision (illustrated in Section 3). Larger samples typically lead to greater precision so even a tiny magnitude can be statistically significant. If you measure enough corn plants, you will eventually obtain a statistically significant difference, even if the fertilizer only increases average height by 1 cm. The insight that statistical significance does not imply practical significance is not new. Deirdre McCloskey has been reminding economists for 35 years (e.g., McCloskey 1983; McCloskey and Ziliak 1996; Ziliak and McCloskey 2004). Unfortunately the term "significant" continues to cause confusion, so the point bears repeating.

To judge magnitude requires paying attention to scale. What are the units of the relevant variables? Does an estimate represent births per month or births per week? The axes on graphs should be clearly labeled, including units. Be careful to distinguish between percent changes and percentage point changes: starting at an unemployment rate of 5%, it makes a big difference whether it falls by 3% (to 4.85%) or by 3 percentage

CONTACT Andrew A. Anderson andrew.anderson@occ.treas.gov Senior Financial Economist, Office of the Comptroller of the Currency, U.S. Department of the Treasury, 400 7th Street SW, Washington, DC 20219.

© 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.

points (to 2%). More generally, when a change is represented in relative terms, consider the base quantity: suppose a new vaccine reduces the incidence of malaria by 50% for a particular population. In absolute terms this could be a huge decrease (if the initial rate were high), but it could also be a small decrease (if the initial rate were low).

In summary, to assess the magnitude of an estimate requires understanding its real-world implications. Be wary of the term "significant." If practical significance is unclear, consult a subject matter expert. Of course, simply attaining practical significance does not ensure a finding is irrefutable. The strength of the evidence depends on sampling and modeling uncertainty.

3. Precision: Quantifying Sampling Uncertainty

A statistic is an approximation to an unknown population parameter based on a random subsample from that population. Statistics generally differ from true population values: the average height of five randomly selected female professional basketball players is unlikely to exactly equal the average height of all female professional basketball players. Given data from the entire population, there would be no sampling uncertainty. Precision is an informal term for how close a statistic is expected to be to the true population value.

Quantifying precision lies at the core of statistics. Popular techniques include credible intervals in the Bayesian framework, and confidence intervals in the frequentist framework. I focus on confidence intervals (CIs) because the target audience for this article is likely to be more familiar with the frequentist approach.

A 95% CI is a range of values to be constructed such that there is a 95% probability that it will contain the true population parameter, assuming the model is correctly specified. Technically, CIs are often misinterpreted (see the Appendix), but people still get the correct idea since a margin-of-error, the half-width of a CI, is an intuitive concept.

3.1. Illustrations

Simple rules, such as the results of null hypothesis significance testing, can cause more serious confusion. I illustrate the use of CIs as well as pitfalls in hypothesis testing using five hypothetical estimates of the change in plant height associated with a corn fertilizer (see Figure 1). Regarding practical importance, suppose a change of 5 cm or more would be notable. Suppose the estimates in Cases 1–3 are considered precise (narrow CIs) while Cases 4 and 5 are not.

In Case 1, the estimate is precise and large enough to be interesting. Regarding statistical significance, consider testing the null hypothesis that the true parameter is zero versus the alternative hypothesis that the true parameter is not equal to zero, at a 5% level. We reject the null if the CI does not include zero. Thus, in Case 1 we reject the null, establishing statistically significant evidence that the parameter is not equal to zero.

In Case 2, we do not reject the null (the CI includes zero). Since the CI is narrow, this outcome is sometimes called a precise zero. Although the result is not statistically significant, it constitutes strong evidence that the parameter

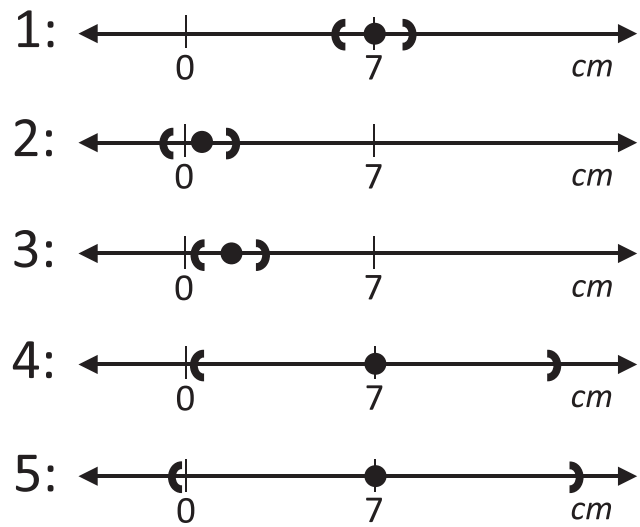


Figure 1. Five estimation outcomes. Dots represent the estimated change in average plant height associated with fertilizer use. Brackets represent the bounds of 95% confidence intervals.

is less than five (the smallest interesting effect size). In terms of magnitude and precision, Cases 2 and 3 provide very similar information despite yielding opposite hypothesis test results. In Case 3, we reject the null even though the estimate is small in magnitude. Here statistical significance does not imply practical significance, like the one centimeter difference in corn height from Section 2.

The estimates in Cases 4 and 5 are significant in practical terms but come with substantial sampling uncertainty. Again, despite providing very similar information, they yield opposite hypothesis test results: we reject the null in Case 4 but not in Case 5. Comparing Case 4 with Case 1, both depict the same point estimate and both are statistically significant, but they differ enormously in precision. Case 1 provides evidence that the parameter is near seven, but in Case 4 the CI extends from just above zero to just under 14.

Regarding Case 5, failure to reject the null is sometimes mistakenly viewed as an acceptance of the null. In other words, finding no statistically significant effect is viewed as evidence that the parameter is zero. In Case 2 that conclusion may be reasonable. But more often, failure to reject simply indicates the results are inconclusive, as in Case 5 where the CI extends from slightly negative to greater than 14. Suppose Case 5 represented an association between a food additive and cancer in laboratory animals. Although we would not reject the hypothesis that there is no relationship, it would be difficult to find a statistician willing to consume the substance.

3.2. Implications

These five cases illustrate how hypothesis test outcomes, in isolation, do not provide all the information required to evaluate a result. Statistical significance is not a proxy for magnitude or precision. CIs provide a useful way to gauge precision while avoiding the potentially confusing terminology of hypothesis testing.

What constitutes sufficient precision? This assessment is tied to the evaluation of practical significance and must also be based

on domain-specific knowledge. One rule-of-thumb is to require the margin-of-error to be less than the smallest magnitude of practical significance. In the corn fertilizer example, an estimate would need a margin-of-error of less than five to be considered precise. The assessment should also account for the fact that it is more difficult to obtain precise estimates in some contexts. So the judgment should consider the level of precision attained in related research. An imprecise estimate may still be informative: suppose the fertilizer were estimated to increase average plant height by 12 cm with a margin-of-error of 6 cm. Then, despite having a wide CI, its lower bound still represents a magnitude of practical importance.

Studies that involve data collection must start by specifying a required level of precision. This specification should then be used to calculate minimum sample sizes. The target level of precision is often stated as a margin-of-error or a level of statistical power. Statistical power is the probability of correctly rejecting a null hypothesis when a specific alternative hypothesis is true. Statistical power is a function of both sample size and the hypothesized parameter values (i.e., effect size). Thus, power analysis requires considering both precision and magnitude. In medical research, grant agencies commonly require sample sizes that yield statistical power of at least 80%.

In summary, statistical significance does not guarantee that an estimate is precise. In contrast, CIs provide an intuitive way to quantify precision. One possible benchmark is to consider an estimate's margin-of-error relative to the smallest parameter value with practical significance. Finally, projects that involve collecting data should start by defining the required level of precision; minimum sample sizes can then be set accordingly.

4. Model Uncertainty

Of the three questions posed in Section 1, whether a model is correctly specified is the most difficult. It is also crucial because point estimates and CIs depend on the validity of the model. Sampling uncertainty is only one component of the overall uncertainty associated with an estimate. I define "model uncertainty" broadly to encompass all other sources of ambiguity. The research process requires making many decisions and the correct choice is often unknown.

Researchers tackle questions like: What explanatory variables should enter the statistical model? What is an appropriate functional form? What are the properties of the error term? Are standard approximations (e.g., asymptotic results for sampling distributions) adequate in a particular research context? Do the observable variables accurately represent the underlying theoretical constructs, for example, are wages an acceptable proxy for worker productivity? What is the proper way to handle data integrity issues such as missing or implausible values? A researcher's answers to these questions are modeling assumptions.

To assess model uncertainty, first identify the modeling assumptions. Second, judge the validity of the assumptions. Third, check how key findings change in response to alternative modeling choices. Let's consider each task in turn.

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. The assumptions most often overlooked may be the formal conditions attached to statistical models. To the benefit of applied researchers, statisticians derive the required conditions for valid estimates and CIs in specific contexts. For example, a method of computing a CI may require that unobserved factors are uncorrelated across observations. Researchers should consult statistical references to ensure they understand the assumptions associated with their models and methods.

Armed with a list of modeling assumptions, the next question is whether they are valid. Since it is rare for every condition to be completely satisfied, asking if a model is *correctly* specified actually means asking whether the model is *adequate* given the research goal. Are any assumptions sufficiently violated to cast doubt on the results? For instance, even after random assignment, the treatment and control groups in an experiment typically differ to some degree in their observable attributes. Are there also differences in unobserved factors? If so, are they substantial enough to confound the results?

To assess the plausibility of modeling assumptions, researchers rely on both theory and evidence. A scatterplot may show that it is reasonable to assume a linear relationship between height and age for elementary school children. Since empirical evidence is not always available, researchers also appeal to theory. Suppose I am estimating the impact of a job training program on wages for high school graduates. Can I simply compare the wages of the graduates who complete the training to the wages of typical high school graduates? In theory, factors such as diligence and motivation influence both earnings and the completion of a job training program. If so, then individuals who finish the training would have earned higher wages even without the program, so the simple difference in average wages would exaggerate the impact of training.

The third step in assessing model uncertainty is to check how estimates change across a range of plausible modeling and data choices. This sensitivity analysis is especially crucial when neither theory nor evidence points to one modeling approach over another. Robustness checks may be conducted using a formal framework such as Bayesian model averaging (Hoeting et al. 1999) or in an ad hoc manner such as using different functional forms and sets of explanatory variables. If a key finding disappears due to a seemingly arbitrary adjustment to the model, then the evidence for the finding is weak.

In summary, the first step when assessing model uncertainty is to identify the modeling assumptions. Assumptions include both formal conditions required by statistical models as well as more judgment-based choices such as sample selection and how to handle data integrity issues. The second step is to assess the validity of the assumptions. Researchers use theory and prior empirical results to support their modeling choices, so subject matter expertise it is especially crucial here. The third step is to assess the degree to which key findings change across model variations.

5. Conclusion

As a method of inquiry, the process of systematic observation, statistical analysis, peer review, and replication is undeniably

effective. Still, the results from individual studies are usually not definitive. Researchers must clearly communicate the sampling and modeling uncertainties associated with their findings. The evaluation of an empirical result should take into account existing evidence on the topic. Is the research community approaching consensus or is there still considerable debate on the issue? When presenting research to audiences who lack subject matter knowledge, it is crucial to provide the context for evaluating the importance and the strength of the evidence.

Appendix: A Common Misinterpretation of Confidence Intervals

Section 3 defines CIs in frequentist statistics. It makes clear that the probability statement applies *before* the sample has been drawn. If a researcher has *already* drawn a random sample, obtained a point estimate and computed its 95% CI, then it is incorrect to say there is a 95% chance the true parameter lies within that specific interval. This is mistaken because the population parameter is not typically considered random, so it has no probability distribution. The realized bounds of a CI are also not random. Before the sample was drawn, the bounds were random variables (functions of a random sample). But after drawing the sample the realized bounds are not random (just as the roll of a die is random, but the result from a particular role is a constant). In short, after drawing the sample, nothing is random so probability statements do not make sense. The population parameter is either inside or outside the realized interval (and since the true parameter value is unknown, we cannot say which is the case).

One way to describe realized bounds is to refer to two implied hypothesis tests. Suppose an estimate has a 95% CI of $[a, b]$. Then, using a two-sided hypothesis test at the 5% level, we would not reject the null hypothesis that the population parameter is equal to a. In a separate test, we would not reject the hypothesis that it is equal to b. In this sense, the data is consistent with true parameter values ranging from a to b.

Acknowledgments

The opinions expressed in this article are those of the author alone, and do not necessarily reflect the views of the Office of the Comptroller of the Currency or the U.S. Department of the Treasury. The author thanks Mike Anderson, Anna Hill, Ron Wasserstein and the referees for helpful comments.

References

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [120]

McCloskey, D. (1983), "The Rhetoric of Economics," *Journal of Economic Literature*, 21, 481–517. [118]

McCloskey, D., and Ziliak, S. (1996), "The Standard Error of Regressions," *Journal of Economic Literature*, 34, 97–114. [118]

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. [118]

Ziliak, S. T., and McCloskey, D. N. (2004), "Size Matters: The Standard Error of Regressions in the American Economic Review," *The Journal of Socio-Economics*, 33, 527–546. [118]

The American Statistician

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

Putting the *P*-Value in its Place

Joachim I. Krueger & Patrick R. Heck

To cite this article: Joachim I. Krueger & Patrick R. Heck (2019) Putting the *P*-Value in its Place, The American Statistician, 73:sup1, 122-128, DOI: 10.1080/00031305.2018.1470033

To link to this article: https://doi.org/10.1080/00031305.2018.1470033

© 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: 11390

View related articles

View Crossmark data

Citing articles: 14 View citing articles

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, 122–128: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1470033

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Putting the *P*-Value in its Place

Joachim I. Krueger[a] and Patrick R. Heck[b]

[a]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, Providence, RI; [b]Geisinger Health System, Danville, PA

ABSTRACT

As the debate over best statistical practices continues in academic journals, conferences, and the blogosphere, working researchers (e.g., psychologists) need to figure out how much time and effort to invest in attending to experts' arguments, how to design their next project, and how to craft a sustainable long-term strategy for data analysis and inference. The present special issue of *The American Statistician* promises help. In this article, we offer a modest proposal for a continued and informed use of the conventional *p*-value without the pitfalls of statistical rituals. Other statistical indices should complement reporting, and extra-statistical (e.g., theoretical) judgments ought to be made with care and clarity.

ARTICLE HISTORY
Received January 2018
Accepted April 2018

KEYWORDS
Bayes' theorem; Inference; Null hypotheses; *p*-values; Statistical significance testing

Dr. Livingstone, I presume!

∼ Stanley to Livingstone, after rejecting the hypothesis that Livingstone was not Livingstone

Walk barefoot on the beach; it gets the toxins out! ∼ A well-meaning neighbor

The most striking pair of facts about significance testing is its immense popularity among researchers, and the intensity of the critical opposition. Since modern sampling techniques and frequentist statistics became *de rigueur* in many fields of scientific study over half a century ago, critics have questioned the wisdom of significance testing in general and the reliance on *p*-values in particular (Gigerenzer 2004; Stigler 1999). Significance tests and *p*-values can be found in many sciences, and thus the context and the nature of the debate vary. In this article, we focus on research typical in experimental psychological science. We discuss statistical inference in the canonical case of the treatment-plus-control experiment designed to test a single, novel, and causal hypothesis with a limited (and often small) sample.

In the prototypical experimental scenario, a researcher who, after having done formal and informal (i.e., intuitive) theoretical work, comes to think that a particular stimulus or treatment might have an effect on participants' experience and behavior. To give a simple example, assume that the researcher thinks that taking a break from creative production increases the eventual total output. In the control condition, respondents generate as many diverse answers to a prompt such as "If all schools were abolished, what would you do to become educated?" as they can without taking a break. In the experimental condition, respondents take a short break during which they engage in a cognitively consuming task before resuming their creative efforts (Ostrowski 2017; see also Gilhooly, Georgiou, and Devery 2012). The statistical question is whether there is a significant difference in creative production between these two conditions.

Many experiments in psychological science take this general form. The properties of the data (e.g., their potential skew, the scale of measurement) help the researcher select an appropriate significance test (e.g., by choosing between parametric and nonparametric tests), compute a test statistic, and ascertain its *p*-value, that is, the probability of the observed test statistic—or any value more extreme than the observed one—assuming that the null hypothesis of no difference is true. Following convention, the researcher rejects the null hypothesis as being a poor fit with the data if $p < 0.05$. The cautious researcher infers that there is some (perhaps limited or preliminary) support for the ordinal and imprecise hypothesis that the treatment has an effect; in our example, taking a break from generating diverse responses increases the eventual total number of responses. In our experience, many researchers choose their words carefully when communicating what they have learned from a significant result. Few insist that they have proven their hypothesis, and many note that the data provide some evidence against the null hypothesis.

Yet, much of the criticism leveled at significance testing, and specifically its most common variant of *null hypothesis significance testing*, NHST, focuses on the limits of many researchers' knowledge of its logic, its proper use, and the meaning of its concepts (Goodman 2008; Greenland et al. 2016). This sort of criticism does not challenge the validity of NHST directly, but it invites the audience to consider alternatives. We prefer a clear distinction between the properties of a method and its (flawed) reception by its practitioners. Yet, while focusing on the former, we hope to educate practicing researchers on the latter.

In the first section of the remainder of this short article, we discuss what we consider the primary contested issue regarding the *p*-values produced by significance testing, which is their

CONTACT Joachim I. Krueger *Joachim@brown.edu* Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, 190 Thayer St., Providence, RI 02912.

Some of the ideas described in this article were presented at a symposium on *p*-values convened by the *American Statistical Association* in Bethesda, MD, October 2017.

© 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.

ability to support reverse inferences about the truth status of the tested hypothesis given the observed data. In the second section, we ask whether significance testing should be replaced *tout court* with Bayesian methods. In the third section, we return to significance testing and discuss its implications for different types of decision error and how these errors are viewed. In the fourth and final section, we explore the psychological and sociological context of the current statistical debate and its implications for the researcher in the lab.

1.1. The Predictive Power of *p*

Chief among the concerns about researchers' ignorance is that they mistake the p-value for the probability of the hypothesis given the data, $p(\mathrm{H}|\mathrm{D})$. An informal—but serviceable—interpretation of the p-value is that it is the probability of the data assuming that the (null) hypothesis is true, $p(\mathrm{D}|\mathrm{H})$ (Wasserstein and Lazar 2016). An answer to the question of whether the p-value is subtly different from $p(\mathrm{D}|\mathrm{H})$ lies beyond this article (Greenland et al. 2016). Many statisticians accept the idea of a close correspondence, while others insist that the differences are deep. Yet, most statisticians agree that to assume without checking that $p(\mathrm{D}|\mathrm{H}) = p(\mathrm{H}|\mathrm{D})$ is to commit a grave reasoning error, namely the fallacy of reverse inference (Krueger 2017). To claim that the null hypothesis is false with a probability of 0.95 because an experiment yielded $p = 0.05$ is to make an irresponsible statement.

The probability of the hypothesis given the data, $p(\mathrm{H}|\mathrm{D})$, is given by Bayes' theorem, which states that $p(H|D) = \frac{p(H)\times p(D|H)}{p(H)\times p(D|H)+p(\sim H)\times p(D|\sim H)}$. The theorem shows that the p-value (or $p(\mathrm{D}|\mathrm{H})$) predicts $p(\mathrm{H}|\mathrm{D})$ but that the prior probability of the hypothesis, $p(\mathrm{H})$, also matters, as does the probability of the data under *any* hypothesis, $p(\mathrm{D})$, which is shown in decomposed form in the denominator. The equality of $p(\mathrm{H}|\mathrm{D}) = p(\mathrm{D}|\mathrm{H})$ may occur, but only for the special case of $p(\mathrm{H}) = p(\mathrm{D})$ (Dawes 1988). Otherwise, the relationship between $p(\mathrm{H}|\mathrm{D})$ and $p(\mathrm{D}|\mathrm{H})$ is imperfect. Using simulation experiments, we mapped the range of possibilities for this relationship (see Krueger and Heck 2017, for details). We can summarize the two main findings thus: first, there is a wide range of positive correlations between the two conditional probabilities; under no circumstance did we find a negative or zero correlation. Second, these correlations are higher inasmuch as the simulated conditions reflect the typical context of empirical work. Specifically, it is typically not the case that $p(\mathrm{H})$ and $p(\mathrm{D}|\mathrm{H})$ are conditionally independent over experiments. If an experiment is very risky, that is, if the probability of the null hypothesis, $p(\mathrm{H})$, is very high, then it is also very unlikely to find a low p-value. The second-order probability of finding a low value for $p(\mathrm{D}|\mathrm{H})$ is low when $p(\mathrm{H})$ is high. As a result, positive correlations between $p(\mathrm{D}|\mathrm{H})$ and $p(\mathrm{H}|\mathrm{D})$ can be high, even in excess of 0.7. In other words, under realistic conditions of the typical experimental research environment, p-values serve as useful heuristic cues for inferences about the posterior probability of the tested hypothesis, $p(\mathrm{H}|\mathrm{D})$.

Our analysis supports the following conclusion: Whereas it is good to warn against equating the p-value with the posterior of the tested hypothesis, it is unwise to suggest that no inference can be made. The absence of a one-to-one correspondence is not the absence of *any* correspondence. Indeed, the application of Bayes' theorem to the logic of NHST reveals that statistical inference is an inference under uncertainty and not a strict computation or logical implication. Hence, we regard the p-value as a useful heuristic cue for estimating the tested hypothesis's posterior probability. In Krueger and Heck (2017), we describe estimation techniques that help researchers identify expected values of $p(\mathrm{H}|\mathrm{D})$ and their associated ranges. We encourage researchers to use p-values for first-pass, heuristic inferences.

1.2. The Bayesian Alternative

The foregoing argument is pragmatic rather than purist. It does not make a clean separation between Bayesian and frequentist approaches. Instead, we treat the observed p-value as input for the estimation of the posterior probability of the tested hypothesis. Absent additional or contravening assumptions, a researcher might favor (if not "accept") a hypothesis if it is more likely than turning up heads in a coin flip ($p(\mathrm{H}|\mathrm{D}) > 0.5$). Similar pragmatic integrations of Bayesian and frequentist operations have been proposed elsewhere (Cohen 1994; Krueger 2001; Nickerson 2000; Trafimow 2003). Much of the current debate, however, is characterized by committed advocacy, which asks researchers to endorse a particular set of principles. The desire for a full conversion of researchers to a particular school of thought is understandable, given that the advocates of these schools take pains to work strictly within their own set of assumptions, and given that mixing approaches can beget confusion and contraction (Gigerenzer 2004; Gigerenzer and Marewski 2015). Yet, in this particular case—estimating $p(\mathrm{H}|\mathrm{D})$ from $p(\mathrm{D}|\mathrm{H})$ using Bayes' theorem—we see little epistemological danger. Instead, we see an opportunity for researchers to better understand the properties of the p-value and its implications for the hypothesis of interest. In the spirit of this special issue on statistical inference, we advise researchers to familiarize themselves with the basic assumptions underlying contemporary schools of statistical thought and to treat available methods as tools in a box, to be used judiciously.

Consider the central assumption about what is perceived to vary and what is perceived to be a fixed condition. Frequentists treat hypotheses as fixed parameters; it may only be the null hypothesis, but it could also be a nonnull or substantive hypothesis, or it could be several hypotheses. Hence, analysis produces probabilities of data—which may vary due to sampling—and which are conditioned on the parameters (i.e., hypotheses). Bayesians, in contrast, treat the data as the conditions given by observation, while treating hypotheses as random variables, or distributed. Hence, Bayesians are interested in posterior probability distributions. If they select two discrete hypotheses, H and ∼H, from these distributions, they can compute odds ratios $p(\mathrm{H}|\mathrm{D})/p(\sim\mathrm{H}|\mathrm{D})$ or ratios of posterior over the prior odds (Ortega and Navarrete 2017). If they consider an entire distribution of hypotheses (i.e., where the hypothesis is a random variable), an infinite number of odds ratios awaits contemplation.

We suspect that many researchers find the Bayesian theory (theory = a way of seeing) of hypotheses as forming a distribution of possibilities counterintuitive—and not only because they have been trained in frequentist statistics. It seems more natural to ask if a *particular* idea or hypothesis is true or false than it

is to consider a set of observations (the data) and represent knowledge and belief as a distribution over an infinite set of hypothesis. The latter does not allow a person to express the strength of belief in any particular idea but only to express how much stronger or weaker the belief is relative to some other particular idea. In contrast, the former (hypothesis-conditioned) way of thinking is aligned with any epistemological framework that begins with core theoretical assumptions, generates testable hypotheses, and ends with experimental data and inferences.

We are inclined—as we foreshadowed in Section 1—to recommend significance testing in small-sample experimentation and amend it with Bayesian inferences to estimate $p(\mathrm{H}|\mathrm{D})$. The p-value, or $p(\mathrm{D}|\mathrm{H})$, is a useful heuristic to infer $p(\mathrm{H}|\mathrm{D})$. The ratio of $p(\mathrm{D}|\mathrm{H})/p(\mathrm{D}|\sim\mathrm{H})$ may be a stronger heuristic, but becomes relevant only if researchers articulate a specific alternative to the null hypothesis. The ratio $p(\mathrm{D}|\mathrm{H})/p(\mathrm{D}|\sim\mathrm{H})$ is a likelihood ratio, LR, when it is computed with the values of the density function. The use of probabilities (areas under the density curve) and use of likelihoods (the value on the y-axis of the density function) makes little difference. Our simulations corroborate what a look at a unimodal distribution reveals, namely that both probabilities and likelihoods become smaller as one moves into the tail of a unimodal distribution. For the normal distribution, for example, the correlation between the log-transformed likelihoods and log-transformed probabilities is nearly perfect (Krueger and Heck 2017).

Now consider the relationship between the p-value (or $p(\mathrm{D}|\mathrm{H})$) and the LR. Suppose the researcher has specified a discrete alternative to the null hypothesis such that $\mathrm{H} \geq \delta = 0$ (i.e., there is no effect between control and treatment on some outcome measure) and $\sim\mathrm{H} \geq \delta = 0.5$ (there is a standardized effect of 0.5 between control and treatment). When more than one experiment has been performed and some variation in the empirical effect size, d, has been observed, the hydraulic relationship between $p(\mathrm{D}|\mathrm{H})$ and $p(\mathrm{D}|\sim\mathrm{H})$ is evident. The p-value now perfectly predicts the LR. If, however, as in our example from experimental psychology, the researcher cannot assign a specific effect size to the alternative hypothesis, and if the alternative (i.e., $\delta \neq 0$) is variable, then $p(\mathrm{D}|\sim\mathrm{H})$ varies, while the observed data and $p(\mathrm{D}|\mathrm{H})$ remain constant. In this scenario, the correlation between p and the LR is not defined. Finally, if there are several experiments, perhaps gleaned from meta-analysis or replication attempts, both $p(\mathrm{D}|\mathrm{H})$ and $p(\mathrm{D}|\sim\mathrm{H})$ are variable, and so is the LR. Here, a positive association between p and LR remains (Krueger and Heck 2018). In other words, the p-value continues to play a useful heuristic role in predicting the LR and $p(\mathrm{H}|\mathrm{D})$.

Our analysis suggests the following conclusion: researchers may wish to supplement p-values with LRs if they can present a rationale for particular alternative hypotheses. In such cases, which remain rare in experimental psychology, they can estimate the posterior probabilities of both hypotheses with greater precision. Either way, it is unwise to discard the p-value entirely as it is generally unwise to ignore the building blocks of composite scores such as ratios or discrepancies (Krueger, Heck, and Asendorpf 2017).

1.3. Errare Humanum Est, Sed Quem Errorem?

One version of the concern that p-values seduce the unwary to infer the posterior probability of the tested hypothesis manifests in warnings about inflated Type I errors (Colquhoun 2015; Ioannides 2005; but see Stroebe 2016). Type I errors are false positives, FP, in decision-theoretic terms (Swets, Dawes, and Monahan 2000). In the context of NHST, FPs refer to mistaken beliefs in something that does not exist. FP aversion directly follows from the worry about the reverse inference fallacy discussed in the first section. If researchers rush to the judgment that $p(\mathrm{H}|\mathrm{D}) < 0.5$ if $p(\mathrm{D}|\mathrm{H}) < 0.05$, they may log many results as discoveries even though $p(\mathrm{H}|\mathrm{D})$ remains greater than 0.5. This is most likely to happen when the work is risky, that is, when $p(\mathrm{H})$ is very high. FP aversion has spawned various recommendations for error control, most of which amount to a call for larger samples. Yet, the primary result of increased sample size is a reduction of Type II errors (i.e., failures to reject false null hypotheses). Although larger samples can raise FP rates, they tend to lower FP *ratios* (when Hits [rejections of false null hypotheses] rise faster than FPs; Krueger and Heck 2017).

An alternative recommendation is to lower the significance threshold. Lowering significance thresholds to $p < 0.005$, for example (Benjamin et al. 2017; or $p < 0.001$, Johnson et al. 2017) will reduce the number of FP errors, but at the cost of increasing the number of Type II errors to an unknown degree (Fiedler, Kutzner, and Krueger 2012). As significance testing becomes more conservative, FP ratios (the proportion of significant findings leading to false inferences) may even increase (Krueger and Heck 2017). We recommend, in the spirit of Neyman and Pearson (1933a, b) that researchers carefully think about the utility they assign to the two types of error before ritualistically endorsing a conventional scheme such as $\delta = 0.05$ and $\delta = 0.2$ (Cohen 1977; 1992; Erdfelder, Faul, and Buchner 1996).

To put these concerns in perspective, consider Meehl's (1978) "strong use" of significance testing as a thought experiment (see also Antonakis 2017). Meehl, being a Popperian falsificationist at the time, asked us to imagine that a specific nonnull hypothesis (which we have referred to as $\sim\mathrm{H}$) is subjected to the ordeal of the significance test. The meanings of the two errors are now reversed. A Type I error (FP) has become the rejection of a true substantive hypothesis, inviting a belief in a false null. In contrast, a Type II error (Miss) is now the failure to reject a false substantive hypothesis, inviting a false belief that the null is false. Consider the implications of these two suggested tactics of error control, increasing power and reducing significance thresholds.

The main consequence of increasing sample size—and thereby power—is the reduction of Type II errors. In the Meehl-Popper strong-inference scenario, a true null hypothesis is more likely to be detected. This prospect might subtly motivate the researcher to limit data collection. From a Popperian perspective, it makes good sense to demand high power. Alternatively, a lowering of the significance threshold to, say, 0.005, will make it harder for researchers to demonstrate invariances (i.e., the absence of an effect). When theory predicts a substantive effect, the conservative researcher may wish to *relax* the criterion of significance, whereas the liberal researcher would tighten it. Again, our recommendation is for researchers to consult relevant theory before deciding whether they can (or should) put a nonnull

hypothesis to the test and how strict this test should be (see also Lakens et al. 2018). If theory or compelling convention does not require unbalanced a priori error rates, we think it prudent to use the same probabilities for both types of error (e.g., $\alpha = \beta = 0.1$; see also Bredenkamp 1980; as well as Neyman and Pearson themselves 1933a, b).

The discourse of error presupposes *decisions* regarding significance and decisions need criteria. The conventional criterion for declaring significance is $p < 0.05$, and much has been made of the presumed rigidity of its application (Greenland 2017; Wasserstein and Lazar 2016). A radical response is to surrender all pretension of decision-making and to limit inference to the task of judgment by, for example, estimating a probable range of values for $p(\mathrm{H}|\mathrm{D})$ given the obtained p-value and auxiliary assumptions (Krueger and Heck 2017).

As experimentalists, we are reluctant to relinquish dichotomous decision-making entirely. The first reason for this traditionalism is that for many questions humans ask of nature, the null (or any particular tested hypothesis) may in fact be true.[1] Nature presents some true-false dichotomies, and it would be unwise to use a decision-making scheme that ignores this. A profusion of pseudo-, para-, and alternative sciences thrive on false claims. They assert the presence of not-nothing when there is nothing (Krueger, Vogrincic-Haselbacher, and Evans, in preparation). The second epigraph of this article may serve as an illustration.

It is easy to see the truth of this claim (this claim itself has a good probability of being true) when considering categorical data. We rightly treat categorical questions to true-or-false responses. In contrast, much of the statistical debate (including this short article) plays out in the world of continuous variables. Here, the claim that the tested hypothesis cannot be literally true is an artifact of allowing each specific point on the scale to be no more than one among an infinity of points. Such a point has no probability, only a likelihood—and as a number, this likelihood is meaningful only when compared with (e.g., divided by) another number.

Returning to our example of the experiment on creativity, the researcher is tasked with assigning respondents randomly to conditions. Comparing the experimental with the control condition, we may wonder if any difference, however small, might eventually yield significance without being an FP. If, however, we compare two control conditions, both created at random and each without a treatment, we would know that $p(\mathrm{H}) = 1$ and that any $p < 0.05$ could only be an FP. We now infer that if the null hypothesis can be true if we ensure that it is, it can also be true when we did not.

The second reason for allowing a judicious use of dichotomous inference is psychological. Categorization is built into perception (Bruner, Goodnow, and Austin 1956). To perceive is to categorize. We cannot see Argos without also seeing a dog, and we cannot see Eumaeos without seeing a man, a Greek, and a swineherd. Categorization affords induction. Recognized as a dog, Argos tells us much about his conspecifics (e.g., love of his master); if we ignored the inductive power of categorical perception, we would be condemned to encounter each dog as a novel creature. Ditto for Eumaeos. Following Bruner and other pioneers, Tajfel (1959; 1969) proposed accentuation theory to formalize the interplay of categorization and perception of stimuli falling on a graded scale. Accentuation theory predicts that if there is a categorical line drawn somewhere on the continuum, stimuli falling to the left (lower) of this line will be seen as having been shifted to the left, while stimuli falling to the right (higher) will be seen as having been shifted to the right. The theory also predicts that this effect is strongest near the line itself, which results in a perceptual narrowing of each category (Krueger and Clement 1994).

Consider the implications of accentuation theory for the perception of p-values. Values < 0.05 will be regarded as small, while values > 0.05 will be regarded as large, with little discrimination among values falling on the same side of the divide. Some may take this perceptual sharpening as evidence of the dangers of dichotomous statistical inference (Gelman and Stern 2006); others will view it as an unavoidable byproduct of categorization. We side with the latter group, thinking that binary statistical inference ultimately stands in the service of action. Neyman and Pearson (1933a, b) explicitly subordinated statistical decision-making to action, and Fisher (1956) did so implicitly—at best. When a fertilizer has been tested, the *agronome* must decide whether to use it. We ask researchers to be mindful of accentuation theory and to consider if and how their work is connected to decisions about action. In this exposition of our findings, we can only mention, but not explore, the differences between inference and decision, and the complexities arising along the road from statistical inference to scientific inference to practical inference. Suffice it to note our agreement with Fisher (1959, p. 100) who wrote that "an important difference is that decisions are final, while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation, but of revision."

1.4. Statistics as a Social Process

The future of statistical practice will not be decided by logical proof or empirical test. Try to imagine what a critical experiment might look like! Would the data be analyzed with significance tests, model-fitting techniques, or subjective Bayesian interpretation? How would you go about managing the two types of error in this experiment? This is a difficult question: if our methods are designed to test the predictions derived from our theories, how might the methods themselves and their parent theories be tested? We submit that the acceptance of a set of methods is largely a matter of social process, rhetoric, and *Zeitgeist*. There surely are innovations, reforms, and refinements of methods that can be strictly justified on logico-mathematical grounds, but such advances are usually made *within* the context of a paradigm or school of thought. A recent example is the introduction of the *p-rep* index, which was meant to replace the conventional p-value (Killeen 2005). Only *p-rep*, but not p, was to reveal the probability of successful replication. It soon turned out that *p-rep* was a mere log-linear transformation of p, and so the index

[1] The argument that any point hypothesis is false (e.g., Cohen 1990; Gelman and Carlin 2017) is a mathematical artifact of the assumption that there is an infinite number of hypotheses (Krueger 2001). It would be more accurate to say that the probability of such a hypothesis to be true is indeterminate; that is, such an hypothesis would be neither true nor false irrespective of the evidence.

went as fast as it had come (Wagenmakers 2007).[2] Other innovations have found traction, however, such as the now widespread use of multi-level regression models (Austin 2017).

When schools of thought whose most central thoughts are incommensurable (Feyerabend 1976; Fleck 1935; Kuhn 1962) vie for the affections of bench scientists, rhetoric and social process become relevant. The result is a cross-purpose debate. In the current climate, Bayesians emphasize the coherence of their methods compared with the ease with which significance tests produce incoherent patterns (e.g., inferences violating the axiom of transitivity). Significance testers, in contrast, emphasize the vulnerability of Bayesian methods to belief bias (Revlin et al. 1980). False priors make bad estimates, unless the data are really big—in which case neither Bayesian nor frequentist methods are needed. Frequentists may turn some of the criticisms leveled at significance testing back on the Bayesians. The dangers of HARKing (Hypothesizing After the Results are Know; Kerr 1998), for example, apply to both schools, as do the dangers of limiting sample size (Krueger and Heck 2017). Other attempts to reform research practice do not affect the analytical methods themselves. Preregistration regimens are intended to protect researchers from one another and themselves (e.g., their implicit biases and *p*-hacking self-deceptions), but are mute on the logic of inductive inference.

When there are no incontestable criteria to settle methodological debates, social processes unfold that might lead to the eventual dominance of one school of thought. These processes are 'soft' in that they cannot leverage irrefutable proof or unquestionable authority. The limits of authority may be found at the level of the journal editor who by way of *Diktat* declares certain methods out of bounds. Reporting of *p-rep* was declared obligatory by the journal *Psychological Science*; its demise was less formal. The editors of *Basic and Applied Psychology* banned *p*-values from their pages (Trafimow and Marks 2015). Professional societies convene task forces to work out recommendations for statistical reporting, but, as part of the social process, the staffing of such task forces must not be partisan (Wasserstein and Lazar 2016; Wilkinson and APA Task Force 1999). As a result, the recommendations tend to be encouragements rather than pre- or proscriptions.

More forceful demands surface when groups of statisticians and researchers band together to issue calls for changes in practice. Recently, Benjamin et al. (2017) proposed a lowering of the conventional significance threshold to 0.005, conceding that this was a patch, and that elegant solutions were being prepared, but that there was no consensus yet. The "et al." of this group were 71 individuals, many of whom are quite distinguished. This is a large number. It dramatizes the claimed consensus that *something* needs to be done. The social psychology of this tactic is to disarm resistance by leveraging the heuristic of "social proof" (Cialdini 1984). If all these statisticians agree—the bench scientist is nudged to infer—then I should shrink my alpha. In a réplique, Lakens recruited 87 co-authors to argue against any privileged alpha level. Reasonable as their arguments are, the use of social proof wears thin when it begins to smack of mimicry and when the ratio of the number of written words to the number of authors becomes distressingly low (here: 22). There ought to be a minimum criterion value for that!

We conclude that the social forces that shape research practice are part and parcel of any evolving science. Instead of wishing them away, we'd do well to understand them, and hope they will move the practice of science forward.

1.5. Conclusion

In 2001, one of us (Krueger 2001) predicted that NHST would outlast many of its challenges. This prediction was partly based on the method's intrinsic value, partly on the beside-the-pointness of some of the critiques, and partly on the fact that by that time, NHST had already shown itself to be resilient. These three reasons are not independent of one another, and there is no reason for complacency. Like others before us (Abelson 1995; Wasserstein and Lazar 2016; Wilkinson and the APA Task Force on Statistical Inference 1999), we wish to impress on research scientists that the *p*-value is a mere heuristic. It has predictive value, but it guarantees that some false inferences will be made. The *p*-value cannot do all of the inductive work; no single method can. We join those who recommend researchers use a toolbox of statistical techniques, employ good judgment, and keep an eye on developments in statistical and data science.

Appendix

To guard against the impression that our appeal to good judgment is mere handwaving and buck-passing, we offer two examples from recent experience and vivid memory. The first example features a trolleyologist[3] explaining the decision to test the significance of 44 successes under the null hypothesis of 25. "I would have done a significance test if I had 50 out of 50 successes," the researcher continued, "because I know the editor would have demanded it." But why perform a significance test to reject the chance hypothesis when the results are so clear that a naked-eye test reveals the effect. To say that a test would have been performed if the result had been 50 out of 50 is to endorse and perpetuate an empty ritual. We advise the use of good judgment to oppose such rituals, and look forward to seeing more articles in which striking results are reported using only descriptive statistics.

The second example involves a correlation between a trait measure of moral orientation and a specific moralistic action. This correlation turned out to be +0.08. With over 900 degrees of freedom, this correlation was "highly significant." When asked whether a correlation of +0.02, if significant in a much larger sample, would be deemed satisfactory as well, the researcher said "yes." What might be done to de-absurdify this approach to data analysis? For this case, we recommend that a plausible substantive hypothesis be selected for strong significance testing in Meehl's sense or for the estimation of a likelihood ratio. To wit, a correlation of +0.3 may be installed to represent ∼H, that is, the baseline expectation of how well general traits predict specific behaviors (Dawes and Smith 1985; Mischel 1968).

ORCID

Patrick R. Heck http://orcid.org/0000-0003-0819-3890

[2] New indices with little or no new informational input occasionally appear in the literature. Bayarri, Benjamin, Berger, and Sellke (2016), for example, suggested that *p*-values be replaced by upper-bound Bayes Factors as $\frac{1}{-ep\log p}$. Here, all the variation is in the *p*-value.

[3] A trolleyologist is someone doing empirical research on philosophical thought experiments such as Philippa Foot's trolley-footbridge dilemma (Foot 1978/1967).

References

Abelson, R. P. (1995), *Statistics as Principled Argument*, Hillsdale, NJ: Erlbaum. [126]

Antonakis, J. (2017), "On Doing Better Science: From Thrill of Discovery to Policy Implications," *The Leadership Quarterly*, 28, 5–21. [124]

Austin, P. C. (2017), "A Tutorial on Multi-Level Survival Analysis: Methods, Models and Applications," *International Statistical Review*, 85, 185–203. [126]

Bayarri, M. J., Benjamin, J., Berger, J. O., and Sellke, T. M. (2016), "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses," *Journal of Mathematical Psychology*, 72, 90–103. [126]

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. (2017), "Redefine Statistical Significance", available at *osf.io/preprints/psyarxiv/mky9j* [124,126]

Bredenkamp, J. (1980), *Theone und Planung psychologischer Experimente* [Theory and Design of Psychological Experiments], Darmstadt, Germany: Stemkopff. [125]

Bruner, J. S., Goodnow, J. J., and Austin, G. A. (1956), *A Study of Thinking*, New York: Wiley. [125]

Cialdini, R. (1984), *Influence: The Psychology of Persuasion*, Boston, MA: Allyn & Bacon. [126]

Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences*, New York: Academic Press. [124]

——— (1990), "Things I Have Learned (So Far)," *American Psychologist*, 45, 1304–1312. [125]

——— (1992), "A Power Primer," *Psychological Bulletin*, 112, 155–159. [124]

——— (1994), "The Earth is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [123]

Colquhoun, D. (2015), "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-Values," *Royal Society Open Science*, 1, 140216. [124]

Dawes, R. M. (1988), *Rational Choice in an Uncertain World*, San Diego, CA: Harcourt, Brace, Jovanovich. [123]

Dawes, R. M., and Smith, T. L. (1985), "Attitude and Opinion Measurement," in *Handbook of Social Psychology* (3rd ed.), eds. G. Lindzey & E. Aronson, New York: Random House, pp. 509–566. [126]

Erdfelder, E., Faul, F., and Buchner, A. (1996), "GPOWER: A General Power Analysis Program," *Behavior Researrch Methods, Instruments & Computers*, 28, 1–11. [124]

Feyerabend, P. K. (1976), *Against Method*, New York: Humanities Press. [126]

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Oxford, UK: Hafner Publishing Co. [125]

Fiedler, K., Kutzner, F., and Krueger, J. I. (2012), "The Long Way from Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate," *Perspectives on Psychological Science*, 7, 661–669. [124]

Fleck, L. (1935), *Entstehung Und Entwicklung Einer Wissenschaftlichen Tatsache. Einführung In Die Lehre Vom Denkstil Und Denkkollektiv*, eds. L. Schäfer and T. Schnelle, Frankfurt, Germany: Benno Schwabe & Co. [126]

Foot, P. (1978/1967), *Virtues and Vices*, Oxford, UK: Blackwell. Originally published in the *Oxford Review*, 5, 1967. [126]

Gelman, A., and Carlin, J. (2017), "Some Natural Solutions to the p-Value Communication Problem—and Why They Won't Work," *Journal of the American Statistical Association*, 112, 899–901. [125]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' is Not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [125]

Gigerenzer, G. (2004), "Mindless Statistics," *The Journal of Socio-Economics*, 33, 587–606. [122,123]

Gigerenzer, G., and Marewski, J. (2015), "Surrogate Science: The Idol of A Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440. [123]

Gilhooly, K. J., Georgiou, G., and Devery, U. (2012), "Incubation and Creativity: Do Something Different," *Thinking & Reasoning*, 19, 137–149. [122]

Goodman, S. (2008), "A Dirty Dozen: Twelve *p*-Value Misconceptions," *Seminars in Hematology*, 45, 135–140. [122]

Greenland, S. (2017), "The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. [125]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, *p*-values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [122,123]

Ioannidis, J. P. A. (2005), "Why Most Published Research Findings are False," *PLoS Medicine*, 2, e124. [124]

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017), "On the Reproducibility of Psychological Science," *Journal of the American Statistical Association*, 112, 1–10. [124]

Kerr, N. L. (1998), "HARKing: Hypothesizing After the Results are Known," *Personality and Social Psychology Review*, 2, 196–217. [126]

Killeen, P. R. (2005), "An Alternative to Null-Hypothesis Significance Tests," *Psychological Science*, 16, 345–353. [125]

Krueger, J. I. (2001), "Null Hypothesis Significance Testing: On the Survival of a Flawed Method," *American Psychologist*, 56, 16–26. [123,125,126]

——— (2017), "Reverse Inference," in *Psychological Science Under Scrutiny: Recent Challenges And Proposed Solutions*, eds. S. O. Lilienfeld & I. D. Waldman, New York: Wiley, pp. 110–124. [123]

Krueger, J., and Clement, R. W. (1994), "Memory-Based Judgments About Multiple Categories: A Revision and Extension of Tajfel's Accentuation Theory," *Journal of Personality and Social Psychology*, 67, 35–47. [125]

Krueger, J. I., and Heck, P. R. (2017), "The Heuristic Value of *p* in Inductive Statistical Inference," *Frontiers in Psychology*, available at *https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00908/full*. [123,124,125,126]

——— (2018), "Testing Significance Testing," [124]

Krueger, J. I., Heck, P. R., and Asendorpf, J. B. (2017), "Self-enhancement: Conceptualization and Assessment," *Collabra: Psychology*, 3, 28. [124]

Krueger, J. I., Vogrincic-Haselbacher, C., and Evans, A. M. (in preparation), "We Need a Credible Theory of Gullibility. To Appear in Forgas, J. P. and Baumeister, R. F. (2019)," in *Homo Credulus: The Social Psychology of Gullibility [The 20th Sydney Symposium on Social Psychology]*, New York: Taylor & Francis. [125]

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press. [126]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018), "Justify Your Alpha," *Nature Human Behavior*, 2, 168–171. [125]

Meehl, P. E. (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, And The Slow Progress of Soft Psychology," *Journal of Consulting and Clinical Psychology*, 46, 806–834. [124]

Mischel, W. (1968), *Personality and Assessment*, Mahwah, NJ: Erlbaum. [124,126]

Neyman, J., and Pearon, E. S. (1933a), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231, 694–706. [124,125]

——— (1933b), "The Testing of Statistical Hypotheses in Relation to Probabilities a Priori," *Mathematical Proceedings of the Cambridge Philosophical Society*, 29, 492–510. [125]

Nickerson, R. S. (2000), "Null hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, 5, 241–301. [123]

Ortega, A., and Navarrete, G. (2017), "Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Signficance Testing (NHST) in Psychology," in *Bayesian Inference*, ed. J. P. Tejedor, Open access, doi:10.5772/intechopen.70230. [123]

Ostrowski, B. (2017), "*Vujà Dé: The Effects of Incubation On Creativity*," unpublished honors thesis in psychology, Providence, RI: Brown University. [122]

Revlin, R., Leirer, V., Yopp, H., and Yopp, R. (1980), "The Belief-Bias Effect in Formal Reasoning: The Influence of Knowledge On Logic," *Memory & Cognition*, 8, 584–592. [126]

Stigler, S. M. (1999), *Statistics on the Table: The History Of Statistical Concepts and Methods*, Cambridge, MA: Harvard University Press. [122]

Stroebe, W. (2016), "Are most Published Social Psychological Findings False?" *Journal of Experimental Social Psychology*, 66, 134–144. [124]

Swets, J. A., Dawes, R. M., and Monahan, J. (2000), "Psychological Science can Improve Diagnostic Decisions," *Psychological Science in the Public Interest*, 1, 1–26. [124]

Tajfel, H. (1959), "Quantitative Judgement in Social Perception," *British Journal of Psychology*, 50, 16–29. [125]

——— (1969), "Cognitive Aspects of Prejudice," *Journal of Social Issues*, 25, 79–97. [125]

Trafimow, D. (2003), "Hypothesis Testing and Theory Evaluation at the Boundaries: Surprising Insights from Bayes's Theorem," *Psychological Review*, 110, 526–535. [123]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [126]

Wagenmakers, E.-J. (2007), "A Practical Solution to the Pervasive Problems of *p*-values," *Psychonomic Bulletin & Review*, 14, 779–804. [126]

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; 21 online expert commentaries available at http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top. [123,125,126]

Wilkinson, L., and the Task Force on Statistical Inference (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist*, 54, 594–604. [126]

The American Statistician

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

Evidence From Marginally Significant *t* Statistics

Valen E. Johnson

To cite this article: Valen E. Johnson (2019) Evidence From Marginally Significant *t* Statistics, The American Statistician, 73:sup1, 129-134, DOI: 10.1080/00031305.2018.1518788

To link to this article: https://doi.org/10.1080/00031305.2018.1518788

© 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: 7371

View related articles

View Crossmark data

Citing articles: 3 View citing articles

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, 129–134: Interpreting and Using p
https://doi.org/10.1080/00031305.2018.1518788

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Evidence From Marginally Significant t Statistics

Valen E. Johnson

Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

This article examines the evidence contained in t statistics that are marginally significant in 5% tests. The bases for evaluating evidence are likelihood ratios and integrated likelihood ratios, computed under a variety of assumptions regarding the alternative hypotheses in null hypothesis significance tests. Likelihood ratios and integrated likelihood ratios provide a useful measure of the evidence in favor of competing hypotheses because they can be interpreted as representing the ratio of the probabilities that each hypothesis assigns to observed data. When they are either very large or very small, they suggest that one hypothesis is much better than the other in predicting observed data. If they are close to 1.0, then both hypotheses provide approximately equally valid explanations for observed data. I find that p-values that are close to 0.05 (i.e., that are "marginally significant") correspond to integrated likelihood ratios that are bounded by approximately 7 in two-sided tests, and by approximately 4 in one-sided tests.

The modest magnitude of integrated likelihood ratios corresponding to p-values close to 0.05 clearly suggests that higher standards of evidence are needed to support claims of novel discoveries and new effects.

ARTICLE HISTORY
Received April 2018
Revised August 2018

KEYWORDS
Bayes factor; Hypothesis test; Posterior odds; Prior odds; Significance level; Uniformly most powerful Bayesian test; Weight of evidence

1. Introduction

In a pair of recent articles (Johnson 2013; Benjamin et al. 2017), my coauthors and I recommended that the threshold for declaring "statistical significance" be changed from 0.05 to 0.005. Criticisms of this proposal have focused on comparisons of Type 1 and Type 2 errors, false negative and false positive rates, and other more sophisticated decision-theoretic-based quantities. There is also a persistent misunderstanding regarding the amount of statistical evidence contained in p-values, and many scientists are unwilling to adjust their interpretation of p-values based on more direct measures of evidence. For example, Lakens et al. 2018 states, "given that the marginal likelihood is sensitive to different choices for the models compared, redefining alpha levels as a function of the Bayes factor is undesirable." Indeed, many nonstatisticians mistakenly interpret p-values as the probability that a null hypothesis is true, and many more are not aware of the relatively arbitrary manner in which the value of 0.05 was chosen to define statistical significance.

In this article, I examine the fundamental question, How much evidence is contained in a t statistic when the p-value is close to 0.05? Ideally, this question would be answered by providing a formula to compute the probability that a null hypothesis is true based on the p-value. That probability is the quantity that scientists are most interested in knowing. Unfortunately, there is no unique mapping from p-values to the probability that a null hypothesis is true, and so this article instead focuses on providing upper bounds on likelihood ratios and integrated likelihood ratios when a p-value of 0.05 is observed. Loosely speaking, a likelihood ratio (LR) represents the ratio of the probability assigned to data under an alternative hypothesis to the probability assigned to data under the null hypothesis. In Bayesian analyses, the LR is directly related to the probability that each hypothesis is true. When the LR is large, the alternative hypothesis provides a better explanation for observed data than the null hypothesis does; when the LR is small, the null provides a better explanation. When the LR is close to 1.0, both hypotheses provide approximately equally valid explanations for observed data. Alternative hypotheses refer to the presence of an effect; the null hypothesis corresponds to no effect. Likelihood ratios can only be computed when all model parameters are completely specified under both hypotheses.

When LR's cannot be computed, integrated likelihood ratios (ILR's) can be computed instead. Like LR's, ILR's reflect the relative probability assigned to the data by alternative and null hypotheses and thus provide a direct measure of evidence regarding the relative validity of two competing hypotheses. The term integrated likelihood ratio (ILR) is used to describe the ratio of marginal densities obtained by integrating out nuisance parameters that define one or both hypotheses. Integrated likelihoods are one of the two main approaches to handling nuisance parameters, the other being maximization (e.g., profile likelihoods). Integrated likelihoods are used in both frequentist and Bayesian settings, and often have desirable properties not possessed by maximization methods (Kalbfleisch and Sprott 1970; Berger, Liseo, and Wolpert 1990). In Bayesian settings, ILRs are called Bayes factors, but due to the data-dependent nature of the alternative hypotheses considered here, resulting ILRs are not consistent with standard Bayesian practice and so the term Bayes factor has been avoided.

CONTACT Valen E. Johnson vjohnson@stat.tamu.edu Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843.
© 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.

Most of the alternative hypotheses examined in this article have been chosen to bias LR's and ILR's in their favor. Similar to earlier analyses in, for example Edwards, Lindman, and Savage (1963), the alternative hypotheses have been chosen to make a p-value of 0.05 look as "significant" as possible. For p-values close to 0.05, I find that LR's and ILR's for two-sided tests are less than 7, and LR's and ILR's for one-sided tests are less than 4. When ILR's are calculated as part of a Bayesian analysis, many statisticians feel that values greater than 10 or 20 are required to provide strong evidence in favor of one hypothesis over another (Jeffreys 1961; Kass and Raftery 1995).

2. One-Sided Tests

To begin, consider one-sided tests of a normal mean. Let $X_1, \ldots, X_n$ denote independent random variables with $N(\mu, \sigma^2)$ distributions. For simplicity, suppose that the null and alternative hypotheses are specified as follows:

$$H_0 : \mu = 0, \qquad H_1 : \mu \sim N(a, g\sigma^2). \tag{1}$$

A normal distribution centered on a with variance g times the observational variance is used to represent the alternative hypothesis. When $g = 0$, the alternative hypothesis becomes a simple hypothesis, that is, a point mass prior centered on a.[1]

The ILR's considered here for composite hypotheses (i.e., $g > 0$) are computed under the assumption that the marginal distribution on the variance parameter σ^2 is proportional to $1/\sigma^2$. This assumption corresponds to an improper, noninformative prior on the variance parameter and is applied to both the null and alternative hypotheses. It also results in certain numerical (but not philosophical) equivalences between standard frequentist and Bayesian analyses. For example, if a noninformative prior density is also imposed on μ, the Bayesian posterior density for μ is a standard t density. Further discussion of noninformative and improper priors on variance parameters can be found in, for example, Berger and Bernardo (1992).

With these assumptions, the marginal density of the data $\mathbf{X} = \{X_1, \ldots, X_n\}$ under the alternative hypothesis, obtained by integrating out μ and the nuisance parameter σ^2, can be expressed as

$$m_1(\mathbf{X}) = \frac{c\,(ng+1)^{-1/2}}{\left[1 + \frac{t_a^2}{(ng+1)(n-1)}\right]^{n/2}}, \tag{2}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad t_a = \frac{\bar{X} - a}{s/\sqrt{n}}, \tag{3}$$

and

$$c = \left[\pi(n-1)s^2\right]^{-n/2}\Gamma\left(\frac{n}{2}\right).$$

The value t_a represents the standard t statistic for testing a hypothesis that $\mu = a$, and s^2 is the usual sample variance.

[1] A simple hypothesis is a hypothesis in which the value of the unknown parameter is completely specified. For composite hypotheses, the value of unknown parameters is only constrained to take values from a specified set, or to be drawn from a specified probability distribution.

The marginal density of the data under the null hypothesis can be obtained by taking $a = 0$ and $g = 0$ in (2), yielding

$$f(\mathbf{X}; a = g = 0) = \frac{c}{\left[1 + \frac{t_0^2}{(n-1)}\right]^{n/2}}. \tag{4}$$

The marginal density of the data under the simple alternative hypothesis $\mu = a$ is similarly obtained by taking $g = 0$ in (2), yielding

$$f(\mathbf{X}; \mu = a, g = 0) = \frac{c}{\left[1 + \frac{t_a^2}{(n-1)}\right]^{n/2}}. \tag{5}$$

For composite alternative hypotheses, it follows that the ILR between the hypotheses specified in (1) can be expressed as

$$\text{ILR} = \frac{m_1(\mathbf{X})}{f(\mathbf{X}; a = g = 0)} = \frac{\left[1 + \frac{t_0^2}{n-1}\right]^{n/2}}{\sqrt{ng+1}\left[1 + \frac{t_a^2}{(ng+1)(n-1)}\right]^{n/2}}. \tag{6}$$

For simple hypotheses, the ILR can be expressed as

$$\text{ILR} = \frac{f(\mathbf{X}; a, g = 0)}{f(\mathbf{X}; a = g = 0)} = \frac{\left[1 + \frac{t_0^2}{n-1}\right]^{n/2}}{\left[1 + \frac{t_a^2}{(n-1)}\right]^{n/2}}. \tag{7}$$

This equation was obtained by integrating out the variance parameter, σ^2, and setting $g = 0$ in (2). Alternatively, (7) can be obtained directly by considering the sampling distribution of the t statistic. Under the null hypothesis, t_0 has a standard t density, while under the alternative hypothesis, t_a has a standard t density. Thus, the ILR defined in (7) can also be regarded from the classical perspective as a simple LR.

2.1. Maximum Integrated Likelihood Ratios

From (2), it follows that the maximum probability that can be assigned to the data under the alternative hypothesis is obtained by taking $a = \bar{X}$ and $g = 0$. For this choice of a and g, the alternative hypothesis becomes a point mass centered on the sample mean, i.e.,

$$H_1 : \mu = \bar{X}. \tag{8}$$

This assumption maximizes the marginal density of the data under the alternative hypothesis. For this choice of alternative, the ILR simplifies to

$$\left[1 + \frac{t_0^2}{(n-1)}\right]^{n/2}. \tag{9}$$

This value represents the maximum value that can be achieved by the ILR for t statistics based on normally distributed data (see Edwards, Lindman, and Savage 1963 for further discussion on maximum likelihood ratios).

In actual scientific practice, sampling variation makes it unlikely that $\bar{X}$ would exactly equal the population mean μ. Nonetheless, Figure 1 depicts the maximum ILR obtained under the alternative hypothesis specified in (8) as a function of the degrees of freedom of the t statistic ν $(= n - 1)$ when t_0 yields a

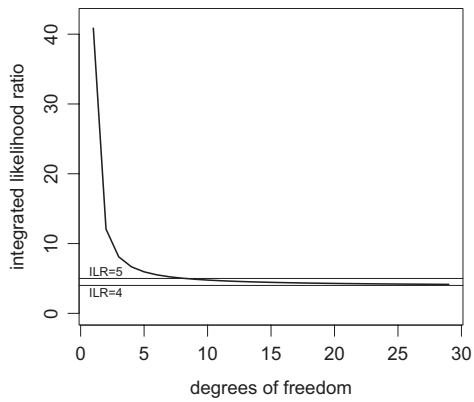


Figure 1. Maximum integrated likelihood ratio for one-sided t-test yielding $p = 0.05$.

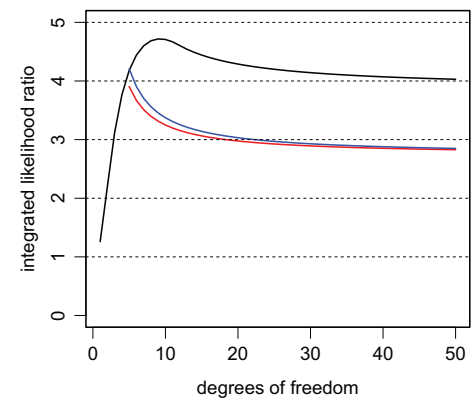


Figure 2. ILRs for one-sided tests. The black curve represents the integrated likelihood ratio for one-sided t-tests yielding $p = 0.05$ under the alternative hypothesis specified in (10). The red curve represents the "average" ILR for a one-sided t-test yielding $p = 0.05$. The red curve was obtained by replacing the marginal density of the t statistic under the alternative hypothesis by its expectation. The blue curve represents the ILR obtained under the alternative hypothesis corresponding to $a = \bar{X}$ and $g = 1/n$ in (1).

p-value of 0.05 (i.e., $t_0 = T^{\nu}_{0.05}$, where T^{ν}_{α} represents the $(1-\alpha)$ quantile of a standard t distribution on ν degrees of freedom). Thus, Figure 1 displays the maximum of the ratio between the marginal probabilities assigned to the data under any alternative hypothesis and the null hypothesis when $t_0 = T^{\nu}_{0.05}$.

From Figure 1, we see that the ILR is less than 5 whenever there are eight or more degrees of freedom. That is, the data is at least 1/5 as likely under the null hypothesis as it is under any alternative hypothesis regarding the value of μ.

For 1 degree of freedom, the ILR can be as high as 40.9. With 1 degree of freedom ($n = 2$), this value is obtained when the t statistic ($\sqrt{n}\bar{X}/s$) is 6.31 and the estimated standardized effect size, $\bar{X}/s$, is 4.46. With 5 degrees of freedom, the maximum ILR of 5.95 is obtained when the estimated standardized effect size is 0.82, and for 12 degrees of freedom the maximum ILR of 4.60 is obtained for an estimated standardized effect size of 0.49.

In many studies in the social sciences, the magnitudes of standardized effect sizes (when present) are often smaller than 1.0. For instance, Cohen (1988) classified standardized effect sizes for differences in means as being "small" when near 0.2, "medium" when near 0.5, and "large" when close to 0.8. Sawilowsky (2009) extended these descriptors to "very large" (1.2) and "huge" (2.0) standardized effect sizes. Large effect sizes are often easy to detect, while very small effect sizes may not be of substantive importance. For this reason, hypothesis tests that attempt to detect small to medium effect sizes typically present the greatest challenge and are often of the most substantive interest. If we modify the alternative hypothesis in (8) to restrict μ to be less than 1/2 of an estimated standardized effect size, then a more realistic alternative hypothesis can be expressed as

$$H_1 : \mu = a = \operatorname{sgn}(\bar{X}) \min\left(|\bar{X}|, \frac{s}{2}\right) \qquad \text{and } g = 0. \tag{10}$$

The black curve in Figure 2 depicts the ILR under this alternative hypothesis. It shows that the maximum constrained ILR occurs at 9 degrees of freedom and is 4.71. For estimated standardized effect sizes known to be less than 0.5 (or medium effect sizes in Cohen's terminology), this figure thus shows that the maximum ILR between the t-statistic under the alternative and null hypotheses is less than 5 whenever $p = 0.05$.

2.2. Accounting for Sampling Variation

2.2.1. Classical Approach

For small to medium estimated standardized effect sizes, the ILRs in the previous section assumed that the true population mean μ under the alternative hypothesis exactly equaled the observed sample mean $\bar{X}$. Based on this assumption, the marginal density of the data under the alternative hypothesis was computed from (2) by taking $a = \bar{X}$ and $g = 0$. Of course, the probability that the sample mean $\bar{X}$ exactly equals the population mean μ is zero.

If, however, the true state of nature was known, then the "true" ILR would be obtained by specifying the alternative hypothesis to be this value. In other words, if the data-generating value of μ was known, we would assume that $a = \mu$ and $g = 0$ in (1). Under this assumption, the ILR would be assigned the value

$$\frac{\left(1 + \frac{t_0^2}{\nu}\right)^{n/2}}{\left(1 + \frac{t_\mu^2}{\nu}\right)^{n/2}}. \tag{11}$$

Unfortunately, the true value of μ is not known, so the quantity in the denominator cannot be computed.

Because we are conditioning on the event $p = 0.05$, we know that $\bar{X} = T^{\nu}_{0.05} s/\sqrt{n}$, or, equivalently, that $t_0 = T^{\nu}_{0.05}$. Under this condition, the numerator in (11) is a fixed and known quantity. However if we ignore the conditioning on the value of $\bar{X}$, then t_μ, evaluated at the true but unknown value of μ, is known to have exactly a t distribution on ν degrees of freedom. This makes it possible to calculate the expected value of

$$\frac{1}{\left[1 + \frac{t_\mu^2}{\nu}\right]^{n/2}}. \tag{12}$$

Simple calculations show that this expectation can be expressed as

$$E_\mu\left[\left(1 + \frac{t_\mu^2}{\nu}\right)^{-n/2}\right] = \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{2n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma(n)}. \tag{13}$$

Thus, even though we don't know the true state of nature and the data-generating value of μ, we can compute the expectation of (12) under this value.

It follows that an approximation to the "average ILR" that would be obtained under the true but unknown μ can be expressed as

$$\text{average ILR} \equiv \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{2n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma(n)}\left[1+\frac{t_0^2}{\nu}\right]^{n/2}. \tag{14}$$

Of course, this expression does not equal the expected value of the ILR because the expectation in (13) ignored the condition that $\bar{X} = T^{\nu}_{0.05}s/\sqrt{n}$. Nonetheless, this expression provides an approximation to the average ILR that would be obtained for the "true" alternative hypothesis.

The red curve in Figure 2 depicts the average ILR for $\nu \in (5, 50)$ and t statistics that yield $p = 0.05$. For medium estimated effect sizes (corresponding to more than 5 degrees of freedom), the average ILR is less than 3 when $p = 0.05$. As before, ILRs greater than 3 can be obtained for $\nu < 5$, but these ILRs correspond to comparatively large and easily detectable standardized effect sizes.

2.2.2. Bayesian Approach

A Bayesian approach can also be taken toward evaluating the uncertainty regarding the true value of μ under the alternative hypothesis. For instance, the alternative hypothesis for μ might be assumed to be normally distributed around $\bar{X}$ with variance σ^2/n (i.e., $a = \bar{X}$ and $g = 1/n$). If the variance was known a priori, this assumption would correspond to specifying the alternative hypothesis to be the posterior distribution on μ given the sample mean $\bar{X}$. It leads to the ILRs displayed by the blue curve (Figure 2). This curve produces ILRs that are very similar to the average ILRs obtained in the previous section. Of course, a genuine Bayesian analysis would not be premised on a prior centered on the sample mean, but the similarity between the average ILR and this pseudo-Bayes factor is revealing.

3. Two-Sided Tests

3.1. Bayesian Approach

From a Bayesian perspective, the conduct of a two-sided test suggests that values of μ above and below the null value are possible, which, in turn, suggests that only alternative hypotheses that are symmetric around the null hypothesis should be considered (Berger and Sellke 1987; Sellke, Bayarri, and Berger 2001). Under this constraint, an alternative hypothesis of the following form approximately maximizes the ILR against the null hypothesis (Berger and Sellke 1987, p.116):

$$H_1 : \mu = \begin{cases} \bar{X} & \text{with probability } \frac{1}{2}, \\ -\bar{X} & \text{with probability } \frac{1}{2}. \end{cases} \tag{15}$$

For this alternative hypothesis, the ILR can be expressed as

$$\left(1+\frac{t_0^2}{\nu}\right)^{n/2}\left[\frac{1}{2}+\frac{1}{2}\left(1+\frac{4t_0^2}{\nu}\right)^{-n/2}\right], \tag{16}$$

where t_0 now refers to $T^{\nu}_{0.025}$.

A plot of this ILR against degrees of freedom ν appears as the green curve in Figure 3. This figure suggests that the maximum ILR for marginally significant t statistics under the constraint of a symmetric alternative hypothesis is less than 5 when there are

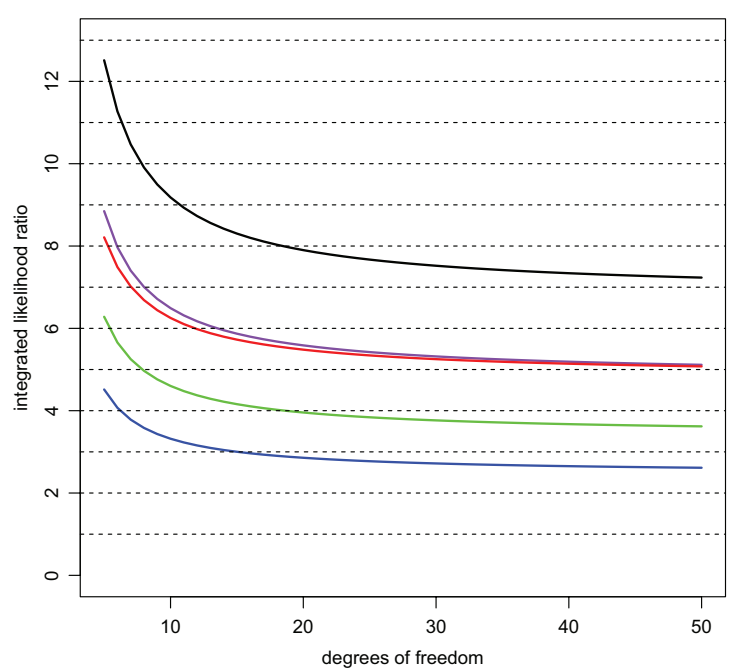


Figure 3. ILRs for two-sided tests. The black curve depicts the maximum ILR for a two-sided t-test yielding $p = 0.05$. The alternative hypothesis underlying this curve assumes that $\mu = T^{\nu}_{0.025}s/\sqrt{n}$, the value of the sample mean that produces a two-sided p-value of 0.05. The green curve represents the ILR for a two-sided t-test yielding $p = 0.05$ obtained by setting $a = \pm\bar{X}$, each with probability one-half, and $g = 0$. The blue curve was obtained similarly, except that $g = 1/n$ to account for variation in the sample mean.The purple curve was obtained by taking $a = \bar{X}$ and $g = 1/n$ in (1). The red curve represents the "average" ILR for two-sided t-tests yielding $p = 0.05$. The marginal likelihood for this curve was obtained by replacing the marginal density of the data under the alternative hypothesis with its expected value at the true value of μ.

8 or more degrees of freedom (small to medium estimated effect sizes).

As in the case of one-sided tests, the alternative hypotheses used to define the ILRs in the Bayesian test can be revised to account for sampling variability in the value of $\bar{X}$. One approach toward accounting for this variability is to assume a symmetric alternative in which 1/2 mass is assigned to two normal densities centered on $\pm\bar{X}$ and variance σ^2/n. This assumption roughly corresponds to taking one-half of the posterior density centered on $\bar{X}$ and re-centering it on $-\bar{X}$. The integrated likelihood ratio that results from this alternative model is

$$\frac{1}{2\sqrt{2}}\left(1+\frac{t_0^2}{\nu}\right)^{n/2}\left[1+\left(1+\frac{2t_0^2}{\nu}\right)^{-n/2}\right]. \tag{17}$$

The blue curve in Figure 3 shows the ILR's that result from this assumption on the prior distribution. The values in this curve approximately mimic the values of the blue curve in Figure 2, which were based on a similar Bayesian analysis of one-sided tests.

If the symmetry constraint on the alternative hypothesis is removed and the alternative hypothesis is instead defined by taking $a = \bar{X}$ and $g = 1/n$, then the ILR can be expressed as

$$\frac{1}{\sqrt{2}}\left(1+\frac{t_0^2}{\nu}\right)^{n/2}. \tag{18}$$

Values of the ILR under this assumption are represented by the purple curve in Figure 3 and are approximately twice the value of the blue curve.

3.2. Classical Approach

Finally, let us examine ILRs for two-sided t-tests that are significant at the 5% level. The maximum bounds in this case are identical to the bounds that would be obtained in a one-sided t-test that yielded $p = 0.025$, and are obtained by assuming that the alternative hypothesis specifies that $a = \bar{X}$ and $g = 0$ in (3). The sample mean is assumed to equal $T^{\nu}_{0.025}s/\sqrt{n}$. The black curve in Figure 3 displays the resulting maximum ILR for 5 or more degrees of freedom, or small to medium estimated standardized effect sizes. Because a two-sided test is performed even though the optimal alternative hypothesis is inherently "one-sided," the ILRs in this scenario are larger than they were in previous scenarios.

As for one-sided tests, the assumption that the true population mean exactly equals the sample mean is unrealistic. If we account for the sampling variation in the sample mean and instead use the expected value of the t density under the assumption that μ is known (as in (13)), then the average ILR can be approximated by the red curve in Figure 3. The values depicted in this curve are approximately twice the values of the blue curve, which were obtained by placing one-half mass each on $a = \pm\bar{X}$ and $g = 1/n$, and are very close to the values in the purple curve obtained by taking $a = \bar{X}$ and $g = 1/n$. As noted previously, the factor of 2 in the former arises from the fact that the alternative split the Bayesian posterior distribution into two, re-centering one-half of the posterior distribution on $-\bar{X}$ in order to maintain a symmetric alternative.

The average ILRs in the case of two-sided tests are between 5 and 8 for small to medium estimated standardized effect sizes and p-values near 0.05.

4. Conclusions

Under a variety of assumptions regarding the values of nonzero effects, ILRs in favor of alternative hypotheses are less than 4 for one-sided t tests based on more than 5 degrees of freedom, and are less than 7 for two-sided tests t tests based on more than 7 degrees of freedom. For alternative hypotheses that are constrained to be symmetric around the null hypotheses, ILRs are less than about 5 or 6 for medium estimated standardized effect sizes, and less than about 3 or 4 for small estimated effect sizes in two-sided tests.

This range of ILR values is less conservative than the Bayesian analyses of p-values and Bayes factors presented in Sellke, Bayarri, and Berger (2001), which required alternative hypotheses to be symmetric—and in many cases unimodal—around the null value. That is, Sellke and coauthors estimated ILRs that were even smaller than those exposed here.

The difference in evidence reflected by one-sided and two-sided bounds on ILRs illustrate the importance of properly specifying alternative hypotheses. Indeed, it is quite possible that many journals and regulators implicitly impose significance thresholds of $p < 0.025$ by requiring that two-sided tests be conducted for alternative hypotheses that are inherently one-sided. Of course, this higher standard for declaring statistical significance is only effective when the sign of an effect is known a priori. It offers no additional protection against HARKing (hypothesizing after results are known; Kerr 1998) when the sign of an effect is not specified before data are analyzed.

In my opinion, the best estimate of the evidence provided by t statistics is provided by the average ILR, which is obtained by replacing the marginal density of data under the alternative hypothesis by its (unconditional) expectation. The expectation of the marginal density of the t statistic under the alternative is free of additional assumptions and represents the exact expectation of a t density at the true value of the population mean μ. It is thus insensitive to prior model choices and other modeling assumptions.

For t statistics based on more than 6 degrees of freedom, the average ILR for two-sided tests is less than 6. For one-sided tests with p-values around 0.05, the average IRL is less than about 3. In other words, the data are, on average, only three or six times more likely under the "true" model than they are under the null hypothesis. Importantly, these values are independent of prior assumptions regarding the value of the population mean under the alternative hypothesis, and apply for all hypothesis tests based on t statistics. They clearly suggest that higher standards of evidence are needed to support claims of novel discoveries and new effects.

Acknowledgments

The author thanks an anonymous associate editor for numerous comments that improved this article.

Funding

Financial support was provided by NIH award R01 CA158113.

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., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T.–H., Hoijtink, H., Jones, J. H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E., (2017), "Redefine Statistical Significance," Nature Human Behaviour, available at *https://www.nature.com/articles/s41562-017-0189-z*. [129]

Berger, J. O., and Bernardo, J. M., (1992), "On the Development of Reference Priors" (with discussion), in *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, New York: Oxford University Press, pp. 35–60. [130]

Berger, J. O., Liseo, B., and Wolpert, R. L., (1990), "Integrated Likelihood Methods for Eliminating Nuisance Parameters," *Statistical Science*, 14, 1–28. [129]

Berger, J. O., and Sellke, T., (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [132]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, New York: Routledge. [131]

Edwards, W., Lindman, H., and Savage, L., (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review,* 70, 193–242. [130]

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.) Oxford, UK: Oxford University Press. [130]

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences,* 110, 19313–19317. [129]

Kalbfleisch, J., and Sprott, D. A. (1970), "Application of Likelihood Methods to Models Involving Large Numbers of Parameters," *Journal of the Royal Statistical Society*, Series B, 32, 175–208. [129]

Kass, R., and Raftery, A. E., (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [130]

Kerr, N. (1998), "HARKing: Hypothesizing after the Results are Known," *Personality and Social Psychology Review*, 2, 196–217. [133]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M, Caldwell, A. R., Calster, B. V., Carlsson, R., Chen, S-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C, Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., Harmelen, A., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A, Konat, B., Lao, J., Larsen, E. G., Lodder, G., Lukavsky, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., Oliveira, C. L., Xivry, J., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A, Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Swiatkowski, W., Vadillo, M. A., Van Assen, M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., Zwaan, R. A., (2018), "Justify your Alpha", *Nature Human Behaviour*, 2, 168–171. [129]

Sawilowsky, S. (2009), "New Effect Size Rules of Thumb," *Journal of Modern Applied Statistical Methods*, 8, 467–474. [131]

Sellke, T., Bayarri, M. J., and Berger, J. O., (2001), "Calibration of p Values for Testing Precise Hypotheses," *The American Statistician*, 55, 62–71. [132,133]

The American Statistician

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

The *p*-value Function and Statistical Inference

D. A. S. Fraser

To cite this article: D. A. S. Fraser (2019) The *p*-value Function and Statistical Inference, The American Statistician, 73:sup1, 135-147, DOI: 10.1080/00031305.2018.1556735

To link to this article: https://doi.org/10.1080/00031305.2018.1556735

© 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: 11841

View related articles

View Crossmark data

Citing articles: 6 View citing articles

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, 135–147: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1556735

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The *p*-value Function and Statistical Inference

D. A. S. Fraser

Department of Statistical Sciences, University of Toronto, Toronto, Canada

ABSTRACT

This article has two objectives. The first and narrower is to formalize the *p*-value function, which records all possible *p*-values, each corresponding to a value for whatever the scalar parameter of interest is for the problem at hand, and to show how this *p*-value function directly provides full inference information for any corresponding user or scientist. The *p*-value function provides familiar inference objects: significance levels, confidence intervals, critical values for fixed-level tests, and the power function at all values of the parameter of interest. It thus gives an immediate accurate and visual summary of inference information for the parameter of interest. We show that the *p*-value function of the key scalar interest parameter records the statistical position of the observed data relative to that parameter, and we then describe an accurate approximation to that *p*-value function which is readily constructed.

ARTICLE HISTORY
Received March 2018
Revised November 2018

KEYWORDS
Accept–Reject; Ancillarity; Box–Cox; Conditioning; Decision or judgment; Discrete data; Extreme value model; Fieller–Creasy; Gamma mean; Percentile position; Power function; Statistical position

1. Introduction

The term *p*-value appears everywhere, in journals of science, medicine, the humanities, and is often followed by decisions at the 5% level; but the term also appears in newspapers and everyday discussion, citing accuracy at the complementary 19 times out of 20. Why 5%, and why "decisions"? Is this a considered process from the statistics discipline, the indicated adjudicator for validity in the fields of statistics and data science, or is it temporary?

The *p*-value was introduced by Fisher (1922) to give some formality to the analysis of data collected in his scientific investigations; he was a highly recognized geneticist and a rising mathematician and statistician, and was closely associated with the intellectual culture of the time. His *p*-value used a measure of departure from what would be expected under a hypothesis, and was then defined as the probability p under that hypothesis of as great or greater departure than that observed. This was later modified by Neyman and Pearson (1933) to a procedure whereby an observed departure with p less than 5% (or other small value) would lead to rejection of the hypothesis, with acceptance otherwise.

Quite early, Sterling (1959) discussed the use of this modification for journal article acceptance, and documented serious risks. And the philosopher and logician, Rozeboom (1960) mentioned "(t)he fallacy of the null-hypothesis significance test" or NHST, and cited an epigram from philosophy that the "accept–reject" paradigm was the "glory of science and the scandal of philosophy;" in other words the glory of statistics and the scandal of logic and reason! These criticisms can not easily be ignored, and general concerns have continued, leading to the ASA's statement on *p*-values and statistical significance (Wasserstein and Lazar 2016), and to further discussion as in this journal issue.

Many of the criticisms of *p*-values trace their origins to the use of bright-line thresholds like the 5% that scientists often rely on, as if such reliance had been universally endorsed by statisticians despite statisticians' own reservations about such thresholds. We assert that many of the problems arise because statements of "significance" are based on supposed rules, that are abstract and too-little grounded in the applied context. Accordingly, we suggest that many of these problems can be reduced or circumvented entirely if statisticians report the entire set of possible *p*-values—the *p*-value function—instead of a single *p*-value. In the spirit of the early Statistical Society of London motto *Aliis exterendum*—statisticians should refrain from prescribing a rule, and leave the judgment of scientific significance to subject-matter experts.

The next four sections of this article are structured to illustrate the main aspects of the *p*-value function for inference. Throughout, we rely mainly on examples to illustrate the essential ideas, leaving many of the details and the underlying mathematical theory to cited references as needed.

Section 2 defines and illustrates the *p*-value function for a paradigmatic simple situation, the Normal location problem. And in this we show how the function can be used to find all the familiar quantities such as confidence limits, critical values, and power. We also emphasize that the *p*-value function in itself provides an immediate inference summary.

Of course real problems from applied science rarely fit easily into the simple formats of textbook examples. Fortunately, however, many real problems can be understood in a way that puts the focus on a single scalar interest parameter.

CONTACT D. A. S. Fraser dasfraser@gmail.com Department of Statistical Sciences, University of Toronto, 100 George Street, Toronto, M5S 3G3 Canada
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.
© 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.

Section 3 presents three scientific problems that are reduced in this way. In applications with complex models and many parameters, available theory outlined in Fraser (2017) provides both an identification of a derived variable that measures the interest parameter θ and a very accurate approximation to its distribution. This leads directly to the p-value function $p(\theta)$ that describes statistically where the observed data are relative to each possible value of the scalar interest parameter θ, and thus fully records the percentile or statistical position of the observed data relative to values of the interest parameter. From a different viewpoint, the p-value function can be seen as defining a right-tailed confidence distribution function for θ, and as the power function (conditional) for any particular θ_0 and any chosen test size.

With the focus narrowed to a parameter of interest, it remains necessary to reduce the comparison data for inference analysis to a set appropriate to the full parameter and then to that for the interest parameter. If there is a scalar sufficient statistic for a scalar interest parameter, as in the Normal location problem, the reduction is straightforward, but few problems are so simple. Fisher (1925) introduced the concept of an ancillary statistic, and suggested conditioning on the value of that ancillary. An important feature of the present approach to inference with the p-value function is that, for a broad range of problems, there is a uniquely determined conditioning that leads to an exponential model for the full parameter, and then by Laplace marginalization to a third-order accurate model for an interest parameter; this then provides a well-determined p-value function. And when sufficiency is available this is equivalent to inference based on the sufficient statistic. Section 4 uses an extremely short-tailed model to provide a brief introduction to continuity conditioning. This conditioning has historical connections with Fisher's concept of ancillarity conditioning, but the present development is distinct from that ancillarity approach.

Section 5 outlines the statistical geometry that provides the third-order accurate approximation for the p-value function in a wide range of problems. The approximation is essentially uniquely determined with inference error of order $O(n^{-3/2})$. Examples also illustrate how conditioning leads to the relevant p-value function even when there is no traditional sufficient statistic.

The article concludes with applications (Section 6) and discussion (Section 7).

2. Likelihood and the *p*-value Function

The Normal location problem, though simple, can often serve as a paradigm, an idealized version of much more complicated applied problems. Suppose that the applied context has determined a scalar variable y that measures some important characteristic θ of an investigation; and to be even simpler suppose y is Normal with mean θ and known standard deviation σ_0, say equal to 1 for transparency.

Example 1. Normal location

For simplicity we assume that $y = (y_1, \ldots, y_n)$ is a sample from a Normal with unit variance and unknown mean θ. The sample mean $\bar{y}$ is also Normal so that for all values of θ the pivot $z = n^{1/2}(\bar{y} - \theta))$ is standard Normal; and for further simplicity suppose that $n = 1$ and the observed $y = y^{\text{obs}} = 10$; and that our theory seeks an assessment or testing of values for θ. In Figure 1(a), we record a typical density of the model and also the observed data value. In Figure 1(b), we record the amount of probability at the observed data point under some θ value and do so as a proportion of the maximum possible; this is designated $L(\theta)$ and is called the observed likelihood function; we also record $p(\theta)$ which is the probability to the left of the data point y^{obs}. Here, $\hat{\theta}^{\text{obs}}$ is the value of the parameter that puts maximum probability density at the observed data point y^{obs}; we also indicate by the dotted curve the corresponding density function $f(y; \hat{\theta}^{\text{obs}})$. This maximum likelihood density function happens to be of central interest for bootstrap calculations.

Now for general values of θ, we can record two items of information. The first is the normalized likelihood function

$$L(\theta) = \exp\{\ell(\theta)\} = \frac{f(y^{\text{obs}}; \theta)}{f(y^{\text{obs}}; \hat{\theta}^{\text{obs}})} = cf^{\text{obs}}(\theta).$$

The numerator is the observed value of the density function when the parameter is θ and tells us how much probability lies

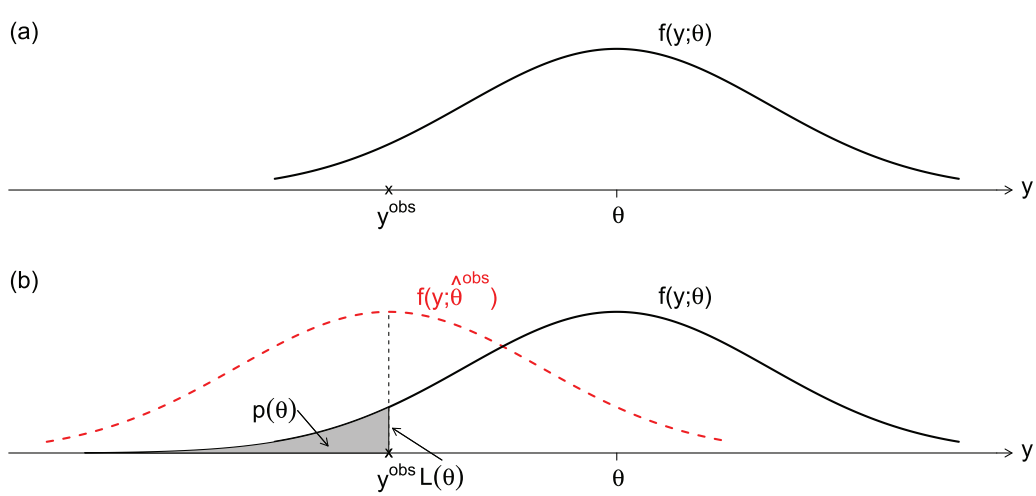


Figure 1. The upper graph presents the model at some typical θ value and the observed data value y^{obs}; the lower graph records in addition the p-value and the likelihood for that θ value, and also in dots the density using the maximum likelihood $\theta = \hat{\theta}^{\text{obs}}$ value for the parameter.

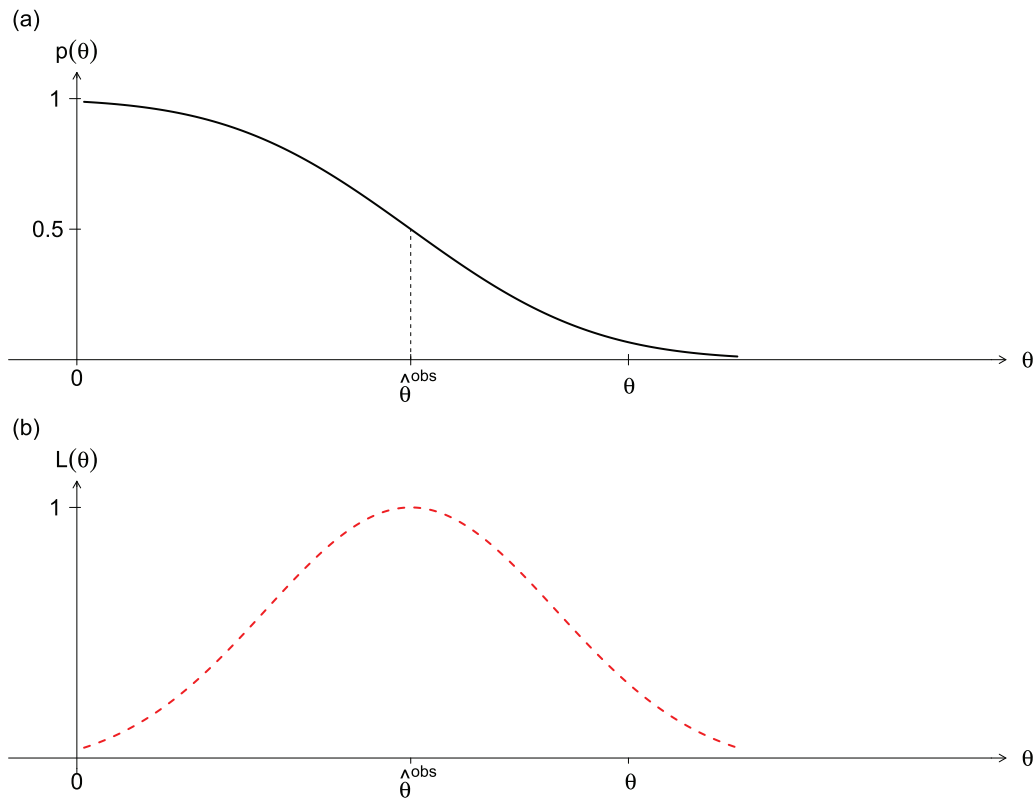


Figure 2. The upper graph presents the p-value function and the lower graph the likelihood function, for the simple Normal example; the median estimate of θ is $\hat{\theta}_{0.50}$ which here is the maximum likelihood value $\hat{\theta}^{\text{obs}} = 10$.

on that observed data point under θ; this is usually presented as a fraction of the maximum possible density at y^{obs}, obtained with $\theta = \hat{\theta}^{\text{obs}}$. For the simple example, see Figures 1(b) and 2.

The second is the p-value function, which is less familiar than the likelihood but its value at any given parameter value is familiar as an ordinary p-value. For given data y^{obs}, the p-value function records the probability left of the data as a function of the parameter θ

$$p(\theta) = F(y^{\text{obs}}; \theta) = F^{\text{obs}}(\theta).$$

See again Figure 2 for this simple Normal example. The p-value function simply records the usual left-sided p-value as a function of a general tested value θ. Thus, the p-value function is the observed value of the distribution function under θ: $p(\theta) = F(y^{\text{obs}}; \theta) = F^{\text{obs}}(\theta)$. As such it records the percentile position of the data for any choice of scalar parameter value θ. Because this function is nonincreasing and $0 \leq p(\theta) \leq 1$, we can regard $1 - p(\theta)$, formally at least, as a cumulative distribution function for θ. The resulting distribution is in the spirit of Fisher's fiducial distribution. As shown in Fraser (2017) and summarized in Section 5, this function is well defined and accurately approximated to the third order, for a wide range of applied problems with vector-valued y and scalar parameter of interest.

These items are what the data and model give us as the primary inference material; they are recorded for our present example in Figure 2(a,b). We recommend the use of these two key inference functions as the inference information.

For Example 1, consider some applications of the p-value function:

(1) With an observed value $y = y^{\text{obs}}$ and for any hypothesized null value θ_0, the value of $p(\theta)$ at θ_0 is the one-sided p-value, and records the statistical position of the observed data.
(2) The endpoint of a left-sided 97.5% confidence interval with data $y^{\text{obs}} = 10$ is the value $\hat{\theta}_{0.975} = p^{-1}(0.975)$ for which the p-value is 97.5%, that is, $Pr\{y \leq 10; \hat{\theta}_{0.975}\} = 0.975$. As seen in Figure 3(a), take the value 0.975 on the vertical $p(\theta)$ axis, run a line horizontally to the p-value curve, and then down to the θ axis; this gives the value $\hat{\theta}_{0.975} = p^{-1}(0.975) = 8.04$, which is the lower 97.5% confidence bound, the value that in repetitions would be less than the true value 97.5% of the time. In the same way, $\hat{\theta}_{0.025} = 11.96$ is the endpoint of a left-sided 2.5% interval. Taken together the two percentiles give a 95% two-sided interval $(\hat{\theta}_{0.975}, \hat{\theta}_{0.025}) = (8.04, 11.96)$. We rely here on confidence intervals as derived by inverting an essential pivot, as distinct from intervals defined by shortest interval with given overall coverage probability. Thus, the 95% interval is the set of parameter values not rejected by either of the 2.5% single-tailed tests. For more on the two different approaches to intervals and the role of conditional inference, together with examples, see Fraser, Reid, and Lin (2018).
(3) Similarly, for the 50% point for which $p(\theta) = 0.50$, $\hat{\theta}_{0.50}$ is the median estimate for θ. First, locate 0.5 on the vertical axis and across to the curve. Then, drop down to the horizontal axis to get the value $\hat{\theta}_{0.50} = 10$.
(4) For any given test size α, the inverse function $p^{-1}(\alpha)$ gives the corresponding parameter value say θ_0 that is being tested with the observed data value as critical point. Although the view taken here argues against fixed-level testing, there

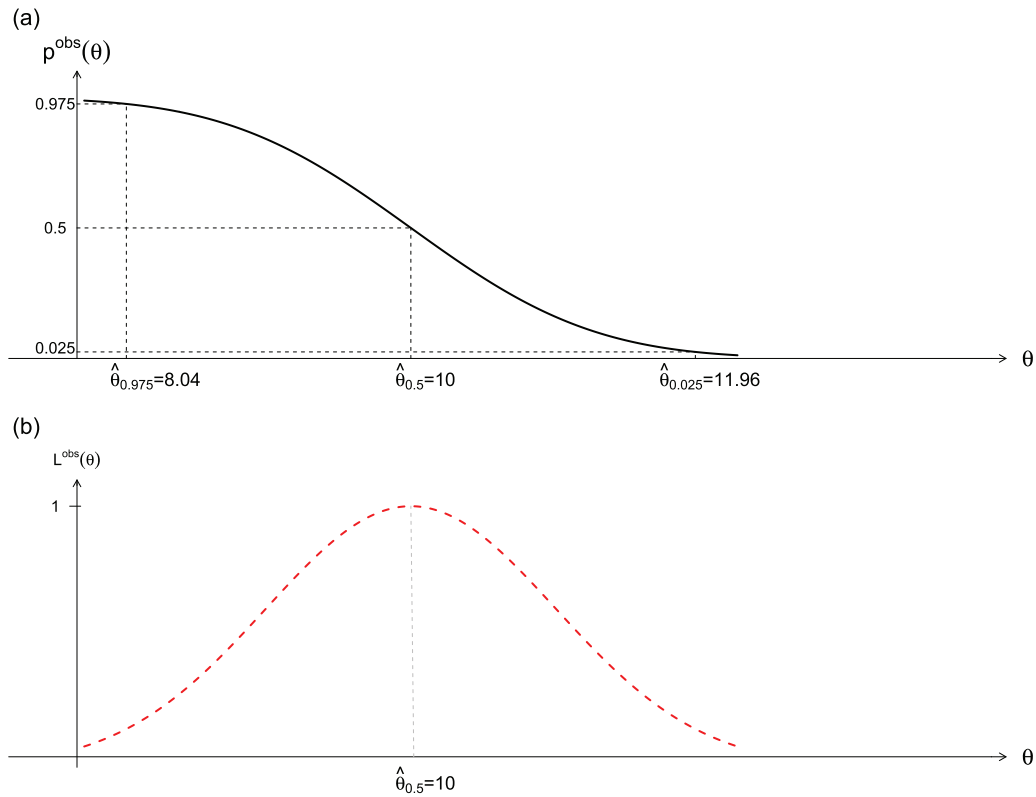


Figure 3. The upper graph (a) indicates the median estimate and the one-sided 0.975 and 0.025 confidence bounds; the lower graph (b) records the observed likelihood function.

are occasions where the notions of size and power can be convenient. To find the value θ_0 being tested at level α relative to the observed data as critical point locate the particular value of α on the vertical axis, then read across to the p-value curve, and then down to the θ_0 for testing.

(5) For the size α test with the data point as critical value, the p-value function gives the power of that test at any other parameter value θ; this is conditional power given model information identified from the data, but as such it is also marginal power: the power at θ is $p(\theta)$. First be aware of the critical value for the test of size α as given in (4). Then for any other value of θ the probability that the α-level test rejects the null hypothesis is $p = \Pr\{y \leq y^{\text{obs}}; \theta\} = p(\theta)$, which is, the value of the p-value function at θ. Perhaps more importantly the p-value function describe the sensitivity of the procedure in distinguishing parameter values in the broad range around the maximum likelihood value.

(6) In addition, if for some value on the θ axis, say $\theta = 8$, we found that the corresponding p-value was large or very large we would infer that the true value was large or much larger than the selected value 8. And similarly if the p-value was small or very small, say at $\theta = 12$ we would infer that the true value was small or much smaller than the selected 12. Thus, the observed p-value function not only shows extremes among θ values but also the direction in which such extremes occur; it thus provides inference information widely for different users who might have concern for different directions of departure.

(7) Based on the properties above, it follows that the location and shape of the p-value function provides a clear way to assess the evidence about the parameter θ. Figure 4 shows three p-value functions corresponding to $n = 1, n = 4$, and $n = 100$.

(8) The p-value function is central to a widely available approach to inference based on approximate conditioning, as discussed in Section 4 and 5. When a complete sufficient statistic is available, the present approach coincides with the usual inferences based on sufficiency. When a reduction by sufficiency is not available as is all too frequently the case, the present approach is widely available and gives an accurate approximation to a well defined p-value function; and as illustrated in Section 5 is immediately available. The theory uses conditioning derived from model continuity. The Normal location problem of this section was chosen as extremely simple, partly to keep technical details to a minimum, and partly to emphasize that the general case is as simple but with slightly less familiar computation. The next three sections show how the ideas of the p-value function for a single, scalar parameters are available quite generally for models with multi-dimensional parameters that have no reduction via a sufficient statistic. And as the next section shows it is often the case that for problems of genuine scientific interest, it is reasonable to focus attention on a single parameter of interest and the present p-value function.

3. Narrowing the Focus to the Essential Parameter: Three Illustrations From Science

The examples in this section are here to illustrate three instances where scientific problems of considerable import lend them-

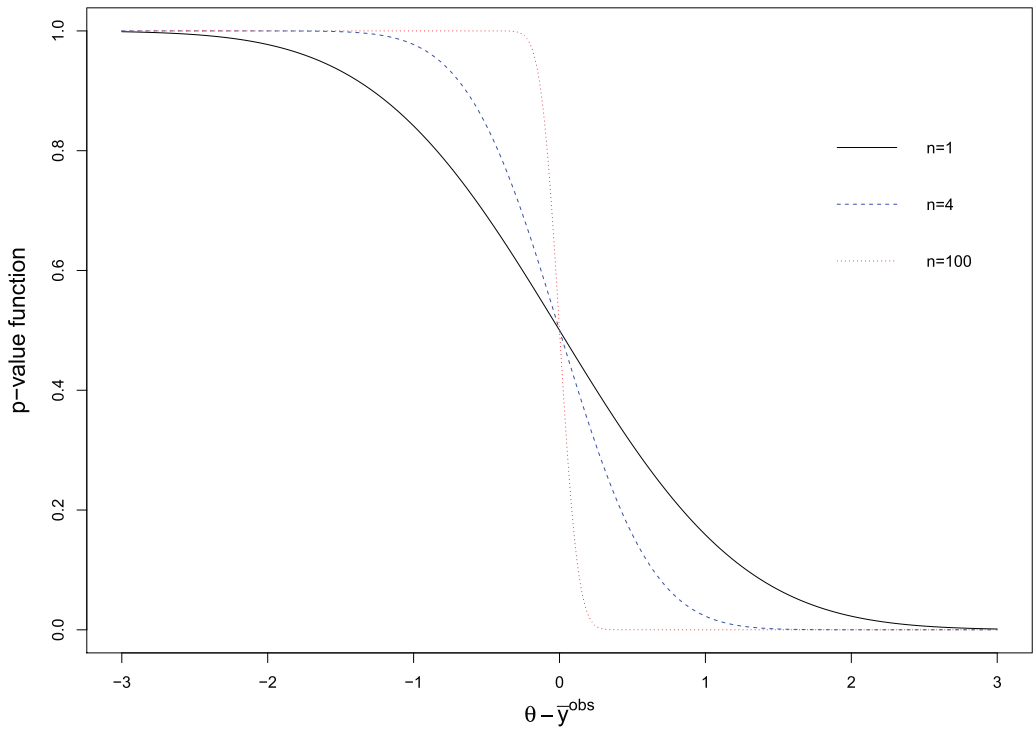


Figure 4. The p-value function for n=1 (solid), n=4 (dashed), n=100 (dotted).

selves to an analysis that puts focus on a scalar parameter of interest and a related scalar variable. These examples discuss such a parameter of interest, but we wait until later sections to address the role of nuisance parameters and the reduction of data observations to a single dimension.

3.1. Trajectory of Light

A theory being floated in the early part of the last century indicated that the trajectory of light would be altered in the presence of a large mass like the sun. An opportunity arose in 1919 when the path of an anticipated total eclipse was expected to transit portions of Africa. During the eclipse an expedition was able to measure the displacement of light from a star in the Hyades cluster whose light would pass very close to the sun, and whose position relative to nearby visible stars was then seen to be displaced, by an amount indicated by the theory. This was an observational study but in many details duplicated aspects of an experiment. In this example, the intrinsic variable relevant to the theory was the displacement on the celestial sphere of the prominent star relative to others whose light trajectory was at some greater distance from the sun, and was calculated away from the sun.

3.2. Search for the Higgs Boson

In 1994, Abe (1994) and coauthors reported on the search for the top quark. High energy physicists were using the collider detector at Fermilab to see if a new particle was generated in certain particle collisions. Particle counts were made in time intervals with and without the collision conditions; in its simplest form this involves a Poisson variable y and parameter θ, and whether the Poisson parameter is shifted to the right from a background radiation level θ_0, thus increased under the experimental intervention. This is an experimental study and such typically provide more support than an observational study.

The statistical aspects of the problem led to related research, for example, Fraser, Reid, and Wong (2004), and to a collaborative workshop with statisticians and high energy physicists (Banff International Research Station 2006). This in turn led to an extensive simulation experiment launched by the high energy physicists, in preparation for the CERN investigation for the Higgs boson. The p-value function approach to this problem is detailed in Davison and Sartori (2008), with the measure of departure recommended in Fraser, Reid, and Wong (2004), and shown to give excellent coverage for the resulting confidence intervals. The p-value reached in the CERN investigation was 1 in 3.5 million, somewhat different from the widely used 5%.

3.3. Human–Computer Interaction

A researcher in human–computer interaction (HCI) is investigating improvements to the interface for a particular popular computer software package. The intrinsic variable could, for example, be the reduction in learning time under the new interface, or a qualitative evaluation of the ease of learning. In this experimental investigation, advice is sought on the appropriate model and appropriate approximation for the distribution of the intrinsic variable.

3.4. Overview

In each of these examples, the methodology for the Normal example in Section 2 is immediately available and follows methods described in Fraser (2017). The resulting p-value and likelihood functions are readily computed, and inference is then of

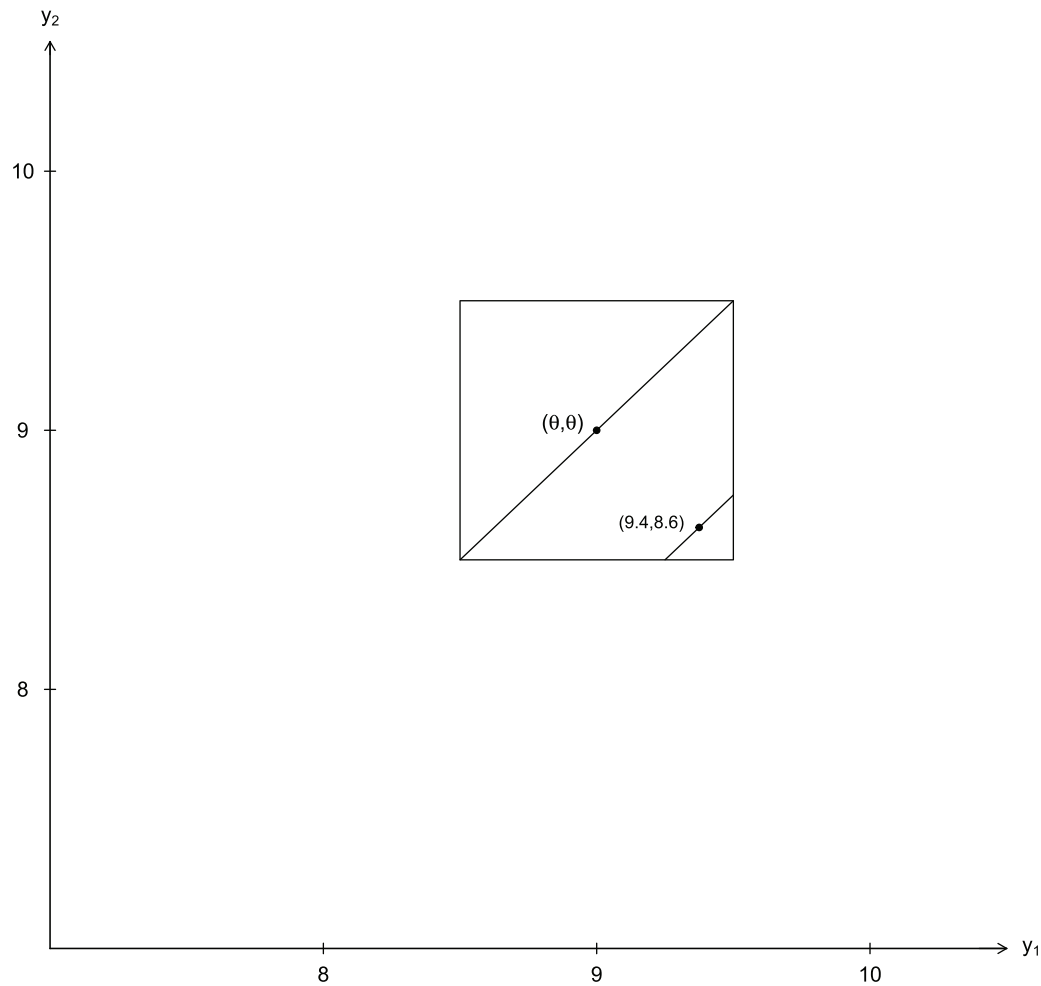


Figure 5. The sample space for a sample of 2 from the uniform distribution on $(\theta - 1/2, \theta + 1/2)$. The observed data point is (9.4, 8.6); the square is the sample space for $\theta = \hat{\theta}^{\text{obs}} = \bar{y}^{\text{obs}} = 9$. Only θ-values in the range 8.9–9.1 can put positive density at the observed data point. The short line through this point illustrates the small data range consistent with the observed data.

the same type as that in the simple example but computational details differ. Extensions from this are also available for testing vector parameters and use recently developed directional tests (Sartori, Fraser, and Reid 2016). The next two sections discuss the conditioning and then the use of the approximations.

4. Conditioning for Accurate Statistical Inference

Fisher (1930) introduced the notion of an ancillary statistic, to formalize the use of conditioning for inference based on features of the observed data. An ancillary statistic is a function of the data whose distribution does not depend on the parameter or parameters in the model for the observed response. As a simple example, if the sample size n in an investigation is random, rather than fixed, but its distribution does not depend on the parameters governing the model for the response y, then the appropriate model for inference is $f(y \mid n; \theta)$, not $f(y, n; \theta, \phi) = f(y \mid n; \theta) f(n; \phi)$, where ϕ might be present as a parameter for the distribution for n. The weighing machine example in Cox (1958) is another example of an obvious conditioning; in this case a random choice is made between a precise measurement of y and an imprecise measurement, and the resulting inference based on y should be conditional on this choice.

Example 2. A pair of uniformly distributed variables

Consider a response y that is uniformly distributed on the interval $(\theta - 1/2, \theta + 1/2)$ and suppose we have a sample of $n = 2$ with data (9.4, 8.6), Figure 5 presents the sample space for some θ value with observed data (9.4, 8.6); the actual sample space indicated is that for $\theta = \hat{\theta}^{\text{obs}} = 9$ which is just $\bar{y}^{\text{obs}}$.

A quick conventional approach might use approximate normality and the pivot $t = (\bar{y} - \theta)/2^{-1/2}(1/12)^{1/2}$, where the denominator is the standard deviation 0.20 of the sample mean. Figure 6 shows for some θ value the corresponding distribution of $\bar{y}$ with its standard deviation 0.20. The corresponding p-value function $\Phi\{(9 - \theta)/0.20\}$ from the data is plotted in Figure 7. This provides an approximate 95% interval for θ as $(9 - 1.96 \cdot 0.20, 9 + 1.96 \cdot 0.20) = (8.61, 9.39)$, but not all is well.

In Figure 5, we show the sample space centered at θ, and indicate the line parallel to the 1-vector through the data point (9.4, 8.6), and related lines would be parallel to this. Suppose we think of changing the θ value: the corresponding square domain will shift to the lower left or to the upper right, but only θ-values from 8.9 to 9.1 can put positive density at the observed data point (9.4, 8.6). The short line through the observed data point records the corresponding range and the parameter can only be at most 0.1 from the data. Any statistic describing these

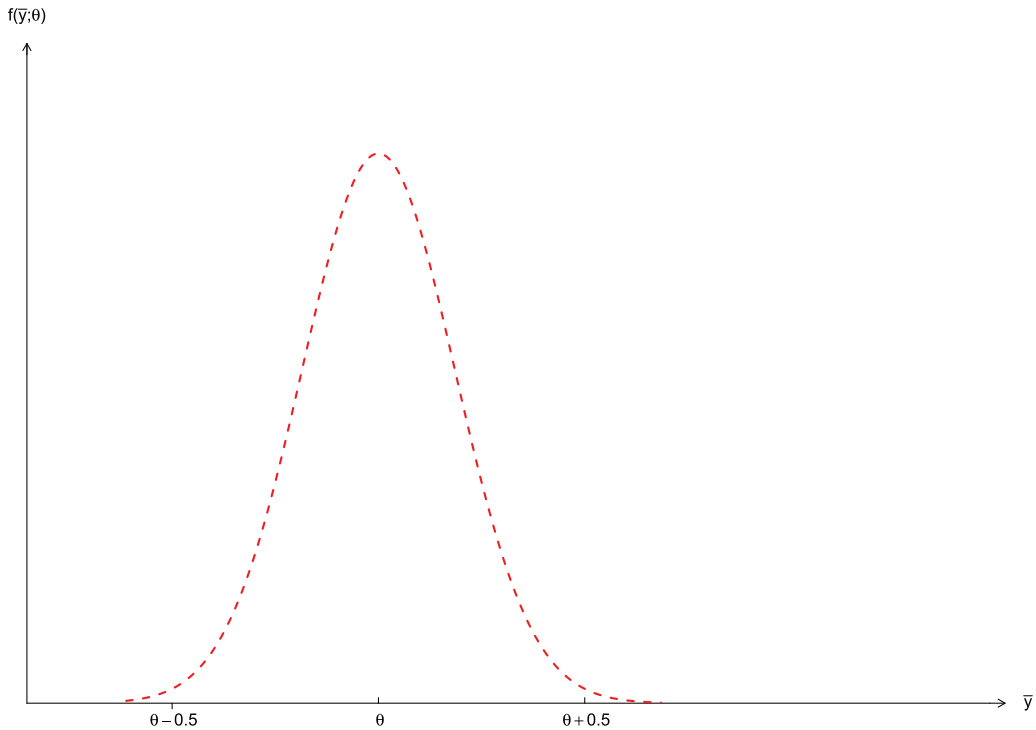


Figure 6. The approximate density of $\bar{y}$ for some θ value.

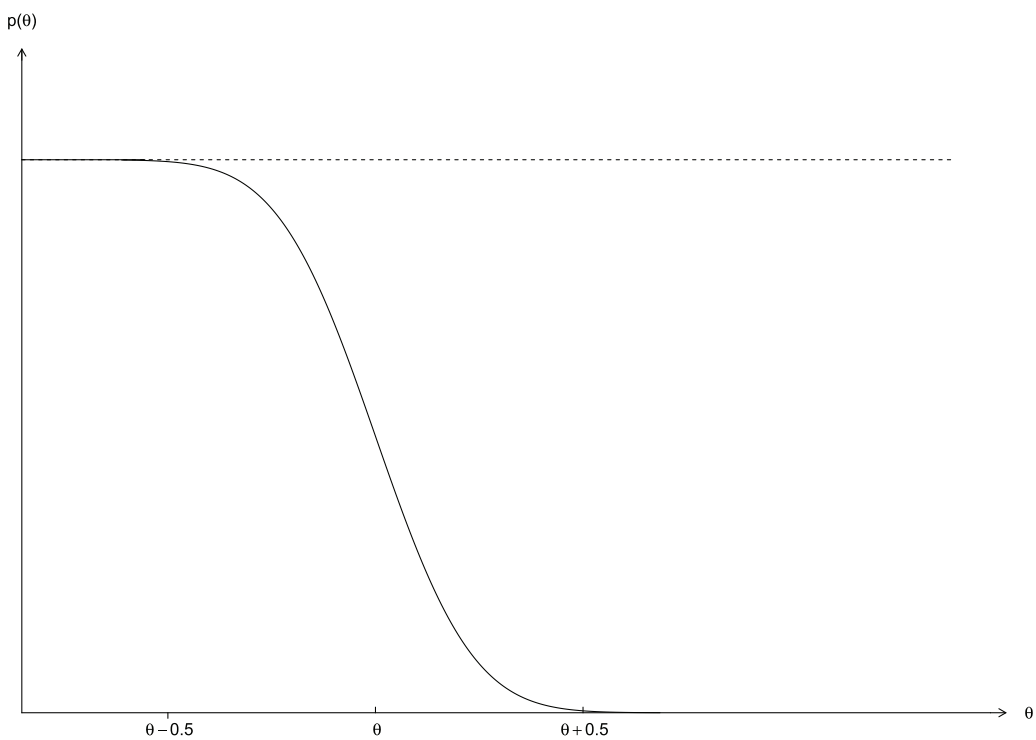


Figure 7. The observed p-value function from the approximate pivot $t = (\bar{y}^{\text{obs}} - \theta)/2^{-1/2}(1/12)^{1/2} = 0.20(\bar{y}^{\text{obs}} - \theta)$ as a function of θ.

diagonal lines has a distribution that cannot depend on θ and is thus ancillary. The conditional model on the line through the data point is $U(\theta - 0.1, \theta + 0.1)$ with half-range 0.1: the parameter can be at most 0.1 from the data and the data can only be at most 0.1 from the parameter. An exact 95% confidence interval for θ is $(9 - 0.095, 9 + 0.095) = (8.905, 9.095)$, which is radically different from the approximate result $(8.61, 9.39)$ in the preceding paragraph.

Indeed, the initial confidence interval includes θ values that are not possible, and also those for which the actual coverage is 100%. Of course, the first interval uses an extreme approximation but conditioning on $y_1 - y_2$ makes clear how the observed data restricts the range of possible parameter values. In Figure 8, we record the exact observed likelihood function from the data $(9.4, 8.6)$; it shows the range of possible values for θ and demonstrates the restrictions on the parameter.

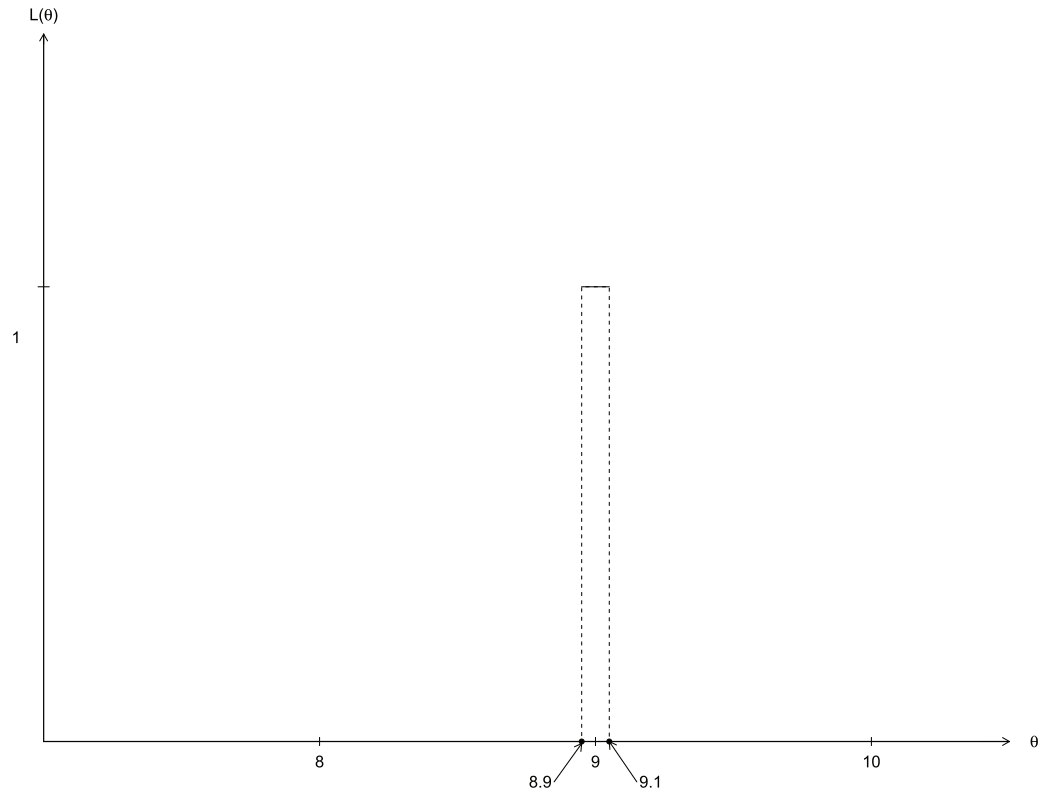


Figure 8. The exact observed a likelihood function from the data (9.4, 8.6).

In many cases for statistical inference, there is not such an obvious function of the data that is exactly distribution constant, However, there are many models in which there is an essentially unique quantile function or data-generating function $y_i = y(\theta, z_i)$, where the distribution of z is known. For example, if y_i follows a location model with density $f(y_i - \theta)$, we can write $y_i = \theta + z_i$, with the distribution of z_i known to be $f(\cdot)$. Similarly with independent sampling from a normal model with $E(y_i) = x_i'\beta$ and $\text{var}(y_i) = \sigma^2$, the model can be written $y_i = x_i'\beta + \sigma z_i$, with z_i following a $N(0, 1)$ distribution.

This quantile function provides the information needed to construct conditional inference, by examining how parameter change affects variable change for fixed quantile. This is summarized by the $n \times p$ matrix of derivatives

$$V = (v_1, ..., v_p) = \frac{\partial y}{\partial \theta}\Big|_{\text{fixed pivot } z}$$

evaluated at the data y^{obs} and maximum likelihood estimate $\hat{\theta}^{\text{obs}}$. The p vectors $v_1, \ldots, v_p$ are tangent to the contours of an approximate ancillary statistic, and the model conditional on these contours can be expressed in exponential family form. Recent asymptotic theory as summarized in Fraser (2017) shows that this conditional model provides a third-order inference information with extremely simple methods of analysis. If the given model happens to be a normal location model, then this will also reproduce the simple analysis above.

In Example 2, the quantile expression of the model is $y_i = \theta + u_i$, where u_i follows a $U(-1/2, 1/2)$ distribution. The 2×1 matrix V is simply $(1, 1)'$, and this defines the sloped lines in the sample space in Figure 5.

In the linear regression model, the matrix V is $n \times (p + 1)$, where p is the dimension of β. The leading $n \times p$ submatrix is simply X with ith row x_i', and the final column of V is the observed standardized residual vector $(\hat{z}_1, \ldots, \hat{z}_n)$.

Example 3. Ratio of Normal variables

This example has appeared widely in the literature and is often called the Fieller (1954)-Creasy (1954) problem. In its simplest form, we have two Normal variables with common variance, and we are interested in the ratio $\psi = \mu_1/\mu_2$ of their means; a nearly equivalent parameter is the direction $\psi = \alpha$ of the vector mean (μ_1, μ_2) as an angle, say negative, from the positive μ_1 axis. The problem also has close connections to regression calibration methods. Various routes to conditioning have been discussed in the past. Fraser, Reid, and Lin (2018) examine the model from the present viewpoint, but we do not attempt to reproduce the discussion here; other related examples are also examined there.

The next section outlines the automatic conditioning based on model continuity: this is preliminary to the widely available accurate approximations for the p-value function. We also outline the route to the accurate p-value function approximations.

5. Accurate *p*-value Functions

Two main features make the p-value function central to inference. The first of these is its multiple uses, as set out in Section 2. The second is the broad availability of an asymptotic approximation to the p-value function. This approximation is uniquely determined to third order and is conditional on an approximately ancillary statistic. The matrix V described in Section 4 enables construction of an approximating exponential family

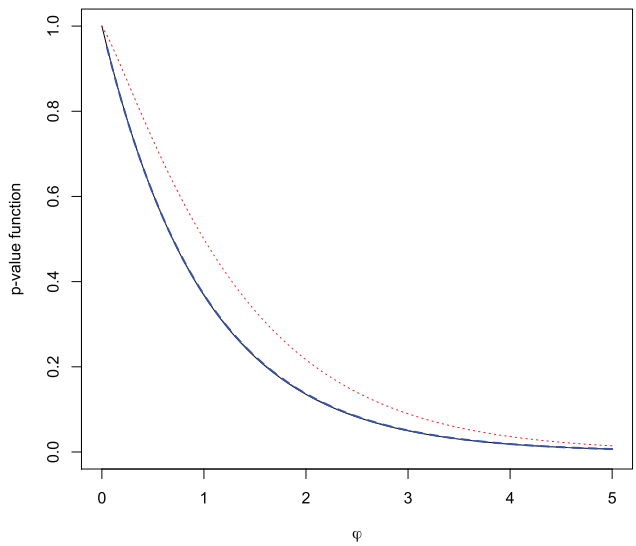


Figure 9. For the simple exponential model, the p-value function is plotted using the Normal approximation for r (dotted), using the third-order approximation (dashed), and compared to the exact p-value function (solid).

model but avoids explicit construction of the approximately ancillary statistic. This exponential model provides a route to inference for a scalar parameter of interest that is accurate to third order, that is, the inference approximation errors are $O(n^{-3/2})$. We illustrate this approximation route in this section; details on the construction of the tangent exponential model, if needed, are provided in Fraser (2017).

Example 4. Exponential life model

As a simple and transparent example consider the exponential life model $f = \exp\{-\varphi y + \log\varphi\}$ with failure rate φ and observed life y, both on the positive axis; and for data, take $y^{\text{obs}} = 1$ with little loss of generality. Note that increasing φ corresponds to decreasing life y. The observed likelihood function is $\ell(\varphi) = -\varphi + \log\varphi$ and the maximum likelihood estimate is $\hat{\varphi} = 1/y^{\text{obs}} = 1$. An approximate p-value function for φ can be obtained from standard asymptotic theory, for example, using the standardized maximum likelihood estimate $q = \hat{\varphi} - \varphi$ we have $p(\varphi)$ approximated by $\Phi(q)$, where Φ is the standard normal cdf. The signed square root of the log-likelihood ratio statistic also follows a normal distribution in the limit, so another approximate p-value function is $p_L(\varphi) = \Phi(r)$, where $r = \text{sign}(\hat{\varphi}-\varphi)\{2(\varphi-\hat{\varphi}-\log\varphi)\}^{1/2}$. As described below, the saddlepoint approximation leads to a pivotal quantity $r^* = r + r^{-1}\log(q/r)$ that follows the standard normal distribution to third order. The approximations $p_L(\varphi)$ and $p_{3rd}(\varphi) = \Phi(r^*)$ are compared to the exact p-value function $p_{Exact}(\varphi) = \exp\{-\varphi\}$ in Figure 10. Although $n = 1$, the third-order approximation is extremely accurate.

The steps in obtaining this approximation are as follows:

1. The exponential model

   Exponential models have a simple form as an exponential tilt of some initial density $h(s)$

$$f(s;\varphi) = \exp\{\varphi' s - \kappa(\varphi)\}h(s), \tag{1}$$

   where the p-dimensional variable $s = s(y)$ and parameter $\varphi = \varphi(\theta)$ are called canonical and can in turn be functions of some background variable y and parameter θ.

2. Saddlepoint accuracy

   The saddlepoint approximation came early to statistics (Daniels 1954) but only more recently has its power for analysis been recognized. The saddlepoint approximation to the density of s can be expressed in terms of familiar likelihood characteristics; in particular, the log-likelihood ratio $r^2/2 = \ell(\hat{\theta};y) - \ell(\theta;y)$ and the standardized maximum likelihood departure $q = (\varphi - \hat{\varphi})\jmath_{\varphi\varphi}^{1/2}$, with the latter calculated in the canonical parameterization. These measures are illustrated in Figure 10. The saddlepoint approximation has the form

$$f(s;\varphi)ds = \frac{\exp(k/n)}{(2\pi)^{1/2}}\exp(-r^2/2)\frac{r}{q}dr \tag{2}$$

   using the just defined familiar statistical quantities r and q.

3. Accurate p-value approximation

   If $p = 1$, then the distribution function can be computed from (2) by noting that $rdr = (\hat{\varphi} - \varphi)ds$ and then integrating by parts, leading to the Barndorff-Nielsen (1983) formula for the p-value function

$$p(\varphi) = \Phi\{r + r^{-1}\log(q/r)\} = \Phi(r^*), \tag{3}$$

   still in terms of the simple r and q statistical quantities. This was used above for the exponential life model.

4. With a nuisance parameter

   Suppose in (1) that the parameter of interest ψ is a component of φ, with the remaining components λ, say treated as nuisance parameters. A relatively simple adjustment to r^* above can be used to compute an accurate approximation to the p-value function for ψ. The arguments use the general theory of conditioning in Section 4, followed by a Laplace approximation for the required integration. The result is an approximation to $p(\psi)$ given by (3) but with r computed from the profile log-likelihood function: $r^2/2 = \ell(\hat{\varphi};s) - \ell(\hat{\varphi}_\psi;s)$ and q replaced by

$$Q = q\left(\frac{|\tilde{\jmath}_{\lambda\lambda}|}{|\hat{\jmath}_{\lambda\lambda}|}\right)^{-1/2}, \tag{4}$$

   where $q = (\hat{\psi} - \psi)\jmath_p(\hat{\psi})^{1/2}$ is the standardized maximum likelihood departure as before, $\jmath_p = -\ell_p''(\psi)$ is the observed information in the profile log-likelihood function, and $\jmath_{\lambda\lambda}$ is the nuisance parameter submatrix of the full observed information matrix, evaluated in the numerator at the constrained maximum likelihood estimate $\hat{\varphi}_\psi$, and evaluated in the denominator at the full maximum likelihood estimate $\hat{\varphi}$.

   Often the parameter of interest is a nonlinear function of the canonical parameter, which we write $\psi(\varphi)$, and the nuisance parameter $\lambda = \lambda(\varphi)$. The arguments leading to (3) can be developed by building on the conditioning of Section 4 and the Laplace integration mentioned above, and the result is an approximation to the p-value function again given by (3), but with q replaced by a quantity similar to Q in (4)

$$Q = \tilde{q}\left(\frac{|\tilde{\jmath}_{(\lambda\lambda)}|}{|\hat{\jmath}_{(\lambda\lambda)}|}\right)^{-1/2}, \tag{5}$$

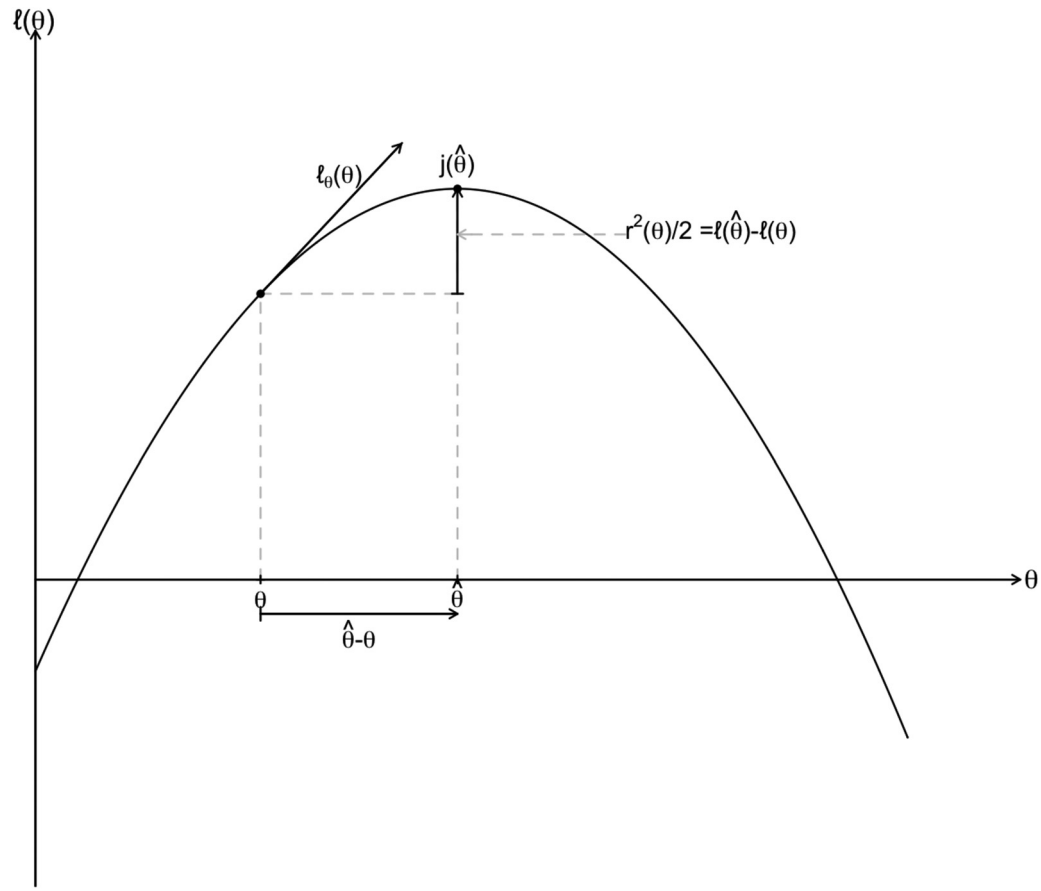


Figure 10. A log-likelihood function with the log-likelihood ratio $r^2/2$ and the standardized maximum likelihood departure $q = (\hat{\theta} - \theta) j_{\theta\theta}^{1/2}$ identified.

where the notation $j_{(\lambda\lambda)}$ means that λ is re-expressed in terms of a nuisance parameterization that accommodates the non-linearity in the dependence of the parameter of interest on the canonical parameter of the exponential model. The quantity $\tilde{q}$ is a standardized maximum likelihood departure, as above, but in a parameterization that takes account of the nonlinearity of ψ as a function of φ; see, for example, Fraser, Reid, and Wu (1999).

In the next section we give an example of the approximation of the p-value function in this setting.

5. General parametric model for y

If the statistical model of interest is not an exponential model, then a preliminary step based on Section 4 is used to construct what we have called the tangent exponential model, which implements conditioning on an approximate ancillary statistic using the matrix V. The canonical parameter of the tangent exponential model is $\varphi(\theta) = \partial\ell(\theta; y)/\partial V$. It can be shown that this approximation is sufficient to obtain an approximate p-value function to third order, and the steps are essentially the same as those for an exponential model described above. The argument is summarized in Reid and Fraser (2010) and Fraser, Reid, and Wu (1999).

One feature of these third-order approximations is that the pivotal quantity used to generate the p-value function is determined at the same time as its distribution is approximated, an astonishing result!

6. Applications

Many examples are available in the literature; see, for example, Fraser, Wong, and Wu (2004) and Fraser, Reid, and Wong (2009) and the book by Brazzale, Davison, and Reid (2007).

Example 5. Mean of a gamma density

As an illustration of the formulas in the previous section, we now consider the analysis of the lifetime data in Gross and Clark (1975) using a gamma model. The observations are (152, 152, 115, 109, 137, 88, 94, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, 69) being 20 survival times for mice exposed to 240 rads of gamma radiation. The gamma exponential model for a sample $(y_1, \ldots, y_n)$ is

$$f(y; \alpha, \beta)\Pi dy_i = \Gamma^{-n}(\alpha)\beta^{n\alpha} \exp\{\alpha s_1 - \beta s_2\}(\Pi y_i)^{-1}\Pi dy_i,$$

where $(s_1, s_2) = (\Sigma \log y_i, \Sigma y_i)$ records the canonical variable and (α, β) records the canonical parameter. Inference for α or β would use (4) for Q in approximation (3); and now we illustrate inference for the mean $\mu = \alpha/\beta$, a nonlinear function of the canonical parameter.

Grice and Bain (1980) considered such inference for the mean and eliminated the remaining nuisance parameter by "plugging in" the nuisance parameter estimate and then using simulations to adjust for the plug-in approach. Fraser, Reid, and Wong (1997) investigated the use of r_μ^* based on (5) and verified its accuracy; the results for the Gross and Clark (1975) data are recorded in Figure 11. For example,

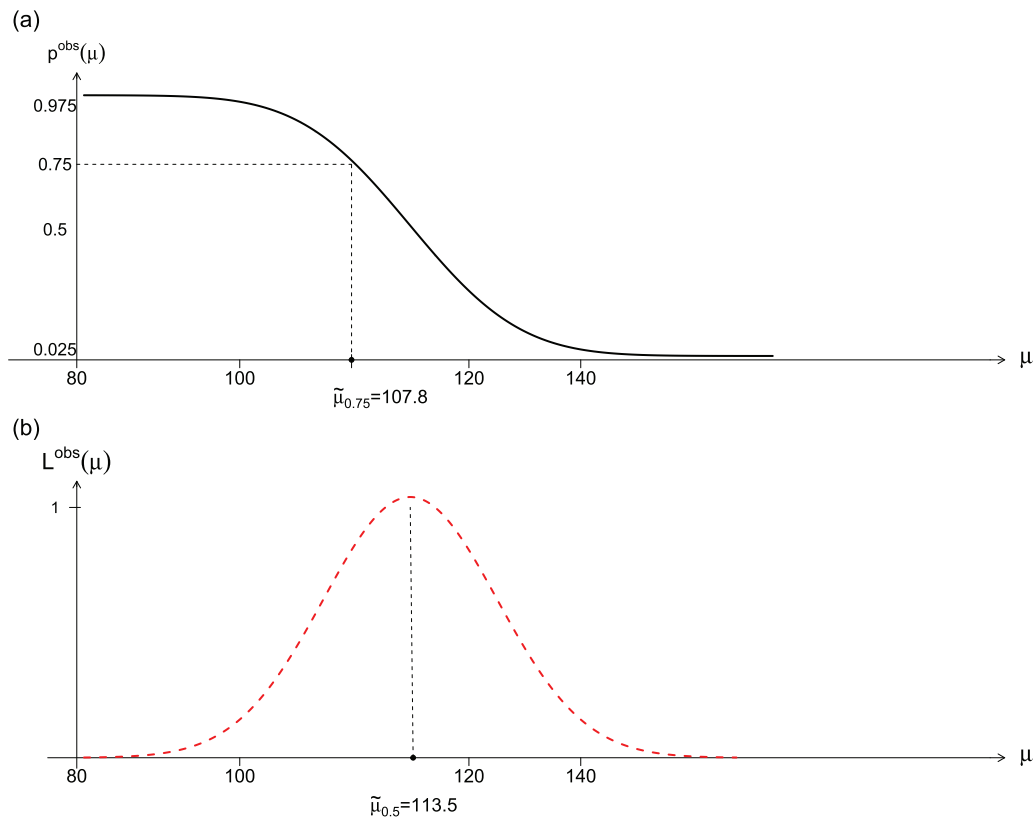


Figure 11. The p-value function and log-likelihood function for the mean of the Gamma model data from Gross and Clark (1975).

the 75% lower confidence bound is 107.8 which tells us that the given data provide 75% confidence that the true μ is in the interval $(107.8, \infty)$. Any one-sided or two-sided confidence interval is thus available immediately from the p-value graph.

Example 6: Model selection: Box and Cox

Location-scale and regression models have an important role in applications. Instances, however, do arise where the investigator is unsure of the mode of expression for his primary variable: Should he make a transformation of some initial variable, say a log or power transformation? Or should he take a family of such transformations and thus have a spectrum of regression models? A common concern is whether curvature should be applied to an initial response variable to achieve the linearity expected for a regression model. Box and Cox (1964) proposed a likelihood analysis, and a Bayesian modification. We consider here the present p-value function approach.

The regression model with the power transformation can be written as $y^{\lambda} = X\beta + e$ where e is a vector of say Normal$(0, \sigma^2)$ errors. For our notation, it seems preferable to write it as

$$y = (X\beta + e)^{1/\lambda},$$

where the power is applied coordinate by coordinate and the data and model coordinates are assumed to be positive. The power or curvature parameter λ is typically the primary parameter; while the regression parameters β only have physical meaning relative to a particular choice of the transformation parameter value λ. The third-order analysis is discussed in Fraser, Wong, and Sun (2009) using an example from Chen, Lockhart, and Stephens (2002). The resulting p-value function for the parameter λ is recorded there as Figure 1(e) on page 12, and the 95% central confidence interval for λ is $(0.694, 2.378)$ on page 14. This shows that it is extremely hard to pin down possible curvature even with the large sample size $n = 107$; linearity corresponds to $\lambda = 1$.

Example 7: Extreme value model

The extreme value model has immediate applicability to the modeling of extremes, say for the largest or smallest, in a sample or sequence of available data values, perhaps collected over time or space; it is of particular importance for the study say of climate extremes and has generalizations in the form of the Weibull model. The model has the density form

$$f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}.$$

For our purposes here, it provides a very simple example where the likelihood and the p-value functions are immediately available, in exact and in the third-order form. Suppose we have a single data value $y^{\text{obs}} = 21.5$ as discussed in Fraser, Reid, and Wong (2009). The log-likelihood and p-value functions have the simple form

$$\ell(\theta) = \theta - e^{\theta - 21.5}, \quad p(\theta) = \exp(-e^{\theta-21.5}).$$

For the third-order approximation, we have r and q as the signed likelihood ratio and the Wald departure

$$r = \text{sign}(\hat\theta - \theta)[2\{\ell(\hat\theta - \ell(\theta)\}]^{1/2}, \quad q = \text{sign}(\hat\theta - \theta)|\hat\varphi - \varphi|\hat{\jmath}_{\varphi\varphi}^{1/2}$$

where the exponential parameterization $\varphi(\theta) = \exp(\theta - 21.5) - 1$ is obtained by differentiating $\ell(\theta; y)$ with respect to y where the V from Section 4 is just a one-by-one "matrix" given as $V = 1$. The likelihood and p-value function are plotted on page 4 of

Fraser, Reid, and Wong (2009); in particular, the plot of the p-value function also records the third-order approximation we have been recommending; it is recorded as a dotted curve that is essentially indistinguishable from the exact as a solid curve. And of course any confidence interval or bound is immediately available.

Example 8: Logistic regression and inference for discrete data

Discrete data, often frequency data such as with the binomial, two by two tables, contingency tables, Poisson data, and logistic regression are commonly analyzed by area specific methods and some times by likelihood methods. Davison, Fraser, and Reid (2006) examine such discrete data analyses and demonstrate the straightforward steps and accuracy available for such problems, using the present p-value function approach and the related third-order accurate approximations; the discreteness, however, does lower the effective accuracy to second order. The general methods are also discussed there together with the minor modifications required for the discreteness. Binary regression is discussed in detail and the data from Brown (1980) on 53 persons with prostate cancer is analyzed. The methods are also applied to Poisson counts and illustrated with data on smoking and lung cancer deaths (Frome 1983).

7. Discussion

The p-value function reports immediately: if the p-value function examined at some parameter value is high or very high as on the left side of the graph, then the indicated true value is large or much larger than that examined; and if the p-value function examined at some value is low or very low as on the right side of the graph, then the indicated true value is small or much smaller than that examined. The full p-value function arguably records the full measurement information that a user should be entitled to know!

We also note that the p-value function is widely available for any scalar interest parameter and any model-data combination with just moderate regularity, and has available computational procedures.

In addition to the overt statistical position, the p-value function also provides easily and accurately many of the familiar types of summary information: a median estimate of the parameter; a one-sided test statistic for a scalar parameter value at any chosen level; the related power function; a lower confidence bound at any level; an upper confidence bound at any level; and confidence intervals with chosen upper and lower confidence limits. The p-value reports all the common inference material, but with high accuracy, basic uniqueness, and wide generality.

From a scientific perspective, the likelihood function and p-value function provide the basis for scientific judgments by an investigator, and by other investigators who might have interest. It thus replaces a blunt yes or no decision by an opportunity for appropriate informed judgment. In the high energy physics examples very small p-values are widely viewed as evidence of a discovery: Abe (1994) obtained $p = 0.0026$, and the LHC collaboration obtained 1 in 3.5-million before they were prepared to claim discovery. This is not a case of statisticians choosing a decision break point for others; but rather the provision of full inference information for others to make judgments. The responsibility for decisions made on the basis of inference information would rest elsewhere.

Acknowledgments

Sincere thanks to the Editor and Associate Editor for extraordinarily helpful suggestions and proposed content; we feel very deep appreciation. And deep thanks to Nancy Reid for counsel and deep insight. And special thanks to Wei Lin for preparing the graphs and for assistance with the manuscript preparation, to Mylène Bédard for calculations and simulations for the gamma mean problem, to Augustine Wong for assistance with examples.

Funding

The author gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada and The Senior Scholars Fund at York University.

References

Abe, F. (1994), "Evidence for Top Quark Production in pøp Collisions at Ãs =1.8 tev," *Physics Review Letters*, 69, 033002. [139,146]

Banff International Research Station (2006), "Statistical Inference Problems in High Energy Physics and Astronomy," available at *https://www.birs.ca/events/2006/5-day-workshops/06w5054*. [139]

Barndorff-Nielsen, O. E. (1983), "On a Formula for the Distribution of the Maximum Likelihood Estimate," *Biometrika*, 70, 343–365. [143]

Box, G., and Cox, D. (1964), "An Analysis of Transformations" (with discussions), *Journal of Royal Statistical Society*, Series B, 26, 211–252. [145]

Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics Case Studies in Small-Sample Statistics*, Cambridge, UK: Cambridge University Press. [144]

Brown, B. (1980), "Prediction Analysis for Binary Data," in *Biostatistics Casebook*, eds. B. W. B. R. G. Miller, B. Efron, and L. E. Moses, New York: Wiley, pp. 3–18. [146]

Chen, G., Lockhart, R., and Stephens, M. (2002), "Box–Cox Transformations in Linear Models: Large Sample Theory and Tests of Normality," *Canadian Journal of Statistics*, 30, 177–234. [145]

Cox, D. R. (1958), "Some Problems Connected With Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–372. [140]

Creasy, M. (1954), "Limits for the Ratios of Means," *Journal of Royal Statistical Society B*, 16, 186–194. [142]

Daniels, H. E. (1954), "Saddlepoint Approximations in Statistics," *Annals of Mathematical Statistics*, 46, 21–31. [143]

Davison, A., Fraser, D., and Reid, N. (2006), "Improved Likelihood Inference for Discrete Data," *Journal of Royal Statistical Society*, Series B, 68, 495–508. [146]

Davison, A. C., and Sartori, N. (2008), "The Banff Challenge: Statistical Detection of a Noisy Signal," *Statistical Science*, 23, 354–364. [139]

Fieller, E. (1954), "Some Problems in Interval Estimation," *Journal of Royal Statistical Society*, Series B, 16, 175 –185. [142]

Fisher, R. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transaction of Royal Society A*, 222, 309–368. [135]

——— (1930), "Inverse Probability," *Proceedings of the Cambridge Philosophical Society*, 26, 528–535. [140]

Fraser, D., Reid, N., and Lin, W. (2018), "When Should Modes of Inference Disagree? Some Simple But Challenging Examples," *Annals of Applied Statistics*, 12, 750–770. [137,142]

Fraser, D., Wong, A., and Sun, Y. (2009), "Three Enigmatic Examples and Inference From Likelihood," *Canadian Journal of Statistics*, 37, 161–181. [145]

Fraser, D., Wong, A., and Wu, J. (2004), "Simple Accurate and Unique: The Methods of Modern Likelihood Theory," *Pakistan Journal of Statistics*, 20, 173–192. [144]

Fraser, D. A. S. (2017), "The *p*-value Function: The Core Concept of Modern Statistical Inference," *Annual Review of Statistics and its Application*, 4, 153–165. [139,142,143]

Fraser, D. A. S., Reid, N., and Wong, A. (1997), "Simple and Accurate Inference for the Mean of the Gamma Model," *Canadian Journal of Statistics*, 25, 91–99. [144]

——— (2004), "On Inference for Bounded Parameters," *Physics Review D*, 69, 033002. [139]

——— (2009), "What a Model With Data Says About Theta," *International Journal of Statistical Science*, 3, 163–177. [144,145,146]

Fraser, D. A. S., Reid, N., and Wu, J. (1999), "A Simple General Formula for Tail Probabilities for Frequentist and Bayesian Inference," *Biometrika*, 86, 249–264. [144]

Frome, E. L. (1983), "The Analysis of Rates Using Poisson Regression Models," *Biometrics*, 39, 665–674. [146]

Grice, J., and Bain, L. (1980), "Inference Concerning the Mean of the Gamma Distribution," *Journal of American Statistical Association.*, 75, 929–933. [144]

Gross, A., and Clark, V. (1975), *Survival Distributions: Reliability Applications in the Biomedical Sciences*. New York: Wiley. [144,145]

Neyman, J., and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transaction of Royal Society A*, 231, 694–706. [135]

Reid, N., and Fraser, D. A. S. (2010), "Mean Likelihood and Higher Order Approximations," *Biometrika*, 97, 159–170. [144]

Rozeboom, W. (1960), "The Fallacy of the Null-Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428. [135]

Sartori, N., Fraser, D., and Reid, N. (2016), "Accurate Directional Inference for Vector Parameters," *Biometrika*, 103, 625–639. [140]

Sterling, T. D. (1959), "Publication Decisions and Their Possible Effects on Inferences Drawn From Tests of Significance—or Vice Versa," *Journal of American Statistical Association*, 54, 30–34. [135]

Wasserstein, R., and Lazar, N. (2016), "The ASA's Statement on *p*-values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [135]

The American Statistician

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

p-Values, Bayes Factors, and Sufficiency

Jonathan Rougier

To cite this article: Jonathan Rougier (2019) *p*-Values, Bayes Factors, and Sufficiency, The American Statistician, 73:sup1, 148-151, DOI: 10.1080/00031305.2018.1502684

To link to this article: https://doi.org/10.1080/00031305.2018.1502684

© 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: 5763

View related articles

View Crossmark data

Citing articles: 4 View citing articles

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, 148–151: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1502684

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

p-Values, Bayes Factors, and Sufficiency

Jonathan Rougier

School of Mathematics, University of Bristol, Bristol, UK

ABSTRACT
Various approaches can be used to construct a model from a null distribution and a test statistic. I prove that one such approach, originating with D. R. Cox, has the property that the *p*-value is never greater than the Generalized Likelihood Ratio (GLR). When combined with the general result that the GLR is never greater than any Bayes factor, we conclude that, under Cox's model, the *p*-value is never greater than any Bayes factor. I also provide a generalization, illustrations for the canonical Normal model, and an alternative approach based on sufficiency. This result is relevant for the ongoing discussion about the evidential value of small *p*-values, and the movement among statisticians to "redefine statistical significance."

ARTICLE HISTORY
Received February 2017
Revised April 2018

KEYWORDS
Embedding model;
Exponential tilting;
Generalized Likelihood Ratio (GLR)

1. Outline

This is a brief contribution to the ongoing discussion about the evidential import of a small p-value (see, e.g., Wasserstein and Lazar 2016). Let $X \in \mathcal{X}$ be a set of observables, and $H_0 : X \sim f_0$ be a null distribution. A "significance procedure" for H_0 is any statistic $p_0 : \mathcal{X} \to \mathbb{R}$ such that $p_0(X)$ under H_0 stochastically dominates a uniform distribution. If p_0 is a significance procedure for H_0, then $p_0(x^{\text{obs}})$ is a "p-value" for H_0, where x^{obs} are the observations of X. The usual way to construct a significance procedure is to propose a test statistic $t : \mathcal{X} \to \mathbb{R}$. Then

$$p_0(x) := Pr_0\{t(X) \geq t(x)\} \tag{1}$$

is a significance procedure according to the Probability Integral Transform, where Pr_0 is the probability under H_0. For more on these definitions see, for example, Casella and Berger (2002, sec. 8.3) and Lehmann and Romano (2005, chap. 9). The distinction between a "procedure" and a "value," which I have taken from Morey et al. (2016), is very useful in practice.

The critical issue is whether it is advisable to dismiss the null distribution on the basis of a small p-value, without explicitly considering any alternatives. The article addresses this issue by producing and justifying an "embedding model" based on the null distribution and the test statistic, in which the null distribution is at one end of the parameter space of the embedding model (Section 2). Within this embedding model

$$p_0(x) \leq G_{01}(x) \leq B_{01}(x), \tag{2}$$

where G_{01} is the Generalized Likelihood Ratio (GLR) and B_{01} is the Bayes factor. It follows from these inequalities that evidential thresholds for Bayes factors or GLRs translate into evidential thresholds for p-values. For example, if we accepted Harold Jeffreys's threshold that $B_{01}(x) = 10^{-3/2} \approx 0.032$ separates "strong" from "very strong" evidence against the null distribution, then $p_0(x) \leq 0.032$ would be the most lenient possible threshold for p-values designed to detect "very strong" evidence against the null distribution. This is less than the conventional threshold of $p_0(x) \leq 0.05$, but not by much; although even a small difference would have a substantial impact in some fields (Masicampo and Lalande 2012).

On the other hand, for a specific null distribution and test statistic, we can construct the embedding model and evaluate the exact relationship between the p-value and the GLR. In the canonical case where the embedding model is Normal (Section 3), a p-value of 0.05 corresponds to a GLR of 0.259, and a GLR of 0.032 corresponds to a p-value of 0.004. So in this case, accepting Jeffreys's threshold would lead to $p_0(x) \leq 0.004$ as the most lenient possible threshold for p-values. This is close to the suggestion of $p_0(x) \leq 0.005$, made by Johnson (2013). This threshold of $p_0(x) \leq 0.005$ has recently been advocated by a large group of statisticians (Benjamin et al. 2018), and questioned by another large group of statisticians (Lakens et al. 2018).

Section 4 provides a different justification for (2) via the sufficiency of the test statistic in the embedding model, which holds when the components of X are independent and identically distributed (IID).

2. D. R. Cox's Embedding Model

The attraction of a significance procedure is that it does not require a model for X within which the null distribution is a single element. Any attempt to link p-values with GLRs and Bayes factors must produce such a model based, as far as possible, only on the ingredients to hand: the null distribution and the test statistic. Clearly these two components are insufficient, and some additional principle must be used to justify any particular choice.

One principle is to assume that the test statistic t was carefully chosen to reflect the question of interest. This suggests an embedding model for X in which t is an unambiguously good choice for testing H_0 versus "not H_0," as was originally proposed

CONTACT Jonathan Rougier j.c.rougier@bristol.ac.uk School of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

© 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.

by D. R. Cox, in Savage et al. (1962, p. 84) and Cox (1977). Cox proposed the exponentially tilted embedding model

$$f(x;\theta) = \frac{f_0(x)\cdot e^{\theta\cdot t(x)}}{M_T(\theta)}, \qquad \theta \geq 0, \tag{3}$$

where M_T is the moment generating function of $t(X)$ under H_0. This model has a monotone likelihood ratio in $t(x)$, and hence the test statistic t is uniformly most powerful (UMP) in testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ (see, e.g., Casella and Berger 2002, sec. 8.3).

This is a "sufficient" argument for (3) as the embedding model; that is, were (3) the model, then t would be the analyst's unambiguous choice of test statistic for H_0 versus "not H_0." But it is also hard to imagine a simpler way to create an embedding model out of just f_0 and t, and this might be a more practical justification for (3). However, Section 4 presents another justification with strong intuitive appeal. I will refer to (3) as the "ET" (exponentially tilted) embedding model.

Initially, consider the Bayes factor for H_0 versus H_1,

$$B_{01}(x) := \frac{f_0(x)}{\int_0^\infty f(x;\theta)\,\pi(\theta)\,d\theta}, \tag{4}$$

where π is some prior distribution on $\theta \in (0,\infty)$. Adopting the approach originally proposed by Edwards, Lindman, and Savage (1963, p. 228), the Bayes factor can be bounded below over the set of all possible priors,

$$\begin{aligned} B_{01}(x) &\geq \inf_\pi \frac{f_0(x)}{\int_0^\infty f(x;\theta)\,\pi(\theta)\,d\theta} \\ &= \frac{f_0(x)}{\sup_{\theta>0} f(x;\theta)} =: G_{01}(x), \end{aligned} \tag{5}$$

where G_{01} is the GLR. This simple result is true for every embedding model. But then, using the ET embedding model in (3),

$$\begin{aligned} G_{01}(x) &= \inf_{\theta>0} \frac{f_0(x)}{f(x;\theta)} \\ &= \inf_{\theta>0} e^{-\theta\cdot t(x)}\, M_T(\theta) \\ &\geq Pr_0\{t(X) \geq t(x)\} \\ &= p_0(x) \end{aligned} \tag{6}$$

according to Chernoff's inequality (e.g., Whittle 2000, chap. 15). Chernoff's inequality is an application of Markov's inequality, and therefore in principle it is tight, but in practice an equality would be very unusual for a statistical model. One exception is when the components of X are IID and $t(x) = x_1 + \cdots + x_n$, in which case (6) is asymptotically exact; this is a result from Large Deviation Theory (see, e.g., Whittle 2000, chap. 18).

Putting the inequalities (5) and (6) together, (3) implies (2). Thus, if the embedding model is the ET embedding model, then the p-value for H_0 is never greater than the GLR for H_0 versus H_1, which is never greater than the Bayes factor for H_0 versus any alternative in the embedding model. It is superficially puzzling that two constructions which seem fundamentally different can be ordered by their values. But, on the one hand, the modern definition of a significance procedure p_0 implies that $p_0(y) \in (0,\infty)$, just like a Bayes factor. On the other hand, the ET embedding model ensures that $B_{01}(y) \in [0,1]$, just like a probability.

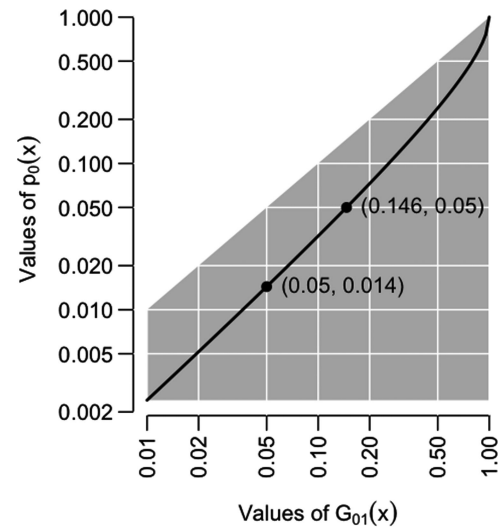


Figure 1. The GLR and Wilks's p-value. The gray region shows the possible values of p_0 under the ET embedding model given in (3), and the solid line the values of Wilks's p-value, based on the asymptotic distribution of $-2\log G_{01}(X)$ under the null distribution.

Curious readers will be wondering how close Wilks's p-value is to its upper bound of $G_{01}(x)$, under the ET embedding model. Wilks's theorem states that

$$-2\log G_{01}(X) \xrightarrow{\text{D}} \chi^2_1 \qquad \text{under } H_0 \text{ and (3)}, \tag{7}$$

if the components of X are IID (plus other technical conditions: see, e.g., Casella and Berger 2002, sec. 10.3). Thus, each value of $G_{01}(x)$ can be mapped to a p-value for H_0:

$$p_0(x) = Pr\{\chi^2_1 \geq -2\log G_{01}(x)\}, \tag{8}$$

where p_0 is an approximate significance procedure for finite n. Figure 1 shows the result: a Wilks's p-value of 0.05 corresponds to a GLR of 0.146. In other words, the IID and $n \to \infty$ conditions on X have reduced the p-value by as much as 10 percentage points. As this example illustrates, it is always pertinent to ask whether it is the observations or the conditions which produce a small p-value.

Finally, note that the ET embedding model can be generalized to

$$f(x;\phi) \propto f_0(x)\cdot e^{\phi\cdot h(t(x))}, \qquad \phi \geq 0, \tag{9}$$

for any increasing h without t losing its UMP property, for which (6) still holds, and (2) likewise. H replaces T, and the final step is

$$Pr_0\{h(t(X)) \geq h(t(x))\} = Pr_0\{t(X) \geq t(x)\} = p_0(x) \tag{10}$$

because h is increasing. Therefore, the ET embedding model is more properly thought of as a class of embedding models, and the inequalities in (2) hold for every embedding model in the class.

3. Illustration: The Normal Model

To illustrate both inequalities in (2), consider the canonical statistical model, first analyzed in this context by Edwards, Lindman, and Savage (1963, p. 228). Let the null distribution be $X \sim N(0,\sigma^2)$ for known σ, where I take $\sigma = 1$ for simplicity

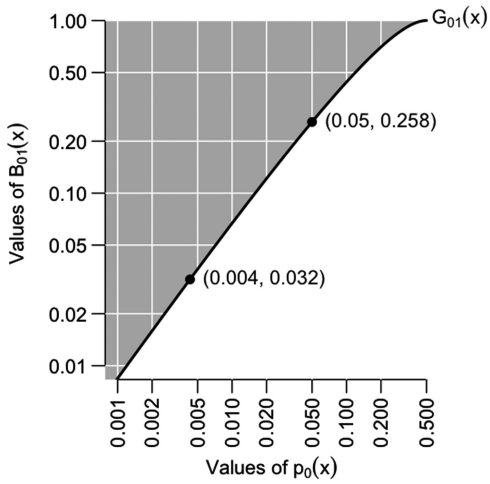


Figure 2. The Generalized Likelihood Ratio G_{01} and possible Bayes factors B_{01} (gray region), as functions of the p-value p_0, for the null distribution $X \sim N(0, 1)$ and the test statistic $t(x) = x$, which is a UMP one-tailed test for location.

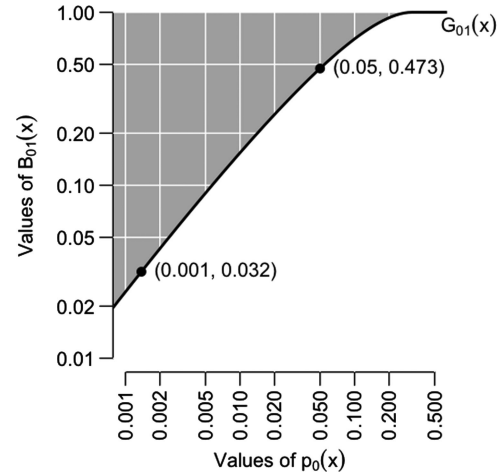


Figure 3. Similar to Figure 2, except based on the "two-tailed" test statistic $t(x) = x^2$, which is a UMP one-tailed test for dispersion.

and without loss of generality. Let the test statistic be $t(x) = x$. Then the ET embedding model is $X \sim N(\theta, 1)$, for which $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ is a conventional one-tailed test for location. The GLR is

$$G_{01}(x) = \begin{cases} 1 & x \leq 0 \\ e^{-\frac{1}{2}x^2} & x > 0. \end{cases} \tag{11}$$

The p-value is a deterministic function of x, from which it is possible to plot G_{01}, the lower bound for B_{01}, as a deterministic function of p_0, as shown in Figure 2.

Figure 2 also shows some specific values: the lower bound on $B_{01}(x)$ when $p_0(x) = 0.05$, and the value of $p_0(x)$ corresponding to a lower bound of $B_{01}(x) = 10^{-3/2} \approx 0.032$, which is the boundary between "strong" and "very strong" evidence against H_0 in the scheme of Jeffreys (1961, see Appendix B). In this example, a p-value at the conventional threshold of 0.05 corresponds to a lower bound on the Bayes factor of 0.259: "Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest" (Edwards, Lindman, and Savage 1963, p. 228). From the other direction, the necessary condition for satisfying Jeffreys's boundary is $p_0(x) \leq 0.004$. Jeffreys's boundary is only a convention, but the sizable absolute discrepancy between the two points in Figure 2, on either scale, casts doubt on the advisability of dismissing the null distribution for a p-value of about 0.05.

In this illustration, the null distribution, test statistic, and ET embedding model combined can give a UMP one-tailed test for location. It is natural to ask whether a different choice of test statistic can give a two-tailed test for location, but the answer must be negative because there is no UMP two-tailed test for location (see, e.g., Casella and Berger 2002, sec. 8.3). It follows that "two-tailed" test statistics might have unexpected ET embedding models.

To illustrate, if $t(x) = x^2$, which is large in both tails of the null distribution, then the ET embedding model is equivalent to

$$X \sim N(0, 1 + \theta), \tag{12}$$

showing that x^2 is a UMP one-tailed test for dispersion. The GLR for this model is

$$G_{01}(x) = \begin{cases} 1 & |x| \leq 1 \\ |x| \cdot e^{-\frac{1}{2}(x^2-1)} & |x| > 1. \end{cases} \tag{13}$$

The relationship between the p-value and the GLR is shown in Figure 3. This figure carries a similar message to Figure 2, which is that there is a large difference between the p-value and the Bayes factor. A p-value at the conventional threshold of 0.05 corresponds to a lower bound on the Bayes factor of 0.473, which is very weak evidence against the null distribution. The necessary condition for satisfying Jeffreys's condition of $B_{01}(x) \leq 0.032$ is $p_0(x) \leq 0.001$, to three decimal places.

In both of these illustrations, we could have chosen nonlinear increasing functions of the test statistic to use in the ET embedding model. For example, x^3 in the first illustration, and x^4 in the second. This changes the ET embedding model, and therefore the implicit hypothesis test for which $t(x)$ is UMP. It also changes the GLR. But it does not change the p-value, and it does not change the result that the p-value is never greater than the GLR (see the end of Section 2). The absolute size of the gap between the p-value and the GLR can only be assessed with respect to a specific choice of test statistic.

4. Justification via Sufficiency

The weakness of the argument in Section 2 is that it relies on exponential tilting to construct the embedding model given in (3), or its generalization in (9), and it loses its force when the analyst does not think that exponential tilting is appropriate. There is an another argument which can be applied in the case where the components of X are IID. The crux of this argument is to arrive at (9) using a sufficiency principle.

We require a one-dimensional version of the Pitman–Koopmans–Darmois (PKD) theorem, which was originally sketched in Fisher (1934), with a modern proof in Schervish (1995, sec. 2.2.3). This theorem validates the following result (plus some technical conditions not given here). If

1. the components of X are IID,
2. the support of the embedding model is constant, and
3. the test statistic is sufficient in the embedding model,

then

$$f(x;\phi) \propto f_0(x) \cdot e^{\phi \cdot h(t(x))}, \tag{14}$$

where h is invertible, and the boundary condition $f(x;0) = f_0(x)$ has been imposed. To orient this model so that large values of the test statistic challenge the null distribution, we take $\phi \geq 0$ and h increasing, similar to (9). Then the argument in Section 2 goes through exactly as before.

Acknowledgments

The author thanks Patrick Rubin-Delanchy and Christian Robert for their helpful comments on previous versions of this article; and a TAS reviewer, whose detailed comments on two versions of this article resulted in many improvements.

Funding

This research was supported by the EPSRC SuSTaIn Grant, reference EP/D063485/1.

References

Benjamin, D. et al., (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [148]

Casella, G., and Berger, R. (2002), *Statistical Inference* (2nd ed.), Pacific Grove, CA: Duxbury. [148,149,150]

Cox, D. (1977), "The Role of Significance Tests" (with discussion and rejoinder), *Scandinavian Journal of Statistics*, 4, 49–70. [149]

Edwards, W., Lindman, H., and Savage, L. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242. [149,150]

Fisher, R. (1934), "Two New Properties of Mathematical Likelihood," *Proceedings of the Royal Society*, Series A, 144, 285–307. [150]

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, UK: Oxford University Press. [150]

Johnson, V. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110, 19313–19317. [148]

Lakens, D., et al. (2018), "Justify Your Alpha," *Nature Human Behaviour*, 2, 168–171. [148]

Lehmann, E., and Romano, J. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Springer. [148]

Masicampo, E., and Lalande, D. (2012), "A Peculiar Prevalence of p Values just Below .05," *The Quarterly Journal of Experimental Psychology*, 65, 2271–2279. [148]

Morey, R., Hoekstra, R., Rouder, J., Lee, M., and Wagenmakers, E.-J. (2016), "The Fallacy of Placing Confidence in Confidence Intervals," *Psychonomic Bulletin & Review*, 23, 103–123. [148]

Savage, L., et al. (1962), *The Foundations of Statistical Inference*, London, UK: Methuen. [149]

Schervish, M. (1995), *Theory of Statistics*, New York: Springer (corrected 2nd printing, 1997). [150]

Wasserstein, R., and Lazar, N. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [148]

Whittle, P. (2000), *Probability via Expectation* (4th ed.), New York: Springer. [149]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

Limitations of *P*-Values and *R*-squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

Sherri Rose & Thomas G. McGuire

To cite this article: Sherri Rose & Thomas G. McGuire (2019) Limitations of *P*-Values and *R*-squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment, The American Statistician, 73:sup1, 152-156, DOI: 10.1080/00031305.2018.1518269

To link to this article: https://doi.org/10.1080/00031305.2018.1518269

© 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: 9089

View related articles

View Crossmark data

Citing articles: 10 View citing articles

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, 152–156: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518269

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Limitations of *P*-Values and *R*-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

Sherri Rose and Thomas G. McGuire

Department of Health Care Policy, Harvard Medical School, Boston, MA

Abstract

Stepwise regression building procedures are commonly used applied statistical tools, despite their well-known drawbacks. While many of their limitations have been widely discussed in the literature, other aspects of the use of individual statistical fit measures, especially in high-dimensional stepwise regression settings, have not. Giving primacy to individual fit, as is done with *p*-values and R^2, when group fit may be the larger concern, can lead to misguided decision making. One of the most consequential uses of stepwise regression is in health care, where these tools allocate hundreds of billions of dollars to health plans enrolling individuals with different predicted health care costs. The main goal of this "risk adjustment" system is to convey incentives to health plans such that they provide health care services fairly, a component of which is not to discriminate in access or care for persons or groups likely to be expensive. We address some specific limitations of *p*-values and R^2 for high-dimensional stepwise regression in this policy problem through an illustrated example by additionally considering a group-level fairness metric.

ARTICLE HISTORY
Received March 2018
Revised August 2018

KEYWORDS
Fairness; Health policy; Risk adjustment; Stepwise regression; *p*-Value

1. Introduction

Many statistics textbooks, including those in graduate-level statistical science and applied health courses, instruct students to perform stepwise regressions for practical data analyses. These methods have been deployed in the context of building prediction functions, where input variables are used to estimate outcomes, and effect estimation, where the parameter of interest is instead the effect of an exposure or intervention on the outcome. Stepwise procedures have varied flavors and may start with an empty or partial model adding variables, a full model with all variables followed by deletion steps, or some combination of these two processes (Efroymson 1966; Draper and Smith 1966; Mantel 1970; Hocking 1976; Miller 2002; Sinisi and van der Laan 2004). After each variable addition or deletion step, an evaluation metric is used to assess whether that step impacted fit. Metrics can include *p*-values, R^2, F tests, and additional choices depending on the procedure.

However, there are numerous widely discussed limitations of stepwise methods, such as misleadingly small *p*-values not adjusted to account for the iterative fitting and biased R^2 measures, among others (Hocking 1976; Harrell 2001). Post-selection inference is an active area of statistical methods research, but most off-the-shelf statistical software does not incorporate these advances. Despite cautioning applied researchers about the shortcomings of traditional stepwise procedures, they are commonly used in practice.

One empirical prediction setting where stepwise regressions are currently applied is risk adjustment in health policy, which spans many areas, including public reporting, quality measurement, and payment models for health insurance plans. This paper focuses on the problem of stepwise regression building for the latter case, the prediction of payments to health plans, as a demonstrative example. We summarize how stepwise regression works in the context of plan payment risk adjustment to predict individual spending as well as some central limitations of *p*-values and R^2 for high-dimensional prediction. We then consider aspects of "fairness" and implement a metric to assess fair predictions for individuals in vulnerable (i.e., undercompensated) groups. Groups, such as individuals with mental health and substance use disorders (MHSUD), are considered according to the means a health plan has at their disposal to discourage enrollment in their health plans and restrict health services.

2. Problems with *p*-Values and R^2 in Stepwise Regression for Risk Adjustment

Risk adjustment in health plan payment as translated into a statistical estimation framework is a familiar, classical prediction question, which we define formally here. Consider a continuous spending outcome Y, a vector of demographic variables $\boldsymbol{D} = (D_1, D_2, \ldots, D_K)$, and a vector of health variables $\boldsymbol{H} = (H_1, H_2, \ldots, H_J)$. A typical risk adjustment model would be given by

$$E(Y|\boldsymbol{D},\boldsymbol{H}) = \alpha\boldsymbol{D} + \beta\boldsymbol{H},$$

where E represents an expectation, $E(Y|\boldsymbol{D},\boldsymbol{H})$ is the conditional mean of Y given $(\boldsymbol{D},\boldsymbol{H}), \alpha$ are coefficients for $\boldsymbol{D}$ demographic variables, and β are coefficients for the $\boldsymbol{H}$ health variables.

CONTACT Sherri Rose rose@hcp.med.harvard.edu Department of Health Care Policy, Harvard Medical School, Boston, MA 02115.
© 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.

The goal is the estimate $E(Y|\boldsymbol{D},\boldsymbol{H})$ our regression function, in order to predict spending. These predictions feed into the determination of the payment the health plan receives in exchange for accepting responsibility for paying for the individual's health care costs. Reasonably accurate prediction is necessary to ensure that health plans, many of which are for-profit, will be willing to accept and provide good care to enrollees with high costs.

Demographic variables generally take the form of age–sex cell indicator variables. Most of the effort in the development of plan payment risk adjustment formulas is associated with the creation and selection of the health variables. Thus, as a simplified introductory example, stepwise regression in this setting could take the following form:

1. Start with a baseline formula containing only demographic variables $\boldsymbol{D} : E(Y|\boldsymbol{D}) = \alpha\boldsymbol{D}$.
2. Assign a rule that will keep added health variables $\boldsymbol{H} = (H_1, H_2, \ldots, H_J)$ in the formula if they have significant p-values and also increase R^2.
3. Sequentially add health variables $\boldsymbol{H}$, beginning with H_1, assessing p-values and R^2 metrics after each addition.

This illustration could be criticized for many established reasons, including the choice to introduce variables one at a time, that it will not evaluate all possible combinations of variables, and that we may not end up with an optimal or near-optimal selection of variables in the final formula (Mantel 1970; Hocking 1976). We choose to focus on the issue of fairness metrics vs. p-values and R^2 as yet another consideration.

Although researchers and regulators recognize a number of criteria for inclusion of variables in risk adjustment formulas, in practice, the most influential statistical criterion appears to be whether a given variable improves performance, as measured by p-value or R^2 statistic. These metrics are automatically produced by statistical programs (oftentimes without appropriate correction for the repeated model fitting), but no metric is automatically generated to capture performance in other, potentially more important dimensions. Simply put, using a p-value or R^2 as the statistical criterion to decide variable inclusion in the prediction function can lead to mistakes, in the sense that improving fit at the person level may not improve fit in the ways that may have even more impact, at the group level.

We do not reiterate many of the arguments on the disadvantages of p-values, for example, as enumerated in Wasserstein and Lazar (2016). However, before defining our group-level fit metric to evaluate fairness, we do highlight two features of p-values that are particularly relevant for stepwise regression building in risk adjustment, which is a high-dimensional prediction problem. First, as is well-known, large sample sizes will yield significant p-values for variable coefficients that are of trivial magnitude (e.g., Chatfield 1995; van der Laan and Rose 2010). Many plan payment risk-adjustment formulas are created using millions of observations; assessing a variable's importance for predicting health spending using p-values is effectively useless. The second is regarding the relationship between R^2 and p-values. We keep in mind that substantial improvements in R^2 are accompanied by significant p-values for the added variable, but also that miniscule improvements in R^2 may result from the addition of a variable with a nonsignificant *or* significant p-value.

3. Additional Metrics for Risk Adjustment

The most important criteria for evaluating a risk-adjustment scheme follow from the efficiency or fairness problem risk adjustment is trying to fix (Layton et al. 2017). In the United States, Medicare Advantage (for older and some disabled adults), Medicaid Managed Care (for people with lower incomes), and the Marketplaces (for otherwise uninsured, created as part of the Affordable Care Act) all operate as individual health insurance markets with competing health plans. In these systems, health plans are prohibited by law from discriminating in enrollment or services against *individuals*. For example, plans must accept any individual who applies for membership. However, plans can and do discriminate against *groups* of individuals, such as those with MHSUD, by: (*a*) limiting provider networks treating these disorders, (*b*) setting low provider payments to mental health providers to discourage supply, (*c*) providing less favorable coverage of drugs, and other means. Consequently, individual-level fit is secondary to group-level fit as a metric for alternative plan payment schemes.

Group-level fit in U.S. plan payment risk adjustment formulas is often measured by predictive ratios, equal to the ratio of predicted values over actual values for a group. A predictive ratio less than 1 indicates that the prediction function underpredicts and will therefore underpay for the group. Layton et al. (2017) found that underpayment as measured by the predictive ratio in the Marketplaces was most severe for the mental illness group, among the four groups they examined. This is concerning because, as noted, although a Marketplace plan must accept all applicants, the plan can provide poor care for MHSUD, discouraging individuals with these conditions from seeking enrollment in the first place.

In Europe, it is more common to measure group fit by net compensation, equal to the average difference between predicted values for a particular group and actual values:

$$\text{Net Compensation} = \frac{\sum_{i\in g} \hat{Y}_i}{n_g} - \frac{\sum_{i\in g} Y_i}{n_g},$$

where $\hat{Y}_i$ is predicted spending for individual i, Y_i is observed spending for individual i, and n_g is the sample size for the group of interest (Layton et al. 2017). The sums are taken over all individuals in the group. Net compensation measures incentives to a plan to provide good service to a group. Groups need not be mutually exclusive, and net compensation can be defined with respect to each health condition group. We use this fairness metric in our demonstration as it is on the same scale as our outcome of interest, and therefore has an easy interpretation.

Other metrics of health plan performance have been implemented or proposed in the case of ensuring equal access to mental health care, but these are recognized as incomplete and of doubtful effectiveness as a basis for monitoring plan services (McGuire 2016). While many fairness metrics focus on classification problems, individual vs. group-based notions of fairness have been studied (e.g., Zemel et al. 2013; Hu and Chen 2017), as well as general frameworks that include nonclassification problems (e.g., Kusner et al. 2017). We refer to Mitchell (2017) for a didactic summary of fairness metrics in the machine learning and computer science literature.

4. Demonstration of the Metrics

We include an instructive example using the Truven Marketscan data for plan payment risk adjustment, which mirrors the approach implemented in the Marketplaces with some modification. Plan payment in the Marketplaces is complex, involving adjustments for geographic factors and premiums the plans collect from enrollees, among other factors (Layton et al. 2018). At the core of the payment scheme, however, is the Department of Health and Human Services Hierarchical Condition Category (HSS-HCC) linear least squares regression prediction model, which determines the base payment for each individual (Centers for Medicare and Medicaid Services 2016).

Each HCC variable is the result of a mapping from a subset of the thousands of five-digit International Classification of Disease and Related Health Problems (ICD) diagnoses reported on claims to a much smaller number of categories. For example, the HCC for "major depressive, bipolar, and paranoid disorders" is generated from over 50 ICD-9 flags. It is important to note that not all ICD-9 (or ICD-10, adopted in 2015) codes map to an HCC used for payment. This has been shown to be problematic with respect to accurate payments, especially for MHSUD in the Marketplaces (Montz et al. 2016). Montz et al. (2016) found systematic underpayment of enrollees with MHSUD on average; a major contributor was that 80% of individuals with MHSUD were not recognized by the Marketplace system because their MHSUD-related ICD-9 codes did not map to an HCC used in the risk adjustment formula.

The HHS-HCC model undergoes regular evaluation, as does its progenitor, the Centers for Medicare and Medicaid (CMS) version used for Medicare Advantage, including consideration of the HCC diagnostic adjustors. The HCCs used in the HHS-HCC model are a subset of the full 264 HCCs defined in the full system mapping. Adding or subtracting HCCs from the right-hand side of the risk-adjustment formula is a component in the evaluation of the prediction function. Government reports lay out the criteria used in defining HCCs, with the first two being the HCC should be "clinically meaningful" and "predictive" (Pope et al. 2004; Ellis et al. 2018). Other risk-adjustment formulas consider additional variable types for special populations, such as measures of functional status in the CMS frailty model (Kautter and Pope 2004).

Our Marketscan sample contained all those continuously enrolled from January 2015 to December 2016 who had prescription drug coverage and mental health coverage. We excluded enrollees with missing geographic region or claims information, as well as those with negative claims. With these restrictions, we took a sample of 4,000,000 enrollees that we used in our analysis. The outcome, total annual expenditures in 2016, was calculated by summing all inpatient, outpatient, and drug payments. Mean total spending for an individual adult was $6,619.

The baseline formula included 75 HCCs from 2015, as well as age and sex, as predictor variables in a main terms parametric regression. We additionally considered two mental health HCCs and two substance use disorder HCCs. These 79 HCCs followed the CMS-HCC risk adjustment formula in place for Medicare Advantage (Pope et al. 2011). We defined the MHSUD group for calculation of net compensation using Clinical Classification Software (CCS) categories, a more comprehensive set of variables compared to the HCCs. Each ICD flag maps to a CCS category, unlike the mapping from ICD to HCC described above. Therefore, our calculation of net compensation will capture the impact of the risk adjustment formula for those with MHSUD recognized *and* unrecognized by the formula. (We emphasize here that individuals with MHSUD but no MHSUD ICD flags will not be captured by HCCs or CCS categories.) The MHSUD group contains 13.8% of the sample, compared to the 2.6% of the sample identified using the four HCCs. Mean total spending for an individual in the MHSUD group was $11,346, which was 71% higher than mean spending for an individual in the total sample.

We describe possible iterative stepwise regression decision-making processes in Figure 1, guided by different metrics (that could be defined a priori and within cross-validation, e.g., to try to avoid cherry picking and overfitting). As noted earlier, with a sample size of four million enrollees, p-values are a largely ineffective metric. All but three variables in the baseline formula were significant, even those with small event rates, such as "pressure ulcer of the skin, with necrosis through to muscle, tendon, bone" with 109 enrollees. The most prevalent HCC was "diabetes without complications," occurring among 5.0% of individuals in the sample. This baseline formula had an adjusted R^2 of 13.1%, which is similar to the 12% R^2 typically achieved with the same specification in Medicare. Net compensation for MHSUD was a nontrivial underpayment, reflected in the negative value −$2,822, which was 25% of the mean total spending in the MHSUD group. Therefore, enrollees with MHSUD are vastly underpaid relative to those without MHSUD, giving insurers a strong incentive to distort their plan offerings to avoid these enrollees.

A stepwise regression formula building exercise that aims to maximize R^2 *or* optimize MHSUD net compensation toward 0 might proceed from the baseline formula (formula 1 in Figure 1) to formula 2, which added two mental health HCCs ("schizophrenia" and "major depressive, bipolar, and paranoid disorders"). This provided a small 2% increase in R^2 but a 31% increase in MHSUD net compensation; p-values for both variables were statistically significant. Because MHSUD net compensation calculated based on formula 1 was a large negative number, *increases* in MHSUD net compensation that did not lead to a value greater than 0 were improvements. Moving from formula 2 to formula 3, we added two substance use HCCs ("drug/alcohol psychosis" and "drug/alcohol dependence"). The R^2 did not change; a stepwise regression maximizing based on R^2 (and parsimony) would revert back to formula 2. However, there was a continued increase in MHSUD net compensation (this time of 11%), thus a stepwise regression optimizing this metric would keep formula 3 and continue.

Moving from formula 3, which contained the complete set of 79 HCCs used in the Medicare Advantage risk adjustment formula, the stepwise regression might then consider deletion of steps of statistically significant variables to improve fit and fairness. Formula 4a removed the HCCs associated with liver conditions with no impact on R^2, but a 2% decrease in MHSUD net compensation (a movement in the wrong direction). A stepwise regression driven by R^2 and parsimony would prefer such a formula, despite being worse than formula 3 for MHSUD net

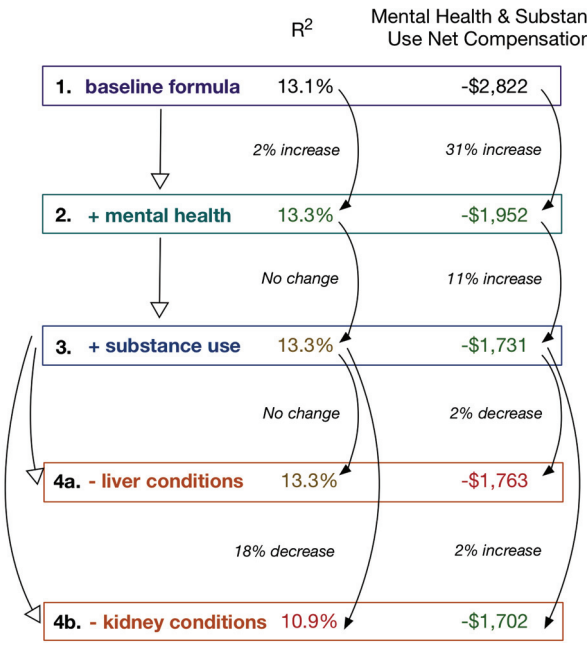


Figure 1. Decision-making flow chart for plan payment formula with differing metrics.

compensation (and likely worse for those with liver conditions). An alternative deletion step from formula 3 might be to remove kidney condition HCCs, represented in formula 4b. Here, R^2 dropped by 18%, yet MHSUD net compensation improved by 2%, and was indeed the best MHSUD net compensation of all the formulas in Figure 1. A stepwise regression that is searching based solely on improving MHSUD net compensation selects formula 4b, despite its overall poorer global fit (as assessed by R^2) and impact on those with kidney conditions.

5. Discussion

Beyond the well-known drawbacks of stepwise regression procedures, building an "appropriate" stepwise risk adjustment formula in a policy environment is not as simple as defining overall fit metrics for evaluation. Overreliance on statistical measures of global fit, such as p-values and R^2, in observational individual-level data does not consider the inequalities created or exacerbated in potentially vulnerable groups. Fair formulas are especially critical in healthcare systems where risk adjustment may have a direct or indirect impact on human health. In this context, p-values and R^2 may not be "morally neutral," to quote Professor Unsworth's use of this phrase for the subjectivity and ethics of statistical algorithms (Reyes 2016).

The massive size of the health care sector in the United States and its salience for personal and social welfare make risk adjustment arguably one of the most consequential applications of stepwise regression for social policy. As another example, valid evaluation of the relative performance of hospitals with respect to quality and outcomes requires risk adjustment for the health of each hospital's patient population, often referred to as "case-mix" (Shahian and Normand 2008). Hospitals with sicker patients along dimensions not included in the risk-adjustment formula will perform poorly in these assessments, leading to payment sanctions in public health insurance programs and possible closure by state regulators. Some approaches to case-mix adjustment for hospital quality explicitly use p-values to select variables in more parsimonious models (e.g., O'Malley et al. 2005).

We argue that there is a pressing need to consider a formal ensemble of metrics for evaluation of plan payment risk adjustment that balances both global fit and multiple fairness metrics. This is a likely scenario in other applied settings. It is well-accepted that evaluation of risk-adjusted models for plan payment involves numerous criteria. While regulators can and do make good faith efforts to examine vulnerable groups using predictive ratios and other measures (Layton et al. 2018), most of these procedures are ad hoc. Our stepwise regression risk adjustment demonstration represented a simplification of a deeply difficult policy problem and considered only one vulnerable group—those with MHSUD. Even in this simplified example with only one fairness metric, the "best" MHSUD net compensation was still a large underpayment relative to mean total spending, further highlighting the challenge of deploying a more comprehensive system with multiple groups. While unlikely to solve the totality of the fairness issues plaguing plan payment risk adjustment, there are numerous planned changes for the 2019 Medicare Advantage plan payment risk adjustment formula, including adding additional MHSUD HCCs (Centers for Medicare & Medicaid Services 2017). Fairness will also not be addressed by simply using machine learning rather than stepwise regression to estimate the risk adjustment formula or perform variable selection (e.g., Rose 2016; Shrestha et al. 2018) if standard statistical fit metrics are still the basis of evaluation for those tools.

The fairness issues facing plan payment risk adjustment are not entirely unique given the pervasive use of stepwise regressions and similar tools in many applications. Consideration of what makes a formula fair for the lives these health care algorithms touch has thus far been comparatively underdeveloped. Important work has been done in the areas of predictive policing (e.g., Lum and Isaac 2016), recidivism (e.g., Chouldechova 2017), and hospital ratings (e.g., Phillips 2018), for example. Leveraging and adapting these vital advances while expanding fairness approaches for the distinctive needs of health plan payment is a crucial issue moving forward. The Association for Computing Machinery recently issued a statement on automated decision-making describing seven principles for algorithmic transparency and accountability: awareness, access and redress, responsibility, explanation, data provenance, auditability, and validation and testing (Association for Computing Machinery 2017). It will be especially fruitful to bring these principles to bear in a context where the algorithms have a real impact on the welfare of vulnerable groups in the health care system.

Acknowledgements

The authors thank Monica Farid for data preparation.

Funding

This work was supported by NIH grant DP2-MD012722 and the Laura and John Arnold Foundation.

References

Association for Computing Machinery (2017), "Statement on Algorithmic Transparency and Accountability," available at *https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf*. [155]

Centers for Medicare & Medicaid Services (2016), "HHS-Operated Risk Adjustment Methodology Meeting Discussion Paper," available at *https://www.cms.gov/CCIIO/Resources/Forms-Reports-and-Other-Resources/Downloads/RA-March-31-White-Paper-032416.pdf*. [154]

Centers for Medicare & Medicaid Services (2017), "Advance Notice of Methodological Changes for Calendar Year (CY) 2019 for the Medicare Advantage (MA) CMS-HCC Risk Adjustment Model," available at *https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/Advance2019Part1.pdf*. [155]

Chatfield, C. (1995), "Model Uncertainty, Data Mining and Statistical Inference," *Journal of the Royal Statistical Society*, Series A, 158, 419–466. [153]

Chouldechova, A. (2017), "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, *5*, 153–163. [155]

Draper, N., and Smith, H. (1966), *Applied Regression Analysis*. New York: Wiley. [152]

Efroymson, M. (1966), "Stepwise Regression—a Backward and Forward Look," *Presented at the Eastern Regional Meetings of the IMS*, Florham Park, NJ: Institute of Mathematical Statistics. [152]

Ellis, R., Martins, B., and Rose, S. (2018), "Risk Adjustment for Health Plan Payment," in eds. T. G. McGuire, and R. C. Van Kleef, *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*, New York: Elsevier. [154]

Harrell, F. (2001), *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*, New York: Springer-Verlag. [152]

Hocking, R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–49. [152,153]

Hu, L., and Chen, Y. (2017), "Fairness at Equilibrium in the Labor Market," in *Proceedings of Fairness, Accountability, and Transparency in Machine Learning*, pp. 1–5. [153]

Kautter, J., and Pope, G. C. (2004), "CMS Frailty Adjustment Model," *Health Care Financing Review*, *26*, 1. [154]

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017), "Counterfactual Fairness," in *Advances in Neural Information Processing Systems*, pp. 4069–4079. [153]

Layton, T. J., McGuire, T. G., and Van Kleef, R. C., (2018), "Deriving Risk Adjustment Payment Weights to Maximize Efficiency of Health Insurance Markets," *Journal of Health Economics*, 61, 93–110. [155]

Layton, T. J., Ellis, R. P., McGuire, T. G., and Van Kleef, R. C. (2017), "Measuring Efficiency of Health Plan Payment Systems in Managed Competition Health Insurance Markets," *Journal of Health Economics*, 56, 237–255. [153]

Layton, T. J., Montz, E., and Shepard, M. (2018), "Health Plan Payment in U.S. Marketplaces: Regulated Competition with a Weak Mandate," in eds. T. G. McGuire, and R. C. Van Kleef, *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*, New York: Elsevier. [154]

Lum, K., and Isaac, W. (2016), "To Predict and Serve?" *Significance*, 13, 14–19. [155]

Mantel, N. (1970), "Why Stepdown Procedures in Variable Selection," *Technometrics*, 12, 591–612. [152,153]

McGuire, T. G. (2016), "Achieving Mental Health Care Parity Might Require Changes in Payment and Competition," *Health Affairs*, 35, 1029–1035. [153]

Miller, A. J. (2002), *Subset Selection in Regression*, Norwell, MA: CRC Press. [152]

Mitchell, S. (2017). "Fairness: Notation, Definitions, Data, Legality," available at *https://speak-statistics-to-power.github.io/fairness/old.html*. [153]

Montz, E., Layton, T. J., Busch, A., Ellis, R., Rose, S., and McGuire T.G. (2016), "Risk Adjustment Simulation: Plans May Have Incentives to Distort Mental Health and Substance Use Coverage," *Health Affairs*, 35, 1022–1028. [154]

O'Malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L., and Cleary, P. D. (2005), "Case-Mix Adjustment of the CAHPS®Hospital Survey," *Health Services Research*, 40, 2162–2181. [155]

Phillips, D. (2018), "At Veterans Hospital in Oregon, a Push for Better Ratings Puts Patients at Risk, Doctors Say," *New York Times*, available at *https://www.nytimes.com/2018/01/01/us/at-veterans-hospital-in-oregon-a-push-for-better-ratings-puts-patients-at-risk-doctors-say.html*. [155]

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Ingber, M. J., Levy, J. M., and Robst, J. (2004), "Risk Adjustment of Medicare Capitation Payments Using The CMS-HCC Model," *Health Care Financing Review*, 25, 119–141. [154]

Pope, G. C., Kautter, J., Ingber, J. J., Freeman, S., Sekar, R., and Newhart, C. (2011), "Evaluation of the CMS-HCC Risk Adjustment Model," available at *http://www.nber.org/risk-adjustment/2011/Evaluation2011/Evaluation_Risk_Adj_Model_2011.pdf*. [154]

Reyes J. (2016), "Technologists Must Do Better: Drexel Prof on the Ethics of Algorithms," *Technical.ly Philly*, available at *https://technical.ly/philly/2016/09/30/kris-unsworth-ethics-algorithms*. [155]

Rose, S. (2016), "A Machine Learning Framework for Plan Payment Risk Adjustment," *Health Services Research*, 51, 2358–2374. [155]

Shahian, D.M., and Normand, S.L. (2008), "Comparison of 'Risk-Adjusted' Hospital Outcomes," *Circulation*, 117, 1955–1963. [155]

Shrestha, A., Bergquist, S., Montz, E., and Rose, S. (2018), "Mental Health Risk Adjustment with Clinical Categories and Machine Learning," *Health Services Research*, 53, 3189–3206. [155]

Sinisi, S.E., and van der Laan, M.J. (2004), "Deletion/Substitution/Addition Algorithm in Learning with Applications In Genomics," *Statistical Applications in Genetics and Molecular Biology*, 3, 1–38. [152]

van der Laan, M. J., and Rose, S. (2010), "Statistics Ready for a Revolution: Next Generation of Statisticians Must Build Tools for Massive Data Sets," *Amstat News*, 399, 38–39. [153]

Wasserstein, R. L., and Lazar, N. (2016), "The ASA's Statement on p-values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [153]

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013), "Learning Fair Representations," in *International Conference on Machine Learning*, Atlanta, GA, pp. 325–333. [153]

The American Statistician

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

An Introduction to Second-Generation *p*-Values

Jeffrey D. Blume, Robert A. Greevy, Valerie F. Welty, Jeffrey R. Smith & William D. Dupont

To cite this article: Jeffrey D. Blume, Robert A. Greevy, Valerie F. Welty, Jeffrey R. Smith & William D. Dupont (2019) An Introduction to Second-Generation *p*-Values, The American Statistician, 73:sup1, 157-167, DOI: 10.1080/00031305.2018.1537893

To link to this article: https://doi.org/10.1080/00031305.2018.1537893

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 13078

View related articles

View Crossmark data

Citing articles: 22 View citing articles

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, 157–167: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1537893

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

An Introduction to Second-Generation *p*-Values

Jeffrey D. Blume[a,b], Robert A. Greevy[a], Valerie F. Welty[a], Jeffrey R. Smith[c,d], and William D. Dupont[a,e]

[a]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN; [b]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN; [c]Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN; [d]Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN; [e]Department of Health Policy, Vanderbilt University School of Medicine, Nashville, TN

ABSTRACT

Second generation *p*-values preserve the simplicity that has made *p*-values popular while resolving critical flaws that promote misinterpretation of data, distraction by trivial effects, and unreproducible assessments of data. The second-generation *p*-value (SGPV) is an extension that formally accounts for scientific relevance by using a composite null hypothesis that captures null and scientifically trivial effects. Because the majority of spurious findings are small effects that are technically nonnull but practically indistinguishable from the null, the second-generation approach greatly reduces the likelihood of a false discovery. SGPVs promote transparency, rigor and reproducibility of scientific results by a priori identifying which candidate hypotheses are practically meaningful and by providing a more reliable statistical summary of when the data are compatible with the candidate hypotheses or null hypotheses, or when the data are inconclusive. We illustrate the importance of these advances using a dataset of 247,000 single-nucleotide polymorphisms, i.e., genetic markers that are potentially associated with prostate cancer.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Likelihood ratios; Null hypothesis; *p*-Value; Statistical evidence

1. Introduction

By now everyone has seen the little red badges that appear on mobile phone apps and computer icons when the software wants attention. Besides stimulating serotonin production, those little red badges serve an important role: they notify the user when their attention is required. Perhaps not surprisingly, *p*-values have become the little red badges of applied statistics; they are assumed to indicate when the observed data are sufficiently informative to warrant the reader's attention. The results, after all, have been deemed "significant."

Having a gross indicator for when a set of data are sufficient to separate signal from noise is not a bad idea. The problem is that *p*-values perform poorly in this role. They confound effect size and precision, blurring the natural emphasis on meaningful effect sizes. They depend on the planned experimental design, even if the plan was not followed. Their proper interpretation is awkward. They are not consistent with the more intuitive measures of statistical evidence that arise in likelihood (Barnard 1949; Royall 1997; Blume 2002), Bayesian (Edwards, Lindman, and Savage 1963; Kass and Raftery 1995), and information theory (Good and Osteyee 1974) paradigms. In practice, these flaws often overshadow the *p*-value's positive attributes (Berger and Sellke 1987; Blume and Peipert 2003; Greenland et al. 2016; Goodman 1993; Cornfield 1966; Royall 1986; Dupont 1983).

For decades now, the statistical community has discouraged researchers from perseverating on the significance level of new findings (Wasserstein and Lazar 2016). Instead, the community seems to favor a more holistic approach that surveys the available evidence, modeling choices, potentially impactful effect sizes, and scientific context. While this emphasis is generally welcomed by practitioners of statistics, it is hard to condense it into a single metric that can be uniformly applied across different disciplines. Efforts to replace the *p*-value with something else—for example, a likelihood ratio (LR, Royall 1997; Blume 2002), Bayes factor (Bayarri et al. 2016), or posterior probability (Spiegelhalter, Abrams, and Myles 2004; Edwards, Lindman, and Savage 1963)—have not yet garnered a large enough following to alter the daily practice of statistics in applied disciplines (e.g., Cristea and Ioannidis 2018). *p*-Value "improvements" have been limited to revising the threshold for significance (Johnson 2013; Benjamin et al. 2018; Lakens et al. 2018) and tweaking the interpretation (Ioannidis 2018). Unfortunately, these efforts, resurrected every few decades since the 1960s (Berkson 1942; Cornfield 1966; Cohen 1994; Savage 1962; Morrison and Henkel 1970), have not led to changes in the daily practice of statistics.

The need for a reliable descriptive summary of the statistical evidence at hand remains unmet. Science looks to statistics for a global assessment—a single number summary—of whether the data favor the null hypothesis, the alternative hypothesis or whether the data are inconclusive. More nuanced assessments are, of course, encouraged. But this does not eliminate the need for a quick and coarse assessment of what the data say that is uniformly applicable across disciplines.

The second-generation *p*-value (SGPV) was developed with this need in mind (Blume et al. 2018). The idea was to improve on the *p*-value, rather than discard it. This meant keeping

CONTACT Jeffrey D. Blume *j.blume@vanderbilt.edu* Department of Biostatistics, Vanderbilt University School of Medicine, 2525 West End Ave, Set 11000, Nashville, TN 37203.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.

© 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.

familiar characteristics, such as the bounds of zero and one, while also adding new capabilities, such as the ability to indicate when the data support the null hypothesis. By construction, the SGPV retains many of the desirable properties of the p-value, incorporates often-wished-for properties, retains excellent control over error rates, operates with low false discovery rates (FDRs), and is readily interpretable by non-statisticians. In a broad sense, the SGPV simply indicates when experimental data support only the null premise or scientifically relevant alternative hypotheses.

The purpose of this note is to introduce the SGPV in a nontechnical manner and explain how it differs from the classical p-value. We will take the phrase "statistical inference" to mean measuring the strength of evidence in observed data about scientific hypotheses. Other inferential domains, such as belief theory ("What do I believe now that I have seen these data?") and decision theory ("What should I do now that I have seen these data?") will not be directly addressed. The focus here is on the most obvious and most frequently asked question in science: "Which scientific hypotheses are supported by the data?" Before introducing and discussing the SGPV, it is important to consider the conceptual framework that allows us to properly evaluate the statistical properties of the SGPV and to examine the manner in which we typically construct the null hypothesis.

2. Evidential Metrics

The majority of scientific activity consists of reporting and interpreting data as scientific evidence. Although there is no generally accepted mathematical framework for conducting an evidential analysis, there are three key metrics that comprise an evidential framework (Blume 2011):

1. A numerical assessment of the strength of evidence in a given body of observations
2. The probability that the numerical assessment will be misleading in a given setting
3. The probability that an observed assessment—one computed from observed data—is mistaken

The first metric is the scale of evidence, the second is the error rate, and the third is the FDR. These three metrics, obviously distinct, must be clearly defined for the evidential framework to be coherent. Failure to accomplish this creates a fatal flaw. Let's take a moment to consider each metric's role in the scientific process.

The first evidential metric is the numerical assessment of the strength of evidence in a given body of data. It is the researcher's essential tool for understanding what the data say. This tool is typically justified by axiomatic or intuitive means. Typically, they are single number summaries, but there is no restriction that this always be the case. Suppose we use $\mathcal{M}$ to measure the strength of evidence. It might be a metric that indicates absolute support for, or against, a single hypothesis such as a p-value or posterior probability. Or, it might be a measure that indicates relative support for one hypothesis over another such as a LR, Bayes factor, or divergence measure. A simple point estimate, such as a mean difference, relative risk, or odds ratio, is not itself an evidential metric because the connection between the point estimate and candidate hypothesis is not formalized. After collecting data, we compute and report $\mathcal{M}$.

Sometimes the data, by way of its interpretation through $\mathcal{M}$, indicate support for a false hypothesis or indicate that the data are inconclusive. Neither outcome is desirable, so studies are designed to minimize the rate at which they generate misleading or inconclusive evidence. Accordingly, the second evidential metric is the propensity to observe undesirable outcomes such as misleading or inconclusive evidence. In statistical lingo, these are often called "error" rates.

Suppose we have two competing hypotheses H_A and H_B. The classical frequency properties of a study design would be $P(\mathcal{M}$ supports $H_A|H_B)$, $P(\mathcal{M}$ supports $H_B|H_A)$, $P(\mathcal{M}$ yields inconclusive evidence $|H_B)$, and $P(\mathcal{M}$ yields inconclusive evidence $|H_A)$. A good study design minimizes these probabilities to the best extent possible. There are, of course, many different statistical and experimental strategies for doing this. The key point is that these probabilistic measures are properties of the experiment, not properties of the data that result from the experiment. For this we need another tool.

Once the data are collected, we compute and report $\mathcal{M}$. We will know if the evidence is strong or if it is inconclusive from the observed value of $\mathcal{M}$. For example, the observed LR might be 100, which would represent very strong evidence favoring one hypothesis over another. However, we will not know if $\mathcal{M}$ is misleading or not. The third evidential metric is the propensity for observed evidence to be misleading; it is an FDR. The FDR is an essential complement to the first evidential metric. We report the observed metric to describe the strength of evidence in the observed data, and we report the FDR to describe the chance that this result is mistaken. Note that, unlike the second evidential metric, the FDR is a property of the observed data.

Continuing with the example, the FDRs would be $P(H_B|\mathcal{M}$ supported $H_A)$ and $P(H_A|\mathcal{M}$ supported $H_B)$. Bayes' theorem must be used to compute these probabilities, so assumptions are made about the likelihood of the candidate hypotheses being true. *The key insight is that once data are observed, it is the FDRs that are the relevant assessments of uncertainty.* The error rates of the study design, critical at the design phase, are no longer relevant for the interpretation of observed data as statistical evidence. Rather it is the FDRs that convey the magnitude of uncertainty for the observed data.

The FDR was not formalized in the statistical literature until fairly recently (Benjamini and Hochberg 1995; Storey 2002, 2003). This is largely why statisticians and applied researchers have been slow to adopt this quantity in its natural role; preferring instead to rely on familiar but flawed interpretations of the Type I and Type II error rates. While it is true that the FDRs depend on the error rates (Wacholder et al. 2004), the FDRs are clearly distinct quantities and they should be treated as such.

A very successful real-world example of this structure comes from diagnostic medicine. Consider using CD4 count to determine if a patient, already infected with HIV, has developed AIDS. The CD4 count is an estimate of the number of active Helper T cells, per cubic millimeter, that can fight an infection. Normal CD4 counts range from 500 to 1500. When the CD4 count falls below 200, it is an indicator that the patient may have developed AIDS. Suppose we take $\mathcal{M}$ to be a LR and patient X is found to have a CD4 count of 150. Is this CD4 count, which supports a diagnosis of AIDS, strong evidence that patient X has AIDS? To answer this question, we compute the

metric, say $\mathcal{M}_{\text{obs}} = 50$, and interpret. The answer is yes; patient X's LR is large and favors the hypothesis that the patient has AIDS. For simplicity here, let's suppose the LR is a smoothly decreasing function of the CD4 count so we can forgo the actual computation of $\mathcal{M}$ and just use CD4 count as a surrogate for $\mathcal{M}$.

The second evidential metric—the properties of the diagnostic test—consist of sensitivity $P(\text{CD4} \leq 200|\text{AIDS}) = 0.9$ and specificity $P(\text{CD4} > 200|\text{No AIDS}) = 0.96$ (the numerical values were chosen for illustration purposes and do not reflect actual performance). Good diagnostic tests have high sensitivity and high specificity. The classical Type I and Type II error rates are the complements of these numbers, so $\alpha = 0.04$ and $\beta = 0.1$. When a physician receives patient X's test results, she reports the observed CD4 count (e.g., CD4 = 150) and the probability that the patient has AIDS given their CD4 count is below 200. This latter quantity is called the positive predictive value of the test result, $P(\text{AIDS}|\text{CD4} \leq 200) = \text{PPV}$. We use Bayes' theorem to compute the PPV from the test's sensitivity and specificity if we know the disease prevalence, say 20% for illustration. Then, we see that $\text{PPV} = 0.9 \times 0.2/(0.9 \times 0.2 + .004 \times 0.8) = 0.85$. The FDR is simply 1 – PPV or $P(\text{No AIDS}|\text{CD4} \leq 200) = 0.15$; it is the probability that patient X does not have AIDS when their CD4 count is less than 200.

In this context, it makes little sense for the physician to report to patient X the complement of the observed specificity, say $P(\text{CD4} \leq 150|\text{No AIDS}) = 0.02$. This is because the patient already knows their CD4 count. Moreover, the patient really wants to know how likely it is that they have AIDS or, if that is not possible, whether their chances of having AIDS has increased now that their CD4 count is known. The rub is that the complement of specificity is the p-value. And it is not at all clear what the p-value adds beyond the observed CD4 count (150) and FDR (15%). Perhaps the physician should report the p-value and the PPV? Or perhaps the CD4 count and the p-value? Neither is quite right. This example illustrates why statistical inference based on p-values can be so confusing. Imagine what would happen if the p-value had to be adjusted for all the HIV tests run at the doctor's office that day. Once data have been observed, it is the FDR that provides the correct assessment of uncertainty in the evidence; not the p-value.

Failure to distinguish between the evidential metrics leads to circular reasoning and irresolvable confusion about the interpretation of data as statistical evidence (Blume and Peipert 2003; Blume 2011). This is the fundamental dilemma that belies the p-value. Is the tail area probability a measure of the strength of evidence against the null hypothesis, or is it the study's effective error rate? Nearly a century after Fisher and Neyman first argued this very point, it remains unclear with divergent opinion on the matter. Fisher would argue for the former and Neyman for the latter, but neither view has carried the day (Royall 1997; Berger 2003). The problem is that there is only one number, the tail area probability, and this one number is used to represent two different concepts. The conflation of these concepts, as they pertain to the tail area probability, is the genesis of the multiple comparisons/multiple looks paradoxes in statistics (Blume and Peipert 2003; Royall 1997). Fisher understood this, writing "*In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence*" (Fisher 1959, p. 93). But this warning seems to have been long forgotten.

The SGPV framework respects the conceptual distinctions critical to a coherent evidential framework (Blume et al. 2018). As we will see, the SGPV is a proportion, not a tail area probability. This helps differentiate the summary of the strength of evidence in a given body of observations—the SGPV—from the design's error rates and the data's FDRs. The mere separation is a substantial inferential advance.

3. Specifying the Null Hypothesis

The SGPV is dependent on an expanded null hypothesis. The idea is to use a composite null hypothesis that reflects the limits on physical or experimental precision in outcome measurements, measurement error, clinical significance, and/or scientific relevance. Interval null hypotheses are constructed by incorporating this information into statistical hypotheses that are stated a priori. For example, when follow-up intervals are limited to weeks, it makes little sense to ponder survival differences of 7 days or less; that level of resolution for the outcome is too high given that the data were collected on a weekly basis. So, the natural interval null hypothesis consists of survival differences of 7 days or less. The interval null should contain, in addition to the precise point null hypothesis, all other point hypotheses that are practically null and would maintain the scientific null premise. While the hypotheses in the interval null may be mathematically distinct, they are all considered scientifically equivalent to the null premise. Examples include:

- $H_0 : 0.95 \leq \text{OR} \leq 1.05$ instead of $H_0 : \text{OR} = 1$,
- $H_0 : -0.1 \leq P(\text{Survival}|\text{Group 1}) - P(\text{Survival}|\text{Group 2}) \leq 0.1$ instead of $H_0 : P(\text{Survival}|\text{Group 1}) - P(\text{Survival}|\text{Group 2})$, and
- $H_0 : -0.5\sigma \leq \mu_1 - \mu_2 \leq 0.5\sigma$ instead of $H_0 : \mu_1 = \mu_2$.

By using an interval null hypothesis, we focus the evidential assessment on scientifically relevant effects. As a consequence, the Type I error rate, held constant in classical frequentist inference, naturally converges to zero as the sample size grows (see Section 5). The disadvantage is mainly procedural; it takes real forethought to specify the width of the interval in advance (Remark A). However, the extra effort does have a payoff. Findings that rule out this interval are both statistically significant and scientifically impactful. They are generally more reliable (the FDR is lower), and they are more likely to reproduce in subsequent studies. The reason for this is that the vast majority of Type I errors occur close to the point null hypothesis (2–4 standard errors away). Ergo, establishing a "buffer zone" translates into a substantial reduction in the rate of false discoveries.

In effect, the switch to an interval null hypothesis eliminates the common problem that classical statistical significance does not imply scientific relevance. This has been known in the statistical literature for some time. For example, Lehmann (1986, secs. 4.5 and 5.2; Good 2007) point out that the infinite precision of a point null hypothesis is rarely required (or justified) in practice. The scientific translation of this is that rejecting the null hypothesis that two effects are identical is largely not helpful, because they could still be nearly identical for all practical purposes. It is more helpful to establish that the effects differ by a meaningful amount. Unfortunately, current practice is to make

this assessment after the data are analyzed, when it is easier to be influenced by the observed results. The SGPV approach requires that this assessment be determined before the data are collected and therefore justified by scientific reasoning independent of the observed data, adding an important layer of rigor to the analysis.

4. The Second-Generation *p*-Value

The SGPV seeks to measure the fraction of data-supported hypotheses that are also scientifically null hypotheses. We will denote the SGPV by p_δ to signal its dependence on the interval null and distinguish it from the classical p-value (Blume et al. 2018). To identify the collection of "data-supported hypotheses," we use an interval estimate such as a confidence interval (CI), a likelihood support interval (SI), or a credible interval. Any type of interval may be used, but the choice impacts the frequency characteristics of the SGPV (see Remark B). In this article, we will use 1/8 likelihood SIs, which have numerous favorable inferential properties (Blume 2002; Royall 1997). Briefly, a 1/8 SI often corresponds to the traditional 95.9% Wald CI. As such, 1/8 SIs can be thought of as slightly conservative 95% CIs that do not need to be adjusted for sample space considerations (Blume 2002). Remark C provides a formal definition of SIs and an example.

Suppose that we are interested in the value of some parameter θ. Let $I = [\theta_l, \theta_u]$ be the interval estimate of θ whose length is given by $|I| = \theta_u - \theta_l$. Let the interval null hypotheses be denoted by H_0 and its length by $|H_0|$. The SGPV is

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2|H_0|}, 1\right\} = \begin{cases} \frac{|I \cap H_0|}{|I|} & \text{when } |I| \le 2|H_0| \\ \frac{1}{2}\frac{|I \cap H_0|}{|H_0|} & \text{when } |I| > 2|H_0| \end{cases} \quad (1)$$

where $I \cap H_0$ is the intersection or overlap of the two intervals. When the data are sufficiently precise, the SGPV is the fraction of I that is in H_0, i.e., $p_\delta = |I \cap H_0|/|I|$. Here, "sufficiently precise" means that the interval estimate is not more than twice the width of the interval null, that is, when $|I| \le 2|H_0|$.

When the interval estimate is very wide with $|I| > 2|H_0|$, I often extends on either side of H_0. In these cases, the quantity $|I \cap H_0|/|I| = |H_0|/|I|$ tends to be small and does not properly reflect the inconclusive nature of the data. The correction term $\max\left\{\frac{|I|}{2|H_0|}, 1\right\}$ replaces the denominator $|I|$ with $2|H_0|$ so that $p_\delta = 0.5 \times |I \cap H_0|/|H_0|$, which is bounded by 1/2. Note that p_δ can still be small despite being shrunk to 1/2 if the overlap $|I \cap H_0|$ is small. The correction factor maps inconclusive data toward a SGPV of 1/2, reserving magnitudes near 1 for data that support the null premise. This behavior is different than in traditional p-values, where large p-values result when data are inconclusive (i.e., when the CI is wide) and when the data support the null hypothesis (i.e., when the CI is tight around the null hypothesis).

Figure 1 illustrates how SGPVs work (reproduced from Blume et al. 2018). The overlap between the interval estimate (here a CI, but easily imagined as a SI) and the interval null is the essence of the SGPV. When the interval estimate is contained within the null interval, the data support only null hypotheses and $p_\delta = 1$. When the interval estimate and null set do not

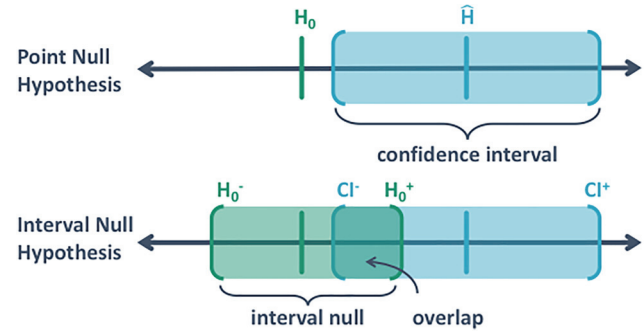


Figure 1. Illustration of a point null hypothesis, H_0; the estimated effect that is the best supported hypothesis, $\hat{H} = \hat{\theta}$; the confidence interval (CI) for the estimated effect [CI^-, CI^+]; and the interval null hypothesis [H_0^-, H_0^+].

overlap, the data are said to be incompatible with the null and $p_\delta = 0$. When the null set and CI partially intersect, $0 < p_\delta < 1$ and the data are inconclusive. In this last case, p_δ communicates the degree of inconclusiveness. For example, when $p_\delta \approx 1/2$, the data are said to be strictly inconclusive. But different levels of inconclusiveness are possible. For example, when $p_\delta \approx 0.2$, the data might be interpreted as trending in support of certain alternative hypotheses. When $p_\delta \approx 0.1$, the data might be interpreted as suggestive of a scientifically meaningful effect but not definitively. When $p_\delta \approx 0.01$ or $p_\delta \approx 0.99$, the data are close to fully supporting some alternative hypotheses or the null premise, respectively. While the descriptors of SGPV magnitude are helpful as communicators, they are not as essential as they are to traditional p-values because the natural ending states are well defined as $p_\delta = 0$ or 1 (Remark D).

To illustrate, consider a study of 100 smokers and 100 nonsmokers, where 65 smokers and 50 nonsmokers developed lung cancer. These data yield an odds ratio of 1.86 for the association of smoking and lung cancer. It is generally thought that odds ratios between 0.9 and 1.1 are too small to lead to meaningful associations. These data result in a 1/8 SI (95.9% CI) for the odds ratio of 1.03–3.36. The SGPV would then be $p_\delta = 0.175 = \left(\frac{1}{2}\right)\frac{(1.1-1.03)}{(1.1-0.9)}$ because the width of the CI is more than twice that of the null interval. The natural overlap measure is adjusted from $\frac{(1.1-1.03)}{(3.36-1.03)} = 0.03$ to 0.175 to reflect that the data are relatively imprecise when the goal is to learn about an interval null as tight as 0.9 to 1.1. With a $p_\delta = 0.175$, we would report that the study yielded inconclusive results. However, if 70 rather than 65 smokers developed lung cancer the odds ratio would be 2.33 with a 1/8 SI of 1.27–4.27. Now the intervals do not overlap so $|I \cap H_0| = 0$. Hence, $p_\delta = \left(\frac{1}{2}\right)\frac{(0)}{(1.1-0.9)} = 0$ and we would report that the data support a scientifically meaningful association between smoking and lung cancer.

The SGPV can be viewed as a formalization of today's standard practice of using CIs to assess the potential scientific impact of new findings. In our view, it is much better to make these judgment calls about scientific impact before looking at the data. SGPVs are intended as summary statistics that indicate when a study has yielded a CI that supports only the null premise or meaningful alternative hypotheses. By focusing only on results that are scientifically meaningful, the SGPV changes the relative importance of statistically significant findings. This means that the findings associated with the smallest p-value(s) will not necessarily have a corresponding SGPV of 0, nor will they

Table 1. Comparison between the classical and second-generation *p*-value.

Property	Classical *p*-value	Second-generation *p*-value
Null hypothesis	Point	Interval
Scale	Number between 0 and 1, excluding zero	Number between 0 and 1, inclusive
Able to indicate support for alternative hypotheses	When small or near 0	When 0
Able to indicate data are inconclusive	When larger than the stated cutoff for significance	When between 0 and 1; near $^1/_2$ is strictly inconclusive
Able to indicate support for null hypotheses	No	When 1
Accounts for practical importance	No	Yes, via specification of the null
Comparable across datasets	Yes, when the sample size is equal	Yes, when the width of the interval null is equal
Interpretation	Awkward, conditions on null hypothesis	Straightforward, conditions on observed data
Assumptions needed for computation	Probability model, sharp null hypothesis, sample space	Probability model, interval null hypothesis
Requires threshold for interpretation	Generally, yes	No, intended as a descriptive statistic
Adjustments required for multiple comparisons/looks?	Yes, if considered a Type I error rate; no, if considered a measure of statistical evidence	No

necessarily correspond to the top ranked SGPVs. This is illustrated nicely in our high-dimensional genetics example (Section 6).

Table 1 provides a side-by-side comparison of the properties of classical p-values and SGPVs. The take home message is that the SGPV is essentially an upgraded classical p-value.

Lastly, two endpoints with a SGPV of zero ($p_\delta = 0$) are compared on the basis of their *delta-gap* (Blume et al. 2018). The delta-gap is the distance between the null interval and the likelihood SI in units of δ, where δ is the half-width of the interval null hypothesis. For example, when the SI is shifted to the right of the null interval, the δ-gap is $(\mathrm{CI}^- - H_0^+)/\delta$. The scaling by δ makes it unit free and therefore easy to compare. Comparisons of δ-gap favor larger effect sizes. We use the delta-gap to rank SGPV findings in Section 6.

5. Frequency Properties

Blume et al. (2018) shows why the frequency properties of SGPVs can be controlled through sample size. For convenience, we mention a few key results here. Let $\hat{\theta}_n$ be a consistent estimator of parameter θ and assume an acceptable approximation to its sampling distribution is $\sqrt{n}(\hat{\theta}_n - \theta)\tilde{A}N(0, V)$, where the variance V is either known or can be readily estimated. This setting applies to most maximum likelihood estimators and posterior modes in large samples. When a $(1-\alpha)$ 100% CI is used as the interval estimate, the SGPV Type I error rate analogue is

$$P_{\theta_0}(p_\delta = 0) = P(p_\delta = 0|\theta = \theta_0) = 2\Phi\left[-\frac{\sqrt{n}\delta}{\sqrt{V}} - Z_{\alpha/2}\right]. \tag{2}$$

Note the dependence on the sample size n and on δ. The Type I error rate remains bounded above by $\alpha = 2\Phi[-Z_{a/2}]$. This is subtly different from Neyman–Pearson hypothesis testing with a composite null, where the size of the test is defined as the maximum Type I error rate over the null space (α). The SGPV approach anchors the Type I error at the natural experimental point null and uses the interval around the point null as a buffer. As such, when $\delta > 0, P_{\theta_0}(p_\delta = 0)$ shrinks to 0 as the sample size grows (instead of remaining constant at α). When $\delta = 0$, we recover the usual constant Type I error rate of α. The derivation of $P_\theta(p_\delta = 0)$, along with the probability that the data are compatible with the null, $P_\theta(p_\delta = 1)$, and the probability that the data are inconclusive $P_\theta(0 < p_\delta < 1)$, can be found in Supplement 1 of Blume et al. (2018). Remark E provides an expression for $P_\theta(p_\delta = 0)$, the SGPV power function.

Once data are collected and the SGPV is computed, the relevant uncertainty measure is the probability that the observed results, say $p_\delta = 0$ or 1, are mistaken. These are known as the *FDR*, $P(H_0|p_\delta = 0)$, and the *false confirmation rate* (*FCR*), $P(H_1|p_\delta = 1)$. The hypothesis notation, H_i for $i = 0, 1$, is left flexible on purpose. It can represent either a point hypothesis, $H_i : \theta = \theta_i$, or a composite hypothesis $H_i : \theta \in \Theta_i$, as needed. For example, $P(p_\delta = 0|H_0) = P_{\theta_0}(p_\delta = 0)$ when $H_0 : \theta = \theta_0$. Otherwise, $P(p_\delta = 0|H_0)$ is the average rate over $H_0 : \theta \in \Theta_0$, defined as $P(p_\delta = 0|H_0) = \int_{\Theta_0} P_\theta(p_\delta = 0)g(\theta)\mathrm{d}\theta$, for some probability distribution $g(\theta)$ over Θ_0. Bayes' rule yields expressions for the FDR and FCR

$$\begin{aligned} P(H_0|p_\delta = 0) &= \left[1 + \frac{P(p_\delta = 0|H_1)}{P(p_\delta = 0|H_0)}\mathrm{r}\right]^{-1} \\ P(H_1|p_\delta = 1) &= \left[1 + \frac{P(p_\delta = 1|H_0)}{P(p_\delta = 1|H_1)}\frac{1}{r}\right]^{-1} \end{aligned} \tag{3}$$

where $r = p(H_1)/P(H_0)$ is the prior probability ratio (Blume et al. 2018). These rates depend on the (possibly averaged) design probabilities $P(p_\delta = 0|H_1), P(p_\delta = 0|H_0), P(p_\delta = 1|H_0), P(p_\delta = 1|H_1)$, through the LR given the observed SGPV. Good designs—those with low probabilities of observing misleading evidence—will have low FDRs.

For fixed prior probability ratio r, the LRs $P(p_\delta = 0|H_1)/P(p_\delta = 0|H_0)$ and $P(p_\delta = 1|H_0)/P(p_\delta = 1|H_1)$ will drive the FDR and FCR to zero as the sample size grows. This is an improvement over the FDR from a classical hypothesis test, which remains constant in the limit: $\alpha/(\alpha + r)$. The convergence to zero of the SGPV's FDR and FCR happens because the design probabilities, for example, $P(p_\delta = 0|H_0)$, converge to zero. Thus, while it may not be possible to identify r, the prior's influence can be mitigated through sample size and good study design. Lastly, it is worth repeating that the Type I error rate, $P(p_\delta = 0|H_0)$, and the FDR, $P(H_0|p_\delta = 0)$, are not exchangeable. Their magnitudes can vary substantially. See Blume et al. (2018) and its Supplement 1 for further details.

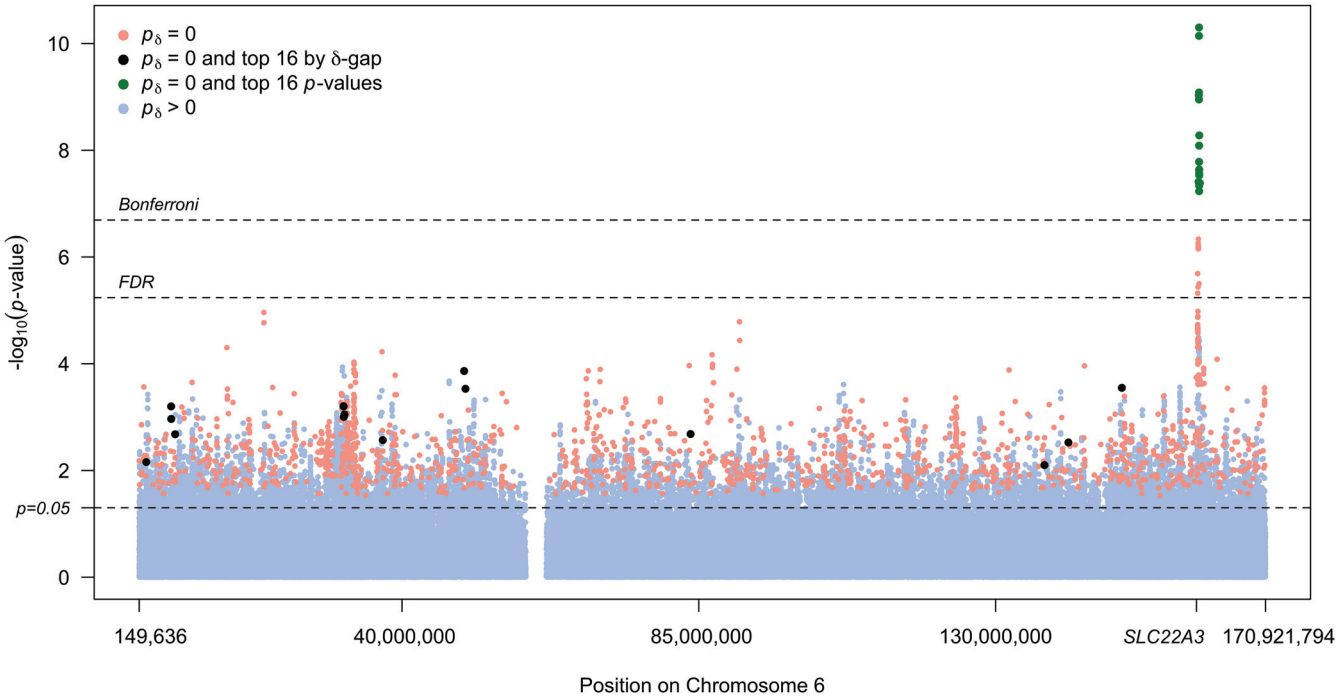


Figure 2. Manhattan plot of SNP associations, colored according to second-generation *p*-value status. The *x*-axis shows SNP position on chromosome 6 and the position of the organic cation transporter gene (SLC22A3). The top 16 associations by *p*-value rank are in green and the top 16 associations by second-generation *p*-value rank are in black. Bonferroni, FDR, and unadjusted significance levels are show by dashed lines.

6. Links to Prostate Cancer in 247,000 SNPs

The International Consortium for Prostate Cancer Genetics (Schaid and Chang 2005; ICPCG 2018) collected data on 3,894 individuals of European descent, each with 4.6 million single-nucleotide polymorphisms (SNPs) available for analysis. Genetic ancestry and independence were confirmed in 2511 cases with prostate cancer and 1383 controls without prostate cancer. SNPs measure the number of variant alleles (0, 1, or 2) at a given position on a chromosome. Alleles with prevalence less than 50% in the population of interest are variant. SNPs are routinely used as genetic markers in studies of disease.

Our analysis focused on 246,563 SNPs from chromosome 6. We chose chromosome 6 because it contains a cluster of genetic variants in or near an organic cation transporter gene (SLC22A3) that are thought to be associated with prostate cancer. We used logistic regressions to obtain 246,563 estimated odds ratios with 1/8 likelihood SIs (95.9% Wald CIs). Odds ratios of small to moderate size are of interest in the context of genetic investigations of common disease. Accordingly, we defined the interval null hypothesis as odds ratios between 0.9 and 1.1111 (= 1/0.9), which is symmetric on the log scale. Remark F records pertinent scientific details.

A Manhattan plot (Figure 2) displays the traditional *p*-values, at sequence chromosome position on chromosome 6 (Remark G), computed under the null hypothesis that the odds ratio is precisely 1. The Bonferroni cutoff ($p = 2.03 \times 10^{-7}$), Benjamini–Hochberg FDR cutoff ($p = 4.77 \times 10^{-6}$), and unadjusted cutoff ($p = 0.05$) are displayed as horizontal dashed lines. The points are colored according to SGPV status. There are 11,844 SNPs with $p < 0.05$, which accounts for 4.8% of all SNPs. Only 27 of these SNPs are statistically significant when controlling the FDR to 5%, and 16 of these meet the Bonferroni criteria. Importantly, all 27 of these SNPs are concentrated in or near the organic cation transporter gene (SLC22A3). These 27 SNPs also have a SGPV of zero ($p_\delta = 0$).

There are 1443 SNPs (0.6% of all SNPs) with $p_\delta = 0$. For comparison with Bonferroni, we identified the top 16 SNPs according to their SGPV *delta-gap* rankings. The δ-gap measures the distance between the interval estimate and the null interval in units of δ (recall δ is the half width of the interval null). When $p_\delta > 0$ the *δ-gap* is defined to be zero because there is no gap between the intervals. Figure 3 shows a modified Manhattan-style plot that displays the delta-gap by its position on chromosome 6. While the 16 Bonferroni SNPs can be seen concentrated around SLC22A3 (in green), the top 16 SNPs by *δ-gap* (black) are instead spread across the chromosome and potentially identify much stronger associations. In fact, the 16 Bonferroni SNPs are ranked between 423rd and 845th by the δ-gap metric, which illustrates that classical methods can disregard a large number of potentially impactful associations.

We can see in Figure 3 that the SGPV approach emphasizes the magnitude of the observed association in addition to classical statistical significance. SNPs with a SGPV of 0 also have a classical *p*-value that is less than 0.041 (the compliment of the coverage probability for the 1/8 SI). However, the converse is not true. Statistically significant findings can have an interval estimate that intersects with the interval null, leading to a SGPV > 0. Moreover, the δ-gap ranking orders findings by the weakest nonnull effect size supported by the data. Hence, the SGPV approach can be thought of as selecting for the potentially strongest nonnull statistically significant effects supported by the data, even when estimated relatively imprecisely.

When comparing top ranked SNPs according to different criteria, it is helpful to examine the 1/8 SIs for the odds ratio relative to the null interval hypothesis. This is done in Figure 4,

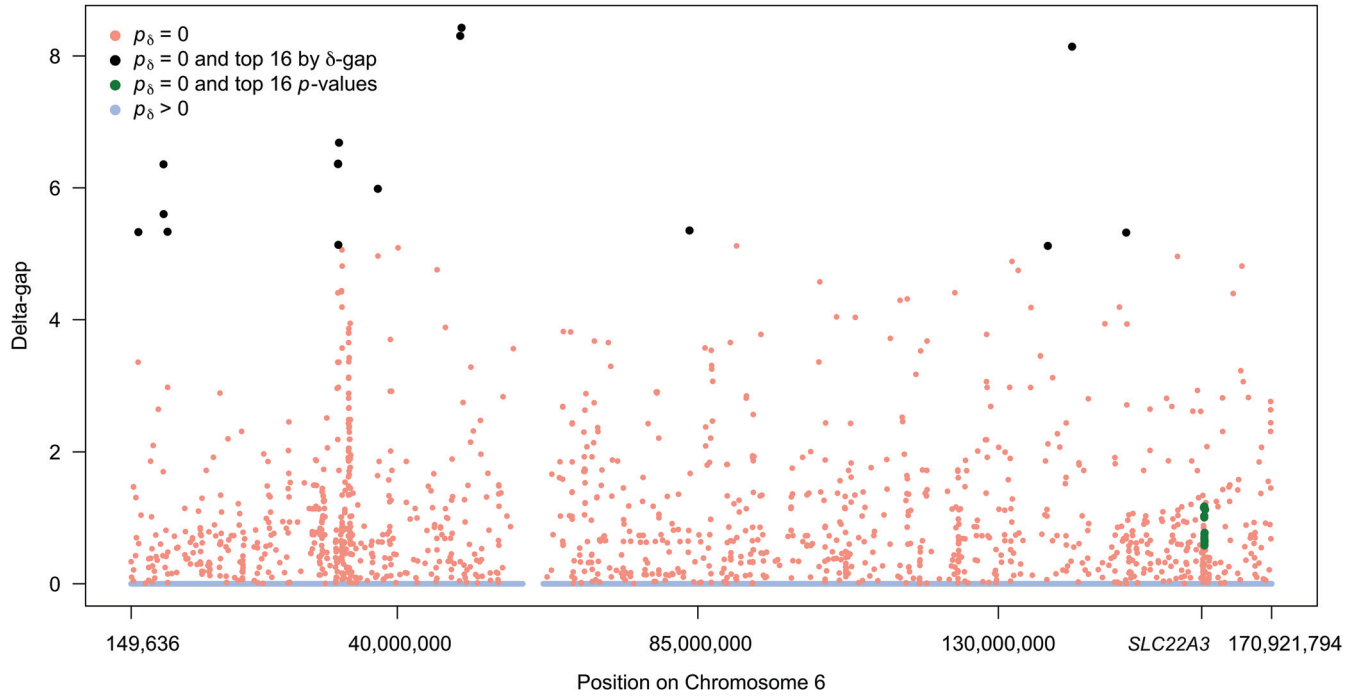


Figure 3. Modified Manhattan-style plot of SNP associations as measured by the delta-gap. Points are colored according to second-generation *p*-value status. The *x*-axis shows SNP position on chromosome 6 and the position of the organic cation transporter gene (SLC22A3). The top 16 associations by *p*-value rank are in green and the top 16 associations by second-generation *p*-value rank are in black.

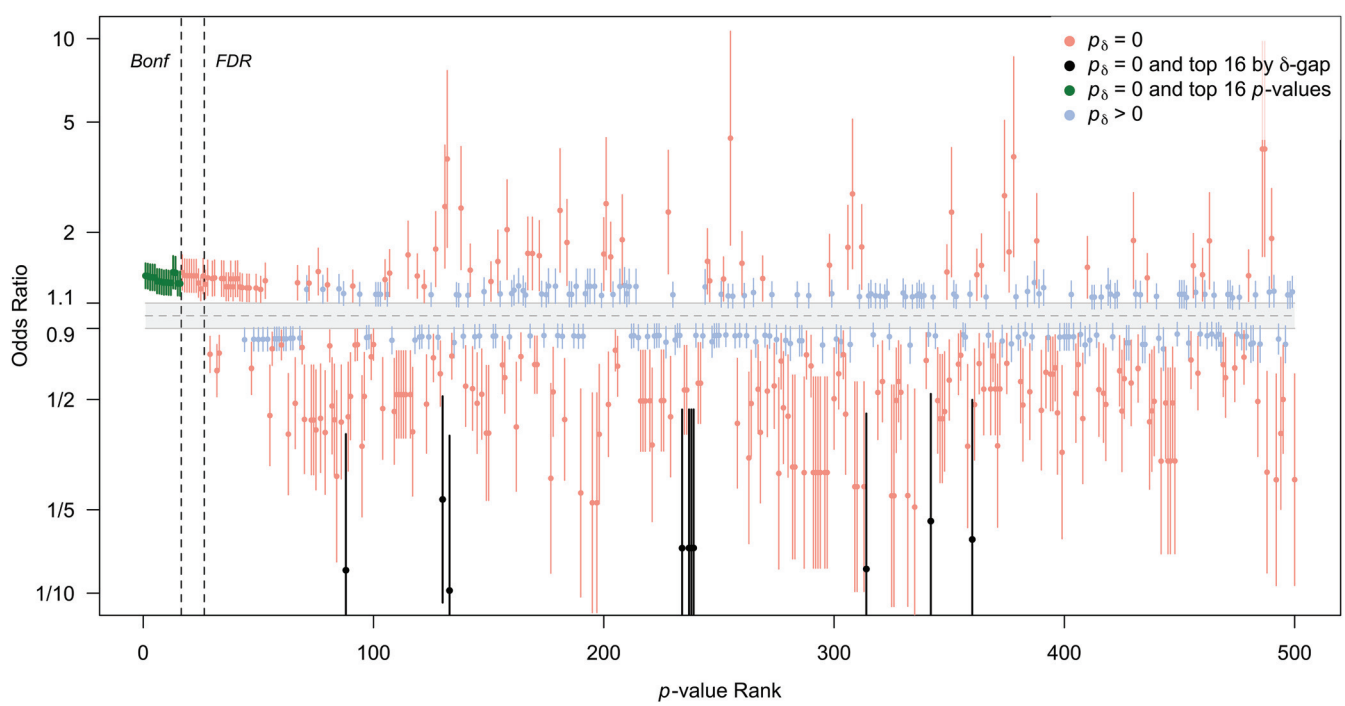


Figure 4. 1/8 likelihood SIs for the estimated odds ratio for SNPs with the smallest 500 *p*-values. The *x*-axis is *p*-value rank. The interval null hypothesis (in gray) ranges from 0.9 to 1.1111 (= 1/0.9). Intervals are colored according to second-generation *p*-value status. The top 16 Bonferroni SNPs are in green and the top 16 SGPV SNPs are in black. Bonferroni ("Bonf") and FDR significance levels are shown by dashed lines.

which displays the SNPs with the 500 smallest classical *p*-values, ranging from 5.01×10^{-11} (rank 1) to 0.0016 (rank 500). The SNPs that met Bonferroni and FDR criteria reflect highly precise estimates of modestly sized odds ratios. Other SNPs that have similar effect sizes and levels of precision, for example, those immediately to the right of the FDR cutoff line, are ignored. Figure 4 also shows that there are many similarly precise SIs that are denoted as inconclusive by the SGPV (blue) because some null effects have not been ruled out.

To further explore the differences in approaches, Figure 5 displays the top 2000 1/8 SIs as ranked by δ-gap when $p_\delta = 0$, and by the SGPV when δ-gap $= 0$ (hereafter referred to as SGPV rank). All 1,443 SNPs with a SGPV of 0 are sorted by their *delta-gap* and displayed in red. Intervals in blue have small but

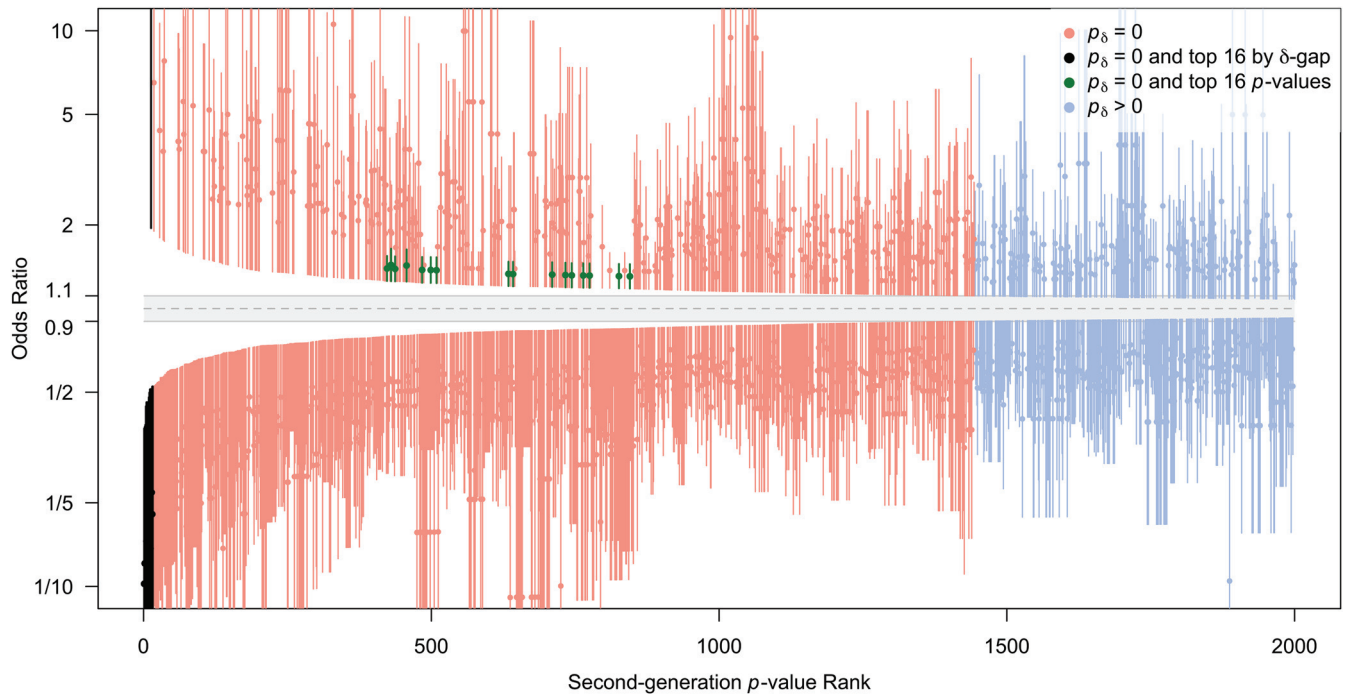


Figure 5. Mirrored waterfall plot. 1/8 likelihood SIs for the estimated odds ratio for SNPs in the top 2,000 second-generation p-value ranking. The x-axis is SGPV rank. The interval null hypothesis (in gray) ranges from 0.9 to 1.1111 (= 1/0.9). Intervals are colored according to second-generation p-value status. The top 16 Bonferroni SNPs are in green and the top 16 SGPV SNPs are in black.

inconclusive SGPVs between 0 and 0.07 (often resulting from a very wide SI that just slightly overlaps the null interval). Notice the 16 Bonferroni SNPs (green) are in the middle of the plot, far from the top findings by SGPV. The estimated effects from the top ranked SGPV SNPs are clearly very large, but not precisely estimated. We also see that the top 16 ranked SGPVs finds associations in both directions, while the Bonferroni and FDR approaches only find associations in one direction. This figure makes it clear that the SGPV prioritizes the observed strength of association in findings that are statistically significant.

It is helpful to see if any correspondence exists between the classical p-value and the SGPV. Figure 6 shows their cross-tabulation. Of the 11,844 SNPs with $p < 0.05$, there are 1,443 that have a SGPV equal to 0. Another 2,622 are effectively inconclusive with SGPVs between 0.4 and 0.6, of which 783 have a SGPV greater than 0.5, indicating very weak support for the null rather than the alternative hypothesis. There are 3,377 SNPs with a SGPV of 1, indicating no association between SNP and cancer status, and they are uniformly dispersed across the chromosome with observed classical p-values of 0.76 or larger. This apparent correspondence should not be over interpreted; it is due to the relatively narrow interval null. As the width of the interval null lengthens, this relationship disappears. For example, when the interval null is expanded to 0.8 to 1.25 (= 1/0.8), there are 83,749 SNPs with $p_\delta = 1$, and the classical p-values for these SNPs are evenly distributed from 0.007 to 1. This is not a typo; there are very precise intervals within the null region that exclude OR = 1 and these SNPs have very small classical p-values. But given their close proximity to OR = 1 they support the null premise better than alternative theories.

Another interesting aspect of Figure 6 is the large number of SGPVs at exactly 1/2. Many of these arise from logistic models with poor fit, often because there are one or more zero cells in the contingency table. Despite this, classical p-values near 1 are often still reported by software, whereas the SGPVs map these cases back to 1/2 (inconclusive). Of the 45,624 classical p-values greater than 0.9, almost half (47.6% or 21,697 cases) correspond to a SGPV of 1/2 because their point estimates were essentially undefined and estimated variances were extremely large. This reinforces the notion that large p-values do not imply the data support the null premise, even in large datasets.

For illustration, we computed several flavors of the FDR for 5 SNPs with a SGPV of 0. These are displayed in Table 2. Three versions of the FDR $= P(H_0|p_\delta = 0)$ were computed. Here θ is the log odds ratio and $\hat{\theta}$ is its MLE. FDR_1 uses the natural point null, $H_0 : \theta = \log(1)$, with the alternative set to the MLE, $H_1 : \theta = \hat{\theta}$. FDR_2 uses a point null hypothesis that is closer to the null boundary, $H_0 : \theta = \log(1.05)$, with a prespecified alternative $H_1 : \theta = \log(2)$. FDR_3 uses the interval null $H_0 : \theta \sim \text{UNIF}[\log(0.9), \log(1/0.9)]$ and the interval alternative $H_1 : \theta \sim \text{UNIF}[\hat{\theta}_l, \hat{\theta}_u]$ where $\hat{\theta}_\text{l}$ and $\hat{\theta}_\text{u}$ are the lower and upper bound of the 1/8 SI. FDR computations follow the formulas in Section 5 and Remark E.

We see from these results that a decent first-order approximation is given by FDR_1, which is also straightforward to compute. All of the FDRs are low, as is expected from a SGPV = 0 finding (Remark H). Some are low because their 1/8 SI is shifted far from the null interval (2nd row), and some are low because their 1/8 SIs is very narrow (4th row, also the top finding by classical p-value) despite being only slightly shifted off the interval null. These calculations also illustrate why using the FDR alone is not best practice for ranking or screening findings. That strategy downplays the scientific importance of the estimated effect size and is sensitive to how the FDR is defined (as evidenced by $\text{FDR}_2 = 30.3\%$ in row 2). Nevertheless, the FDR is essential

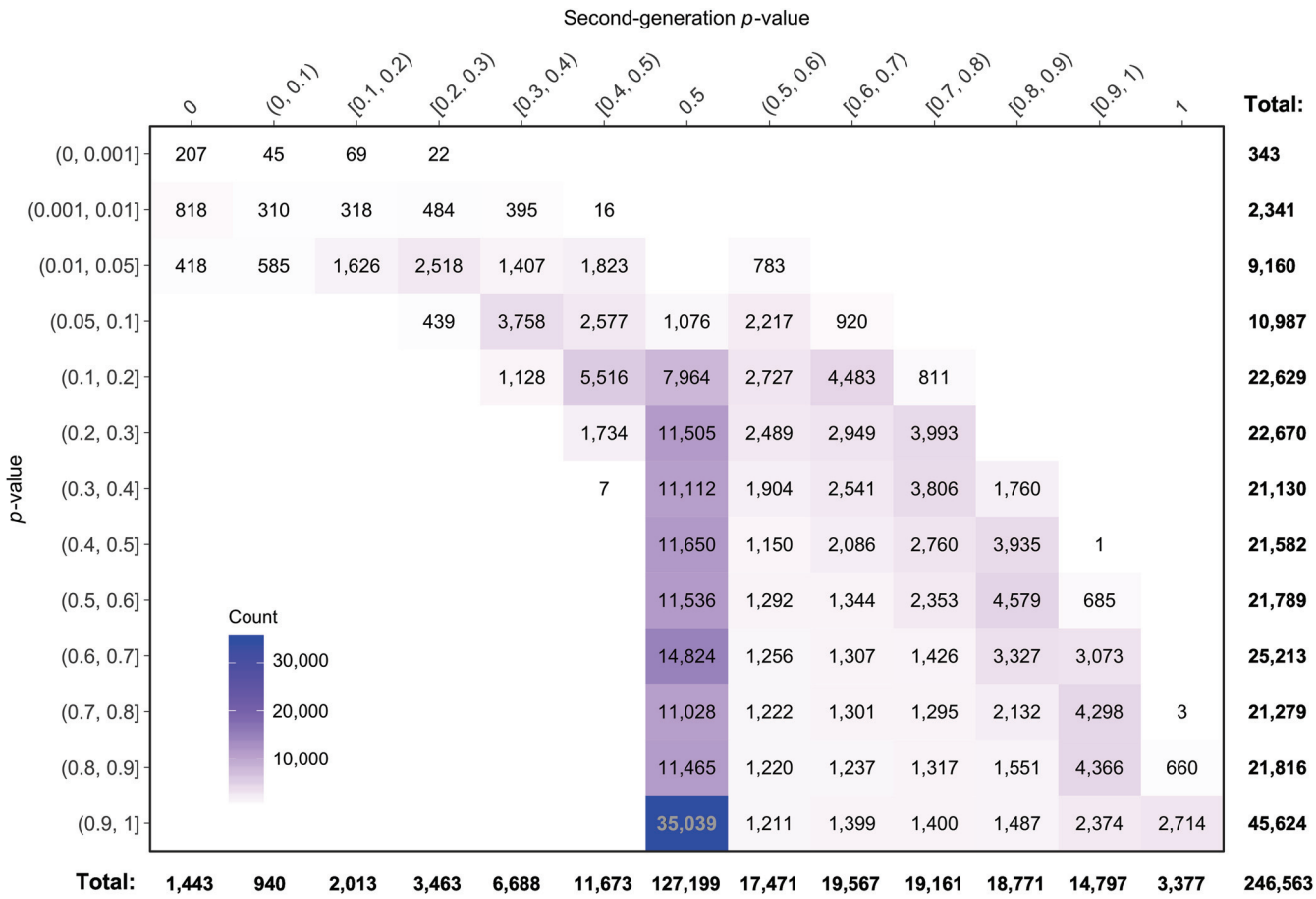


Figure 6. Cross-tabulation of classical *p*-values and second-generation *p*-values in the prostate cancer data example.

Table 2. False discovery rates of 5 SGPV = 0 findings computed under various null and alternative hypothesis configurations.

SNP ID	SGPV rank	*p*-Value rank	Odds ratio	1/8 SI lower limit	1/8 SI upper limit	FDR_1	FDR_2	FDR_3
kgp4568244_C	1	133	0.10	0.03	0.37	2.9%	17.1%	3.3%
kgp8051290_G	13	2002	15.58	1.95	124.68	4.3%	30.3%	4.9%
kgp4497498_A	28	255	4.37	1.80	10.64	2.5%	8.6%	3.1%
rs3123636_G	423	1	1.39	1.26	1.55	0.004%	0.1%	0.4%
kgp7460928_G	1443	3310	1.78	1.11	2.87	2.4%	2.0%	3.0%

NOTE: The prior probability ratio, *r*, is set to 1.

when gauging the potential for a finding to be mistaken or misleading.

Lastly, an example global summary of the results is as follows:

Of the 246,563 SNPs examined in this analysis, 1,443 were found to have strong evidence of a meaningful association with cancer status (data supported ORs > 1.11 *or* < 0.9*, SGPV* $= 0$*). Another 3,377 SNPs were found to have practically no association with prostate cancer (*$0.9 < ORs < 1.11$*, SGPV* $= 1$*). The remaining 241,743 SNPs were inconclusive to varying degrees. There were 940 SNPs in which the data were suggestive of a meaningful association with cancer status but were unable to rule out trivial effects (*$0 < SGPV < 0.1$*). Similarly, there were 14,797 SNPs in which the data were suggestive of no association, but not strong enough to rule out meaningful effects (*$0.9 < SGPV < 1$*). The remaining 226,006 SNPs were effectively inconclusive (*$0.1 < SGPV < 0.9$*).*

A detailed summary for individual SNP findings might be:

SNPs kgp4568244_C (1/8 SI for OR: 0.03–0.37, SGPV = 0) and kgp8051290_G (1/8 SI: 1.95–124.7, SGPV = 0) were the top ranked protective and harmful SNPs. SNPs rs9257135_G (1/8 SI: 0.92–1.09, SGPV = 1) and kgp10695421_A (1/8 SI: 0.91–1.1, SGPV = 1) were found to be unassociated with cancer status. SNP kgp17007117_A (1/8 SI: 1.08–3.6, SGPV = 0.07) was suggestive of a relationship, while findings for SNP kgp8948004_A (1/8 SI: 0.96–1.20, SGPV = 0.65) were inconclusive.

7. Comments and Remarks

The second-generation p-value (SGPV) provides the inferential properties that many scientists wish were attributes of the classic *p*-value. In addition, they are easy to interpret, they retain excellent frequency properties, they provide error rate control,

and they have reduced false discovery rate (FDR). The prostate cancer example shows how SGPVs can be much more informative than classical p-values. Moreover, we saw that the Bonferroni procedure tends to find modest odds ratios for common variants, while SGPVs tend to identify the strongest associations that are both statistically significant and scientifically meaningful. This behavior was also observed in the high-dimensional example used in Blume et al. (2018). Changing culture so that we routinely specify internal null hypotheses before the study is conducted, as opposed to after, will be challenging. However, the rewards would be substantial; manifesting as increasing statistical reliability and increased scientific reproducibility of observed results. Prominent use of the second-generation p-value would encourage that type of culture shift, and both science and statistics would be better off for it.

Remark A. In practice the best we can do is to encourage prespecification of the null premise. Some scientific domains are better at this than others. For example, in clinical trials the study protocol is posted to clincialtrials.gov before the study enrolls patients. Nevertheless, the specification of the null premise is always an important scientific issue and recognizing that this should occur before analyzing the data is an important step forward for reproducibility and rigor.

Remark B. Although the coverage probability of the interval estimate is not represented symbolically on p_δ, it plays a major role in determining the frequency properties of p_δ, that is, the error rates and FDRs. The connection is detailed in later sections. It is a key incentive to use intervals with good frequency properties (unbiasedness, shortest expected length, robust coverage, etc.) and to be transparent about why a certain type of interval was chosen.

Remark C. For parameter θ, likelihood function $L(\theta)$ and MLE $\hat{\theta}$, the $1/k$ likelihood SI is the set $\{\theta : L(\theta)/L(\hat{\theta}) \geq 1/k\}$. A SI is the collection of parameter values under the "hump" of the likelihood function when scaled by its maximum; these parameter values have nearly the same likelihood value as the maximum likelihood estimator (MLE) itself, and are "close" in that sense. When $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ and σ^2 is known, the $1/k$ SI is $\bar{X}_n \pm \sigma\sqrt{2\log(k)/n}$. When $k = 8, \sqrt{2\log(k)} = 2.04$, and the 1/8 SI is equivalent to an unbiased 95.9% CI with shortest expected length. An advantage of SIs is that they often correspond to CIs that are desirable in some sense, for example, shortest expected length or unbiasedness. This discourages the use of nonstandard CIs that have the same coverage probability. See Royall (1997) and Blume (2002) for more on SIs.

Remark D. SGPVs of 0 and 1 are intended to be clear endpoints. They do not imply overwhelming precision, but rather that the accumulated data support only meaningful alternative hypotheses or only practically null hypotheses. The label "suggestive" can be attached to SGPVs between 0 and 0.1 or 0.9 and 1, to describe data that may warrant further attention for ancillary reasons (e.g., limited resources). The key point, The key point is that SGPVs are summary measures to be used for facile communication of results; they should not replace examination of the interval estimate in the context of scientific discussions and policy decisions.

Remark E. The power function for the SGPV is $P_\theta(p_\delta = 0) = \Phi\left[\frac{\sqrt{n}(\theta_0-\delta)}{\sqrt{V}} - \frac{\sqrt{n}\theta}{\sqrt{V}} - Z_{\alpha/2}\right] + \Phi\left[-\frac{\sqrt{n}(\theta_0+\delta)}{\sqrt{V}} + \frac{\sqrt{n}\theta}{\sqrt{V}} - Z_{\alpha/2}\right]$ when the interval null is symmetric about θ_0. When $\theta = \theta_0$, this reduces to $P_{\theta_0}(p_\delta = 0)$ in (2). This expression is needed for computation of FDRs.

Remark F. The 246,563 SNPs investigated here were informative and not in strong local linkage disequilibrium. The Bonferroni correction is overly conservative when there is strong linkage disequilibrium (i.e., high correlation) among SNPs. To reduce this problem slightly, we identified bins of contiguous SNPs with an $R^2 > 0.98$ and eliminated all but one SNP from each bin. The number of variant alleles was treated as a continuous measure in the logistic models, which is standard practice.

Remark G. SNP position was recorded as the number of nucleotides from the start of the p-arm of chromosome 6 based on reference assembly GRCh37/hg19.

Remark H. FDR rates that result from a screening procedure based on p-values will be larger than those computed based on SGPV screening. For example, SNP rs3123636_G has the smallest p-value (4th row, Table 2), but its FDR under classical p-value screening inflates over 10-fold to $\sim$ 3.9% for all three FDR flavors. Inflation occurs even when a multiple comparison correction, such as Bonferroni, is applied. Blume et al. (2018) provide details and an example.

Remark I. A link to a repository of R functions for computing second generation p-values and associated quantities can be found on *www.statisticalevidence.com*, GitHub *https://github.com/weltybiostat/sgpv*, or by contacting the corresponding author.

References

Barnard, G. (1949), "Statistical Inference," *Journal of the Royal Statistical Society*, Series B, 11, 115–149. [157]

Bayarri, M. J., Benjamin, D. J., Berger, J. O., and Selke, T. M. (2016), "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses," *Journal of Mathematical Psychology*, 72, 90–103. [157]

Benjamin, D. J., Berger, J. O. et al. (2018), "Redefine Statistical Significance." *Nature Human Behavior*, 2, 6–10. [157]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [158]

Berger, J. O. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science*, 18, 1–32. [159]

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [157]

Berkson, J. (1942), "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, 37, 325–335. [157]

Blume, J. D. (2002), "Likelihood Methods for Measuring Statistical Evidence," *Statistics in Medicine*, 21, 2563–2599. [157,160,166]

——— (2011), "Likelihood and Its Evidential Framework," in *Handbook of the Philosophy of Science: Philosophy of Statistics*, eds. D. M. Gabbay and J. Woods, San Diego, CA: North Holland, pp. 493–511. [158,159]

Blume, J. D., McGowan, L. D., Greevy, R. A., and Dupont, W. D. (2018), "Second-Generation *p*-Values: Improved Rigor, Reproducibility, & Transparency in Statistical Analyses," *PLoS One*, 13, e0188299. [157,159,160,161,166]

Blume, J. D., and Peipert, J. F. (2003), "What Your Statistician Never Told You About *P*-Values," *Journal of the American Association of Gynecologic Laparoscopists*, 10, 439–444, PMID: 14738627. [157,159]

Cohen, J. (1994), "The Earth Is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [157]

Cornfield, J. (1966), "Sequential Trials, Sequential Analysis, and the Likelihood Principle," *The American Statistician*, 20, 18–23. [157]

Cristea, I. A., and Ioannidis, J. P. A. (2018), "*P*-Values in Display Items Are Ubiquitous and Almost Invariably Significant: A Survey of Top Science Journals," *PLoS One*, 13, e0197440. [157]

Dupont, W. D. (1983), "Sequential Stopping Rules and Sequentially Adjusted *p*-Values: Does One Require the Other?" (with discussion), *Controlled Clinical Trials*, 4, 3–35. [157]

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242. [157]

Fisher, R. A. (1959), *Statistical Methods and Scientific Inference* (2nd ed.), New York: Hafner. [159]

Good, I. J. (2007), "C420. The Existence of Sharp Null Hypotheses," *Journal of Statistical Computation and Simulation*, 49, 241–242. [159]

Good, I. J., and Osteyee, D. B. (1974), *Information, Weight of Evidence, the Singularity Between Probability Measures and Signal Detection*, Lecture Notes in Mathematics, New York: Springer-Verlag. [157]

Goodman, S. N.(1993), "*p*-Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496. [157]

Greenland, S., Senn, S., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [157]

ICPCG (2018), International Consortium for Prostate Cancer Genetics, Genome Wide Association Study of Familial Prostate Cancer, available at *https://www.icpcg.org/*; dbGaP Study Accession: phs000733.v1.p1, available at: *https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000733.v1.p1*. [162]

Ioannidis, J. P. A. (2018), "The Proposal to Lower *P*-Value Thresholds to 0.005," *Journal of the American Statistical Association*, 319, 1429–1430. [157]

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110, 19313–19317. [157]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [157]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., and Buchanan, E. M. (2018), "Justify Your Alpha," *Nature Human Behaviour*, 2, 168–171. [157]

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley [1st ed. (1959)]. [159]

Morrison, D. E., and Henkel, R. E. (1970), *The Significance Test Controversy*, Chicago, IL: Aldine. [157]

Royall, R. M. (1986), "The Effect of Sample Size on the Meaning of Significance Tests," *The American Statistician*, 40, 313–315. [157]

——— (1997), *Statistical Evidence: A Likelihood Paradigm*, London: Chapman and Hall. [157,159,160,166]

Savage, L. J. (1962), "The Foundations of Statistics Reconsidered," in *Studies in Subjective Probability*, eds. H. E. Hyburg Jr. and H. E. Smokler, New York: Wiley. [157]

Schaid, D. J., and Chang, B. L. (2005), "Description of the International Consortium for Prostate Cancer Genetics, and Failure to Replicate Linkage of Hereditary Prostate Cancer to 20q13," *Prostate*, 63, 276–290. [162]

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004), *Bayesian Approach to Clinical Trials and Health-Care Evaluation*, West Sussex, England: Wiley. [157]

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498. [158]

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value," *Annals of Statistics*, 31, 2013–2035. [158]

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004), "Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies," *Journal of the National Cancer Institute*, 96, 434–442. [158]

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. [157]

The American Statistician

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

A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting its Use

William M. Goodman, Susan E. Spruill & Eugene Komaroff

To cite this article: William M. Goodman, Susan E. Spruill & Eugene Komaroff (2019) A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting its Use, The American Statistician, 73:sup1, 168-185, DOI: 10.1080/00031305.2018.1564697

To link to this article: https://doi.org/10.1080/00031305.2018.1564697

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 5711

View related articles

View Crossmark data

Citing articles: 14 View citing articles

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, 168–185: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1564697

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting its Use

William M. Goodman[a], Susan E. Spruill[b], and Eugene Komaroff[c]

[a]University of Ontario Institute of Technology, Ontario, Canada; [b]Applied Statistics and Consulting, Spruce Pine, NC; [c]Graduate School, Keiser University, Fort Lauderdale, FL

ABSTRACT

When the editors of Basic and Applied Social Psychology effectively banned the use of null hypothesis significance testing (NHST) from articles published in their journal, it set off a fire-storm of discussions both supporting the decision and defending the utility of NHST in scientific research. At the heart of NHST is the *p*-value which is the probability of obtaining an effect equal to or more extreme than the one observed in the sample data, given the null hypothesis and other model assumptions. Although this is conceptually different from the probability of the null hypothesis being true, given the sample, *p*-values nonetheless *can* provide evidential information, toward making an inference about a parameter. Applying a 10,000-case simulation described in this article, the authors found that *p*-values' inferential signals to either reject or not reject a null hypothesis about the mean ($\alpha = 0.05$) were consistent for almost 70% of the cases with the parameter's true location for the sampled-from population. Success increases if a hybrid decision criterion, *minimum effect size plus p-value* (MESP), is used. Here, rejecting the null also requires the difference of the observed statistic from the exact null to be meaningfully large or practically significant, in the researcher's judgment and experience. The simulation compares performances of several methods: from *p*-value and/or effect size-based, to confidence-interval based, under various conditions of true location of the mean, test power, and comparative sizes of the meaningful distance and population variability. For *any* inference procedure that outputs a binary indicator, like flagging whether a *p*-value is significant, the output of one single experiment is *not* sufficient evidence for a definitive conclusion. Yet, if a tool like MESP generates a relatively reliable signal and is used knowledgeably as part of a research *process*, it can provide useful information.

ARTICLE HISTORY
Received March 2018
Accepted December 2018

KEYWORDS
NHST; Minimum effect size plus *p*-value criterion; MESP; statistical evidence; meaningful distance; true power; true Type I error rate

1. Introduction

There has been ongoing discussion of the use and misuse of *p*-values ever since R.A Fisher introduced the concept in Statistical Methods for Research Workers (Fisher 1925). However, in February 2015, it seemed a line had been drawn in the sand when the editors of Basic and Applied Social Psychology (BASP) effectively banned the use of null hypothesis significance testing (NHST) in manuscripts submitted for publication (Trafimow and Marks 2015). Blogging websites lit up as the news was spread; some met it with fanfare (Siegfried 2015; Woolston 2015), while others stoically attempted to defend NHST's continued use (Nuzzo 2015; Leek and Peng 2015), albeit with the understanding that *p*-values alone are not adequate and are prone to misuses.

In March 2016, the American Statistical Association released a policy statement on *p*-values outlining their context, the process by which they should be used and their purpose in scientific literature and research (Wasserstein and Lazar 2016a, p. 129). In this statement the authors noted that "the statistical community has been deeply concerned about issues of *reproducibility and replicability* of scientific conclusions," and echoing Peng's (2015, p. 30) concerns, stated that the "misunderstanding or misuse of statistical inference is only one cause of the reproducibility crisis." The ASA's policy statement was met with mixed emotions from the statistical community at large. Shortly following the release of this policy statement, the publishers of *The American Statistician*, Taylor & Francis Group, made available a supplement of statements from 23 well-respected statisticians from around the world (Wasserstein and Lazar 2016b). Even those who still consider *p*-values a tool for separating results that warrant more study from those likely to be due to random chance generally concurred that *p*-values are useless if not assessed in the proper context and derived from properly designed studies. Recently added to this literature have been three multiauthor position articles, whose signatories either recommend (Benjamin et al. 2017) or oppose (Trafimow et al. 2018, Lakens et al. 2018) salvaging *p*-values by lowering the common $P < 0.05$ criterion to $P < 0.005$.

While some researchers favor using alternative statistics, such as confidence intervals, effect sizes or Bayes factors, there is evidence to support that these methods are not that different and often provide similar interpretations of uncertainty (Wetzels et al. 2011). However, there are cases where statistical methods do not clearly align. Closer inspection of these cases

CONTACT William M. Goodman bill.goodman@uoit.ca University of Ontario Institute of Technology, 2000 Simcoe Street North, Ontario, L1H 7K4, Canada.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.
Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/TAS.

© 2019 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.

is likely to reveal problems with sample size, study design or implementation of protocols. Therefore, the problem starts with research design and continues through educating non-statisticians on how to interpret and present the volumes of information that result from analyzing their research data. These requirements do not disappear if effect size measures like Cohen's (1988) *d* are utilized—stripped of references to their "significance." What does this all mean for researchers going forward?

The response of this article is to step back from discussions on theoretical grounds about *p*-values' relative utility (if any) as evidence for a hypothesis, and to seek *empirical evidence* to help address the issue. How often does it occur in practice that a null hypothesis (H_0) that is really consistent with the population parameter is rejected by conventional testing? And what factors affect that error rate? Conceptually, the question "How often are true null hypotheses rejected?" is familiar to statisticians, as asking for a *Type I error rate*. However, that rate is usually not derived from observations but is calculated based on abstract models; that is, the calculations do not compare the true values of the parameters of interest with corresponding *p*-value-based inferences for what those values might be.

This omission of what we have called using empirical evidence to assess *p*-value's success is understandable, because in practice, a researcher seeks the value of a parameter when it is not known and likely not directly observable. He or she can hypothesize, collect data, and compare samples to hypotheses, while other researchers do the same. Yet none of them has privileged access to the objectively real answer, to definitively confirm their own or others' work. Excellent articles by Howard et al. (2000) and Wetzels et al. (2011) compare results by different methods such as conventional NHST, meta-analysis of replicated studies, and Bayesian techniques, when applied to the same sets of data. Their implication is that various method's results could be triangulated to maximize confidence in one's findings. This approach is certainly worth considering, yet those articles still do not directly compare methods' outputs, collectively or individually, to the actual population parameters. To get around the verification problem described here, this article adapts a simulation method proposed by Goodman (2010), which provides a way that the actual values of parameters can be known and directly compared with inferences made about those parameters pursuant to hypothesis testing—or other methods—applied to simulated samples.

Simulations are also used in a recent, thought-provoking article by Krueger and Heck (2017) to assess the success of *p*-value-based inferences. However, their approach does not directly compare inference results for the simulated cases against the cases' corresponding true parameters. They assess what they call "hits" and "misses" (p. 5) through comparing each case's simulated *p*-value-based inference (to reject the null or not) with the corresponding *posterior probability* that the case's null hypothesis is false or not. But the posterior probability itself is an inference, with the *p*-value being an input to it; so, their method does not present a direct comparison of *p*-value-based inferences with the corresponding true values of the parameters. (The authors acknowledge that Krueger and Heck's simulation generates, when setting up each iteration, population values to sample from; but it is not clear in their article if the true parameters remain stored and accessible for making direct comparisons with the inferences.)

The simulation described in this article has three interrelated goals:

(1) The first goal is to explore whether *p*-values can have evidential value. According to Trafimow et al. (2018, p. 2) they cannot: *P* values give the "probability of the finding, and of more extreme findings, given that the null hypothesis and all the other assumptions about the model were correct…., and one would need to make an *invalid inverse inference* [italics added] to draw a conclusion about the probability of the null hypothesis given the finding."

This definitively worded claim is refutable, both logically and empirically. Krueger and Heck's (2017, p. 11) article addresses the logical claim: Under plausible assumptions, the *p*-value does give information that *is* relevant to assessing the posterior probability of the hypothesis which they symbolize as (P(H|D)) While it may be imprudent to rely *only* on *p*-values, ignoring other factors, or to *infer a specific probability value* for P(H|D), it is not logically invalid to incorporate a *p*-value as a premise or step in the process of making an inference about a hypothesis An empirical exploration of the question is described in this article: It is observed that *p*-values can and *do* provide evidential information that is relevant for making an inference—which, again, does not mean their stand-alone application to this end would be fool-proof or advisable, or that there could not be better alternatives.

(2) If *p*-values do have demonstrated value as evidence about hypotheses, then the article's second goal is to explore the nature and limitations of that value. Demonstrating that *p*-values have evidential value is not to endorse a simplistic "bright line" interpretation, whereby just obtaining a *p*-value $< \alpha$ is sufficient information to support a definitive inference. Even assuming ideal test assumptions are satisfied (not a certainty in real research), such as having an unbiased, representative sample and no unacknowledged confounders that impact results, nonetheless the true rates of correct inference of a test can vary considerably depending on test power and other test considerations that this article will explore.

(3) Further, if *p*-values have evidential value, then the article's third goal is to explore how they compare in that respect with possible alternative approaches, including a hybrid approach introduced by the authors. Is one particular approach always the better or the worse one? Might there be trade-offs among methods' advantages and disadvantages, for different contexts or research goals? To implement this goal, the alternative methods' inferences are exposed to the same simulation-based, empirical assessments as are the *p*-value-based inferences. Details of the approaches being compared are elaborated in Section 2; but the intent is to plausibly represent four identifiable *types* of approaches that could be used to provide support for (though not definitively confirm) an inference about the true location of a population mean: A Conventional *p*-value based test (two tailed; "Is *p*-value < 0.05"); a Distance-Only assessment (compare asking: "Is Cohen's *d* large or small?"); an Interval-Based

approach (i.e., focused on the location of a sample based confidence interval); and lastly a proposed hybrid approach, called *minimum effect size plus p-value* (*MESP*), which is a hybrid of the *p*-value- and distance-based approaches. Clearly, the designed simulation could not try to model or assess every variation of the approaches mentioned; so, this article's design choices, and their potential impacts on the generality of its conclusions, are discussed in the *Model Assumptions and Limitations* (Section 4).

Before describing the study's methodology in detail, in Section 2, some preliminary concepts will be clarified: Namely, (in Section 1.1) meaningful (or "minimum practically significant") effect sizes; and (in Section 1.2) distinguishing an inference technique's inference indication or decision, based on one experiment (which is the focus for this article), from a definitive research conclusion about a parameter (which none of the inference techniques compared here can offer as direct output).

1.1. Minimum Practically Significant Distances

For continuous data in particular, it is mathematically impossible for a real parameter to exactly equal a point null hypothesis for that value. So, the question arises: How close to exactly equal must H_0 be to the true parameter to say that H_0 is true, or is consistent with the true parameter?

Some writers suggest that null hypotheses would better be viewed as having thickness or width (Berger and Delampady 1987). In fields such as Psychology, a width-less hypothesis value may sometimes be difficult to formulate (Nunnally 1960; Meehl 1967; Berger and Sellke 1987; Chow 1988; Folger 1989). Blume et al. (2018, p. 3 of 17) suggest replacing the conception of H_0 as an *exact single* value, with the idea of a range of values they call the "null interval" that contains all values that would be considered practically or scientifically equivalent to equaling the null. Proponents of NHST also acknowledge that small differences from the null may not be of practical importance or meaningful; so, for example, the calculations for minimum sample size to detect a significant difference include an input for a minimum distance—smaller distances being considered not meaningfully distinguishable from equaling the null.

There is no fixed answer for how large a difference must be from the null to be considered meaningful; this can depend on context for the particular study. Ellis (2010, p. 35) observes in his *Guide* on effect sizes that "in the right context even small effects may be meaningful," if for example they could "trigger big consequences." Blume et al. (2018, p. 3) suggest that their null interval's bounds should be "constructed by incorporating information about the scientific context—such as inherent limits on measurement precision, clinical significance or scientific significance."

To reflect this issue, a term is introduced in this article called *Minimum Practically Significant Distance* (MPSD). In the Methods section that follows, the MPSD represents, for each pass of the simulation, the value that the simulated study's researchers would deem the smallest observed distance from equaling exactly the null that could be considered meaningfully large. It is presumed in the simulation that the value would be decided upon by the researchers in good faith, and with competence in their field of expertise. (For discussion of possible risks of, for example, MPSD "hacking", see in the *Model Assumptions and Limitations* section (Section 4).)

Note that the inference-assessing role of the MPSD for this simulation study is distinct from the MPSD's role, if any, within any particular inference methods. In the simulation, the true population parameter's value, μ, is considered not meaningfully different from equaling an exactly specified null hypothesis value, H_0, if and only if μ falls in this range:

$$\textit{Thick Null}: \ (\mathbf{H_0} - \mathbf{MPSD}) \leq \mu \leq (\mathbf{H_0} + \mathbf{MPSD})$$

In this article, *thick null* refers to the range of parameter values that would be deemed not meaningfully different from equaling the exactly specified point null value.

A similar range may be constructed by specific inference methods, as a criterion for their inferences. For example, the interval-based and MESP methods make explicit reference to an interval around the null mean, which could be depicted as follows in interval notation:

$$\textit{Null Interval}: \ [(\mathbf{H_0} - \mathbf{MPSD}), \ (\mathbf{H_0} + \mathbf{MPSD})]$$

The *null interval*, unlike the thick null, is not explicitly stating a hypothesis about the true parameter. But a method's *procedure* may compare and draw conclusions from the relative positions of the constructed null interval range and the observed value of the sample mean. A conventional *p*-value-based test, on the other hand, does not formally consider null thickness, though in practice a researcher might be sensitive to that consideration.

1.2. Inference Methods' Outputs as Indicator Signals

A valid concern about the *use* of *p*-values addresses what some call "bright line", binary interpretations of what they signify (Wasserstein and Lazar 2016a, p. 131). The present authors agree that conducting a single test that just barely passes a bright line criterion (such as p-value < 0.05), without replication and without thoroughly checking its design assumptions, is not appropriate to count as significantly confirming a conclusion about a population parameter. In this sense, binary interpretations of *p*-values are problematic.

But that consideration does not mean that any binary output, per se, is necessarily objectionable. While a single result of p-value < 0.05 does not justify drawing a final conclusion, it could still have value as what Krueger and Heck (2017, p. 1) call a "heuristic cue," within a larger process for drawing an inference. An indicator would have such value *if* there is a confirmed, *general* association between its outputs and the corresponding true values of the parameter, such that, obtaining a specific signal would provide relevant but not definitive evidence in support making an inference about the true value.

2. Methods

The simulation applies to just one-sample tests for the mean, but the same methods could be adapted for other testing situations. Key details are described in this section; supplemental discussions relating to design decisions are also found in Sections 1.1, 1.2, 4, and Appendices A1 and A2.

2.1. Set-up Values and Sampling for Each Pass

2.1.1. Set-up Values

Prior to running each iteration (*pass*) of the simulation, set-up values were randomly generated for each of four main elements that impact the next sample and its test results: $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, **n**, and **MPSD**. These elements are described here, along with boundary ranges for their random specifications (shown in italics). *Ranges* of possible values were used when setting up these elements for each pass, so that sensitivity of the test results to these factors and their interactions could be assessed.

- Two true parameters for an actual, normally distributed population: the mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$). These values are henceforth referred to as the *real mean* and *real standard deviation* respectively. ($75 \leq \mu \leq 125$; $4 \leq \sigma \leq 60$)
- Size $\boldsymbol{n}$ for a random sample to be drawn from the population whose true parameters are the real mean μ and the real standard deviation σ ($5 \leq n \leq 100$)
- A value for *Minimum Practically Significant Distance* (**MPSD**). MPSD is the minimum effect size (distance between the exact null mean and the observed sample's mean) that a simulated researcher conducting that pass would consider, reflecting competent professional practice, sufficient to deem the effect size scientifically or practically meaningful or non-trivial. ($2 \leq MPSD \leq 20$)

For each pass of the simulation, the set-up values for each of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, **n**, and **MPSD** were determined independently by randomly selecting an integer from that set-up value's corresponding range of equiprobable integer values that was prespecified. For example, the value **n** determined for any given pass was equally likely be any integer in the range from 5 to 100, inclusive.

The ranges specified for these set-up values were intended to ensure that statistically interesting and realistic scenarios were generated; but this could not be foolproof. For example, the true means were designed to never be farther than 25 from the null hypothesis (always equal 100); yet, for random combinations of large n's and small σ's, even a distance of 25 could represent many standard errors from the null. Would a researcher seriously test a null hypothesis that far off base? Such unusual combinations, however, were found to be relatively infrequent among the simulated cases actually generated (see Appendix A1); and no attempt was made to screen the set-up values or their combinations that were generated.

2.1.2. Sample Statistics

Once the set-up values for $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, **n**, and **MPSD** were generated for a pass, a random sample of data was taken for the pass. The sample was drawn by making $\boldsymbol{n}$ independent random selections of values from a presumed-normally distributed population having the mean $\boldsymbol{\mu}$ and the standard deviation $\boldsymbol{\sigma}$. From that sample, these statistics were calculated: sample mean ($\bar{\mathbf{x}}$) and sample standard deviation (**s**).

2.1.3. Null Hypothesis

For every pass in the simulation, it was presumed that a simulated researcher is assessing a hypothesis that $\mu = 100$. For the conventional tests, this represents a null hypothesis $\boldsymbol{H_0: \mu = 100}$, with the *null mean* equaling exactly 100. For the methods where a *null interval* is considered when making the inference, the null interval for each pass is constructed (for two-tailed tests) as the interval from (100 - MPSD) to (100 + MPSD).

2.1.4. Case Generation

A *case* refers in this article to one unique scenario, with its set-up values and sample statistics, produced by one pass of the simulation, together with its derived results, such as the observed distance of $\bar{x}$ from the null mean and a corresponding p-value. All the results and figures reported in this article are based on a dataset of 10,000 cases generated by the procedures described in this section.

2.2. Inference Indications by the Compared Methods for Each Pass

For each pass of the simulation, the population parameters and other elements generated for the case were recorded, and a corresponding, two-tailed conventional p-value was calculated. Based on the data generated in the pass, the simulation identified the finding that would be returned by each of the several compared test approaches, as to whether a hypothesis that the population mean equals the null mean should be rejected.

For each case generated by the simulation, the true state of affairs (*fact*) about the parameter is compared with the corresponding inference indications (*decisions*) that would be produced by the methods listed below:

- *Fact: The true location of the population parameter relative to the thick null.* The *thick null* refers to the range of possible locations of the real parameter bounded by ((exactly equal 100) $\pm$ MPSD). *A method's decision is deemed consistent with the true location of the parameter (or correct) if the method rejects the null in a case where the real mean's value falls beyond the range of the thick null, or if the method does not reject the null in a case when the real mean's value falls within the range of the thick null*
- *Decision by conventional* p*-value method.* This decision is generated by calculations for a one sample, two-tailed t-test for the mean, with $\alpha = 0.05$. *The method indicates to Reject the null for cases where p-value < 0.05; otherwise the method indicates: Do not reject the null.* This approach's focus is on sampling error: It aims to avoid concluding that an observed distance from null mean is significant if the distance could be plausibly explained just by sampling error.
  - To experiment with NHST variations, *two additional p-value based decisions* were also considered*: For (a) a $P \leq 0.005$ criterion (two-tailed test), and (b) a onetailed test ($P < 0.05$)* (discussed in the Appendix) Separately, for each of these variations of a t-test for the mean, it is noted whether a test conducted by that particular method would indicate to reject or not reject the null.
- *Decision by Minimum Effect Size Plus p-value (MESP) method.* This decision is the output of calculations for a hybrid method proposed and recommended by the authors. *The method indicates to Reject the null for cases which satisfy*

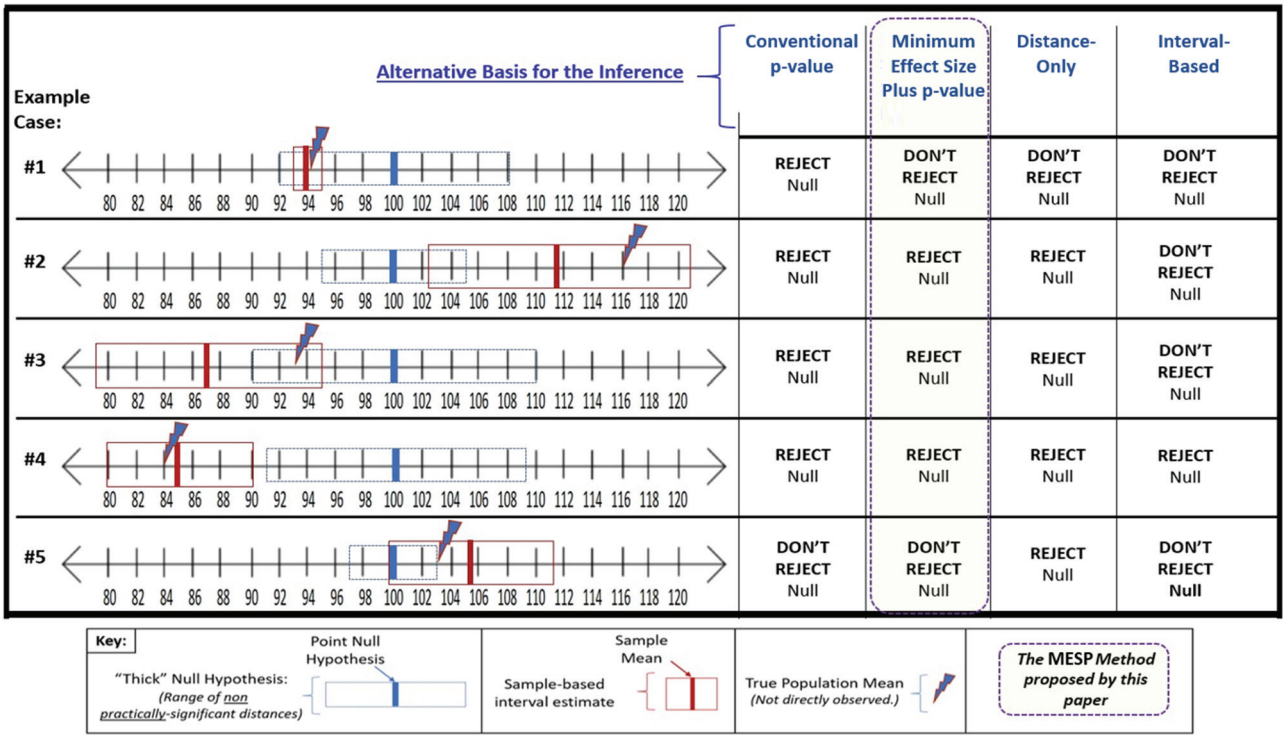


Figure 1. Example comparisons of methods' inferences with the actual parameter.

these two conditions: (1) p-value < α = 0.05 and (2) the observed effect size (i.e., absolute value of (sample mean – null mean)) ≥ MPSD; otherwise the method indicates: Do not reject the null. This method, with a proposed acronym ***MESP*** considers the meaningfulness of the observed effect size while also accounting for sampling-error.

- *Decision by Distance-Only method.* This method considers only the absolute value of the effect size without reference to sampling error. *The method indicates to Reject the null for cases where |observed effect size| ≥ MPSD; otherwise the method indicates: Do not reject the null.*
- *Decision by Interval-Based method.* This approach avoids conventional hypothesis testing, while being sensitive to sampling error, as well as (in some versions) thick null issues. *The method indicates to Reject the null only if there is no overlap between* these two intervals: *(1) the thick null interval bounded by (null mean ± MPSD), and (2) a 95% Confidence Interval, centred around the observed sample mean; otherwise the method indicates: Do not reject the null.*

More details about the algorithms used in this study for these approaches are provided in the *Supplemental file*: "Key Algorithms Implemented." In practice there can be many variations for implementing similar approaches. For example, both conventional and MESP tests could set a larger or smaller value for α ; a distance-only approach could be framed in terms of seeking a large Cohen's *d* value; and an interval-based approach could be implemented with different specific formulas for constructing the intervals around the null mean and/or the sample mean. Nonetheless, patterns very similar to those in Figure 1, illustrating some ways the four basic approaches described can compare with each other and with the (usually not known directly) real value of the parameter, would arise regardless of specific calculation details utilized. The five example cases illustrated in the figure were not generated in the main simulation run but were generated separately by the same algorithms for purposes of this illustration.

The real, but not directly observable, population mean for each example case is shown with a lightning bolt. The blue rectangle for each case (labeled "'Thick' Null Hypothesis" in the key) represents the range of values for the scenario that would be considered not meaningfully different from equaling the null mean, identified as the central blue line. Based on a case's random sample drawn from the true population, the sample mean is shown as a vertical red line, with a surrounding red rectangle to signify a sample-based confidence interval for the estimate. Example Case #4 illustrates what is needed for all four methods to output the same inference indication that the evidence goes against the null: The *p*-value is low, and the sample mean and confidence interval around the sample mean lie beyond the thick null interval. #5 illustrates a scenario where the sample mean is beyond the thick null, which triggers a correct decision to Reject by Distance-only, yet the methods looking for a low enough *p*-value or for no overlap of thick null with the sample confidence interval estimate, do not reject in this instance. On the other hand, Example Case #1 shows a *p*-value-based decision to reject that is inconsistent with the mean's true location (which is within the thick null).

3. Results

Figure 2 highlights a particular weakness of the *p*-value-based inference method compared to some alternatives, when the inference decision is to reject a null hypothesis that the population mean equals a specific value. If the true parameter is

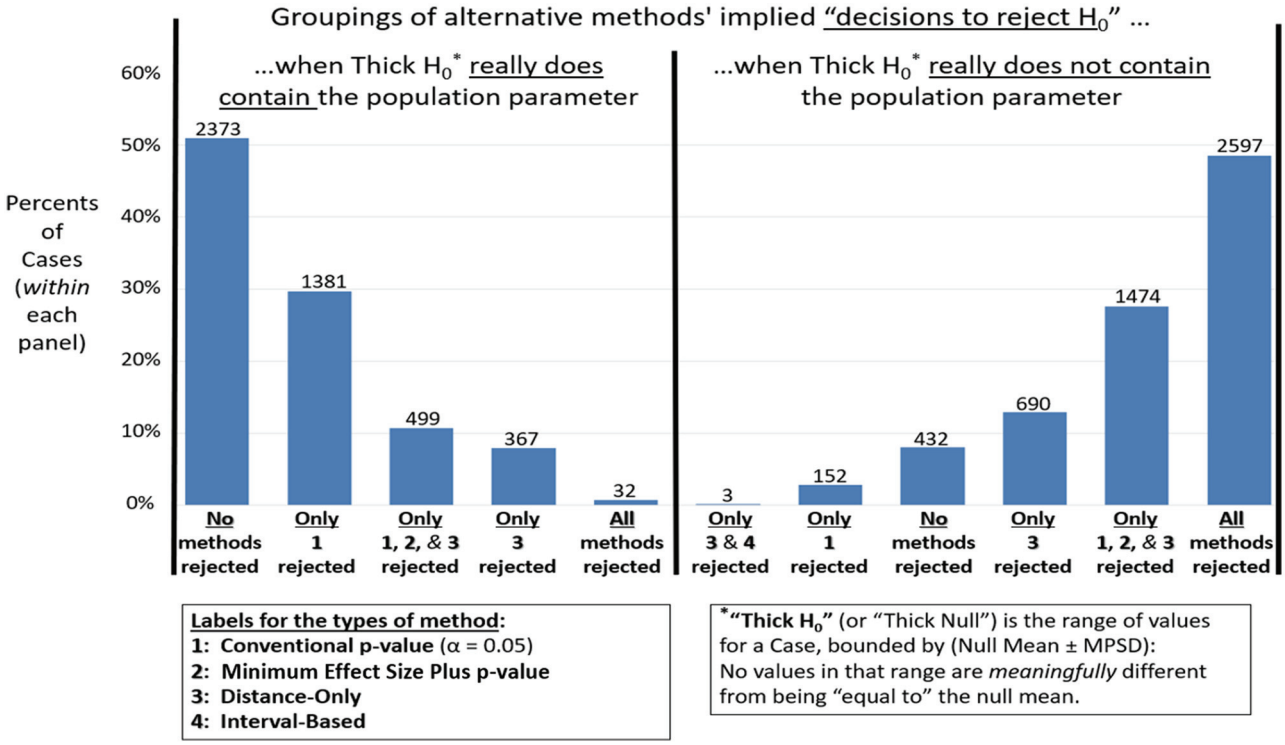


Figure 2. Inference-success rates for four alternative methods.

Table 1. Impacts of power and method on overall inference success rates.

		Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach and power				
Power[a]	Number of simulated cases	Conventional[b]	Small alpha[c]	MESP[d]	Distance-only[e]	Interval-based[f]
≥ 0.80	4214	63.2%	71.1%	91.2%	91.2%	84.3%
0.30 to 0.80	2791	81.5%	78.9%	83.2%	83.2%	69.5%
< 0.30	2995	67.6%	54.0%	67.6%	78.5%	57.8%
Numbers of cases	10,000	6963	6817	8192	8518	7220

[a]Nominal power, as calculated for a z-test for the mean, given these inputs: $\boldsymbol{\alpha} = 0.05$; minimum detectable difference = MPSD for the case; n = the simulated case's sample size; and $\boldsymbol{\sigma}$ = the true population sigma for the case
[b]Two-tailed, p-value based test, $\boldsymbol{\alpha} = 0.05$
[c]Two-tailed, p-value based test, $\boldsymbol{\alpha} = 0.005$
[d]Minimum effect size plus p-value approach proposed in this article.
[e]Directly compares effect size to MPSD
[f]Compares the thick null interval with a sample-based confidence interval for the mean

not meaningfully different from the null value (i.e., if the thick null is true), then a decision to reject the null would be in error. The left panel of Figure 2 shows that when the thick null was really true, the p-value-based method erred by rejecting it in 40% of the cases—often on its own (second bar) and sometimes it erred along with other methods (3rd and 5th bars). The alternative methods all performed better in this regard. The right panel shows that the p-value-based inferences were much more successful when the thick null was actually false: Alone, or with other methods, the conventional p-value's call to reject the null was correct in about 80% of the cases (represented in the 7th, 10th, and 11th bars).

These reported percentages were moderately sensitive to sample size, but much more influential was how large was the thick null range for the inference (determined by the Minimum Practically Significant Distance) relative to the variability in the underlying population. These sensitivities were explored further in Tables 1–3.

Table 1 compares the *overall* success rates of the alternative methods, controlled for the *nominal power* of the cases' inference conditions. Nominal power values were determined for each case in the simulation, based on conventional power calculations for a one-sample z-test for the mean, given: $\alpha = 0.05$, σ equal the case's true population standard deviation, n = the case's sample size, and minimum detectable difference equal to the case's MPSD. These power values are nominal, since even the p-value-based approaches were not z-tests (since researchers would not know the true σ), and the z-test power calculations do not apply directly to the other methods. Nonetheless, the power-based distinctions in the table reflect differences in sample sizes and relative MPSD and variance sizes, etc., that would be relevant in general for making successful inferences.

A fifth method, proposed by some to address the false discovery problem for novel discoveries, is included in the above comparison; this bases its results on a smaller-alpha version of a conventional t-test ($\alpha = 0.005$). (See Appendix A2 for

Table 2. Impact of power, method, and true location of the null on inference success.

Does the true location fall within the bounds of the "thick null"?	Power[a]	Number of simulated cases	Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach, power, and true location of the mean				
			Conventional[b]	Small Alpha[c]	MESP[d]	Distance-Only[e]	Interval-based[f]
Yes (4652 Cases)	≥ 0.80	2436	37.1%	52.7%	90.4%	90.4%	99.7%
	0.30 to 0.80	1325	76.7%	92.9%	82.9%	78.9%	99.5%
	< 0.30	891	92.0%	98.9%	92.0%	56.9%	98.0%
No (5348 Cases)	≥ 0.80	1778	98.9%	96.5%	92.4%	92.4%	63.1%
	0.30 to 0.80	1466	85.8%	66.3%	83.4%	87.2%	42.4%
	< 0.30	2104	57.3%	34.9%	57.3%	87.6%	40.7%
Numbers of cases:		10,000	6963	6817	8192	8518	7220

[a]Nominal power, as calculated for a *z*-test for the mean, given these inputs: $\alpha = 0{,}05$; minimum detectable difference = MPSD for the case; n = the simulated case's sample size; and σ = the true population sigma for the case
[b]Two-tailed, *p*-value based test, $\alpha = 0.05$
[c]Two-tailed, *p*-value based test, $\alpha = 0.005$
[d]Minimum effect size plus *p*-value approach, proposed in this article.
[e]Directly compares effect size to MPSD
[f]Compares the thick null interval with a sample-based confidence interval for the mean

Table 3. Impact of relative MPSD and method on inference success.

Does the true location fall within the bounds of the "thick null"?	Decile for MPSD in population standard deviations[a]	Number of simulated cases	Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach, relative MPSD, and true location of the mean				
			Conventional[b]	Small Alpha[c]	MESP[d]	Distance-Only[e]	Interval-Based[f]
Yes (4652 Cases)	1	184	91.8%	99.5%	91.8%	43.5%	96.2%
	2	259	90.0%	98.5%	90.0%	52.1%	99,2%
	3	359	85.0%	97.2%	85.2%	64.1%	99.4%
	4	429	82.8%	935%	86.2%	79.5%	99.1%
	5	532	75.9%	90.8%	85.2%	78.2%	99.2%
	6	524	66.0%	84.9%	84.9%	81.9%	99.4%
	7	580	56.6%	76.7%	86.2%	83.8%	99.3%
	8	579	49.9%	66.8%	88.1%	87.0%	99.5%
	9	596	35.2%	50.2%	93.0%	92.6%	99.8%
	10	610	16.6%	24.3%	95.2%	95.2%	99.7%
No (5348 Cases)	1	807	54.9%	34.3%	54.9%	91.6%	45.4%
	2	682	66.7%	43.3%	66.7%	90.3%	45.2%
	3	703	71.0%	50.6%	71.0%	85.8%	44.0%
	4	566	78.8%	61.3%	77.0%	86.2%	39.6%
	5	476	82.1%	64.3%	77.5%	85.1%	36.6%
	6	467	86.9%	74.9%	79.7%	83.9%	43.7%
	7	427	90.6%	79.6%	83.8%	88.1%	41.9%
	8	419	96.4%	90.9%	91.4%	92.1%	58.2%
	9	410	97.8%	93.7%	93.2%	93.9%	70.0%
	10	391	100.0%	98.7%	95.7%	95.7%	78.0%
Numbers of cases		10,000	6963	6817	8192	8518	7220

[a]These ranges of values for MPSD/σ correspond to the deciles:

1: 0.000–0.107	*6: 0.349–0.421*
2: 0.107–0.167	*7: 0.421–0.531*
3: 0.167–0.232	*8: 0.531–0.750*
4: 0.232–0.290	*9: 0.750–1.214*
5: 0.290–0.349	*10: 1.214–5.000*

[b]Two-tailed, *p*-value based test, $\alpha = 0.05$
[c]Two-tailed, *p*-value based test, $\alpha = 0.005$
[d]Minimum effect size plus p-value approach, proposed in this article
[e]Directly compares effect size to MPSD
[f]Compares the thick null interval with a sample-based confidence interval for the mean

more discussion on *false discovery rates* (FDRs).) Success rates based just on method can be calculated from Table 1 by dividing each method's case count by the total number of cases: 10,000. Distance-Only appears to perform the best by all these measures, followed by MESP. The authors caution, however, that depending on one's research goals and context, a researcher may be more concerned with the risks of (true) Type I error *or* of (true) Type II error. Table 1 masks those distinctions by averaging the two error risks into a single number.

Table 2 breaks down the results of Table 1, based on whether the true mean falls within the thick null or falls beyond the thick null. Figure 3 displays this information graphically.

In the following discussions, power refers to nominal power unless true power is specifically indicated. A method's *true power* is the proportion of really false thick nulls that the method correctly signals to Reject.

The various methods' true powers are revealed by the success rates in the bottom half of Table 2. In all these cases, rejecting

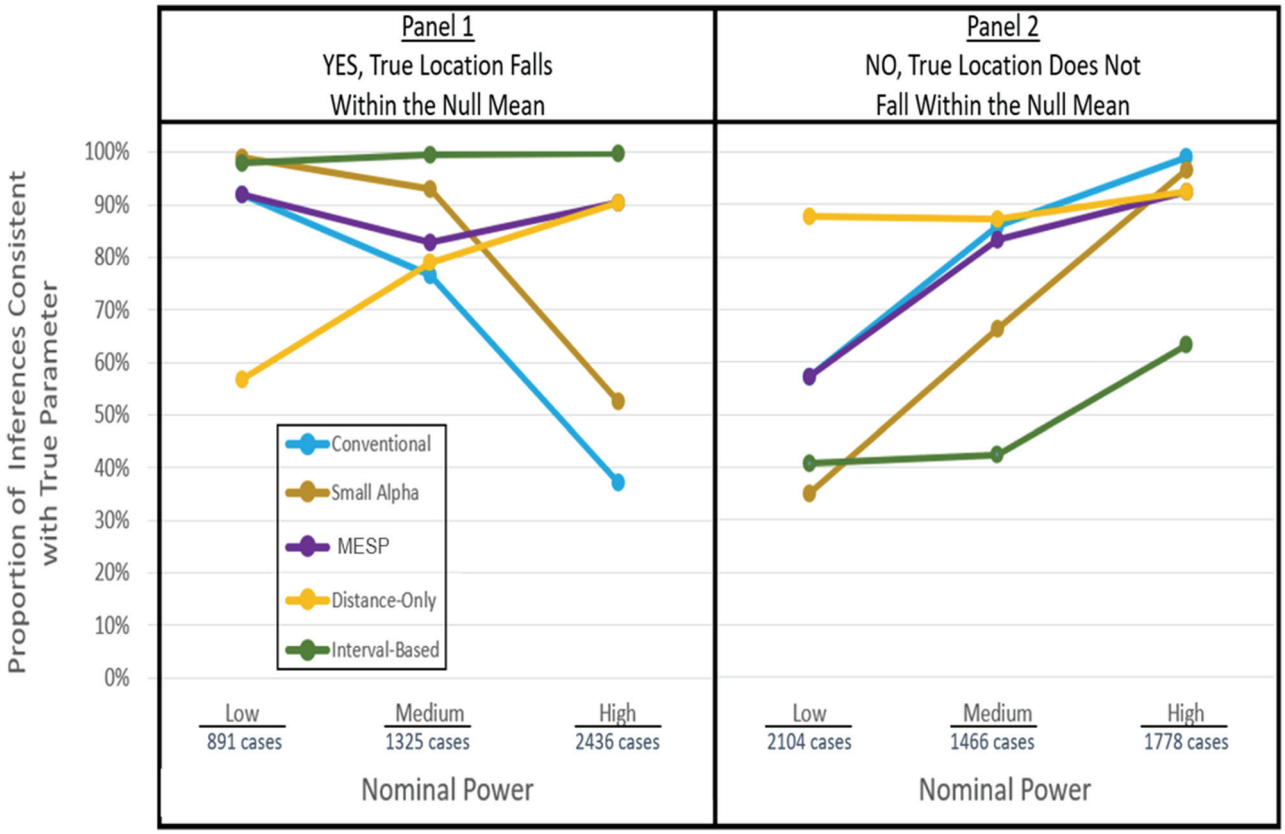


Figure 3. Graph of impact of power, method, and true location of the null on inference success.

the null is the correct inference, since the real mean's location is *outside* the thick null range. The conventional test stands up well against the alternatives in its true power to detect a difference from the null when there is a difference. For high nominal-power cases, it correctly rejected the false null almost 99% of the time and was correct 86% of the time for mid-power cases. When the null should be rejected in low-power cases, the conventional test does falter, as expected; in that scenario, only 57% of its inferences are correct. Interestingly, however, in low power cases, the *only* method that out-performs the conventional, $\alpha = 0.05$ test is the distance-only approach, with a success rate of 88%.

Replacing the $P < 0.05$ criterion with $P < 0.005$ (small-alpha method) shows a considerable cost in true power for lower nominal-power cases. For the lowest-power cases, it only detects 35% of cases where there is really a meaningful difference from the null. However, when interpreting the small-alpha findings, a caution is that Benjamin et al.'s (2017) proposal for $P < 0.005$ is specifically addressed to "novel discoveries," that is, tests for rejecting the null where there is low prior probability for the alternative hypothesis being true. The simulation does not explicitly manipulate or control for that prior probability, when applying methods for comparison; so, the small-alpha results may not model an implementation of Benjamin et al.'s specific proposal with complete fidelity.

In trying to reduce Type I error (discussed next paragraph), the interval-based method pays a considerable price in true power: We see that for all levels of nominal power under the simulated conditions, the interval-based method has the worst or second-worst inference success when the null is really false. It outperforms only the small-alpha method when nominal power is low, with inference success of 41% in those cases compared to small-alpha's 35%. The proposed **MESP** approach also has its main advantages regarding Type I error rather than for true power; yet we see that its benefits have little cost in terms of true power compared to the conventional test. For all nominal powers, MESP's true power is comparable to the conventional approaches.

The top half of Table 2 shows methods' rates of inference success (i.e., decisions to correctly *not* reject the null) when the real parameter is *not* meaningfully different from the null value (i.e., the thick null is true). A *true Type I error* occurs when the thick null is true yet is rejected (or a non-meaningful effect size is deemed meaningful, etc.); so, a method's true Type I error rate (or *true* α) is the complement of its inference success rate for this half of the table. The table confirms the well-known limitation of conventional p-value tests when sample size (hence power) are large: For high-power cases where the thick null is really true, the conventional method's true Type I error rate is a sobering 63% (based on 1— (successful inference rate = 37%)); for mid-power cases, its true α is 1 – 77% = 23%. The "α" in the conventional method's "$\alpha = 0.05$" criterion is clearly nominal; we see that the *true* α—the empirical Type I error rate for the method—is much higher. The small alpha approach predictably improves true Type I error rates, though they are still very poor for high-power cases, where the true α is 1 – 53% = 47%. The distance-only method, on the other hand, performs best under the simulated conditions for higher power cases: Its true α ranges from 1 – 90% = 10% for high-power cases to 1 – 57% = 43% for low-power cases.

The interval-based method implemented for this study outperforms all the displayed alternatives, with respect to true Type I error: never higher than 1 – 98% = 2%. But as mentioned with respect to the bottom half of the table, the method pays a considerable price in terms of true power.

MESP appears to offer a good compromise: Its true α stays comparatively low for all nominal power levels (ranging from 8% to 17%), while we observed its true power remains roughly comparable to conventional tests.

MESP's combined use of a p-value and a distance criterion results in a notable pattern in Table 2: For both true and false thick nulls, MESP and distance-only methods perform identically when power is high, while the MESP and conventional methods perform identically when power is low. The reason is, when nominal power is high, MESP's indication to reject or not-reject essentially hinges on the distance criterion's signal (since getting a low p-value will be relatively easy to attain, and not the deciding issue). So, when the distance criterion leads to error in such cases (either rejecting a true null or not-rejecting a false one), the MESP shares in the error. But when nominal power is low, effect sizes tend to need to be larger to trigger the p-value reject signal (so attaining minimal distance is not the deciding issue); instead, MESP's indication to reject or not reject tends to hinge on the p-value signal—so those methods share the same success and error rates.

A deeper analysis of the power breakdown in Table 2, with more power categories, is provided in Appendix A2, in Table A1.

All methods' success rates are sensitive to *relative MPSD,* that is, the magnitude of the minimum practically significant distance (MPSD), relative to population standard deviation. These sizes ranged from (< 0.04 x σ) up to (5 x σ), and Table 3 breaks this size range into deciles. The discussion below focuses on Deciles 2 to 9 within each column. (Potential reliability issues of the outer deciles are discussed in Section 4.2 and Appendix A1).

In the top entries of the fourth column of Table 3, we see that the true Type I error rate (true α) for the conventional method is not a large concern when relative MPSD magnitudes are in the lowest deciles. In decile 2, true α was only about 10% (i.e., 1—90% successful inference rate for when thick null is true). But as relative MPSD increases, the conventional method's true error rate increases consistently in tandem—reaching (1—35% successful = 65%) by the ninth decile. This occurs because the conventional method does not consider practical importance; a larger relative MPSD results in more cases where a value is rejected based on the null mean alone, yet is getting included in the expanded thick null, so it is an error to reject the null.

The conventional method's true power to correctly reject a false null (bottom half of fourth column), increases consistently as relative MPSD increases. This is because, for subsets of the false-null cases which have wider thick null intervals, the p-value approach is assessing effect sizes that tend to be larger, and so is more likely to correctly reject the null for those cases.

The small-alpha approach's success rates (fifth column) show similar directions of sensitivity to relative MPSD as for the conventional approach, across all deciles, for the same reasons as explained in the previous two paragraphs. The interval-based method that was modeled appears insensitive to relative MPSD size and is almost always correct when the thick null is true (top half of right column). Its generally poor true power to reject a false null improves a bit (about 70%) when MPSD is relatively large compared to σ (perhaps because this method is *also* sensitive to having a small σ, for reasons not related to MPSD.) The distance-only method appears comparatively insensitive to relative MPSD size for true power (bottom half of the seventh column). However, for small relative MPSD's (top of the seventh column), the method's true α rate is very poor: In the second decile, (1 – (52% success rate when the null is true) = 48%. This occurs because small differences that the method deems meaningful are often due simply to sampling error, which this method does not account for.

The **MESP** method generally maintains strong relative performance regardless of the size of MPSD/ σ. The method's true power is comparable, for all relative MPSD levels, with the conventional method's; while it is true α rates (top half of the sixth column) never exceed 15% (in the sixth row, based on 1—85% success).

Table A2 in the appendix shows how the effects of power and relative MPSD size combine. For example, the lowest inference success rate (37%) displayed in Table 2 is for the conventional test, when nominal power is high and the thick null is true. If that context is subdivided based on relative MPSD size, then consistently with Table 3, the lower success rates are more specifically occurring as relative MPSD increases: 62% success for the second quartile but down to 49% for the third quartile, etc. Because some of the simulated combinations of power and relative MPSD in Table A2 would, if they occurred in practice, reflect conditions of poor study design or execution, such cases may be overrepresented in the simulation, which could affect the reported findings.

Figure 4 illustrates how p-values combined with a distance criterion can be informative for inference as *heuristic cues*, as discussed in Section 1.2—even though p-values on their own are noisy as predictors of whether a thick null is really false or true. The figure plots, for each case in the simulation, the true distance of the parameter from the null (Y-axis), versus the two-tailed p-value that was generated for that case (X-axis). For any case shown below the central horizontal line, the distance of the true parameter value from 100 is less than the minimum practically significant distance, consistent with the thick null being correct; above the line the thick null would not be correct. (Logarithmic scales for the axes are simply to make the graph's salient features more visually prominent, with less blank space.) Cases where the MESP inference procedure would signal to reject the null are displayed as the red, horizontal lines; this rejection signal is seen to clearly, though imperfectly, be associated with cases when the thick null really is false. In other words, obtaining the MESP signal-to-reject provides a heuristic, yet nondefinitive, piece of evidence toward inferring the mean's true location.

4. Model Assumptions and Limitations

4.1. Representativeness of the Model

Although the authors believe that the versions of methods included in the simulation model are reasonably generic and representative of the methods discussed, the generality of the model has not been formally tested or demonstrated for

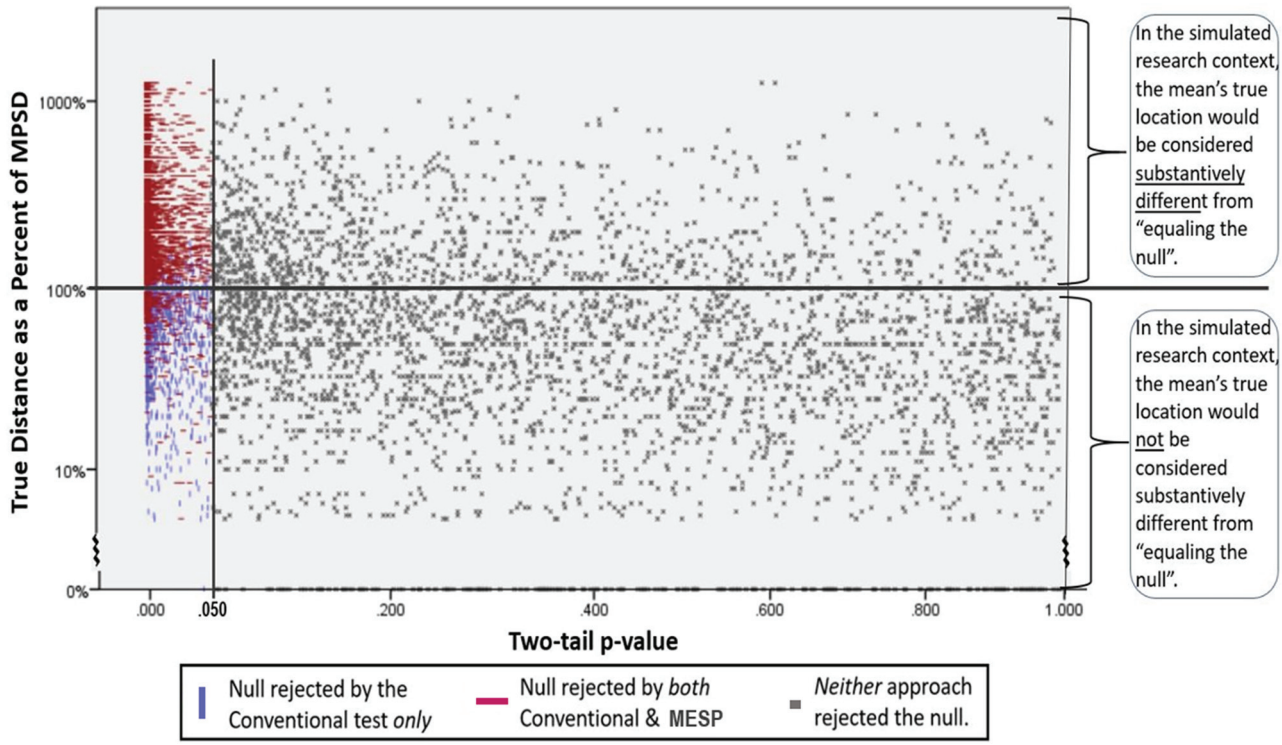


Figure 4. *p*-Value's inference performance, compared to MESP.

nonincluded variations. The applicability of the results for one-tailed tests was checked in Appendix A2, Table A4, and the simulation results there appear comparable to those in Table 2 for two-tailed tests.

Results reported for the interval-based method may not fully extend to a recent variation of the method proposed by Blume et al. (2018). Both versions signal to reject the null when the thick null interval and a sample-based 95% confidence interval (CI) do not overlap. But absent that signal, Blume et al. distinguish (a) when the null interval subsumes the CI completely versus (b) when the intervals' overlap is only partial. The impact if any of that added distinction is not modeled in the present study.

4.2. Generation of Data for Cases

The design choices made in this study for generating set-up values from particular input-ranges can impact the comparative decision results that are reported. This echoes the insight of Bayesians that *prior probabilities* can impact *posterior probabilities*. It is not possible to make absolute inferences based on observed data, without some reference to the context in which the data are drawn. For example, combinations of set-up values generated by the simulation were not assigned different weights for being more or less likely to be encountered, though in some research contexts, perhaps some combinations *are* more or less likely. Therefore, the exact numbers reported in the Results section for true power and true α , etc., cannot be interpreted in absolute terms. In general, more research is needed on the extent and nature of the simulation's sensitivity to unmodeled differences in actual likelihoods, for different combinations of scenarios or extremes of distributions generated by the simulation.

More detailed discussions on the design decisions for the generated set-up values and their impacts on the overall distributions of the test cases, and on the plausibility of those distributions to represent a realistic example set, are provided in Appendix A1.

4.3. Risks of MPSD "Hacking"

An acknowledged risk of including a distance criterion within MESP is that researchers could potentially "hack" results by setting MPSD sizes that are biased or specified-post hoc to their own study's advantage. The authors believe this issue is important, but that it should be viewed as primarily a matter of professional ethics and training, and also that risks for such tainting are not inherently different for MESP than those for *any* inference method—for example, one's selecting a more favorable α or Cohen's d benchmark post hoc. The authors recommended in Section 1.1 that MPSD values should derive from a researcher's reflective professional practice, and be documented along with his or her results. This suggestion is consistent with proposals that research benchmarks of this sort be preregistered in some way before starting one's study. The authors also encourage efforts for the research community at large to become more sensitive to issues of effect size. For example, if the data include strong measures such as amounts of money spent or cigarettes smoked over a period of time, unstandardized regression weights can inform whether the impact of a one unit increase in the independent variable really has an appreciable impact on the dependent variable; such as, is the impact just a small number of extra or fewer dollars spent or cigarettes

smoked over the period, or is the increase or reduction in dozens or hundreds of dollars or cigarettes?

Note that in the simulation, random assignments of MPSDs to cases is not intended to imply that MPSDs' assignments to *actual research* would or should be random—just as sample sizes would not be random in practice but are assigned randomly in the simulation. The goal for the simulation was to generate many different combinations of independently generated set-up values, to observe the factors' effects and interactions. Each simulated case is taken to represent data collected by a specific (simulated) researcher who would have set that study's MPSD value competently and appropriately, in the context of his or her research specialty.

5. Discussion and Conclusions

This study confirms, using simulations, that p-values provide some evidential information relevant to an inference about a population mean, given a sample. It examines p-values' strengths and weaknesses if used as an inference signal in various contexts, and compares these with successes and weaknesses of some alternative methods. Among these, an alternative, hybrid inference method is introduced, which uses a criterion called minimum effect size plus p-value (MESP).

Key findings from the simulation are summarized below:

- For tests that were high nominal powered (e.g., large samples; small population variance), the p-value-only approaches had the worst true Type I error rates. Specifically, for *conventional* $P < 0.05$ tests with high nominal power, true α was 63%. For *small alpha* tests, true α was 47% with nominal high power. For both versions of p-value tests compared, true α increases markedly (i.e., inference success drops) as the thickness of the null increases.
- Both p-value methods have good true power when nominal power is high. But for *small alpha* true power drops to under 40% for low nominal power cases, or when the relative MPSD is small. *Conventional* ($P < 0.05$) performs better in those circumstances.
- The distance-only method falters in true α ($\alpha = 43\%$) for low-power cases, or when relative MPSD is comparatively small. The method has good true power ($> 87\%$) for all nominal power levels and relative thicknesses of the null.
- A generic interval-based method has consistently good true α for all nominal power levels and relative thicknesses of the null—never worse than 2%. But for all nominal powers, its true power is much less than the conventional methods.
- MESP balances reasonably consistent true power (roughly equivalent with the ($P < 0.05$) at all nominal powers and relative thicknesses of the null, with a true α that compares well with other methods—without other methods' problems with some nominal power levels or thicknesses of the null.
- Changing from two- to one-tailed tests does not appreciably change the performance patterns described for the p-value and MESP approaches.
- A method that combines p-values with a distance criterion can generate heuristic, but nondefinitive, evidence toward inferring a mean's location—even though p-values on their own are noisy as predictors of the mean's true location.

The percentages reported could potentially be different under different—yet plausible—simulated conditions.

In summary, all of the compared methods have strengths and weaknesses; and *none* of them generates an automatic final answer for a definitive inference, based on one application. In practice, the researcher is blind to the left columns of Tables 2 and 3, which show the true mean, and this makes a crucial difference: The interval-based method is, for example, the best method under the simulated conditions when the null is true, but the worst method under those conditions when the null is false. But the researcher cannot know which half of that column he or she is really working in.

The authors recommend preferring a tool that (a) is not sensitive to factors that are not knowable to the researcher (such as the true mean, and to a lesser extent, the true size of MPSD relative to population sigma), and (b) performs well in contexts that the researcher *can* check for. MESP performs well regardless of whether the unseen real mean happens to be within the thick null or not. Its true power weakens in low nominal power cases; but the researcher can know when that applies, and can respond accordingly.

Added to those advantages for MESP is the ready availability of the p-value component of its indicator. Statistical software can already flag when a correlation or regression coefficient or other estimate appears significant, in terms of the $p < 0.05$ signal. If a researcher also has in mind a reasoned criterion for a meaningful effect size, MESP can be directly applied to the case, without requiring any new type of calculations. In contrast, data to implement the procedures for the interval-based method may not be visible on standard outputs, and may require additional, unfamiliar calculations.

As noted in Section 4, the exact success ratios displayed in this article's text and tables are specific to the design decisions for the simulations. Choosing a p-value criterion other than 0.05 or 0.005, or designing alternative algorithms for distance-only or interval-based decisions, or revising the bounds of possible inputs for the simulation, etc., could all impact apparent results. Yet, the versions of methods compared in this study, and the ranges of independently varied random set-up factors for cases, were designed to generate a dataset of cases that the authors believe are reasonably representative of real-research cases and methods that could be encountered. Informal, additional checks were also conducted (such as adapting the simulation to assess tests for correlation rather than for the mean); and the broad patterns of the results reported here appear to be generalizable.

The usefulness of any inference method's results depends on the care taken to ensure that all fundamental assumptions are satisfied for using that method—not just assumptions specific to that method, such as distributional assumptions, but also, more generally, having an unbiased, representative sample, with no unacknowledged confounders influencing results, or missing variables causing proxy effects, and so on. The comparisons presented in this article are intended to the compare the findings of the several methods so far as they have been carefully and properly implemented, under circumstances that meet their assumptions.

With that proviso, the authors believe that the NHST model can still have a place in scientific research and in the journalistic reporting of research findings if results are interpreted properly

and not taken as automatically justifying final conclusions. If the p-value criterion is met, it should *also* be assessed whether a meaningful minimum effect size was observed. MESP combines the traditional p-value < 0.05 condition with a minimum effect size criterion. The simulations show that even for the traditional method, "$\alpha = 0.05$" is essentially just a nominal trigger, it is not the *true* α; MESP merely acknowledges this *heuristic* use of the α term and adds the important effect-size criterion. If the two rejection criteria for MESP occur, take that cue and build a convincing story using full disclosure on sampling methods, sample size and what is already known, unknown or hypothesized.

Appendix A. Additional Analyses and Simulation Details

A.1. Distributions of Set-up Values and Their Combinations

How realistic or generalizable the simulation's results are could be impacted by the distributions of the set-up values that were randomly generated over the 10,000 passes of the simulation—or of the relations among those values. This appendix displays and discusses several of those key distributions.

- **Individual Set-up Values**

For each pass of the simulation, set-up values for each of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, **n**, and **MPSD** were independently determined by randomly selecting an integer from a specified range of equiprobable integer values for that element: Those ranges were $75 \leq \boldsymbol{\mu} \leq 125$; $4 \leq \boldsymbol{\sigma} \leq 60$; $5 \leq \boldsymbol{n} \leq 100$; and $2 \leq \mathbf{MPSD} \leq 20$. Figure A1 shows that—allowing for random variation—the distributions are all flat as intended by the study, and bounded within the designed ranges.

- **Combinations of Set-up Values**

Each pass of the simulation is intended to model a realistic example of a hypothesis test for a population mean that might be conducted, with respect to the actual values that $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, **n**, and **MPSD** may have when the test is conducted. The limits bounding an example's being realistic are difficult to define formally; but some potential issues can be intuited and addressed in connection with the following results.

Since H_0 is unvaried in the simulation, the relation between mean and sigma that is relevant for the simulated hypothesis tests is the ratio $|\mu - H_0|/\sigma$, that is, the absolute real distance of the population mean from the null mean, in standard deviations. Figure A2 shows that most of the simulated true distances had a magnitude of less than one population standard deviation, and very few magnitudes exceeded 2.5 standard deviations. Occasionally in practice, researchers may be unaware if they have posited a null value that far from the true μ (perhaps because an unobserved change occurred in the environment), so the few outliers like this generated in the distribution are not necessarily implausible, and are useful for comparing the different inference methods' performance in such cases.

In most cases, the values for relative MPSD—that is, MPSD standardized relative to population standard deviation (MPSD/σ)—were very consistent with effect sizes that, for example, researchers would want to power their studies to detect. Assuming a researcher has obtained a reasonable, preliminary estimate for the population standard deviation σ, then it may be unrealistic that MPSD would be

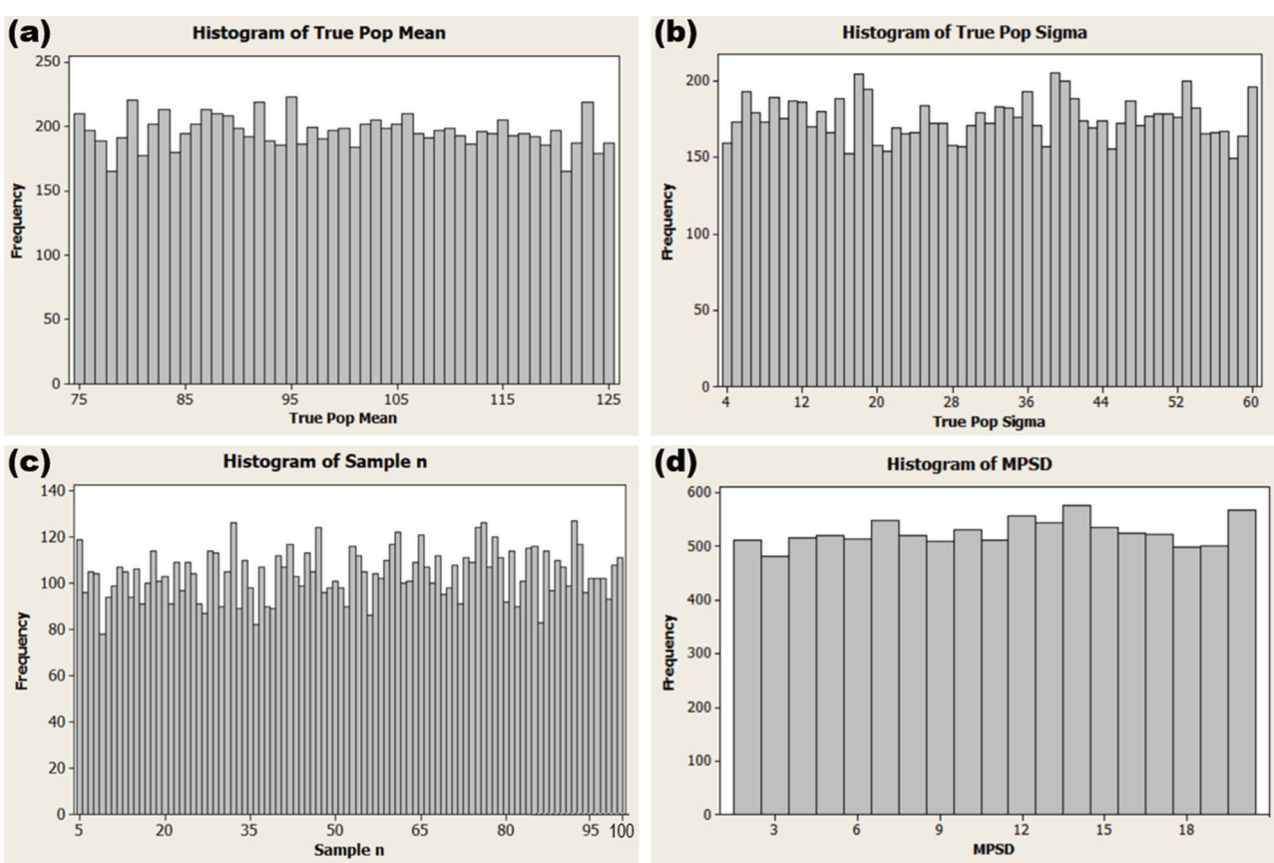


Figure A1. Distributions of the individual set-up values.

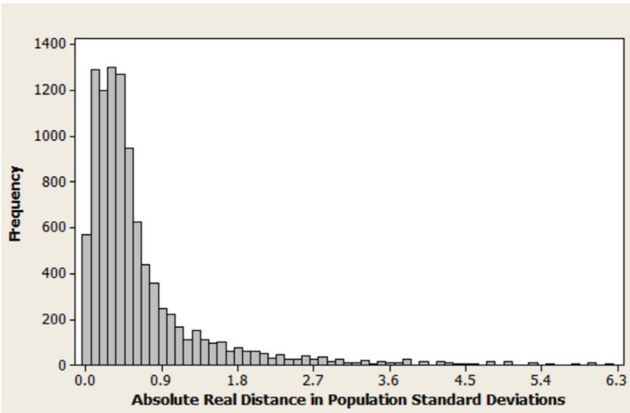


Figure A2. Histogram of absolute real distances in population standard deviations.

set deliberately to large values like two or more standard deviations. However, note that large MPSD/ σ ratios of this sort in Figure A3 occur only within the uppermost decile of the distribution. Also, the lowest decile is of interest, since it roughly corresponds to a smaller setting of MPSD—0.1 standard deviation—that social scientists sometimes employ. Table 3 in the main text groups MPSD/σ values into deciles, in order to observe the comparative impacts of relative MPSD sizes at various levels, including the more extreme categories.

The potential distortion that large sample sizes can have on conventional test results is often discussed. However, the derivation on the right panel of Figure A4 clarifies, more specifically, that the test statistic is inflated (lowering the p-value) by the *ratio* of $\sqrt{n}$ to σ. Even though the simulation has capped the set-up values for n at 100, some cases' results could be distorted if n is large *and* σ is small; however, according to the histogram in the figure's left panel, this "multiplier" ratio for the test statistic does not appear to have an excessive spread for the simulated data. Larger values could realistically occur if a researcher has access to larger sample sizes, and takes advantage of them, while perhaps is not certain of the true sigma value when estimating sample size or power.

Figure A5 shows the distribution of ratios generated by the simulation for absolute real distances (i.e., $|\mu - H_0|$) over MPSD. For cases where the true mean falls within the thick null (top half of Tables 2 and 3), that ratio is ≤ 1; where that ratio is > 1, the true mean falls beyond the thick null (bottom half of Tables 2 and 3).

The simulation did not exclude any cases based on this ratio, since it was not known, a priori, whether it would be realistic for ($|\mu$ - $H_0|$)/MPSD to have a ratio as large as, for example, 6 or greater. This suggests questions for further research: (a) How do actual researchers set the values they would use for MPSD? (b) In particular, do they include in their reasoning some estimate of how far from the null the true mean *could* realistically fall? (c) If so, how do they derive their belief or expectation about how far that could be, and (d) do they generally tend to be accurate in that estimate.

A.2. Additional Analyses of Methods' Inference Success

- **Impact of Power and Method on Inference Success (Expanded)**

Table A1 deepens the analysis in Table 2, by breaking power into six categories instead of the three categories used in the main text. The finer-grain analysis in Table A1 preserves the general value trends displayed in Table 2 for each status of the thick null, when scanning down the alternative methods' columns. However, for the highest or lowest power categories, sudden value jumps do occur in some columns. For example, when the thick null is true, the success rate of the conventional method decreases monotonically as power increases—yet the drop is 30 percentage points from second-highest to the highest power category. And that method's success rate drops 20 percentage points at the lowest power category, when the thick null is incorrect. It would useful in future research to investigate for any factors, unmodeled in the simulation, that may contribute to these sudden jumps.

- **Impact of Combined Power and Relative MPSD on Methods' Inference Success Rates.**

Tables A2 and A3 supplement the stand-alone analyses in the main text for the impacts of Power (Table 2) and relative MPSD (Table 3), respectively, on different methods' inference success rates. Table A2 looks at the impact of the two factors combined. To reduce the complexity of the table, the sub-classes for MPSD/ σ are based on quartiles rather than deciles for the distribution. Table A3 supports using that simplification of classes for relative MPSD: Namely, if values up and down individual columns in Table A3 (Quartile-based) are directly

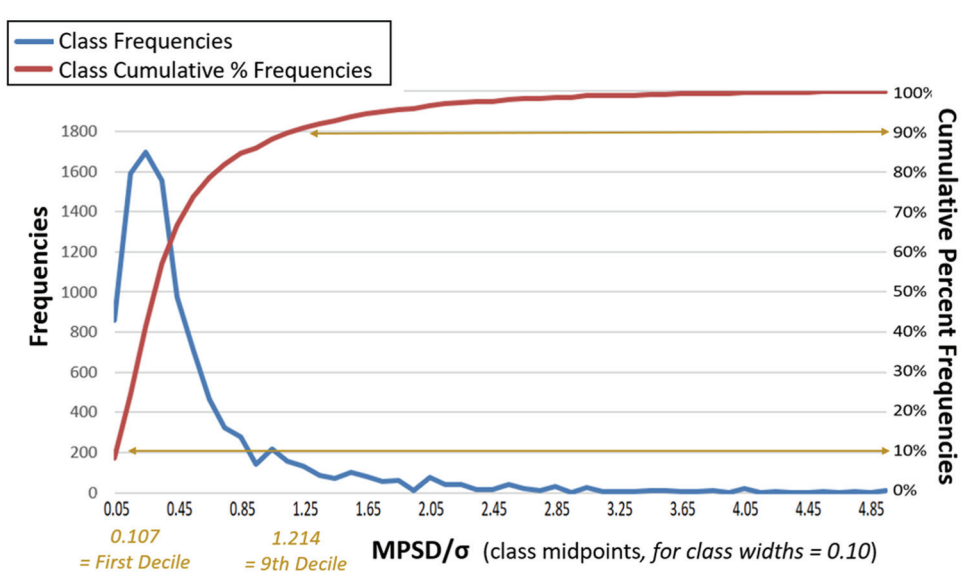


Figure A3. Absolute and cumulative percent frequencies for **MPSD/σ.**

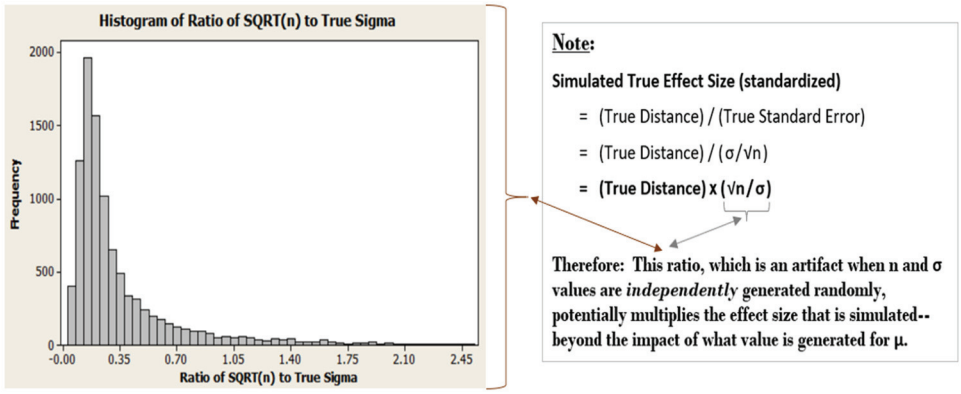


Figure A4. Ratio of $\sqrt{n}$ to σ.

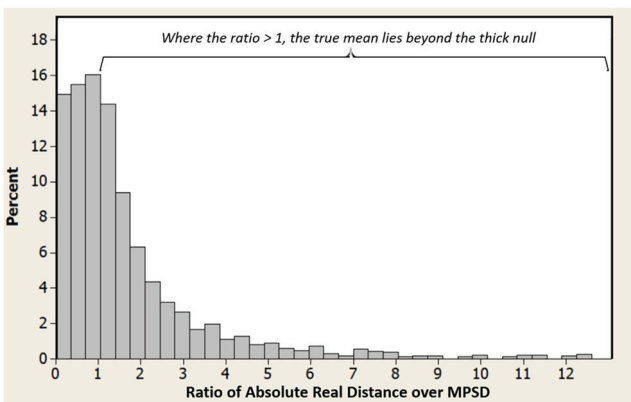


Figure A5. Ratio of absolute real distance to MPSD.

compared with counterparts in Table 3 (Decile-base), it is observed that dividing the data more finely into deciles uncovers no reversals or inconsistency of the general trends that can be observed using quartiles.

- **Impact of Power on Methods' Inference Success Rates—One-tailed Cases**

Table A4 and Figure A6 suggest that the results would be very similar for one-tailed tests, as for two-tailed tests. Only the cases of the main simulation where $\mu > 100$ are evaluated in this section, presuming contexts where the true mean could not realistically be less than that null; and *one-tailed* p-value calculations were applied where applicable. The small-alpha method was not evaluated for this table.

When comparing relative strengths and weaknesses of different methods, broken down by whether the thick null is true, and nominal power levels, Table A4 (for one-tailed cases) is clearly very similar to Table 2 (for two-tailed cases). Apart from small random variations (from reducing the set of relevant cases included), Table A4's differences from Table 2 are as expected: For a given p-value cutoff, a one-tailed test does not require as large an effect size (in the one direction of interest) as a two-tailed test does to reject a null hypothesis. Therefore, at all nominal power levels, the one-tailed conventional test shows a bit more power to reject the null when the thick null is really false yet has poorer success rates (i.e., more true Type I error) when the thick nulls are really true. Because of its extra distance criterion, **MESP** did not display that same increase in true power, for one-tailed versus two-tailed decisions, for the cases with the highest nominal power. For one-tailed tests, for all but the lowest-powered cases, MESP continues to show better success when the thick null is true (smaller true α) compared to the conventional method.

- **FDRs Comparisons and Implications**

Sometimes, many tests are conducted simultaneously, for example, in multiple testing of biological materials on "microarrays." Each of the many tests simulated for this study—even if they were simultaneous—models a *stand-alone* hypothesis test, with different set-up values. Still, this partial analogy with multiple-testing suggests a useful concept to

Table A1. Impact of power and method on inference success rates—expanded version.

Does the true location fall within the bounds of the "thick null"?	Nominal power	Number of simulated cases	Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach, power, and true location of the mean				
			Conventional	Small alpha	MESP	Distance-only	Interval-based
Yes	≥ 0.90	2075	32.8%	47.1%	91.3%	91.3%	99.7%
	0.80–0.90	361	62.0%	84.8%	85.0%	85.0%	100.0%
	0.60–0.80	542	68.8%	88.4%	81.7%	81.5%	99.3%
	0.40–0.60	489	81.8%	95.7%	84.3%	80.4%	99.6%
	0.20–0.40	575	85.7%	97.2%	85.9%	71.0%	99.1%
	< 0.20	610	93.4%	99.3%	93.4%	50.7%	97.7%
No	≥ 0.90	1475	99.7%	98.8%	93.5%	93.5%	66.0%
	0.80–0.90	303	95.4%	85,1%	87.1%	87.1%	48.8%
	0.60–0.80	502	92.4%	76.7%	86.9%	87.3%	43.0%
	0.40–0.60	590	87.1%	65.1%	85.8%	88.6%	41.4%
	0.20– 0.40	840	73.1%	50.8%	73.1%	86.0%	43.9%
	< 0.20	1638	53.2%	31.2%	53.2%	87.8%	39.6%
Numbers of cases		10,000	6963	6817	8192	8518	7220

Table A2. Impact of combined power and relative MPSD on methods' inference success rates.

Does the true location fall within the bounds of the "thick null"?	Power[a]	Quartile for MPSD in population standard deviations[b]	Number of simulated cases	Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach, relative MPSD, nominal power, and true location of the mean. Conventional[c]	Small alpha[d]	MESP[e]	Distance-only[f]	Interval-based[g]
	≥ 0.80	1	–	–	–	–	–	–
		2	126	61.9%	81.7%	81.0%	81.0%	100.0%
		3	879	49.0%	69.6%	86.5%	86.5%	99.7%
		4	1431	27.6%	39.7%	93.6%	93.6%	99.7%
	0.30–0.8	1	87	79.3%	94.3%	79.3%	69.0%	100.0%
Yes		2	745	77.4%	92.S%	83.2%	79.9%	99.7%
(4652 Cases)		3	423	74.7%	92.4%	83.0%	73.7%	99.1%
		4	70	77.1%	95.7%	84,3%	81.4%	98.6%
	< 0.30	1	518	91.5%	99.0%	91.5%	50.4%	98.1%
		2	297	92.9%	98.3%	92.9%	65.3%	97.6%
		3	74	91.9%	100.0%	91.9%	67.6%	98.6%
		4	2	100.0%	100.0%	100.0%	100.0%	100.0%
	≥ 0.80	1	–	–	–	–	–	–
		2	113	96.5%	85.8%	84.1%	84.1%	48.7%
		3	701	98.3%	95.7%	90.2%	90.2%	51.5%
		4	964	99.7%	98.2%	95.0%	95.0%	73.2%
	0.30–0.80	1	185	82.2%	60.5%	82.2%	87.6%	51.9%
No		2	908	881%	71.7%	85.8%	88.3%	42.8%
(5348 Cases)		3	324	83.0%	59.3%	78.7%	84.6 %	36.4%
		4	49	75.5%	34.7%	73.5%	81.6 %	36.7%
	< 0.30	1	1656	60.3%	37.9%	60.3%	90.9%	44.3%
		2	378	47.9%	25.9%	47.9%	76.7%	28.8%
		3	69	37.7%	13.0%	37.7%	68.1%	20.3%
		4	1	100.0%	0.0%	100.0%	100.0%	0.0%
	Numbers of cases		10,000	6963	6817	8192	8518	7220

[a]Nominal power, as calculated for a z-test for the mean, given these inputs: $\boldsymbol{\alpha} = 0.05$; minimum detectable difference_MPSD for the case; n = the simulated case's sample size; and $\boldsymbol{\sigma}$ = the true population sigma for the case.
[b]These ranges of values for MPSD/$\boldsymbol{\sigma}$ correspond to the Quartiles: 1: 0.000 to 0.200 2: 0.200 to 0.349 3: 0.349 to 0.615 4: 0.615 to 5.000
[c]Two-tailed. p-value based test, $\boldsymbol{\alpha} = 0.05$
[d]Two-tailed, p-value based test, $\boldsymbol{\alpha} = 0.005$
[e]Minimum effect size plus p-value approach, proposed in this article.
[f]Directly compares effect size to MPSD.
[g]Compares the Thick Null Interval with a sample-based confidence interval for the mean.

Table A3. Impact of relative MPSD on methods' inference success rates—using quartiles.

Does the true location fall within the bounds of the "thick null"?	Quartile for MPSD in population standard deviations[a]	Number of simulated cases	Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach, relative MPSD, and true location of the mean Conventional[b]	Small alpha[c]	MESP[d]	Distance-only[e]	Interval-based[f]
Yes	1	605	89.8%	98.3%	89.8%	53.1%	98.3%
	2	1168	79.7%	93.0%	85.4%	76.3%	99.2%
	3	1376	59.2%	78.3%	85.7%	83.1%	99.4%
	4	1503	30.0%	42.4%	93.2%	93.1%	99.7%
NO	1	1841	62.5%	40.2%	62.5%	90.5%	45.1%
	2	1399	77.9%	60.5%	75.4%	84.8%	39.5%
	3	1094	89.9%	79.7%	83.5%	87.1%	45.1%
	4	1014	98.5%	95.1%	94.0%	94.4%	71.4%
Numbers of cases:		10,000	6963	6817	8192	8518	7220

[a]"1": $<$0.200 "2": 0.200 to 0.349, "3": 0.349 to 0.615, and "4": 0.615 to 5.000
[b]Two-tailed, p-value based test, $\alpha = 0.05$.
[c]Two-tailed, p-value based test, $\alpha = 0.005$.
[d]Minimum effect size plus p-value approach, proposed in this article.
[e]Directly compare effect size to MPSD.
[f]Compares the thick null interval with a sample-based confidence interval for the mean.

consider, called FDR, which may also suggest a potential limitation of the simulation design.

In the microarrays analogy, each individual test that rejects the null (i.e., each "discovery") could be interpreted as testing *positive* for a condition of interest. The FDR is a way to quantify, for comparisons and quality assessment, the overall rate of false positives across all the simultaneously conducted tests (Colquhoun 2014). FDR is calculated by dividing the number of false-positive test results (when there is really

Table A4. Impact of power and method on inference success, for one-tailed cases.

Does the true location fall within the bounds of the "thick null"?	Power[a]	Number of simulated cases	Proportions of implied inferences that were consistent with the true location of the mean, for each combination of approach, power, and true location of the mean. Conventional[b]	MESP[c]	Distance-only[d]	Interval-based[e]
Yes (2217 Cases)	$\geq$ 0.80	1156	30.3%	90.2%	90.2%	99.7%
	0.30 to 0.80	653	62.6%	76.9%	75.5%	99.2%
	$<$ 0.30	408	87.5%	87.5%	56.9%	98.5%
No (2636 Cases)	$\geq$ 0.80	896	99.2%	91.4%	91.4%	63.2%
	0.30–0.80	710	92.4%	88.7%	89.4%	43.5%
	$<$ 0.30	1030	65.8%	65.8%	87.2%	41.7%
Numbers of cases		4853	3339	4029	4120	3507

[a]Nominal power, as calculated for a *z*-test for the mean, given these inputs: $\boldsymbol{\alpha} = 0.05$; minimum detectable difference = MPSD for the case; n = the simulated case's sample size; and $\boldsymbol{\sigma}$ = the true population sigma for the case
[b]One-tailed, *p*-value based test $\alpha = 0.05$.
[c]Minimum effect size plus *p*-value approach, using the *p*-value for one-tailed.
[d]Directly compares effect size to MPSD.
[e]Compares the thick null interval with a sample-based confidence interval for the mean.

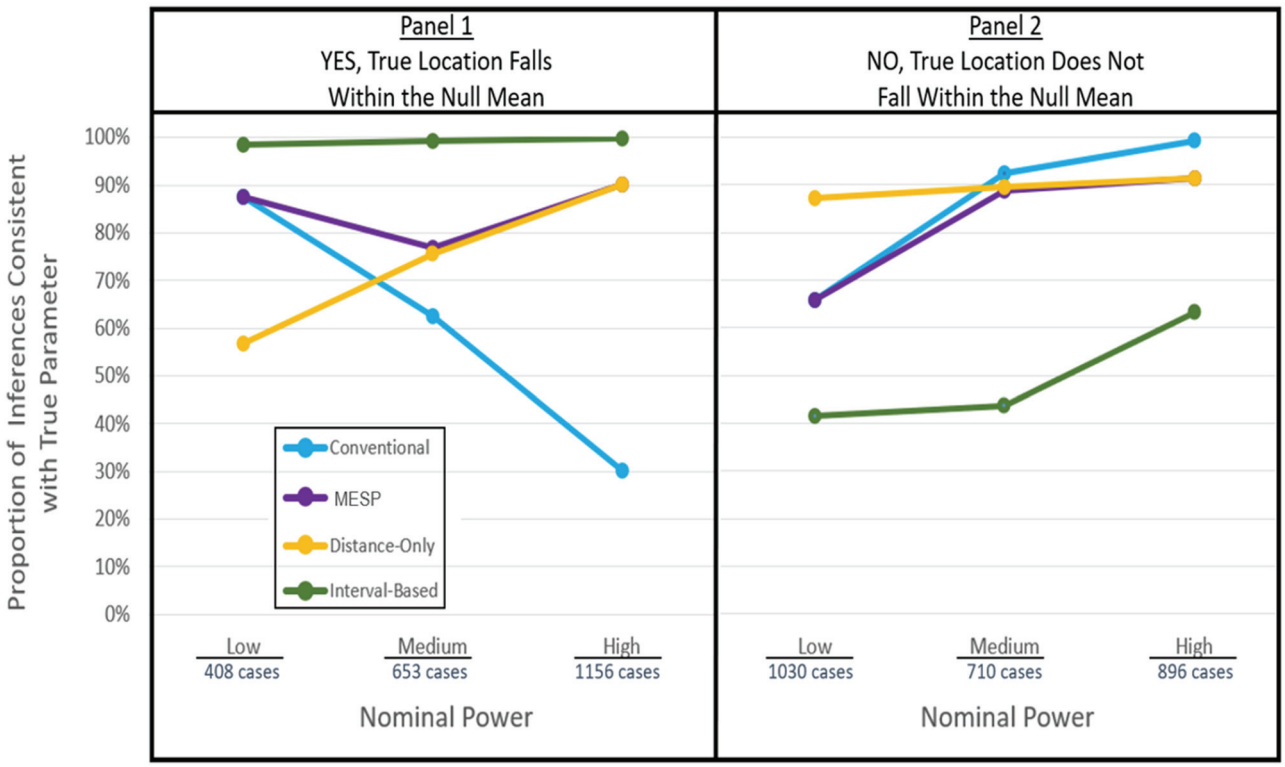


Figure A6. Graph of Impact of power and method on inference success, for one-tailed cases.

no effect) by the sum of all positive test results (the number of false positives + the number of true positives).

Figure A7 compares methods' relative successes by FDR, where a lower value is preferable. There are rough similarities to the results in the upper half of Table 2 (not controlling for nominal power); but FDR factors in information from both halves of Table 2—both real effect cases and no effect cases.

To calculate FDR in practice, the process (depicted in Colquhoun's "Figure 2" (2014, p. 4)) requires the ratio of real effects cases to no effects cases to first be estimated in advance; and, from this *prior* estimate, expected counts of true and false positives to be estimated in turn. In Figure A7, the counts in the boxes are not estimated, but are observed directly from the simulation results. Yet, that said, the counts and ratios of inference successes displayed in Figure A7 do depend on the prior ratio of real to no effects generated by the simulation (in the second column of boxes); and that ratio could have been quite different—if for example, the simulated true means were all far from the null, so real effects were more prevalent. This reflection echoes the observation made with respect to Figure A5, for the simulation's distribution of (absolute real distance)/MPSD. Further study of what distributions and scenarios would be most realistic to embody in the simulation is recommended.

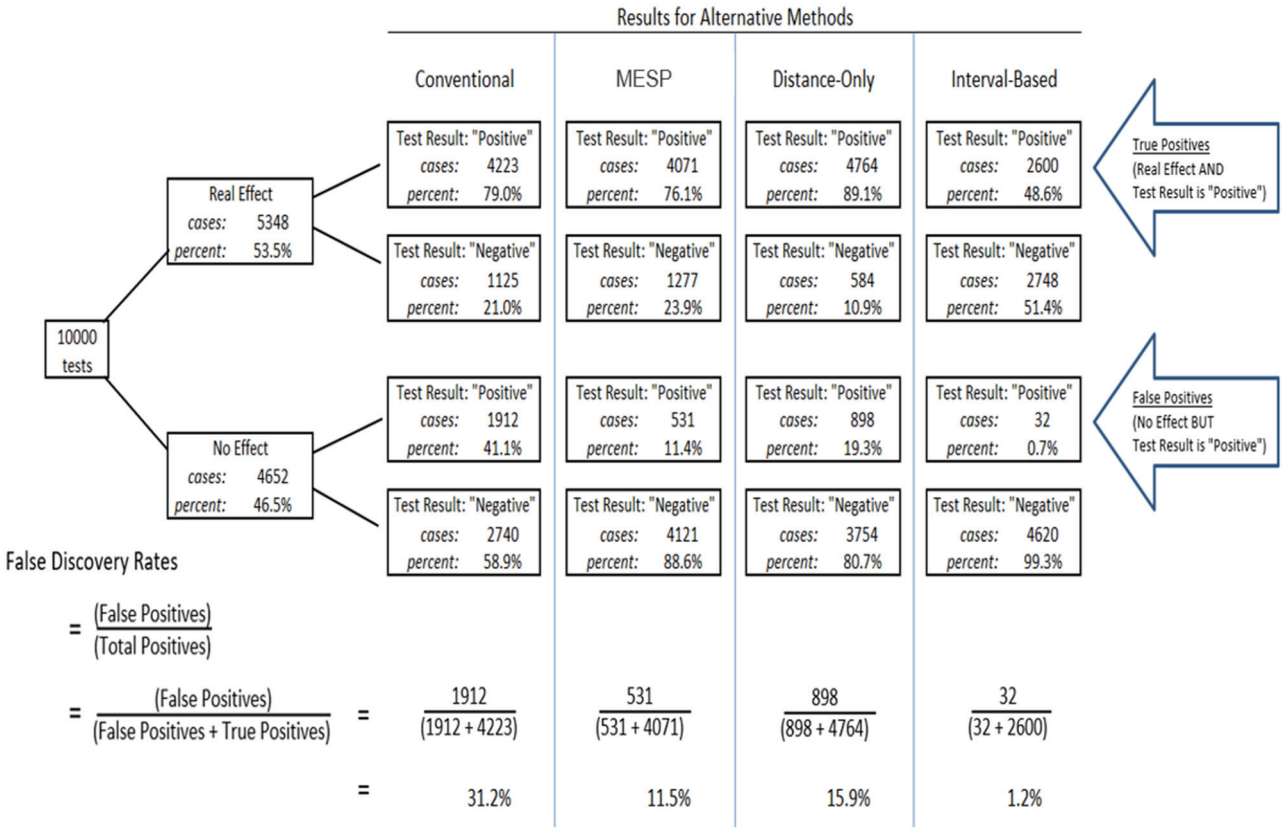


Figure A7. Methods' False-positive rates compared.

Supplementary Material

1. Title: Comparative Inference Experiments_Data and Formulas (Excel file)

Simulation results and formulas. This Excel file contains the full set of outputs (10,000 cases) for the simulation run described in the article. It also includes pages showing (a) all formulas used within the simulation, and (b) explanations of the formulas.

Note that for the output columns *generated during the simulation run*, only the resulting numeric values of the outputs are included in the Results spreadsheet; whereas *for supplemental columns* added to the Results spreadsheet—for example, to calculate nominal power for each case or to calculate what would have been the one-tailed *p*-value for each case—the formulas are retained in the spreadsheet in those extra columns.

2. Title: *Supplemental_Key Algorithms Implemented* (PDF file)

Generic descriptions of algorithms. This PDF document provides brief, generic descriptions of key algorithms and formulas implemented in this article's simulation.

References

Benjamin, D.J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., …, Johnson, V. (2017, July 22). "Redefine Statistical Significance," PsyArXiv Preprints (Online), available at *https://psyarxiv.com/mky9j* [168]

Berger, J.O., and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2(3), 317–352. [170]

Berger, J.O., and Sellke, T. (1987), "Rejoinder to Comments on 'Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence,'" *Journal of the American Statistical Association*, 82(397), 135–139. [170]

Blume J.D., D'Agostino McGowan, L., Dupont W.D., Greevy, R.A. Jr (2018) "Second-Generation *p*Values: Improved Rigor, Reproducibility, & Transparency in Statistical Analyses," *PLoS ONE*, 13(3): e0188299, available at *https://doi.org/10.1371/journal.pone.0188299* [170,177]

Chow, S.L. (1988), "Significance Test or Effect size?" *Psychological Bulletin*, 103, 105–110. [170]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates. [169]

Colquhoun, D. (2014), "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-values," *Royal Society Open Science*, 1: 140216. [182]

Ellis, P.D. (2010), *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*, Cambridge, UK: Cambridge University Press. [170]

Fisher, R. A. (1925), *Statistical Methods for Research Workers* (1st ed.), Edinburgh, Scotland: Oliver & Boyd. [Chapter 3, 5th paragraph.] [168]

Folger, R. (1989), "Significance Tests and the Duplicity of Binary Decisions," *Psychological Bulletin*, 106, 155–160. [170]

Goodman, W.M. (2010), "The Undetectable Difference: An Experimental Look at the 'Problem' of p-Values," in *American Statistical Association JSM Proceedings*, available at *http://www.statlit.org/pdf/2010GoodmanASA.pdf*. [169]

Greenland, S. (2017). "Invited Commentary: The need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645.

Howard, G.S., Maxwell, S.E., and Fleming, K.J. (2000), "The Proof of the Pudding: An Illustration of the Relative Strengths of Null Hypothesis, Meta-Analysis, and Bayesian Analysis," *Psychological Methods*, 5(3), 315–332. [169]

Krueger, J.I., and Heck, P. R. (2017), "The Heuristic Value of p in Inductive Statistical Inference," *Frontiers in Psychology*, available at *http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00908/abstract* [169,170]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., …, Zwaan, R.A. (2018). "Justify your Alpha," *Nature Human Behavior*, 2, 168–171. [168]

Leek, J.T. and Peng, R.D. (2015), "Statistics: p-Values are Just the Tip pf the Iceberg," *Nature*, 520, 612, available at *https://www.nature.com/polopoly_fs/1.17412!/menu/main/topColumns/topLeftColumn/pdf/520612a.pdf* [168]

Meehl, P.E. (1967), "Theory Testing in Psychology and in Physics: A Methodological Paradox," *Philosophy of Science*, 34, 103–115. [170]

Nunnally, J. (1960), "The Place of Statistics in Psychology," *Educational and Psychological Measurement*, 20, 641–650. [170]

Nuzzo, R. (2015), "Scientists Perturbed by Loss of Stat Tool to Sift Research Fudge from Fact," *Scientific American* [online], available at *https://www.scientificamerican.com/article/scientists-perturbed-by-loss-of-stat-toolstosift-research-fudge-from-fact/*. [168]

Peng, R. (2015), "The Reproducibility Crisis in Science: A Statistical Counterattack," *Significance*, 12(3), 30–32. [168]

Siegfried, T. (2015), "P Value Ban: Small Step for a Journal, Giant Leap for Science," *Science News* [online], available at *https://www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leapscience*. [168]

Trafimow, D, Amrhein, V, Areshenkoff, C.N., Barrera-Causil, C.J., Beh, E.J., Bilgiç, Y.K., …, Marmolejo-Ramos, F. (2018) "Manipulating the Alpha Level Cannot Cure Significance Testing," *Frontiers in Psychology*, 9, 699, available at *https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00699/full* [168,169]

Trafimow, D. and Marks, M. (2015), Editorial, *Basic and Applied Social Psychology*, 37(1), 1–2. [168]

Wasserstein, R.L., and Lazar, N.A. (2016a), "The ASA's Statement on p-Values: Context, Process, and Purpose (Editorial, March 7, 2016)," *The American Statistician*, 70(2), 129–133. [168,170]

Wasserstein, R.L., and Lazar, N.A. (2016b), "The ASA's Statement on p-Values: Context, Process, and Purpose. Figshare, *The American Statistician* [online, supplemental materials] (March 7, 2016), available at *https://figshare.com/collections/The_ASA's_statement_on_p_values_context_process_and_purpose/2851090/1*. [168]

Wetzels, R., Matzke, D., Lee, M.D., Rouder, J.N., Iverson G.J., and Wagenmakers, E.J. (2011), "Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests," *Perspectives on Psychological Science*, 6(3), 291–298. [168,169]

Woolston, C. (2015), "Psychology Journal Bans p-Values," *Nature*, 519, 9, available at *http://www.nature.com/news/psychologyjournalbans-p-values-1.17001*. [168]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

Three Recommendations for Improving the Use of *p*-Values

Daniel J. Benjamin & James O. Berger

To cite this article: Daniel J. Benjamin & James O. Berger (2019) Three Recommendations for Improving the Use of *p*-Values, The American Statistician, 73:sup1, 186-191, DOI: 10.1080/00031305.2018.1543135

To link to this article: https://doi.org/10.1080/00031305.2018.1543135

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 18899

View related articles

View Crossmark data

Citing articles: 65 View citing articles

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, 186–191: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543135

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Three Recommendations for Improving the Use of *p*-Values

Daniel J. Benjamin[a,b] and James O. Berger[c]

[a]Center for Economic and Social Research and Department of Economics, University of Southern California, Los Angeles, CA; [b]National Bureau of Economic Research, Cambridge, MA; [c]Department of Statistical Science, Duke University, Durham, NC

ABSTRACT

Researchers commonly use *p*-values to answer the question: How strongly does the evidence favor the alternative hypothesis relative to the null hypothesis? *p*-Values themselves do not directly answer this question and are often misinterpreted in ways that lead to overstating the evidence against the null hypothesis. Even in the "post $p < 0.05$ era," however, it is quite possible that *p*-values will continue to be widely reported and used to assess the strength of evidence (if for no other reason than the widespread availability and use of statistical software that routinely produces *p*-values and thereby implicitly advocates for their use). If so, the potential for misinterpretation will persist. In this article, we recommend three practices that would help researchers more accurately interpret *p*-values. Each of the three recommended practices involves interpreting *p*-values in light of their corresponding "Bayes factor bound," which is the largest odds in favor of the alternative hypothesis relative to the null hypothesis that is consistent with the observed data. The Bayes factor bound generally indicates that a given *p*-value provides weaker evidence against the null hypothesis than typically assumed. We therefore believe that our recommendations can guard against some of the most harmful *p*-value misinterpretations. 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 on statistical significance and *p*-values.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
Bayes factor; *p*-Value, Post-experimental odds

1. Introduction

1.1. The Recommendations

Even as we begin to enter the "post $p < 0.05$ era," we anticipate that *p*-values will continue to be widely reported. In statistical practice, perhaps the single biggest problem with *p*-values is that they are often misinterpreted in ways that lead to overstating the evidence against the null hypothesis. In this article, we recommend three practices, in increasing level of sophistication, that would help safeguard against such misinterpretation.

Recommendation 0.1: If using the current language of "statistical significance" for a novel discovery, replace the 0.05 threshold with 0.005. Refer to discoveries with a *p*-value between 0.05 and 0.005 as "suggestive," rather than "significant."

Recommendation 0.2: When reporting a *p*-value, p, in a test of the null hypothesis H_0 versus an alternative H_1, also report that the data-based odds of H_1 being true to H_0 being true are <u>at most</u> $1/[-e\,p\,\log p]$, where log is the natural logarithm and e is its constant base.

Recommendation 0.3: Determine and report your prior odds of H_1 to H_0 (i.e., the odds of the hypotheses being true prior to seeing the data), and derive and report the final (posterior) odds of H_1 to H_0, which are the prior odds multiplied by the data-based odds. Alternatively, report that the final (posterior) odds are <u>at most</u> the prior odds multiplied by $1/[-e\,p\,\log p]$.

Note that we label these recommendations 0.1, 0.2, and 0.3 to emphasize that each is a small step away from current practice and that, even taken all together, they do not bring us to the ultimate destination of the ideal "post $p < 0.05$ era." In later sections, we discuss and motivate each of these recommendations, although reversing the order of the first two, for easier flow.

1.2. Context

After the influential ASA statement on statistical significance and *p*-values (Wasserstein and Lazar 2016) and the call for papers to air different perspectives on the "post $p < 0.05$ era," we considered the possibility that science might not directly move to this new era that many would like to see. After all, major changes regarding significance testing have been repeatedly advocated for roughly 80 years, and it is not clear that this new initiative will succeed where others have failed.

In this context, our goal is to suggest minimal changes that would require little effort for the scientific community to implement. Motivating this goal are our hope that easy (but impactful) changes might be adopted and our worry that more complicated

CONTACT James O. Berger berger@stat.duke.edu Department of Statistical Science, Duke University, Durham, NC 27708-0251.
© 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.

changes could be resisted simply because they are perceived to be too difficult for routine implementation.

For instance, we agree with the ASA statement (Wasserstein and Lazar 2016) that "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." Such dichotomization is often harmful because it obscures the uncertainty that is intrinsic to any statistical evidence (e.g., Amrhein and Greenland 2018; McShane and Gelman 2017). Moreover, use of a fixed level discourages consideration of crucial elements of the problem, such as the magnitude of the effect, costs of false positives and false negatives (or more generally the loss function in a decision), prior information, etc. (Lakens et al. 2018; Bayarri et al. 2016). Indeed, in our own statistical practice, we strive to fully use all such information. But we need to recognize the possibility that many (most?) investigators will simply find giving up on "statistical significance" too difficult and will end up staying with the current system.

We believe that our recommendations, if adopted, will move the scientific community toward the ideal "post $p < 0.05$ era": if $p < 0.05$ no longer provides a license for treating a conclusion as "true," then seriously thinking about other elements of the problem in the statistical analysis (e.g., transparency of statistical analysis and reporting) will become more appealing. We thus view the recommendations merely as initial, temporary steps on the way to the final destination, but steps that make significant progress immediately. Once our recommendations are adopted, we will advocate for further steps that move us yet closer to the ideal, holistic approach to assessing statistical evidence summarized in the ASA statement (Wasserstein and Lazar 2016).

Finally, note that our recommendations only address the "significance" part of the problem. As it is put in the ASA statement (Wasserstein and Lazar 2016): "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result." Assessing the magnitude of the effect (i.e., "practical significance") is typically at least as important. We have our preferred methods for making such assessments, but we do not discuss them here because our focus in this article is on the use of p-values. Nonetheless, we repeat that it is essential to assess practical importance.

2. The Strength of Evidence Question (SOEQ)

Our recommendations are intended to help researchers, journals, and other parties more accurately answer the question they commonly use p-values to answer: How strongly does the evidence from the data favor the alternative hypothesis relative to the null hypothesis? We call this question the Strength of Evidence Question (SOEQ). For example, in an experiment, researchers want to know how strong the evidence from the data is that some treatment has an effect, as opposed to no effect.

Of the many misinterpretations of p-values (for some examples, see Goodman 2008; Greenland et al. 2016), some are attempts to use the p-value to answer the SOEQ. For example, one possible misunderstanding is that the p-value is the probability that the test statistic would have its observed value under the null hypothesis as opposed to the alternative hypothesis. Thus, a p-value of 0.05 is interpreted as meaning that the observed data only had a 5% chance of occurring if the null hypothesis were true but a 95% chance of occurring if the alternative hypothesis were true. Under this misinterpretation, $p = 0.05$ corresponds to odds of 19:1 in favor of the alternative hypothesis relative to the null hypothesis. This misinterpretation generates a perception that a p-value of 0.05 provides *much* greater evidence against the null hypothesis than it actually does. (As we explain below, $p = 0.05$ actually corresponds to odds of *at most* 2.44:1 in favor of the alternative hypothesis relative to the null hypothesis.)

The correct definition of the p-value is the probability, under the null hypothesis, of observing a test statistic as extreme or more extreme than its observed value. Thus, the p-value cannot provide a direct answer to the SOEQ for two reasons. First, the p-value is evaluated only under the null hypothesis and, hence, cannot say anything directly about the comparison between the null and alternative hypotheses. (For some purposes, this property may be touted as a strength of the p-value, but in terms of providing an answer to the SOEQ, it is not a strength.) Second, it is the probability that the test statistic is *as extreme or more extreme* than its observed value, whereas the SOEQ asks about the likelihood of the data that was obtained, not the likelihood of more extreme data that might have been obtained. In the above example of a misinterpretation, both of these reasons pertain. In our experience, each of these two reasons roughly equally affects the difficulty with the interpretation of p-values.

To a Bayesian, there is a straightforward, direct answer to the SOEQ. This answer is the Bayes factor, defined as

$$\mathrm{BF} \equiv \frac{\text{average likelihood of the observed data under the alternative hypothesis}}{\text{likelihood of the observed data under the null hypothesis}}.$$

(The average in the numerator is taken with respect to an assumed prior distribution of parameter values under the alternative hypothesis. This statement of the definition assumes that the null hypothesis is simple, with the parameter equal to a specific value such as zero; otherwise, the denominator also needs to be an average.) The Bayes factor is the odds of observing the data under the alternative hypothesis to observing it under the null hypothesis. Traditionally, the Bayes factor has been considered unsatisfactory to frequentists because the "average likelihood" in the numerator of BF needs to be computed using some assumed prior distribution for the parameter values under the alternative hypothesis. However, we recently showed, in Bayarri et al. (2016), that BF has a fully frequentist justification for many common situations involving testing a null hypothesis of zero effect versus an alternative hypothesis of nonzero effect. Thus, Bayesians and frequentists can unite in promoting BF as the correct tool for answering the SOEQ.

Compared with the p-value, however, the Bayes factor has two practical disadvantages. First, depending on the assumed prior distribution for the alternative hypothesis, the computation of the numerator of BF may not be as straightforward as running a prepackaged command. Second, specification of the prior distribution (or range of prior distributions) might lead to disagreements, or simply be viewed as too complex. Because of these disadvantages—together with the fact that Bayes factors

are simply less familiar than p-values to most researchers—Bayes factors are unlikely to be adopted quickly for widespread use in answering the SOEQ. In the foreseeable future, it seems likely that researchers will continue misusing p-values for that purpose.

3. Converting p-Values Into Bayes Factors

To extract useful information about the SOEQ from a p-value, we would ideally translate a p-value into a Bayes factor. Unfortunately, there is no unique mapping between p-values and Bayes factors because, unlike calculating the p-value, calculating the Bayes factor requires specifying an alternative hypothesis (more specifically, a prior distribution for the parameter values under the alternative hypothesis).

Fortunately, several methods (Edwards, Lindman, and Savage 1963; Vovk 1993; Johnson 2013; Sellke, Bayarri, and Berger 2001) have been developed for using the p-value to calculate an upper bound on BF, called the Bayes factor bound, which we denote by BFB. The methods for calculating BFB differ from each other in terms of what they assume about the class of alternative hypotheses, but in many relevant applied settings, they generate similar values of BFB (Benjamin et al. 2018). The Bayes factor bound represents the strongest case for the alternative hypothesis relative to the null hypothesis: the highest possible BF consistent with the observed p-value. Because the Bayes factor is justifiable to both Bayesians and frequentists, the Bayes factor bound is, as well.

Held and Ott (2018) provide practical advice about which Bayes factor bound to use depending on the context. For concreteness, we focus here on the simplest formula for calculating a Bayes factor bound

$$\mathrm{BF} \leq \mathrm{BFB} \equiv \frac{1}{-e\,p\log p}\,. \tag{1}$$

The bound was originally derived by Vovk (1993), and Sellke, Bayarri, and Berger (2001) showed that it holds under quite general conditions. This value of BFB can be straightforwardly calculated as a simple function of the p-value, using only the natural logarithm and its base, the constant e. Calculating BFB does not require specifying an alternative hypothesis because it is an upper bound across a large class of reasonable alternative hypotheses.

The following table shows the value of BFB for a wide range of p-values. For those who are more comfortable with posterior probabilities than odds, the table also gives the corresponding upper bound on the posterior probability of H_1. When the prior probabilities of H_0 and H_1 are equal, this upper bound on the posterior probability of H_1 is given by $\Pr^U(H_1 \mid p) = BFB/(1+BFB)$.

p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
BFB	1.60	2.44	8.13	13.9	52.9	400	3226
$\Pr^U(H_1 \mid p)$	0.62	0.71	0.89	0.933	0.981	0.998	0.9997

These calculations illustrate how p-values often point to much weaker evidence against the null hypothesis than researchers typically assume. Indeed, results that just reach conventional levels of significance do not actually provide very strong evidence against the null hypothesis. For example, a p-value of 0.05 corresponds to a Bayes factor of at most 2.44:1. That is, the data imply odds in favor of the alternative hypothesis relative to the null hypothesis of at most 2.44 to 1. So if the null and alternative hypotheses were originally equally likely, there remains, at least, a 29% chance that the null hypothesis is true. A p-value of 0.01—often considered "highly significant"—corresponds to at most 8.13 to 1 odds, hardly overwhelmingly convincing odds. In this case, upon starting from equally likely null and alternative hypotheses, there remains at least an 11% chance that the null hypothesis is true.

4. Justification of the Recommendations

The three recommended practices all involve interpreting p-values in light of their corresponding BFB. Doing so provides some safeguard against overestimating the strength of evidence from the p-value.

4.1. Recommendation 0.2

Recommendation 0.2 codifies the idea of converting a p-value into interpretable odds. To be concrete, in place of the current practice of reporting "$p = 0.05$," we advocate reporting "$p = 0.05$, BFB = 2.44." Since BFB indicates the strongest potentially justifiable inference from the data, reporting it would alert researchers when seemingly strong evidence is actually not very compelling. Its use would therefore help prevent researchers from being misled into concluding too much from the p-value of a finding.

Relative to the crude Recommendation 0.1, Recommendation 0.2 has the virtue of moving beyond a bright-line threshold and interpreting the p-value, correctly, as a continuous variable. Thus, under Recommendation 0.2 (and the follow-on Recommendation 0.3), the language of "significant" and "suggestive" can be dispensed with altogether.

Interestingly, although BFB is only an upper bound on the Bayes factor, we report evidence in Bayarri et al. (2016) that, when calculated from real data from a range of scientific fields, BFB is often not that far from the BF implied by a scientifically reasonable alternative hypothesis.

4.2. Recommendation 0.1

Recommendation 0.1 is a crude follow-up to the idea of interpreting a p-value in terms of its corresponding BFB. The recommendation is that if—despite its inappropriateness—a researcher, journal, or other party insists on using a threshold for "statistical significance" for assessing the credibility of a novel discovery, using 0.005 as that threshold is better than using 0.05. The argument against a 0.05 threshold is clear from the table above: a p-value of 0.05 corresponds to odds of at most 2.44 to 1 against the null hypothesis, which is fairly weak evidence. If researchers treat findings with p-values just below 0.05 as having been established, then many "established" findings will turn out to be false positives and will fail to replicate. This is a purely statistical fact; it will be true even if the original study had no other problems (such as poor study design or multiple hypothesis testing) that could lead to false

positives. We believe that this statistical fact has contributed to the alarmingly high levels of nonreplicability that have been documented in several research communities (e.g., Open Science Collaboration 2015; Camerer et al. 2016; Johnson et al. 2017). It should be emphasized that we are discussing evidence for novel discoveries here; odds of 2.44 to 1 may be adequate for other purposes, such as replicating a previous finding.

While any threshold is necessarily arbitrary, the table above shows that a p-value of 0.005 can correspond to much stronger evidence, with odds of up to 13.9 to 1 against the null hypothesis. We suspect that for most researchers, the term "statistical significance" connotes odds of this order of magnitude. Thus, the redefinition to a threshold of 0.005 would align the term's actual meaning with how it is interpreted. Relabeling findings with p-values between 0.05 and 0.005 as "suggestive" would similarly communicate more accurately the answer to the SOEQ implied by a p-value in that range. The recommendation to redefine statistical significance to the 0.005 threshold has been made several times previously (e.g., Greenwald et al. 1996; Johnson 2013), and we recently joined many other researchers in endorsing it (Benjamin et al. 2018).

We leave it up to the authors and readers of papers to judge whether a finding is a "novel discovery," just as such judgment is left up to the authors and readers of papers today. Researchers have an incentive to claim novelty if they can do so credibly, since novelty is (appropriately) rewarded by the scientific community. Recommendation 0.1, however, would imply that such claims must be accompanied by a stricter threshold for describing findings as "significant."

In Benjamin et al. (2018), we responded to four potential objections to *Recommendation 0.1*. Here, we briefly summarize one of these and our response to it. The objection is that changing the significance threshold from 0.05 to 0.005 would cause an unacceptable increase in the rate of false negatives. Our response is 2-fold. First, failing to reject the null hypothesis does not mean accepting the null hypothesis. Indeed, relabeling p-values between 0.05 and 0.005 as "suggestive" provides a middle ground between acceptance and rejection. Second, the false negative rate will not increase if sample sizes are increased so that statistical power is held constant. For a wide range of common statistical tests, holding statistical power constant would require an increase in sample sizes of roughly 70%. Such an increase is not trivial, but it is achievable in many domains of research, and we believe that the gains in credibility of results are worth the cost.

We caution, however, that even a p-value of 0.005 does not ensure that the evidence against the null hypothesis is strong. The corresponding values of BFB are upper bounds, and the true strength of evidence will sometimes be much weaker. (This would occur if the observed evidence is not only highly inconsistent with the null hypothesis but also highly inconsistent with reasonable alternative hypotheses.) Nonetheless, Recommendation 0.1 would guard against weak evidence being mistaken for strong evidence.

Although we have put forth Recommendation 0.1, we strongly discourage its long-term and widespread use. We fully agree with the ASA statement on statistical significance and p-values (Wasserstein and Lazar 2016) that "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." Recommendation 0.1 should be adopted only temporarily and only in those research fields in which access to statistical expertise is insufficient to implement Recommendations 0.2 or 0.3 (or approaches beyond Recommendation 0.3).

4.3. Recommendation 0.3

While researchers often ask themselves the SOEQ, the question they really want to answer is a Bayesian one: How likely is the alternative hypothesis relative to the null hypothesis? We call this the More Likely Hypothesis Question (MLHQ). In a typical experiment, this question is: how likely is it that there is truly an effect of the treatment, as opposed to no effect? This is the crucial question for understanding what we should conclude from the results of a research study. For instance, does $p = 0.01$ really mean the same thing in a study of a new medical treatment, a test of whether a particular gene is related to a particular disease, and an investigation of extrasensory perception? Clearly, the answer to the MLHQ should depend on the prior odds of the alternative hypothesis relative to the null hypothesis. While it might be reasonable to assign prior odds of 1 to 1 for the medical treatment, prior odds for genetic studies are often chosen to be 1 to 100,000 (Wellcome Trust Case Control Consortium, 2007), and prior odds for extrasensory perception might be extremely small for most scientists.

To a Bayesian, the MLHQ can be answered straightforwardly —without any need for a p-value—by multiplying the Bayes factor by the prior odds, yielding what we have called the "post-experimental odds" (Bayarri et al. 2016). To be concrete about Recommendation 0.3 in cases where the Bayes factor has been calculated, analyses should include a statement such as "the Bayes factor of 4:1, when combined with our prior odds of 1:2, implies post-experimental odds of $\frac{4}{1} \times \frac{1}{2} = \frac{2}{1}$ in favor of the alternative hypothesis."

Recommendation 0.3 also accommodates cases where, instead of the Bayes factor, only a p-value is available. Then, according to our Recommendation 0.3, analyses should include a statement such as "the p-value of 0.05 corresponds to the upper bound on the Bayes factor of 2.44 which, combined with our prior odds of 1:2, implies post-experimental odds of at most $\frac{2.44}{1} \times \frac{1}{2} = \frac{1.22}{1}$ in favor of the alternative hypothesis."

The prior odds can be informed by researchers' beliefs, scientific consensus, and evidence from related research questions. See, for example, Dias, Morton, and Quigley (2018) for a review of methods for eliciting subjective prior odds, Dreber et al. (2015) for an estimate of prior odds based on researchers' beliefs elicited using prediction markets, and Johnson (2013) for an estimate of prior odds based on evidence from replication studies. If there is a range of prior odds, then the corresponding range of post-experimental odds should be calculated and presented. The prior odds should be justified by the researchers, and this justification should be critically evaluated by a paper's referees and readers. To ensure that the prior odds are chosen before the data are seen, it should be expected that the prior odds and their justification be pre-registered.

In *Recommendation 0.1*, the threshold of 0.005 was stated to apply to "novel discoveries." The intent of this language was to

highlight "novel" situations where the prior odds of a discovery are no higher that 1 to 1. Then the post-experimental odds of a discovery can be no higher than the corresponding BFB and, hence, no larger than 13.9 to 1 for a p-value equal to this threshold of 0.005. In some scenarios—such as replicated studies or Phase III clinical trials (where there is considerable prior evidence available from Phase I and Phase II trials)—the prior odds in favor of the alternative hypothesis can be considerably larger than 1 to 1. In general, in such situations, it is strongly preferable to combine the information from all studies on a topic (e.g., via meta-analysis) and assess the evidence as a whole.

We put *Recommendation 0.3* last because we understand that being fully transparent about prior beliefs is not part of the current scientific culture. Nonetheless, we believe it should become part of the culture in the "post $p < 0.05$ era" because the prior odds can matter greatly in evaluating the odds of an alternative hypothesis relative to the null hypothesis (see, e.g., Nuzzo 2014). Note that, if Recommendations 0.2 and 0.3 are adopted, there is no need to worry about whether the discovery is "novel" or not; the issue of its prior probability will be addressed directly.

5. Summary

All three of our recommendations involve interpreting a p-value in terms of its corresponding Bayes factor bound BFB. BFB has both Bayesian and frequentist justification, and it is as simple to calculate as the p-value. Moreover, it has a straightforward interpretation as an answer to the SOEQ, which we believe is the question that researchers often aim to answer when they report p-values. While we would encourage researchers to adopt all our recommendations, we have purposely given a hierarchy of recommendations according to simplicity, to pre-empt the claim that it is too difficult to change current practice.

In addition, we admit to having an ulterior motive in promoting use of the Bayes factor bound. If our recommendations are adopted, they may have the useful side effect of making researchers more comfortable with Bayesian tools and ideas, in particular the Bayes factor and prior odds. We hope that this greater familiarity and comfort could help facilitate interest in a much wider range of useful statistical tools that go beyond p-values.

Indeed, regardless of which recommendation is initially adopted in a particular field, we strongly urge continued and rapid progress toward adopting each higher numbered recommendation and then beyond, toward a holistic interpretation of statistical evidence that fully conforms to the principles laid out in the ASA statement on statistical significance and p-values (Wasserstein and Lazar 2016).

Acknowledgments

The authors thank the referees, associate editor, and especially the editor, Allen Schirm, for extensive and helpful comments that considerably improved the article.

Funding

This research was supported by the National Science Foundation, under grants DMS-1007773, DMS-1407775, and BCS-1521855.

References

Amrhein, V., and Greenland, S. (2018), "Remove, Rather Than Redefine, Statistical Significance," *Nature Human Behaviour*, 2, 4. [187]

Bayarri, M. J., Benjamin, D., Berger, J., and Sellke, T. (2016), "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses," *Journal of Mathematical Psychology*, 72, 90–103. [187,188,189]

Benjamin, D., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Eferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfeld, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A, Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [188,189]

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016), "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436. [189]

Dias, L. C., Morton, A., and Quigley, J. (2018), *Elicitation: The Science and Art of Structuring Judgement*, Cham: Springer. [189]

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015), "Using Prediction Markets to Estimate the Reproducibility of Scientific Research," *Proceedings of the National Academy of Sciences*, 112, 15343–15347. [189]

Edwards, W., Lindman, H., and Savage, L. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242. [188]

Goodman, S. (2008), "A Dirty Dozen: Twelve *P*-Value Misconceptions," *Seminars in Hematology*, 45, 135–140. [187]

Greenland, S., Senn, S., Rothman, K., Carlin, J., Poole, C., Goodman, S., and Altman, D. (2016), "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [187]

Greenwald, A. G., Gonzales, R., Harris, R. J., and Guthrie, D. (1996), "Effect Sizes and *p* Values: What Should Be Reported and What Should Be Replicated?," *Psychophysiology*, 33, 175–183. [189]

Held, L., and Ott, M. (2018), "On *P*-Values and Bayes Factors," *Annual Review of Statistics and Its Application*, 5, 393–419. [188]

Johnson, V. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110, 19313–19317. [188,189]

Johnson, V., Payne, R., Wang, T., Mandal, S., and Asher, A. (2017), "On the Reproducibility of Psychological Science," *Journal of the American Statistical Association*, 112, 1–10. [189]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., van Calster, B., Carlsson, R., Chin Chen, S., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., Lino de Oliveira, C., Orban de Xivry, J.-J., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smit, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., and Zwaan, R. A. (2018), "Justify Your Alpha," *Nature Human Behaviour*, 2, 168–171. [187]

McShane, B. B. and Gelman, A. (2017), "Abandon Statistical Significance," *Nature*, 551(7682), 582. [187]

Nuzzo, R. (2014), "Scientific Method: Statistical Errors," *Nature News*, 506, 150. [190]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [189]

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of *p* Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71. [188]

Vovk, V. G. (1993), "A Logic of Probability, With Application to the Foundations of Statistics," *Journal of the Royal Statistical Society*, Series B, 55, 317–351. [188]

Wasserstein, R., and Lazar, N (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [186,187,189,190]

Wellcome Trust Case Control Consortium (2007), "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls," *Nature*, 447, 661–678.

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

The False Positive Risk: A Proposal Concerning What to Do About *p*-Values

David Colquhoun

To cite this article: David Colquhoun (2019) The False Positive Risk: A Proposal Concerning What to Do About *p*-Values, The American Statistician, 73:sup1, 192-201, DOI: 10.1080/00031305.2018.1529622

To link to this article: https://doi.org/10.1080/00031305.2018.1529622

© 2019 American Statistical Association

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 13254

View related articles

View Crossmark data

Citing articles: 52 View citing articles

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, 192–201: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529622

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The False Positive Risk: A Proposal Concerning What to Do About *p*-Values

David Colquhoun

University College London, London, UK

Abstract

It is widely acknowledged that the biomedical literature suffers from a surfeit of false positive results. Part of the reason for this is the persistence of the myth that observation of $p < 0.05$ is sufficient justification to claim that you have made a discovery. It is hopeless to expect users to change their reliance on *p*-values unless they are offered an alternative way of judging the reliability of their conclusions. If the alternative method is to have a chance of being adopted widely, it will have to be easy to understand and to calculate. One such proposal is based on calculation of false positive risk(FPR). It is suggested that *p*-values and confidence intervals should continue to be given, but that they should be supplemented by a single additional number that conveys the strength of the evidence better than the *p*-value. This number could be the minimum FPR (that calculated on the assumption of a prior probability of 0.5, the largest value that can be assumed in the absence of hard prior data). Alternatively one could specify the prior probability that it would be necessary to believe in order to achieve an FPR of, say, 0.05.

ARTICLE HISTORY
Received February 2018
Revised July 2018

KEYWORDS
Bayes; False positive; False positive report probability; False positive risk; FPR; Likelihood ratio; Point null; Positive predictive value

1. Introduction

It is a truth universally acknowledged that some areas of science suffer from a surfeit of false positives.

By a false positive, I mean that you claim that an effect exists when in fact the results could easily have occurred by chance only.

Every false positive that gets published in the peer-reviewed literature provides ammunition for those who oppose science. That has never been more important. There is now strong pressure from within science as well as from those outside it, to improve reliability of the published results.

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.

In the meantime, improvements in statistical practice can help.

In this article, I am referring to the practice of experimentalists, and to the statistical advice given by journal editors, when they do not have a professional statistician has part of the team. The vast majority of published tests of significance are in this category.

The aim is to answer the following 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? In this article, we expand on the proposals made by Colquhoun (2017, 2018) and compare them with some other methods that have been proposed.

2. How Can Statistical Practice by Users and Journals Be Changed?

In the biomedical literature, it is still almost universal to do a test of statistical significance, and if the resulting *p*-value is less than 0.05 to declare that the result is "statistically significant" and to conclude that the claimed effect is likely to be real, rather than chance. It has been known for many decades that this procedure is unsafe. It can lead to many false positives.

Many users still think that the *p*-value is the probability that your results are due to chance (Gigerenzer, Krauss, and Vitouch 2004). That, it seems, is what they want to know. But that is not what the *p*-value tells you.

There can be no doubt that the teaching of statistics to users has failed to improve this practice.

> The standard approach in teaching, of stressing the formal definition of a *p*-value while warning against its misinterpretation, has simply been an abysmal failure. (Sellke, Bayarri and Berger 2001, p. 71)

The ASA statement (Wasserstein and Lazar 2016) gave a cogent critique of *p*-values. But it failed to come to any conclusions about what should be done. That is why it is unlikely to have much effect on users of statistics. The reason for the lack of recommendations was the inability of statisticians to agree. In the opening words of Kristin Lennox, at a talk she gave at Lawrence Livermore labs,

> Today I speak to you of war. A war that has pitted statistician against statistician for nearly 100 years. A mathematical

CONTACT David Colquhoun d.colquhoun@ucl.ac.uk University College London, London WC1E 6BT, UK.
*https://arxiv.org/abs/1802.04888

© 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.

> conflict that has recently come to the attention of the "normal" people. (Lennox 2016)

I suggest that the only hope of having the much-needed improvement in statistical practice is to adopt some sort of compromise between frequentist and Bayesian approaches. A good candidate for this is to convert the observed p-value to the false positive risk (FPR). This should appeal to users because the FPR is what most users still think that p-values give them. It is the probability that your result occurred by chance only, a very familiar concept.

There is nothing very new about "recalibrating" p-values, such that they mean what users mistakenly thought they mean. Various ways of doing this have been proposed, notably by Berger and Sellke (1987), Sellke et al. (2001), Goodman (1993, 1999a,b) and by Johnson (2013a,b).

The misinterpretation of p-values is, of course, only one of the widespread misunderstandings of statistics. It is relatively easy to teach users about the hazards of multiple comparisons and the problem of p-hacking. It seems to be much harder to teach what a p-value tells you and what it does not tell you. The abuse of $p < 0.05$ is still almost universal, at least in the biomedical literature. Journals still give ghastly advice like

> a level of probability (p) deemed to constitute the threshold for statistical significance should be defined in Methods, and not varied later in Results (by presentation of multiple levels of significance). Thus, ordinarily $p < 0.05$ should be used throughout a paper to denote statistically significant differences between groups. (Curtis et al. 2014)

Many people have advised culture change. Desirable though that is, such pleading has had very little effect. Others have said that authors should ask a professional statistician. That is obviously desirable too, but there are nowhere near enough statisticians to make that possible, and especially not enough statisticians who are willing to spend time analyzing your data.

The constraints on improvement are as follows.

1. Users will continue to use statistical tests, with the aim of trying to avoid making fools of themselves by publishing false positive results. That is entirely laudable.
2. Most users will not have access to professional statistical advice. There just is not enough of it (in a form that users can understand, and put into practice).
3. The only way to stop people claiming a discovery whenever they observe $p < 0.05$ is to give them an alternative procedure.
4. Full Bayesian analysis will probably never be adopted as a routine substitute for p-values. It has a place in formal clinical trials that are guided by a professional statistician, but it is too complicated for most users, and experimentalists distrust (rightly, in my view) informative priors. As Valen Johnson said

> …subjective Bayesian testing procedures have not been - and will likely never be- generally accepted by the scientific community. (Johnson 2013a,b)

A special case of Bayesian analysis, the false positive risk *approach*, is simple enough to be widely understood and allows the p-value to be supplemented by a single number that gives a much better idea of the strength of the evidence than a p-value alone.

3. The False Positive Risk Approach

I will consider here only the question of how to avoid false positives when interpreting the result of a single unbiased test of significance. All the other problems of multiple comparisons, inadequate randomization or blinding, p-hacking, etc. can only make matters still worse in real life.

To proceed, we need to define more precisely "the probability that your results are due to chance." This should be defined as the probability, in the light of the p-value that you observe, you declare that an effect is real, when in fact, it is not. It is the risk of getting a false positive. This quantity has been given various names. For example, it has been called the false positive report rate (FPRP) by Wacholder et al. (2004) though they used the *p-less-than* approach (see Colquhoun 2017, sec. 3) which is inappropriate for interpretation of single tests. It was called the false discovery rate by Colquhoun (2014, 2016a,b)— this was a foolish choice because it risks confusion with the use of that term in the context of multiple comparison problems (e.g., Benjamini and Hochberg, 1995). A better term is the false positive rate, or, better still, false positive risk. The latter term was recommended by Prof. D. Spiegelhalter (personal communication), because we are talking about interpretation of a single experiment: we are not trying to estimate long-term error rates.

It should be noted that corrections for multiple comparisons aim to correct only the type 1 error rate. The result is essentially a corrected p-value and is, therefore, open to misinterpretation in the same way as any other p-value.

In order to calculate the FPR, it is necessary to invoke Bayes' theorem. This can be expressed in terms of odds as follows:

$$\text{posterior odds on } \mathrm{H}_1 = \text{Bayes factor} \times \text{prior odds} \tag{1}$$

Or, in symbols,

$$\frac{P(\mathrm{H}_1|\text{data})}{P(\mathrm{H}_0|\text{data})} = \frac{P(\text{data}|\mathrm{H}_1)}{P(\text{data}|\mathrm{H}_0)} \times \frac{P(\mathrm{H}_1)}{P(\mathrm{H}_0)} \tag{2}$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis.

The bad news is that there are many ways to calculate the Bayes factor (see, e.g., Held and Ott, 2018). The good news is that many of them give similar results, within a factor of 2 or so, so whichever method is used, the conclusions are not likely to differ much in practice. In order to simulate t-tests (e.g., as in Colquhoun 2014), it is necessary to specify an alternative hypothesis,for example, that the true effect size is 1 standard deviation. As we shall see below (Section 5, Figure 1), this does not reduce the generality of the conclusions as much as it might appear at the first sight. Thus, we choose to test

$$\mathrm{H}_0 : \theta = 0 \text{ against } \mathrm{H}_1 : \theta = \theta 1 \neq 0$$

One advantage of using a simple alternative hypothesis is that the Bayes factor becomes a simple likelihood ratio, and there is no need to postulate a prior *distribution*, but rather a simple prior probability suffices: $P(\mathrm{H}_1) = 1 - P(\mathrm{H}_0)$ is the probability

that there is a real (nonzero) effect size. Thus, Equation (2) can be written as

$$\frac{P(H_1|\text{data})}{1-P(H_1|\text{data})} = L_{10} \times \frac{P(H_1)}{1-P(H_1)} \tag{3}$$

where the likelihood ratio in favor of H_1, relative to H_0, is defined as L_{10} (following the notation of Held and Ott 2018).

$$L_{10} = \frac{P(\text{data}|H_1)}{P(\text{data}|H_0)} \tag{4}$$

It has often been argued that the likelihood ratio alone is a better way than a p-value to represent the strength of evidence against the null hypothesis, for example, Goodman (1993, 1999a,b).

It would indeed be a big advance if likelihood ratios were cited, instead of, or as well as, p-values. One problem with this proposal is that very few nonprofessional users of statistics are familiar with the idea of likelihood ratios as evidence. Another problem is that it is hard to judge how big the likelihood ratio must be before it is safe to claim that you have good evidence against the null hypothesis. It seems better to specify the result as an FPR, if only because that is what most users still think, mistakenly, that that is what the p-value tells them (e.g., Gigerenzer, Krauss, and Vitouch 2004). The idea is already familiar to users. That gives it a chance of being adopted. The FPR is

$$P(H_0|\text{data}) = 1 - P(H_1|\text{data})$$

so, from Equations (3) and (4), we get

$$\text{FPR} = \frac{1}{1 + L_{10}\frac{P(H_1)}{1-P(H_1)}}. \tag{5}$$

Thus, in the case where the prior probability, $P(H_1) = 0.5$, the FPR is seen to be just another way of expressing the likelihood ratio:

$$\text{FPR} = \frac{1}{1+L_{10}}.$$

Notice that we have chosen to write the likelihood ratio as the odds on H_1. The bigger the value, the more probable is the data given H_1, relative to H_0. Often the likelihood ratio is written the other way up, as $L_{01} = 1/L_{10}$, so large values favor H_0. References to minimum values of L_{01}—the smallest evidence for H_0—are, therefore, equivalent to maximum values for L_{10}, the greatest evidence for there being a real nonzero effect that is provided by the data.

In the case of a simple alternative hypothesis, the likelihood ratio can be calculated in two different ways: the *p-equals* approach or the *p-less-than* approach (e.g. Goodman 1993, 1999b, and sec. 3 in Colquhoun, 2017). Once the experiment is done, the p-value that it produces is part of the observations. We are interested only in the particular p-value that was observed, not in smaller p-values that were not observed. Since the aim of this work is to interpret a single p-value from a single experiment (rather than trying to estimate the long term error rate), the *p-equals* method is the appropriate way to calculate likelihood ratios.

For small samples, for which a t-test is appropriate, the likelihood ratio can be calculated as a ratio of probability densities (see the appendix). The effect of sample size on FPR was investigated by Colquhoun (2017).

Bayes' theorem requires that we specify a prior probability and in the vast majority of cases, we do not know what it is. But I don't think that that means we have to give up on the FPR.

There are two ways to circumvent the inconvenient fact that we cannot give a value for the prior probability that there's a real effect. One is to assume that the prior probability of a real effect, $P(H_1)$, is 0.5, that is, that there is a 50:50 chance of there being a real effect before the experiment was done. The other way is to calculate the prior that you would need to believe in order to achieve an FPR of, say 5%. This so-called reverse Bayes approach was mentioned by Good (1950) and has been used, for example, by Carlin and Louis (1996), Matthews (2001, 2018), Held (2003) and Colquhoun (2017, 2018). In both cases, it would be up to you to persuade your readers and journal editors that the value of the prior probability is reasonable. Inductive inference inevitably has an element of subjectivity (Colquhoun 2016).

The proposal to specify the FPR is likely to be more acceptable to users than most because it uses concepts which are familiar to most users. It starts by calculating p-values and uses the familiar point null hypothesis. The idea of prior probability can be explained via diagnostic screening tests and the results can be explained to users by simulation of tests (e.g., Colquhoun 2014).

Exact calculations can be done using R programs (Colquhoun 2017), or with a web calculator (Colquhoun and Longstaff 2017). The variables involved are the observed p-value, the FPR, and the prior probability that there is a real effect. If any two of these are specified, the third can be calculated. The calculations give the same FPR if the number of observations and the normalized effect size are adjusted to keep the power constant (Figure 1 and Colquhoun 2017), so when doing the calculations these should be chosen to match the power of your experiment: see Section 5.

4. The Point Null

Some people do not like the assumption of a point null that is made in this proposed approach to calculating the FPR, but it seems natural to users who wish to know how likely their observations would be if there were really no effect (i.e., if the point null were true). It is often claimed that the null is never exactly true. This is not necessarily so (just give identical treatments to both groups). But more importantly, it does not matter. We are not saying that the effect size is exactly zero: that would obviously be impossible. We are looking at the likelihood ratio, that is, at the probability of the data if the true effect size were not zero relative to the probability of the data if the effect size were zero. If the latter probability is bigger than the former then clearly we cannot be sure that there is anything but chance at work. This does *not* mean that we are saying that the effect size is exactly zero. From the point of view of the experimenter, testing the point null makes total sense. In any case, it makes little difference if the null hypothesis is a narrow band centered on zero (Berger and Delampady 1987).

5. The Simple Alternative Hypothesis

It was stated above that the need to specify a fixed alternative hypothesis is not as restrictive as it seems at first. This is the

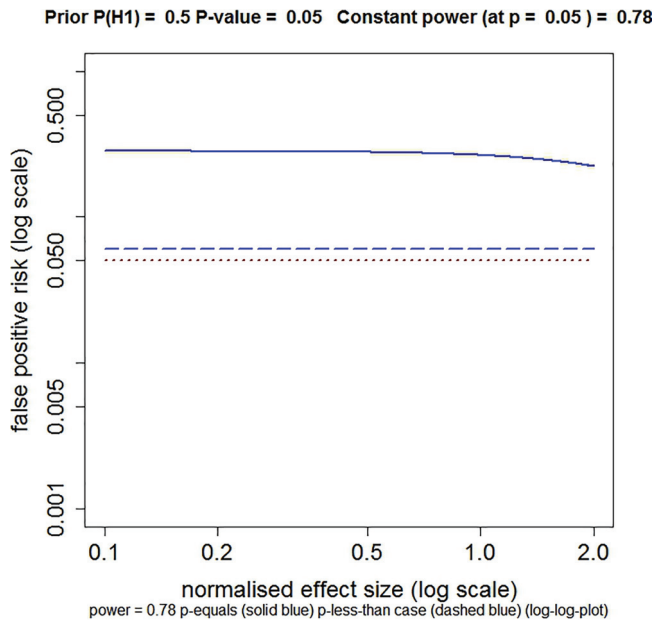


Figure 1. Plot of FPR against normalized effect size, with power kept constant throughout the curves by varying n. The dashed blue line shows the FPR calculated by the *p-less-than* method. The solid blue line shows the FPR calculated by the *p-equals* method. This example is calculated for an observed p-value of 0.05, with power kept constant at 0.78 (power calculated conventionally at $p = 0.05$), with prior probability $P(\mathrm{H}_1) = 0.5$. The sample size needed to keep the power constant at 0.78 varies from $n = 1495$ at effect size $= 0.1$, to $n = 5$ at effect size = 2.0 (the range of plotted values). The dotted red line marks an FPR of 0.05, the same as the observed p-value. Calculated with *Plot-FPR-v-ES-constant-power.R*, output file: *FPR-vs-ES-const-power.txt* (supplementary material).

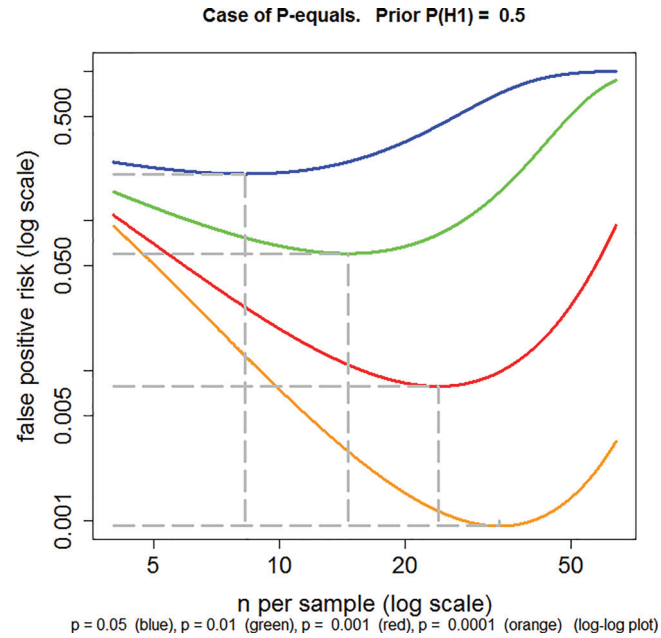


Figure 2. FPR plotted against n, the number of observations per group for a two independent sample t-test, with normalized true effect size of 1 standard deviation. The FPR is calculated by the *p-equals* method, with a prior probability $P(\mathrm{H}_1) = 0.5$ (Equation A6). Log–log plot. Calculations for four different observed p-values, from top to bottom these are: $p = 0.05$ (blue), $p = 0.01$ (green), $p = 0.001$ (red), $p = 0.0001$ (orange). The power of the tests varies throughout the curves. For example, for the $p = 0.05$ curve, the power is 0.22 for $n = 4$ and power is 0.9999 for $n = 64$ (the extremes of the plotted range). The minimum FPR (marked with gray-dashed lines) is 0.206 at $n = 8$. This may be compared with the FPR of 0.27 at $n = 16$, at which point the power is 0.78 (as in Colquhoun, 2014, 2017). Values for other curves are given in the print file for these plots (Calculated with *Plot-FPR-vs-n.R*. Output file: *Plot-FPR-vs-n.txt*. supplementary material).

case because FPR is essentially independent of the effect size if the power of the experiment is constant. This is illustrated in Figure 1, which shows the result of plotting FPR against the true normalized effect size. In this graph, the power is kept constant by changing the sample size, n, for each point. The power is calculated in the standard way, for the normalized effect size and $p = 0.05$. When the FPR is calculated by the *p-less-than* method, it is exactly independent of the effect size. When the FPR is calculated by the *p-equals* method, as is appropriate for our problem, the FPR is essentially independent of effect size up to an effect size of about 1 (standard deviation) and declines only slightly for larger effects. This is the basis for the recommendation that the inputs for the calculator should match the power of the actual experiment.

6. The Jeffreys–Lindley "Paradox"

It may be surprising, at first sight, that the FPR approaches 100% as the sample size is increased when it is calculated by the *p-equals* method, whereas the FPR decreases monotonically with sample size when it is calculated by the *p-less-than* method. The former case is illustrated in Figure 2. As sample size increases, the evidence in favor of the null hypothesis declines at first but then increases.

This phenomenon is an example of the Jeffreys–Lindley "paradox" (Lindley 1957; Bartlett 1957).

It can be explained intuitively as follows.

It is convenient to use the web calculator at *http://fpr-calc.ucl.ac.uk/*. The third radio button shows that for $p = 0.05$, and normalized effect size of 1, a sample of 16 observations in each group gives a power of 0.78. That is usually considered to be a well-powered experiment. Many published experiments have much lower power. For an observed p-value of 0.05, the FPR for $n = 16$ and a prior of 0.5 is 27% when calculated by the *p-equals* method. This is one way of looking at the weakness of the evidence against the null hypothesis that is provided by observing $p = 0.05$ in a single well-powered experiment. Another way is to note that the likelihood ratio in favor of there being a real effect (null hypothesis being untrue) is only about 3 to 1 (2.76, see Colquhoun 2017, sec. 5.1). This is much less than the odds of 19:1 that are often wrongly inferred from the observed $p = 0.05$.

Now to explain the apparent paradox. If the sample size is increased from 16 to 64, keeping everything else the same, the FPR increases from 0.78 (at $n = 16$) to 0.996 (at $n = 64$) when calculated by the *p-equals* method, that is, observation of $p = 0.05$ when $n = 64$ provides strong evidence *for* the null hypothesis. But notice that this sample size corresponds to a power of 0.9999, far higher than is ever achieved in practice. The reason for the increased FPR with $n = 64$ can be understood by looking at the corresponding likelihood ratios. If the calculation for $n = 64$ is done by the *p-less-than* method, the likelihood ratio is 20.0—quite high because most p-values will be less than 0.05 in this situation. But the likelihood ratio found by the *p-equals* method has fallen from 2.76 for $n = 16$, to 0.0045 for $n = 64$. In words, the observation (of $p = 0.05$) would be 222 times as likely (1/0.0045) if the null hypothesis were true

than it would be if the null hypothesis were false. So observing $p = 0.05$ provides *strong evidence for the null hypothesis* in this case. The reason for this is that with $n = 64$ it is very unlikely that a p-value *as big as* 0.05 would be observed: 99.99% of p-values are less than 0.05 (that is the power), and 98.7% of p-values are less than 0.001 (found using the R script *two_sample-simulation-+LR.R*: supplementary material). With a power of 0.9999, it would be very suspicious indeed to observe a p-value *as big as* 0.05.

The resolution of this problem is given in Appendix A4 of Colquhoun (2017), and in the Notes tab of the calculator. Make sure that the power matches that of your experiment. If you use a silly power, you will get a silly answer.

7. Some Alternative Proposals

The suggestion made here is one of several that have been made with the aim of expressing evidence in ways that do not suffer from the same problems as p-values. They are all based on normal distributions apart from the present proposal (Colquhoun 2014, 2017) which is based on Student's t distribution so that its results are appropriate for the small samples that typify most lab work.

1. Many authors have proposed replacement of null hypothesis testing with full Bayesian analysis. These have their place in, for example, large clinical trials with a professional statistical advisor. But they are unlikely to replace the p-value in routine use for two reasons. One is that experimenters will probably never trust the use of informative prior distributions that are not based on hard evidence. The other is that the necessity to test a range of priors makes statement of the results too lengthy and cumbersome for routine use in papers that often cite many p-values.
2. Senn (2015) has produced an example in which the FPR is close to the p-value. This uses a smooth prior distribution, rather than the lumped prior used here, and has, so far, been demonstrated only for one-sided tests. This result differs from the results obtained by simulation of repeated t-tests (Colquhoun 2014) and is in contrast with the general view that the FPR is bigger, often much bigger, than the p-value.
3. It has been proposed, as an interim way to reduce the risk of false positives, that the threshold for significance should be reduced from 0.05 to 0.005 (Benjamin et al. 2017). This proposal has the huge disadvantage that it perpetuates the dichotomization into "significant" and "nonsignificant." And even $p = 0.005$ is not safe for implausible hypotheses (see Colquhoun 2017, sec. 9 and Table 1). The value of $p = 0.005$ proposed by Benjamin et al. (2017) would, in order to achieve an FPR of 5%, require a prior probability of real effect of about $P(H_1) = 0.4$ (from the web calculator, with power = 0.78). It is, therefore, safe only for plausible hypotheses. If the prior probability were only $P(\mathrm{H}_1) = 0.1$, the FPR with $p = 0.005$ would be 24% (from the web calculator: Colquhoun and Longstaff 2017). It would still be unacceptably high even with $p = 0.005$. In order to reduce the FPR to 5% with a prior probability of $P(\mathrm{H}_1) = 0.1$, we would need to observe $p = 0.00045$. This conclusion differs from that of Benjamin et al. (2017) who say that the $p = 0.005$ threshold, with prior = 0.1, would reduce the FPR to 5% (rather than 24%). This is because their calculation is based on the *p-less-than* interpretation which, in my opinion, is not the correct way to look at the problem (see Colquhoun 2017, sec. 3).
4. Matthews (2018) proposed a more sophisticated version of the reverse Bayesian proposal made here. He shows that, in the case where there is no objective prior information, the observation of $p = 0.013$ implies 95% credibility, in the Bayesian sense (though Matthews does not suggest that this should be used as a revised threshold value). This can be compared with the approach of Colquhoun (2017) which says that when you observe $p = 0.013$ that you would need a prior probability of there being a real effect of 0.61 to achieve an FPR of 0.05 (with the default inputs, power = 0.78). Or alternatively (using the third radio button on the web calculator), that the FPR is 7.7% for a prior of 0.5.

   In practice there is not much difference between a prior of 0.61 and 0.5 (or between an FPR of 7.7% and 5%), and to that extent both approaches give similar answers. But my approach would give a much bigger FPR for implausible hypotheses (those with priors, $P(\mathrm{H}_1)$, less than 0.5), but Matthews (2018) would deal with implausible hypotheses in a different way.
5. Bayesian approaches which use a prior distribution that is most likely to reject the null hypothesis. These include Berger and Sellke (1987) and Sellke et al. (2001)—(see the appendix), and V. Johnson's uniformly most powerful Bayesian tests (Johnson 2013a,b), For example, if we observe $p = 0.05$, Berger et al. find an FPR of 22%–29% and Johnson finds 17%–25%. These are comparable with the FPR of 27% found here for a well-powered study with a prior probability of 0.5.
6. Goodman (1999b) has given an expression for calculating the maximum likelihood ratio in favor of H_1 in the case where samples are large enough that the normal distribution, rather than Student's t, is appropriate (see, Equation A9).

Three of these approaches are compared in Figure 3 and in Table 1.

Figure 3 shows the FPR plotted against the observed p-value. The solid blue line is calculated as in Colquhoun (2017), for a well-powered experiment ($n = 16$, normalized effect size = 1 SD), using the *p-equals* method, and prior probability $P(\mathrm{H}_1) = 0.5$ (Equations A5 and A6). The dashed blue line shows the FPR calculated by the Sellke–Berger approach (Equations A8 and A6), and the dotted blue line is calculated by the Goodman approach (Equations A9 and A6). In all cases, the FPR is a great deal bigger than the p-value (equality is indicated by the dotted red line). The Goodman approach agrees quite well with that of Colquhoun (2017) for observed p-values below 0.05, the range of interest. The Sellke–Berger approach gives larger FPR s over this range, but does not differ enough that the conclusions would differ much in practice.

Table 1 shows values calculated for five different observed p-values, shown in column 1. The second column shows the corresponding likelihood ratios, calculated as in Colquhoun (2017), and Equation (A5). These are calculated from the known effect size and standard deviation, but calculations based on the observed effect size and sample standard deviation are similar, especially for p-values below 0.01 (see Colquhoun 2017,

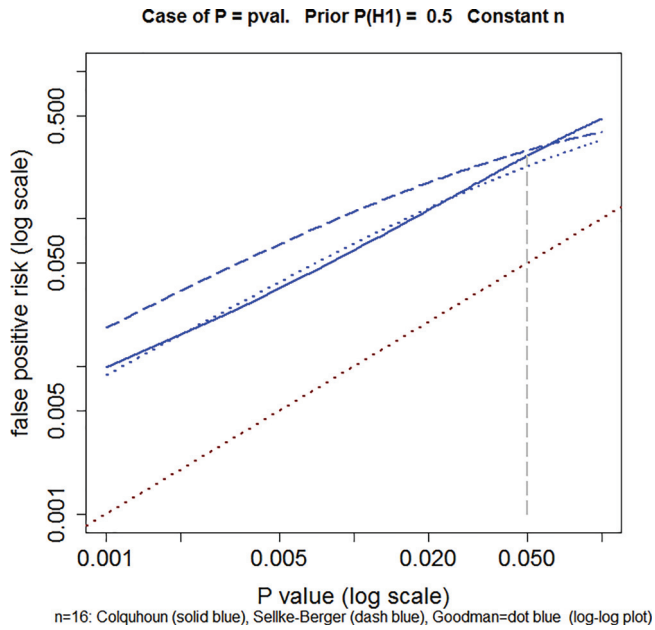


Figure 3. Comparison of three approaches to calculation of FPR in the case of a simple alternative hypothesis. Solid blue line: the *p-equals* method for *t*-tests described in Colquhoun (2017), and Equations (A5) and (A6). Dashed blue line: the Sellke–Berger approach, using Equations (A8) and (A6). Dotted blue line: the Goodman approach, calculated using Equations (A9) and (A6). Dotted red line is where points would lie if FPR was equal to the *p*-value. This example is calculated for $n = 16$, normalized effect size = 1 and prior probability $P(\mathrm{H}_1) = 0.5$. Log–log plot. Calculated with *Plot-FPR-vs-Pval+Sellke-Goodman.R* (supplementary material).

sec. 5.1). For example, the last row of Table 1 shows that if you observed $p = 0.001$, the likelihood ratio in favor of H_1 would be 100, which would normally be regarded as very strong evidence against the null hypothesis. However, this is not necessarily true. If the prior probability $P(\mathrm{H}_1)$ were only 0.1, then Table 1 shows that the FPR would be 0.08, still bigger that the conventional 5% value. In this case, you would need to observe $p = 0.00045$ in order to achieve an FPR of 0.05 (e.g., using radio button 2 on the calculator at *http://fpr-calc.ucl.ac.uk/*)

The third column of Table 1 shows the result of the reverse Bayesian approach, proposed by Matthews (2001) and by Colquhoun (2017), calculated with Equation (A7). If you observed $p = 0.05$, then in order to achieve an FPR = 0.05 you would need to persuade your readers that it was reasonable to assume a very high prior probability, $P(\mathrm{H}_1) = 0.87$, that the observed effect was not merely chance. It is unlikely that anyone would believe that in the absence of hard evidence. If you observed $p = 0.005$ (penultimate row of Table 1) the prior needed to achieve FPR = 0.05 would be 0.4, so you would be safe if the hypothesis was reasonably plausible.

The remaining columns of Table 1 show the FPR s that correspond to each of the *p*-values in column 1. They are calculated as described in the appendix, for prior probabilities, $P(\mathrm{H}_1) = 0.5$ and for $P(\mathrm{H}_1) = 0.1$.

It is gratifying that all three methods of calculating the FPR give results that are close enough to each other that much the same conclusions would be drawn in practice. All of them suggest that if you observe $p = 0.05$ in an unbiased, well-powered experiment, and you claim a discovery on that basis, you are likely to be wrong at least 20–30% of the time, and over 70% of the time if the hypothesis is a priori implausible: ($P(\mathrm{H}_1) = 0.1$).

8. Real-Life Examples

Example 1. The data used by Student (1908).

The results in Table 2 are observations made by Cushny and Peebles (1905) on the sleep-inducing properties of (–)-hyoscyamine (drug A) and (–)-hyoscine (drug B). These data were used W. S. Gosset as an example to illustrate the use of his *t*-test, in the paper in which the test was introduced (Student, 1908). The conventional *t*-test for two independent samples is done as follows (more detail in Colquhoun 1971, sec. 9.4).

Number per sample: $n_A = 10, n_B = 10$.

Mean response: drug A: +0.75; drug B + 2.33.

Standard deviations: drug A $s(y_A) = 1.78901$; drug B $s(y_B) = 2.002249$.

Effect size (difference between means): $E = 2.33 - 0.75 = 1.58$.

We wish to test whether drug B is really better than A or whether the results could have plausibly arisen if the null hypothesis (true difference is zero) were true.

Table 1. Calculations done with web calculator *http://fpr-calc.ucl.ac.uk/*, or with R scripts (see the appendix).

			Minimum FPR (mFPR) prior = 0.5			False positive risk, FPR with prior = 0.1		
Observed *p*-value	Likelihood ratio: ($\mathrm{H}_1/\mathrm{H}_0$)	Prior for FPR = 0.05	Colquhoun value	Sellke–Berger value	Goodman value	Colquhoun value	Sellke–Berger value	Goodman value
0.05	2.8	0.87	0.27	0.29	0.227	0.76	0.79	0.725
0.025	6.1	0.76	0.14	0.20	0.140	0.60	0.69	0.593
0.01	15	0.55	0.061	0.11	0.068	0.37	0.53	0.395
0.005	29	0.40	0.034	0.067	0.037	0.24	0.39	0.259
0.001	100	0.16	0.0099	0.018	0.0088	0.082	0.14	0.074

NOTES: Column 1: The observed *p*-value.
Column 2: Likelihood ratios (likelihood for there being a real effect divided by likelihood of null hypothesis). Calculated for a well-powered (0.78) experiment, as in Equation (A5) and Colquhoun (2017).
Column 3: The prior probability of there being a real effect, $P(\mathrm{H}_1)$, that it would be necessary to postulate in order to achieve an FPR (FPR) of 5%. Calculated for a well-powered (0.78) experiment as in Equation (A7). (The reverse Bayes approach: Colquhoun 2017).
Columns 4–6: The minimum FPR, that is, the FPR that corresponds to prior probability of $P(\mathrm{H}_1) = 0.5$, calculated by three methods, Colquhoun value (calculated for a well-powered (0.78) experiment as in Equations (A5) and (A6) and Colquhoun (2017). Sellke–Berger value: calculated as in Equations (A8) and (A6). Goodman value, calculated as in Equations (A9) and (A6).
Columns 7–9. As columns 4–6, but for an implausible hypothesis with prior probability, $P(\mathrm{H}_1) = 0.1$.

Table 2. Response in hours extra sleep (compared with controls) induced by (–)-hyoscyamine (A) and (–)-hyoscine (B). From Cushny and Peebles (1905).

Drug A y_A	Drug B y_B
+0.7	+1.9
−1.6	+0.8
−0.2	+1.1
−1.2	+0.1
−0.1	−0.1
+3.4	+4.4
+3.7	+5.5
−0.8	+1.6
0.0	+4.6
+2.0	+3.4

The pooled estimate of the error within groups is

$$s(y) = \sqrt{\frac{(n_A - 1)s(y_A)^2 + (n_B - 1)s(y_B)^2}{(n_A - 1) + (n_B - 1)}} = 1.899$$

Normalized effect size $= E/s(y) = 0.83218$.

Standard deviation of effect size $s_E = \sqrt{\frac{s(y)^2}{n_A} + \frac{s(y)^2}{n_B}} = 0.84909$.

Degrees of freedom: $(n_A - 1) + (n_B - 1) = 18$.

Student's t: $t_{18} = \frac{E}{s_E} = 1.8608$.

$P = 0.07918$

The p-value exceeds 0.05 so the difference between means would conventionally be described as "non-significant." A very similar p-value (0.08) is obtained from the same data by the randomization test (Colquhoun 1971, sec. 9.2); or, in more detail, in Colquhoun (2015), in which all 184,756 ways of selecting 10 observations from 20 are plotted.

This conventional p-value result can now be compared with the FPR result. This can be done most conveniently by use of the web calculator (*http://fpr-calc.ucl.ac.uk/*). The inputs for the calculator are the p-value (0.07918), the normalized effect size (0.83218), and the sample size ($n = 10$). If the prior probability of a real effect were 0.5, then the minimum FPR (mFPR) would be 0.28, a lot bigger than the p-value (and it would be bigger still if the hypothesis were thought to be implausible, that is, if its prior probability were less than 0.5). Another way to express the uncertainty is to note that in order to make the FPR 5%, we would have to postulate a prior probability of there being a real effect, $P(\mathrm{H}_1) = 0.88$. Without strong prior evidence, the existence of a real effect is very doubtful. A third way to put the result is to say that the likelihood ratio is only 2.54 so the evidence of the experiment makes the existence of a real effect only 2.54 times as likely as it was before the experiment.

The calculator shows also that the power of this study, calculated for $p = 0.05$ using the observed response size and standard deviation is only 0.42.

Example 2. A study of transcranial electromagnetic stimulation.

A study of transcranial electromagnetic stimulation (TMS), published in *Science*, concluded that TMS "improved associative memory performance," $p = 0.043$ (Wang et al. 2014).

The authors of this article said only "The increase in performance for baseline to Post-Tx was greater for stimulation than for sham $[T_{(15)} = 2.21, p = 0.043]$." The mean raw scores (numbers of words correctly recalled to associated face cues) are given in their Table S2 in the supplementary material. The calculations use the change in score between baseline and post-treatment. The effect size is the difference in this quantity between stimulation and sham. From Table S2, the effect size is 1.88 with standard deviation 1.70 ($n = 8$). The normalized effect size is, therefore, 1.88/1.70 = 1.10, and the post hoc power is 0.54. Rather than saying only "$p = 0.043$," better ways to express the result of this experiment would be as follows (found from the web calculator, radio buttons 3, 1, and 2, respectively). In all cases, the effect size and its confidence interval should be given in the text, not relegated to the supplementary material.

The increase in memory performance was 1.88 ± 0.85 (stand deviation of the mean, i.e., standard error) with confidence interval 0.055–3.7 (extra words recalled on a baseline of about 10). Thus, $p = 0.043$.

This should be supplemented as follows:

(a) The observation of $p = 0.043$ implies a minimum FPR (that for prior probability, $P(\mathrm{H}_1) = 0.5$), of at least 18%. This is the probability that the results occurred by chance only, so the result is no more than suggestive.

or

(b) The increase in performance gave $p = 0.043$. In order to reduce the FPR to 0.05, it would be necessary to assume that we were almost certain (prior probability, $P(\mathrm{H}_1) = 0.81$) that there was a real effect before the experiment was done. There is no independent evidence for this assumption so the result is no more than suggestive.

or

(c) The increase in performance gave $p = 0.043$. In order to reduce the minimum FPR to 0.05, it would have been necessary to observe $p = 0.0043$, so the result is no more than suggestive.

Of these options, the first is probably the best because it is the shortest, and because it avoids the suggestion that FPR = 0.05 is being suggested as a new "bright line," in place of $p = 0.05$.

9. Conclusions

In my proposal, the terms "significant" and "nonsignificant" would not be used at all. This change has been suggested by many other people, but these suggestions have had little or no effect on practice. A p-value and confidence interval would still be stated but they would be supplemented by a single extra number that gives a better measure of the strength of the evidence than is provided by the p-value.

(a) This number could be the minimum FPR, that is, that calculated on the basis that the prior probability of a real effect is 0.5 (the largest value that it is reasonable to assume without hard evidence for a higher value: see Colquhoun, 2017, fig. 3). This has the advantage that it is easy to understand. It has the disadvantage that it is safe only for plausible hypotheses. Nonetheless, it would be a lot better than giving only the p-value and confidence interval.

(b) Or the additional number could be the prior probability that you would need to believe in order to achieve an FPR of, say, 0.05. This has the disadvantage that most users are unfamiliar with the idea of prior probabilities, which are, in most cases subjective probabilities. It also runs the risk that FPR = 0.05 will become a new "threshold," in place of $p = 0.05$.

Examples of improved ways to express the strength of the evidence are given above, at the end of Section 8.

I appreciate that my proposal is only one of many ways to look at the problem of false positives. Other approaches can give different answers. But it has the virtue of being simple to understand and simple to calculate. Therefore it has a chance of being adopted widely.

All the results in this article have been derived in the context of p-values found from t-tests of the difference between means of two independent samples. The question of the extent to which they apply to p-values found in other ways remains to be investigated. Since the logical flaw in the use of p-values as a measure of the strength of evidence is the same however they are derived, it seems likely that the FPRs will be similar in other situations.

Of course any attempt to reduce the number of false positives will necessarily increase the number of false negatives. In practice, decisions must depend on the relative costs (in reputation, or in money, or both) that are incurred by wrongly claiming a real effect when there is none, and by failing to detect a real effect when there is one.

Appendix

In this appendix, we give details of three methods of calculating the likelihood ratio, hence the FPR, for tests of a point null hypothesis against a simple alternative.

Appendix A1. The *p-equals* approach as used by Colquhoun (2017)

The critical value of the t statistic is calculated from the observed p-value as

$$t_{crit} = qt\left(1 - \frac{p}{2}, df, ncp = 0\right) \tag{A1}$$

where $qt()$ is the inverse cumulative distribution of Student's t statistic (the notation is as in R). The arguments are the observed p-value, p, the number of degrees of freedom, df, and the noncentrality parameter, ncp (zero under the null hypothesis)

Numerical values given in this section all refer to the example shown in Figure A1 (reproduced from Colquhoun, 2017). In this case $p = 0.05$, with 30 degrees of freedom so $t_{crit} = 2.04$.

Under the null hypothesis the probability density that corresponds to the observed p-value is

$$y_0 = dt\left(tcrit, df, ncp = 0\right) \tag{A2}$$

where $dt()$ is the probability density function of Student's t distribution with degrees of freedom df and noncentrality parameter = zero under the null hypothesis. This is the value marked y_0 in Figure A1, and in that example its value is 0.0526.

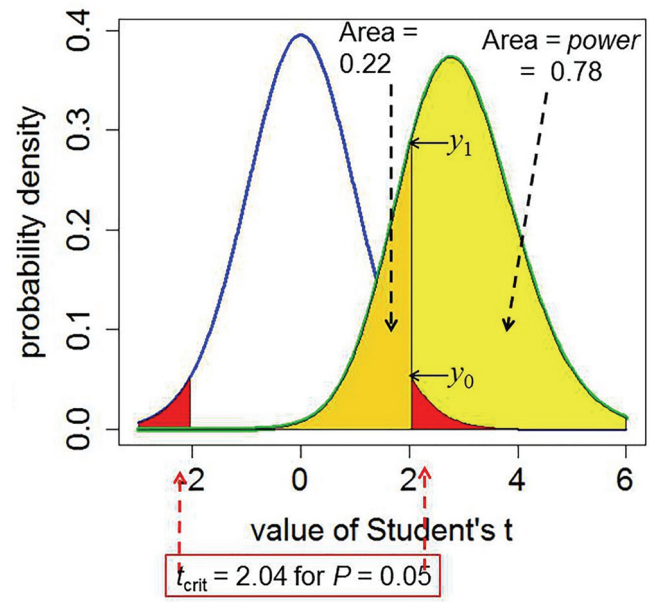


Figure A1. Definitions for a null hypothesis significance test: reproduced from Colquhoun (2017). A Student's t-test is used to analyze the difference between the means of two groups of $n = 16$ observations. The t value, therefore, has $2(n-1) = 30$ d.f. The blue line represents the distribution of Student's t under the null hypothesis (H_0): the true difference between means is zero. The green line shows the noncentral distribution of Student's t under the alternative hypothesis (H_1): the true difference between means is 1 (1 SD). The critical value of t for 30 d.f. and $p = 0.05$ is 2.04, so, for a two-sided test, any value of t above 2.04, or below -2.04, would be deemed "significant." These values are represented by the red areas. When the alternative hypothesis is true (green line), the probability that the value of t is below the critical level (2.04) is 22% (gold shaded): these represent false negative results. Consequently, the area under the green curve above $t = 2.04$ (shaded yellow) is the probability that a "significant" result will be found when there is in fact a real effect (H_1 is true): this is the *power* of the test, in this case 78%. The ordinates marked y_0 ($= 0.526$) and y_1 ($= 0.290$) are used to calculate likelihood ratios for the *p-equals* case.

Under the alternative hypothesis, we need to use the noncentral t distribution. The noncentrality parameter is

$$ncp = \frac{E}{s_E} \tag{A3}$$

where E is the difference between means (1 in the example) and s_E is the standard deviation of the difference between means, 0.354, so $ncp = 2.828$. The probability density that corresponds with $t_{crit} = 2.04$ is

$$y_0 = dt\left(tcrit, df, ncp = 2.828\right) \tag{A4}$$

In the example in Figure A1, this is $y_1 = 0.290$.

The likelihood of a hypothesis is defined as a number that is directly proportional to the probability of making the observations when the hypothesis is true. For the *p-equals* case, we want the probability that the value of t is equal to that observed. This is proportional to the probability density that corresponds to the observed value. (More formally, the probability is the area of a narrow band centered on the observed value, but we can let the width of this band go to zero.) For a two-sided test, under the null hypothesis the probability occurs twice, once at $t = -2.04$ and once at $+2.04$. Thus, under the *p-equals* interpretation, the likelihood ratio in favor of H_1 is

$$L_{10} = \frac{L(H_1)}{L(H_0)} = \frac{\text{Prob}(\text{data}|H_1)}{\text{Prob}(\text{data}|H_0)} = \frac{y_1}{2y_0} \tag{A5}$$

In the example in Figure A1, the likelihood ratio is $0.290/(2 \times 0.0526) = 2.76$. The alternative hypothesis is 2.76 times more likely than the null hypothesis.

The likelihood ratio can be expressed as an FPR via Equation (5) in the text

$$\mathrm{FPR} = \frac{1}{1 + L_{10}\dfrac{P(\mathrm{H}_1)}{1 - P(\mathrm{H}_1)}} \tag{A6}$$

The reverse Bayes approach expresses the result as the prior probability, $P(\mathrm{H}_1)$, that would be needed to get a specified FPR. This can be found by solving Equation (A6) for $P(\mathrm{H}_1)$

$$P(\mathrm{H}_1) = \frac{(1 - \mathrm{FPR})}{(1 - \mathrm{FPR} + L_{10}\mathrm{FPR})} \tag{A7}$$

Appendix A2. The Sellke–Berger approach

Sellke et al. (2001) proposed to solve the problem of the unknown prior distribution by looking for an upper bound for the likelihood ratio for H_1 relative to H_0. Expressed as a function of the observed p-value, they suggest

$$L_{10} = \frac{1}{-ep\log(p)} \tag{A8}$$

(this holds for $p < 1/e$, where $e = 2.71828\ldots$). This is the largest odds in favor of rejection of the null hypothesis, H_0, that can be generated by any prior distribution, whatever its shape. It is the choice that most favors the rejection of the null hypothesis. It turns out that nevertheless, it rejects the null hypothesis much less often than the p-value. As before, the likelihood ratio from Equation (A8) can be converted to an FPR via Equation (A6), or to the reverse Bayes approach with Equation (A7).

For the case where the prior probability of having a real effect is $P(\mathrm{H}_1) = P(\mathrm{H}_0) = 0.5$, this can be interpreted as the minimum false discovery rate (Sellke et al. 2001). It gives the *minimum* probability that, when a test is 'significant', the null hypothesis is true, that is, it is an estimate of the minimum FPR.

The FPR can be calculated by substituting the likelihood ratio from Equation (A8) into Equation (A6), and the reverse Bayesian calculation of $P(H_1)$ can be found by substituting this FPR into Equation (A7).

Appendix A3. The Goodman approach

For large samples, such that the t distribution is well approximated by the normal distribution, Goodman (1999b) gives the maximum likelihood ratio in favor of H_1 as

$$L_{10} = \frac{1}{2\exp(-z^2/2)} \tag{A9}$$

(This is the likelihood ratio calculated at the sample mean, the maximum likelihood estimate of the true effect size.)

In Equation (A9), z is the observed standard normal deviate, which is related to the observed p-value, according to

$$z = q\mathrm{norm}\left(1 - \frac{p}{2}, 0, 1\right)$$

Here, $q$norm is the inverse cumulative probability distribution function for a standard normal distribution (the notation is as in R). The factor of 2 on the bottom line of Equation (A9) is needed since we are talking about a two-sided test: see Held and Ott, 2018, eq. 13).

The FPR can be calculated by substituting L_{10} from Equation (A9) into Equation (A6). The reverse Bayesian calculation of $P(\mathrm{H}_1)$ can be found by substituting this FPR into Equation (A7).

Supplementary Materials

R scripts (and output files) used for calculations and plots are provided as online supplementary materials.

Acknowledgments

I am especially grateful to the following people: Dr Leonhard Held Epidemiology, Biostatistics and Prevention Institute, University of Zurich, CH-8001 Zurich, for helpful discussions; Dr R.A.J. Matthews for helpful discussions about his paper (Matthews 2018); Prof. Stephen Senn (Competence Center in Methodology and Statistics, CRP-Santé, Luxembourg); and Prof. D. Spiegelhalter for reading an earlier version of Colquhoun (2017) and suggesting the term "false positive risk."

References

Bartlett, M. S. (1957), "A Comment on D.V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534. [195]

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., et al. (2017), "Redefine Statistical Significance," in *PsyRxiv*, available *https://psyarxiv.com/mky9j/* [196]

Benjamini, Y. A., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300, available at *http://www.math.tau.ac.il/~ybenja/MyPapers/benjamini_hochberg1995.pdf* [193]

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–352. [194,196]

Berger, J. O., and Sellke, T. (1987), "Testing A Point Null Hypothesis – the Irreconcilability of P-Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [193]

Carlin, B. P., and Louis, T. A. (1996), "Identifying Prior Distributions that Produce Specific Decisions, with Application to Monitoring Clinical Trials," in *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, Chaloner and J. Geweke, New York: Wiley, pp. 493–503. [194]

Colquhoun, D. (1971) *Lectures on Biostatistics*, London, UK: Oxford University Press, available at *http://www.dcscience.net/Lectures_on_biostatistics-ocr4.pdf*. [197,198]

——— (2014), "An Investigation of the False Discovery Rate and the Misinterpretation of p-values," *Royal Society Open Science*, 1, 140216, available *http://rsos.royalsocietypublishing.org/content/1/3/140216* [193,194,195,196]

——— (2015), "The Perils of *p*-values," *Chalkdust Magazine*, available at *http://chalkdustmagazine.com/features/the-perils-of-p-values/* [198]

——— (2016), "The Problem with *p*-values," *Aeon Magazine*, available at *https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant* [193,194]

——— (2017), "The Reproducibility of Research and the Misinterpretation of *P* Values," *Royal Society Open Science*, 4, 171085, available at *http://rsos.royalsocietypublishing.org/content/4/12/171085*. [192,194,195,196,197,198,199,200]

——— (2018), "The False Positive Risk: A Proposal Concerning What to Do About p-values (version 2)," available at *https://www.youtube.com/watch?v=jZWgijUnIxI* [192,194]

Colquhoun, D., and Longstaff, C. (2017), "False Positive Risk Calculator", available *http://fpr-calc.ucl.ac.uk/* [193,194,196]

Curtis, M. J., Bond, R. A., Spina, D., Ahluwalia, A., Alexander, S. P. A., Giembycz, M. A., Gilchrist, A., Hoyer, D., Insel, P. A., Izzo, A. A., et al. (2014), "Experimental Design and Analysis and their Reporting: New Guidance for Publication in BJP," *British Pharmacological Society*, available at *http://onlinelibrary.wiley.com/doi/10.1111/bph.12856/full* [193]

Cushny, A. R., & Peebles, A. R. (1905), "The Action of Optical Isomers: II. Hyoscines," *The Journal of Physiology*, 32, 501–510, available at *https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1905.sp001097* [197,198]

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004), "The Null Ritual. What You Always Wanted to Know About Significance Testing but Were Afraid to Ask," in ed. D. Kaplan, *The Sage Handbook of Quantitative Methodology*

for the Social Sciences, Thousand Oaks, CA: Sage Publishing, pp. 391–408. [192,194]

Good, I. J. (1950), *Probability and the Weighing of Evidence*, London: Charles Griffin and Co. Ltd. [194]

Goodman, S. N. (1993), "p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496. [193,194]

——— (1999a), "Toward Evidence-based Medical Statistics. 1: The P value Fallacy," *Annals of Internal Medicine*, 130, 995–1004. [193,194]

——— (1999b), "Toward Evidence-based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013. [194,196,200]

Lennox, K. (2016), "All About that Bayes: Probability, Statistics, and the Quest to Quantify Uncertainty," available at *https://www.youtube.com/watch?v=eDMGDhyDxuY* [193]

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192. [195]

Held, L. (2013), "Reverse-Bayes Analysis of Two Common Misinterpretations of Significance Tests," *Clinical Trials (London, England)*, 10, 236–242. [194]

Held, L., and Ott, M. (2018), "On p-Values and Bayes Factors," *Annual Review of Statistics and Its Application*, 5, 393–419 [193,194,200]

Johnson, V. E. (2013a), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19313–19317. [193,196]

——— (2013b), "Uniformly Most Powerful Bayesian Tests," *The Annals of Statistics*, 41, 1716–1741. [2]

——— (2018), "Beyond 'Significance': Principles and Practice of the Analysis of Credibility," *Royal Society Open Science*, 5, 171047, available *http://rsos.royalsocietypublishing.org/content/5/1/171047* [194,196,200]

Matthews, R. A. J. (2001), "Why Should Clinicians Care About Bayesian Methods?," *Journal of Statistical Planning and Inference*, 94, 43–58. [194,197]

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of *p* Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71. [192,193,196,200]

Senn, S. J. (2015), "Double Jeopardy?: Judge Jeffreys Upholds the Law (Sequel to the Pathetic P-value)," in *Error Statistics Philosophy*, available at *https://errorstatistics.com/2015/05/09/stephen-senn-double-jeopardy-judge-jeffreys-upholds-the-law-guest-post/2017* [196]

Student (1908), "The Probable Error of a Mean," *Biometrika*, 6, 1–25. [197]

Wacholder, S., Chanock, S., Garcia-Closas, M., El, G. L., and Rothman, N. (2004), "Assessing the Probability that a Positive Report is False: An Approach for Molecular Epidemiology Studies," *Journal of the National Cancer Institute*, 96, 434–442. [193]

Wang, J. X., Rogers, L. M., Gross, E. Z., Ryals, A. J., Dokucu, M. E., Brandstatt, K. L., Hermiller, M. S., and Voss, J. L. (2014), "Targeted Enhancement of Cortical-hippocampal Brain Networks and Associative Memory," *Science*, 345, 1054–1057. [198]

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. [192]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility

Robert A. J. Matthews

To cite this article: Robert A. J. Matthews (2019) Moving Towards the Post $p<0.05$ Era via the Analysis of Credibility, The American Statistician, 73:sup1, 202-212, DOI: 10.1080/00031305.2018.1543136

To link to this article: https://doi.org/10.1080/00031305.2018.1543136

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 7136

View related articles

View Crossmark data

Citing articles: 17 View citing articles

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, 202–212: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543136

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility

Robert A. J. Matthews

Department of Mathematics, Aston University, Birmingham, UK

ABSTRACT

It is now widely accepted that the techniques of null hypothesis significance testing (NHST) are routinely misused and misinterpreted by researchers seeking insight from data. There is, however, no consensus on acceptable alternatives, leaving researchers with little choice but to continue using NHST, regardless of its failings. I examine the potential for the Analysis of Credibility (AnCred) to resolve this impasse. Using real-life examples, I assess the ability of AnCred to provide researchers with a simple but robust framework for assessing study findings that goes beyond the standard dichotomy of statistical significance/nonsignificance. By extracting more insight from standard summary statistics while offering more protection against inferential fallacies, AnCred may encourage researchers to move toward the post $p < 0.05$ era.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
Analysis of credibility; Bayesian inference; Null hypothesis Significance testing; *p*-Values

1. Introduction

The process of turning data into insight is central to the scientific enterprise. It is therefore remarkable that the most widely used approach—null hypothesis significance testing (NHST)—has been subjected to devastating criticism for so long to so little effect. The emergence of the so-called replication crisis has catalyzed widespread debate about NHST in general and the use of *p*-values in particular (e.g., Nuzzo 2014; Baker 2016). This debate reached a turning-point in March 2016 with the publication of an unprecedented statement by the American Statistical Association (ASA) stating that *p*-values are "commonly misused and misinterpreted" (Wasserstein and Lazar 2016) and calling for researchers to move toward a "post $p < 0.05$ era."

However, the statement stopped short of giving specific recommendations on how this transition should be achieved, stating only that statisticians "often supplement or even replace *p*-values" through, for example, emphasizing estimation, Bayesian methods, and false discovery rates. This absence of clear guidance has led to concern that the statement will become yet another failed attempt to deal with the threat posed by NHST to reliable scientific inference (see, e.g., Goodman 2016a, Matthews, Wasserstein, and Spiegelhalter 2017).

In what follows, I examine the potential of the Analysis of Credibility (AnCred, Matthews 2018) to encourage researchers to embrace more reliable forms of inference. Its origins lie in a Bayesian technique for assessing clinical trial outcomes in the context of existing knowledge (Matthews 2001a, 2001b; Spiegelhalter, Abrams, and Myles 2004). Its underlying theory has since found applications in health care evaluation (Spiegelhalter 2004), epidemiology (Rothman et al. 2008), risk assessment (Greenland 2011), and interpretive issues in inference (Held 2013).

In the current context, I examine the two features of AnCred that make it potentially attractive to researchers in any discipline looking to move beyond NHST:

- Its ability to extract more insight from standard summary statistics, allowing researchers to go beyond the simplistic true/false dichotomization encouraged by *p*-values toward a more nuanced assessment of new findings.
- The enhanced protection it offers against common inferential misconceptions promoted by NHST and highlighted by the ASA statement.

I begin by outlining the inferential concepts underpinning AnCred.

2. The Conceptual Basis of the Analysis of Credibility

NHST dichotomizes findings according to whether or not they possess statistical significance, as reflected by $p < 0.05$. Notoriously, this has become the de facto criterion by which researchers judge whether a finding is worth taking seriously or not. As the ASA statement made clear, however, this is an inappropriate role for the *p*-value, which bears no simple relationship to effect size or weight of evidence, and offers no means of setting findings in context. Furthermore, it encourages the dangerously false perception that statistically significant results are in some sense "true," while nonsignificant results are "null" or "negative" outcomes.

These deficiencies have led to calls for the research community to transition from NHST and *p*-values to greater use of estimation via confidence intervals (CIs). The latter are widely recognized as concise but informative summary statistics (e.g., Rothman 1978; Gardner and Altman 1986). The resulting

CONTACT Robert A. J. Matthews rajm@physics.org Department of Mathematics, Aston University, Birmingham B4 7ET, UK.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.
© 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.

inferential benefits are impressive: the bounds of a CI give insight into the best-supported magnitude of an effect, together with its direction and associated uncertainty—none of which flows directly from a p-value. CIs are also better suited to the aggregation of insight and the replication of findings, being more readily combined than p-values, and giving a more intuitive measure of the probability of replication (e.g., Cumming 2008).

While the use of estimation methods is growing, such is the grip of NHST that CIs are routinely converted back into p-values or simply inspected to check their bounds exclude no effect. As a result, most of their inferential content is thrown away. In addition, the meaning of CIs under the NHST paradigm is routinely misunderstood in ways similar to that of p-values (see, e.g., Hoekstra et al. 2014; Greenland et al. 2016).

AnCred is based on the pragmatic view that if the research community is to move beyond NHST, it requires inferential tools that are widely applicable, simple to use, and less prone to misinterpretation. To achieve these goals, AnCred shifts the focus of inference away from null hypothesis testing and toward effect size estimation relevant to the substantive hypothesis, with findings assessed via the concept of *credibility*. With its origins in Bayesian inference, credibility reflects both the evidential weight of a finding, as captured by the so-called likelihood, and the level of support for the finding provided by existing knowledge as reflected by a prior distribution. Bayesian inference then usually combines this prior information with the likelihood using Bayes' theorem leading to a posterior distribution giving the updated level of evidence for the claimed effect, commonly summarized as a 95% *credible interval*. Unlike conventional CIs, this has a straightforward interpretation, in that there is a 95% probability this interval contains the true effect size. Thus, if this posterior credible interval excludes no effect, evidence for a nonzero effect size is said to be "credible at the 95% level."

With its use of credible intervals, AnCred thus provides a bridge between the dominant but flawed NHST paradigm and the less familiar but more informative methods of Bayesian estimation. It also allows the familiar summary statistic of CIs to be retained but with an interpretation in line with the basic research question: do the data provide credible evidence for the substantive hypothesis of a nonzero effect?

Any inferential method based on Bayesian techniques must confront the so-called problem of priors: the vulnerability of inferences to choices of prior distributions biased—unconsciously or otherwise—to give a desired outcome. This has long been seen as a major challenge to the general acceptance of Bayesian methods, and has prompted a variety of methods for dealing with it (e.g., Spiegelhalter, Abrams, and Myles 2004). The approach used in AnCred is to obviate the need to specify a prior by *deducing* it from the data (Good 1950; Carlin and Louis 1996). To do this, Bayes' theorem is inverted and the prior deduced subject to the requirement that when combined with the likelihood it leads to a posterior distribution that excludes no effect. The resulting prior can then be compared with existing evidence to see if it can be justified in the light of what is known. If this is the case, the new finding can be deemed credible evidence of a nonzero effect.

The comparison required by AnCred can be performed in a variety of ways: using CIs to summarize prior evidence from meta-analyses or systematic reviews of previous studies, or ranges based on plausibility arguments. AnCred facilitates the comparison by virtue of maintaining a consistent format for the summary statistics throughout the analysis; for example, if the finding is stated as a 95% CI for an odds ratio (OR), AnCred gives the necessary prior in the same terms. The outcome is a principled and quantitative assessment of the new finding in the context of what is known—in marked contrast to the cursory and qualitative discussion sections typically accompanying NHST-based studies.

As AnCred allows the prior needed for credibility to extracted objectively from the likelihood, concerns about priors being hand-picked to achieve a desired outcome do not arise. Simply put, if a finding is shown to require a level of prior support not found in the existing knowledge base, that finding does not provide credible evidence for a non-zero effect. While selective use of the knowledge base remains possible, AnCred readily exposes such practices and allows the impact of less favorable supporting evidence to be easily determined.

In the case of entirely unprecedented findings, however, comparison with appropriate prior insight may not be possible. In such instances, AnCred assesses the credibility of the finding using the only evidence available: the finding itself. Such *intrinsic credibility* is determined using a criterion given in the next section, whose basis is explained in the Appendix.

With credibility as its defining concept, AnCred encourages a transition away from unthinking dichotomization based on statistical significance. This practice reflects the widespread, but erroneous, belief that $p < 0.05$ is a universally valid criterion for the reality of an effect, regardless of context. AnCred, in contrast, assesses findings on the basis of their credibility, which depends not only on evidential weight, but also support from prior insight. As such, it discourages the belief that findings in isolation can be relied on to demonstrate the truth or otherwise of a hypothesis. The assessment of credibility is necessarily context-dependent and thus open to quantitative challenge and debate by the relevant research community, rather than simple declaration by those responsible for the finding.

The theory and modelling assumptions underpinning the above features of AnCred are summarized in the Appendix. As the remit of this Special Issue is to explore *practical* suggestions for transitioning to the post $p < 0.05$ era, the rest of this article will focus on assessing the extent to which AnCred addresses the real-life challenges involved in making that transition.

3. How AnCred Is Applied in Practice

Any proposal for encouraging changes in inferential practice must accept the ubiquity of NHST. From the training of new researchers through to the stipulations of leading journals, the concepts of p-values, statistical significance and nonsignificance look set to remain widely used for the foreseeable future, whatever their failings. 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.

The formalism of AnCred is aligned with this view. Operationally, it proceeds as follows:

1. The study finding is summarized using standard methods as a 95% CI (L, U); this represents the likelihood, with point estimate M.
2. This likelihood is then used to deduce the range of prior effect sizes which, when combined with the likelihood, lead to a posterior range that just excludes no effect at the 95% level. The resulting *critical prior interval* (CPI) is used to assess the credibility of a finding, according to whether it is statistically significant or not. The appropriate formulas for calculating the CPI in each case are found in the Appendix. For example, if the effect size is stated as a difference in means or proportions, the CPI for statistically significant findings is (-SL, +SL), where SL is the *skepticism limit* calculated from the likelihood (L, U) via Equation (A1). For statistically *non*significant findings, the CPI is (0, AL), where AL is the *advocacy limit* (AL) given by Equation (A3).
3. To assess the credibility of the finding under AnCred, this CPI is compared with effect sizes supported by prior evidence, in the form of CIs or plausible ranges (see Appendix). The criteria used depend on whether the finding is statistically significant, nonsignificant or is unprecedented:

   - For statistically significant results: if prior evidence supports effect sizes at least as large as the calculated SL, the finding is both statistically significant and provides credible evidence for a nonzero effect.
   - For statistically nonsignificant results: if prior evidence supports effect sizes lying within the advocacy CPI, the finding provides credible evidence of a nonzero effect size, despite being nonsignificant.
   - For unprecedented results for which no appropriate prior evidence exists, a finding is both statistically significant and *intrinsically* credible if the point value M of the 95% CI lies outside the calculated CPI.

The examples that follow show how these criteria are applied in practice. As we shall see, one striking feature of the use of AnCred it that statistically significant findings can nevertheless lack credibility, for example, because their evidential weight is too weak to compensate for their implausibility as indicated by prior evidence. Similarly, statistically *non*significant results can still provide evidential support for nonzero effects, and provide useful constraint on plausible effect sizes. This highlights the feature of NHST so often overlooked by researchers, and highlighted by the ASA statement: statistical significance and nonsignificance are unreliable indicators of the presence or absence of genuine effects.

To illustrate the ability of AnCred to bring more inferential power to bear on research questions, I now give some representative worked examples. To fix ideas, these are in the form of prototypical findings whose effect size is summarized by a 95% CI for an OR, the substantive hypothesis being that the OR exceeds unity.

Example 1. OR = 1.57 (1.12, 2.21). This is a statistically significant finding ($p = 0.01$). To assess whether it also provides *credible* evidence for a genuine effect, we turn to AnCred and calculate the appropriate CPI from the summary CI. For statistically significant ORs, we use (A2) to calculate the SL, which in this case is 1.47. Thus, for this finding to be deemed credible, AnCred requires that there is prior evidence for plausible ORs at least as large as this; for example, if the prior evidence for a positive effect is summarized by a CI of (L_o, U_o), then we require $L_o > \text{SL}$. This might be entirely reasonable in some contexts, for example, the outcome of a new educational intervention; in such cases we may conclude that the finding provides both statistically significant *and* credible evidence for a nonzero effect at the 95% level. However, in other contexts—for example, an increase in environmental cancer risk—such an SL may be highly implausible. Either way, AnCred has moved the assessment beyond the standard simplistic true/false dichotomization to a discussion based on both the weight of evidence from the new finding and the context as set by prior knowledge.

Example 2. OR = 1.57 (1.03, 2.40). This finding has the same central effect size as before and is again statistically significant ($p = 0.04$). This much is clear simply by noting the point estimate and also that the range excludes no effect. However, this common practice overlooks a crucial difference from the previous example: the CI is considerably wider and its lower bound runs closer to no effect; as such it represents a weaker level of evidence for a positive effect. AnCred quantifies this, with Equation (A2) leading to an SL of 3.0. Thus, the finding is now only credible evidence for a positive effect if there is prior evidence that plausible ORs are at least as large as 3.0. So, for example, if the prior evidence is summarized by a CI of (L_o, U_o), then $L_o > 3.0$. Intuitively, the more demanding SL reflects the fact that this study provides weaker evidence than the previous example, and thus needs more support from prior evidence to achieve credibility.

If the finding concerned an entirely novel effect for which no clearly relevant prior evidence exists, then the concept of *intrinsic credibility* applies. In this case, we see that the point value M of the CI—and thus best-supported effect size—is an OR of 1.57 which is lower than the SL, and lies within the skepticism CPI of (1/3.0, 3.0). Thus, for an entirely novel finding, this level of evidence would be statistically significant but not intrinsically credible support for a genuine effect.

Example 3. OR = 2.03 (1.03, 4.00). This has the same level of statistical significance as the previous example ($p = 0.04$), but a higher point estimate. Many researchers would thus regard it to be a more impressive finding. However, this again highlights the failings of the common practice of focusing solely on the point estimate and whether the CI excludes no effect. With its upper bound of 4.00 the CI is considerably wider than before, reflecting its lower evidential weight. AnCred reflects this, with (A2) now showing that, despite having the same p-value as before, the wider CI leads to a much more demanding SL: prior evidence that ORs exceed SL ~ 10 is now needed to achieve credibility. AnCred has here highlighted the fact that findings with the same p-value do not necessarily provide the same evidential weight.

Example 4. OR = 1.57 (0.93, 2.65). This result has the same point value as the first two examples, but is formally *non*significant. Furthermore, its p-value of 0.09 lies in the range $0.05 \leq p \leq 0.10$ often characterized by euphemisms such as "approaching significance," "marginally significant," etc. Such findings often prompt researchers to try to claim support for their substantive

hypothesis by blurring the normally sharp cut-off for statistical significance. In contrast, AnCred provides a principled approach to such cases, unpacking the evidential content of the CI and showing the implications. In this example, it confirms the intuition that despite being nonsignificant, there *is* some evidence for a positive effect here, as reflected by its point estimate exceeding unity. AnCred goes further, however, and reveals the weakness of this evidence. Putting the 95% CI into (A4) shows that the advocacy CPI is (1, 1100). This is extremely broad, implying that essentially *any* prior evidence for a positive effect will suffice to make the finding credible support for such an effect, despite its nonsignificance. That is, the width of the CPI is warning that the finding adds essentially nothing to what is already known from prior evidence. As such, claims that its "marginal" significance adds to the evidence for a positive effect are questionable. Even less justifiable, however, would be the common interpretation of any such nonsignificant finding as indicating the absence of any effect at all. Only more evidential weight from additional data can bring the required clarity.

This is in marked contrast to Example 2, whose 95% CI has the same point estimate but is both tighter and entirely excludes no effect, reflecting its much greater evidential weight. This leads to a CPI putting much greater constraint on those seeking to challenge its credibility as evidence of a positive effect.

Example 5. OR = 1.10 (0.80, 1.5). With a p-value of 0.57, this strongly nonsignificant result would typically be dismissed by many researchers as powerful evidence of the *absence* of an effect. As such, it may be given only scant attention in a paper, or indeed remain unpublished. AnCred shows that both responses are inappropriate. Hints of this can be seen simply by inspection of the CI: the central effect is in the direction of a positive effect, and the CI is tighter than in the previous example, suggesting it still carries evidential weight for a positive effect. AnCred confirms this: from Equation (A4) the stated 95% CI leads to AL = 1.5. Thus, despite its nonsignificance, the finding provides credible evidence of a positive effect if prior evidence supports effect sizes lying in the advocacy CPI of (1.0, 1.5). This is much tighter than the previous CPI, despite the similarity of effect size and p-value, essentially because the 95% CI is tighter and the lower bound is further in the direction of a *negative* effect (ORs < 1.0). It thus puts more constraint on prior effect sizes able to produce credible evidence of a positive effect. Specifically, prior evidence for effect sizes wholly within the advocacy CPI of (1.0, 1.5)—for example, summarized by a 95% CI of, say, (1.2, 1.4)—will make the finding credible evidence of a positive effect at the 95% level, despite being nonsignificant. More diffuse prior evidence lacks the "heft" needed to achieve this, even if its upper bound substantially exceeds AL. For example, prior evidence summarized by (1.2, 4.0) is too diffuse to make the finding of a positive effect credible.

AnCred is, however, based on conservative criteria which guarantee credibility when met. Findings may thus still be rendered credible by prior evidence lying outside the AL if it carries sufficient heft—that is its range is sufficiently tight, for example a 95% CI of (1.8, 5.0). Such situations may encourage researchers to perform a specific Bayesian calculation to make their case. However, in doing so, they must also justify going outside the conservative framework of AnCred. Either way, in this example AnCred highlights the inadvisability of simply dismissing nonsignificant findings as "negative." Once set in the context of prior knowledge, even strongly nonsignificant findings can provide credible evidence for nonzero effects.

Example 6. OR = 0.84 (0.45, 1.55). This has the same nonsignificant p-value as before (0.57), but now the point estimate—0.84—contradicts the substantive hypothesis of OR > 1. In such cases, there is no CPI capable of turning this into credible evidence for a positive effect. However, AnCred confirms the intuition that such findings do support the existence of a *negative* effect. To see this, the advocacy CPI is simply reversed to (AL, 1), consistent with a substantive hypothesis of OR < 1, with AL now calculated using the reciprocal of (A4). We then find a CPI of (0.46, 1), implying that the finding is credible evidence for OR < 1 if prior evidence exists for effect sizes lying wholly in this range. AnCred has here highlighted the fact that while a study may not support the expected substantive hypothesis, this does not imply it supports the complete absence of any effect, and/or should therefore be discarded.

As with the previous example, if the conservative criteria of AnCred are relaxed, prior evidence for effect sizes lying beyond the AL may also render the finding credible; this can be investigated via a standard Bayesian analysis.

These prototypical examples illustrate how AnCred goes beyond the standard dichotomy of NHST, extracting additional insight from findings and providing a basis for quantitative discussion of their implications. I now examine the potential for AnCred to offer enhanced protection against the inferential traps highlighted by the ASA in its 2016 statement.

4. AnCred as a Means of Avoiding Inferential Fallacies

The misconceptions surrounding NHST and p-values are so numerous there have been several attempts to curate them. Goodman (2008) focuses on a "Dirty Dozen," while the authors of the Supplement to the ASA's 2016 statement identified more than 20 (Greenland et al. 2016). Many of these misconceptions can be avoided by moving from NHST to estimation; however, as we have seen, this alone is not enough. The use of AnCred provides extra protection against inferential fallacies by reducing the opportunities for them to arise in the first place.

4.1. Misinterpretation of p-Values and CIs

The most notorious inferential error in NHST is the interpretation of a p-value as the probability that the null hypothesis is true. This and related fallacies have been promulgated even by statistical texts (for a review, see Hubbard 2016a). It is widely held that estimation and CIs are less prone to such misinterpretations. However, they are certainly not immune, with standard CIs often being interpreted as the range within which the true effect size lies with specified probability. This interpretation is valid only within a Bayesian framework, under

which CIs become credible intervals with uninformative priors. By explicitly using the Bayesian framework, the use of AnCred obviates this interpretative issue.

4.2. Misappropriation of CIs in Dichotomization

Despite their role in estimation, CIs are commonly hijacked for use in NHST, in which the focus is simply on whether the CI includes values corresponding to no effect. This in turn leads to the notorious practice of taking the resulting dichotomy of statistical significance/nonsignificance as equivalent to the truth or falsity of the substantive hypothesis. In contrast to this inferentially wasteful and misleading practice, AnCred makes full use of the content of the CI, generating SLs and ALs that allow researchers to make more nuanced assessments of the credibility of new findings in the context of existing insight.

4.3. Misuse of Power in the Interpretation of Study Findings

Researchers commonly believe that the bigger the sample, the more compelling the resulting inference. While intuitively appealing, the limitations of this maxim are less widely known. This may explain why the concept of statistical power, which is directly related to sample size, is the source of some of the most recalcitrant inferential fallacies. Power is frequently used retrospectively to interpret study findings, despite being a probabilistic concept that lacks meaning once the results are known (Goodman and Berlin 1994). Researchers hoping to find a genuine effect from a small study but confronted with a statistically nonsignificant result frequently ascribe it to "inadequate" power. Conversely, a nonsignificant outcome from a study with high retrospective power is often seen as strong evidence against the existence of the effect.

Such practices have long been criticized without obvious effect (e.g., Hoenig and Heisey 2001). This suggests that even if retrospective power is an invalid concept, there is a demand for ways of assessing the weight of evidence once the findings are known. AnCred meets this demand via the concept of credibility. Findings based on small samples typically have wider CPIs than those from large samples, making claims of statistical significance/nonsignificance more vulnerable to challenge, as intuition would suggest. Furthermore, this approach to assessing evidential weight is free from the issues that undermine retrospective power. However, the expressions (A1)–(A4) also show that sample size alone is not enough to determine the CPI in any specific case: both the location of the point estimate of the CI and the relationship to its width also play a role. As such, the use of AnCred also offers protection against the counterintuitive phenomenon known as the power fallacy: the belief that findings from small studies are necessarily less compelling than large studies (Wagenmakers et al. 2015). To see this, consider the following real-life example.

Example: Adverse effects of reboxetine. In their systematic review of adverse effects of this anti-depressant, Eyding et al. (2010) include two double-blinded, randomized controlled trials with 510 and 210 patients, respectively. The former gave an OR for adverse effects relative to placebo of 1.73, with 95% CI of (1.07, 2.79); the smaller study gave a larger point estimate of 4.30, but was also substantially less precise, with a 95% CI of (1.85, 9.99). Being based on fewer than half the number of patients, intuition suggests the smaller study is necessarily less compelling. AnCred shows such reasoning to be a case of the power fallacy. The application of (A2) shows that the smaller study has a SL of 1.8, compared with 2.4 for the larger study. Thus, surprisingly, the smaller study actually provides *more* compelling evidence of adverse effects, in the sense of putting more constraint on skeptics seeking to dismiss the evidence as not credible. In essence, this is because AnCred makes use of *both* the uncertainty surrounding the effect size (i.e., the width of the CI) and the margin by which it excludes no effect (i.e., the location of the bounds of the CI relative to no effect) in making its assessment. AnCred has here highlighted the unreliability of focusing solely on the width of the CI in assessing the evidential value of a finding.

4.4. Misuse of Significance as a Marker of "Discordant" Results

The replication crisis which motivated the ASA statement has rekindled long-standing concerns over the use of p-values as guides to the probability of successful replication. For example, if an initial study obtained the true effect size along with $p = 0.05$, there is only a 50% probability of getting so impressive a p-value even from a perfect replication (Goodman 1992). For the "intuitive" replication probability of 95% to hold, the initial study must achieve $p \leq 3 \times 10^{-4}$.

The use of CIs does less violence to intuition: under ideal conditions, the probability that the mean of a replication falls within the 95% CI of the initial study is 83% (Cumming, Williams, and Fidler 2004). Nevertheless, CIs remain vulnerable to serious misconceptions concerning replications. Of these, the most prevalent is that studies whose CIs encompass no effect are necessarily discordant with those which exclude it (and vice versa).

This fallacy is commonly the result of the hijacking of the CI to serve as a proxy for a p-value. The following case-study shows how AnCred addresses this fallacy.

Example: The effect of statins on glioma risk. Widely prescribed for controlling cholesterol, statins appear to have a protective effect against brain cancers (see, e.g., Gaist et al. 2014). To investigate this link, Seliger et al. (2016) conducted a large case-control comparison involving 2469 cases of glioma and a 10-fold matched control group. The point estimate for cancer incidence among those receiving statins was an OR of 0.75, consistent with the protective effect found in earlier studies. Yet despite this, the 95% CI of (0.48, 1.17) encompassed no effect, leading the authors to conclude their findings were discordant with these previous studies.

As Greenland noted in a rejoinder (2017) "Such statistical misinterpretations have been decried for at least four decades," adding that the authors' error was clearly unintentional, but nevertheless "undermined their considerable and respectable effort." The use of AnCred would have averted this unfortunate exchange, and also provided more insight into the research question at issue. Instead of interpreting the findings simply

via the dichotomy of whether or not the CI includes no effect, AnCred subjects them to challenge using the CI's entire inferential content. As the substantive hypothesis under test is for a *reduction* of risk, the appropriate CPI is (AL, 1), where AL is the advocacy limit calculated from (A4). This leads to the conclusion that Seliger et al.'s interpretation of their nonsignificant finding can be successfully challenged if there exists prior evidence of a protective effect in the range (0.14, 1.0). The authors themselves cite two prior studies with 95% CIs of (0.52, 1.00) and (0.59, 0.98), respectively, both within the required range. Thus, far from being discordant, Seliger et al.'s finding adds to the growing evidence *in favor* of the protective effect of statins.

4.5. The Inability of p-Values to Incorporate Plausibility

Lacking any direct means of taking account of plausibility, the NHST paradigm is notoriously capable of lending support to patently spurious claims (e.g., Hines 1998; Austin et al. 2006; Bennett, Miller, and Wolford 2009). While often blamed on practices such as data-dredging, failure to take account of plausibility is more pernicious as it can undermine entire areas of research (e.g., Bracken 2009; Ioannidis 2013).

By setting new findings in context via Bayes' theorem, AnCred captures in quantitative terms the well-known maxim that "extraordinary claims require extraordinary evidence" (Sagan 1980). This is especially valuable when assessing startling claims whose statistical significance conceals an underlying lack of evidential support.

Example: The benefits of prompt administration of clot-busters. In 1992, a UK-based study suggested that death rates from heart attacks could be reduced by 50% if the so-called clot-buster drug anistreplase was given at the scene rather than after arrival in hospital (GREAT Group 1992). So large a benefit was surprising: prior experience suggested the reduction in mortality risk was likely closer to 15–20% (Pocock and Spiegelhalter 1992). However, the fact that the findings—summarized by an OR of 0.47 with a 95% CI of (0.23, 0.97)—were also statistically significant ($p = 0.04$) aroused much interest.

The unreliability of inferences based solely on p-values without reference to plausibility is made clear by the application of AnCred. This leads, via (A2), to a CPI of (0.1, 1.0), implying the finding is only credible if there exists prior evidence that prompt administration leads to *at least* a 90% reduction in mortality. This has no basis in reality. AnCred has thus cut through the statistical significance to focus on the startling lack of evidential weight behind the headline figure, due to both the width of the CI and how close it comes to encompassing no effect. However, the study's failure to provide credible evidence for an effect does not imply it is worthless; as with the nonsignificant finding in the glioma/statin study, it still contains some evidential weight. As Pocock and Spiegelhalter showed, this can be extracted using conventional Bayesian analysis, which pointed to a much more modest effect size similar to that found in a subsequent meta-analysis (Morrison et al. 2000). In this example, AnCred serves to alert researchers to the weakness of the original evidence, the need to avoid using statistical significance as a proxy for the truth of the substantive hypothesis, and the value in carrying out a full Bayesian analysis.

4.6. Significance as an Unsafe Basis for Scientific Conclusions

Goodman (2008) argues that the overarching misconception about NHST is that statistical significance forms a suitable basis for scientific conclusions or policy. This has led to the bizarre situation of each new study often being seen as in some sense superseding all that went before. While flattering for researchers and journal editors, this is a parody of the notion of scientific insight flowing from the accumulation of reliable evidence. It also threatens trust in the scientific process among policy-makers and the public alike (Devlin 2017). Nevertheless, the ubiquity of NHST continues to feed "insights" into the academic and public arena based solely on statistical significance/nonsignificance. In a final illustration of the use of AnCred, I examine a recent and controversial example of the fallacy of using nonsignificance as a proxy for the falsity of the substantive hypothesis, in breach of the well-known maxim "Absence of evidence is not evidence of absence" (Altman and Bland 2004).

Example: The effectiveness of stents in angina patients. Introduced in the 1970s and implanted in over 500,000 such patients each year, stents have long been regarded as plausibly and observably beneficial to patients with stable angina. This view has recently been challenged by the outcome of ORBITA, the first blinded randomized controlled trial of the effectiveness of stents (Al-Lamee et al. 2018). The primary endpoint was increased exercise time, which was 28.4 sec for those receiving a stent, compared to only 11.8 sec for those in the control group; the summary statistic was the difference of +16.6 sec, with a 95% CI of (−8.9, 42.0) sec. Despite the point estimate being consistent with benefit, the nonsignificance of the finding ($p = 0.2$) led the researchers to conclude that stents do not increase exercise time, with positive results from previous studies possibly being a placebo effect. An accompanying commentary (Brown and Redberg 2017) went even further, described the implications as "profound and far-reaching," with the findings showing "unequivocally" the absence of benefit compared to other therapies, and requiring "all cardiology guidelines to be revised."

Many cardiologists expressed surprise at the findings. Some questioned the wisdom of overturning decades of practice on the basis of a study of 200 patients; others raised technical issues. Remarkably few commentators noted, however, that the point estimate of benefit of +16.6 sec is clearly consistent with genuine efficacy, albeit of questionable value given the risks involved in stent operations. In addition, the finding was not set in context, despite its implausibility in light of prior experience and the relatively broad CI, both of which suggest the nonsignificance is potentially misleading.

By exploiting the full inferential content of the summary statistic, AnCred circumvents the "null effect" fallacy encouraged by nonsignificant results. It also allows the findings to be put in quantitative context. Applying (A3) to the summary CI leads to an AL of +115 sec. So high a level shows that the

ORBITA evidence, while valuable, is too weak to undermine the credibility of claims that stents provide clinically meaningful benefit.

Thus, contrary to the impression created by the standard misinterpretation of nonsignificance, ORBITA provides no compelling justification for revising cardiology guidelines. Rather, it points to a small positive level of benefit. This is of course hinted at by the point value of the 95% CI, but the NHST paradigm led both the researchers and commentators to focus solely on the nonsignificance of the finding—that is the location of just one of the CI bounds—and then misinterpret this as evidence of no benefit. In contrast, by making full use of the inferential content of the CI, AnCred allows the result to be in the context of existing insight. In the process, it counters the misguided practice of regarding a single, modest but valuable contribution to knowledge as a dichotomously definitive "fact."

5. Conclusion

In his commentary to the ASA statement, Goodman (2016a) states:

> "*What follows this statement is as or more important than the statement itself…We need to formulate a vision of what success looks like, and how we will get there. If not, we can start drafting the language of the 2116 ASA statement tomorrow.*"

If the statement is to succeed where previous attempts at reform have failed, the reasons for those past failures must be addressed. The "publish or perish" research milieu encourages the use of inappropriate inferential practices (see, e.g., Smaldino and McElreath 2016) as "poor methods get results" (Horton 2015). Other barriers include a reluctance to move beyond basic statistical training (Schwab et al. 2011) and incredulity that so prevalent a technique as NHST can be seriously flawed (Hubbard 2016b).

This article has examined the extent to which the use of AnCred alongside current practices can overcome these barriers. Using real-life case-studies, I have highlighted the following practical benefits offered by AnCred:

- It provides researchers with a principled means for going beyond familiar NHST-based methods, extracting more insight from them while requiring only modest changes to statistical training.
- The Bayesian framework underpinning AnCred shifts the focus of inference to the substantive hypothesis rather than the null, and leads to straightforward interpretations of findings based on familiar summary statistics.
- This in turn provides enhanced protection against common but recalcitrant inferential fallacies resulting from using NHST in isolation.
- Under AnCred, simplistic assessment of evidential weight based on width of CI or sample size is replaced by a more nuanced quantitative assessment based on SLs and ALs.
- By "unpacking" the weight of evidence in CIs in this way, AnCred identifies situations where statistically significant evidence for a nonzero effect requires an unjustifiable level of support from prior evidence to achieve credibility.
- AnCred reduces the incentive for practices such as data dredging and p-hacking by extracting valuable insight from nonsignificant findings, often dismissed as "null" outcomes.
- This in turn reduces the risk of such findings going unreported, a notorious cause of publication bias, waste of research resources, and scientific misconduct.

At the same time, AnCred addresses broader concerns about the misuse of NHST by the research community:

- By setting new findings in the broader context of existing insight, AnCred combats the common but false perception of statistical significance/nonsignificance as synonymous with the "truth" or otherwise of a substantive hypothesis.
- This in turn encourages the view that, regardless of their statistical significance/nonsignificance, findings from well-conducted studies are all contributions to the accumulation of insight, rather than definitive, standalone conclusions that supersede all that came before.
- The use of AnCred ameliorates the so-called replication crisis by extending the assessment of findings beyond simplistic dichotomization as statistically significant/nonsignificant, while bringing more clarity to the concept of "discordant" studies.
- The concepts of SL and AL eliminate the need to use such flawed notions as "borderline" significance and post-hoc power to explain disappointingly "negative" study outcomes.
- The combination of evidential weight and prior evidence used in AnCred subjects weak and implausible claims to the maxim "extraordinary claims require extraordinary evidence."

The above features constitute a checklist for comparing AnCred with other proposals for moving toward a post $p < 0.05$ era (e.g., Goodman 2016b; Benjamin et al. 2018; Colquhoun 2018). Like any such proposal, however, AnCred is based on principles and assumptions that can be challenged. Notably, it models definitions of fair-minded skepticism and advocacy using normal distributions, and uses 95% CIs to produce assessments of credibility at the 95% level. These choices are pragmatic, in that they lead to computational simplicity and retain a familiar standard whose ubiquity suggests widespread acceptability. They are not mandatory, however; as noted in the Appendix, the principles behind AnCred can be extended to other distributions and credibility levels.

Perhaps the most obvious potential criticism of AnCred is that the concept of credibility simply replaces statistical significance as a means of dichotomizing findings. However, it should be stressed that dichotomization per se has never been the problem with NHST; it is the actions that flow from it. The categorization of a finding as statistically significant depends solely on the finding itself, takes no account of context or plausibility and is based on a universal (if widely misinterpreted) cut-off. As such, statistical significance will sustain a sharp and incontrovertible dichotomy—a feature which has encouraged the belief that it is equivalent to the dichotomy of the reality or otherwise of an effect (Fisher 1929). In contrast, categorization of a finding as credible is both conditional *and* context dependent. It depends on the size of effect found, its uncertainty, and its relationship to existing knowledge—which is itself open

to discussion in terms of selection and aggregation. As such, it offers no basis for a sharp and incontrovertible dichotomization based solely on the finding itself. Instead, the credibility of any given finding is open to debate and challenge by the research community. Credibility thus reflects the realities of the scientific enterprise, which approaches truth asymptotically, not peremptorily.

It may be argued that AnCred does not go far enough, in that it stops short of subjecting every finding to a full Bayesian analysis. The merits of having a standardized conservative framework for assessing new findings should not be under-estimated, however—not least because it gives continuity with conventional practice, whereby researchers perform an initial assessment of their findings. In any case, AnCred does not preclude more sophisticated analysis: indeed, it encourages it via its use of CIs, which are far more informative than p-values. There is no shortage of guidance on how to perform such analysis—for those motivated to do so. It is hoped that AnCred will provide such motivation by showing how even relatively elementary techniques beyond the NHST paradigm bring major benefits such as extracting more inferential content from standard summary statistics, and reduced risk of misinterpretation of the implications.

AnCred is not merely "quick and dirty Bayes," however. Its use of the inversion of the central theorem addresses the long-standing debate over setting appropriate priors that bedevils even the most sophisticated Bayesian analyses. The inversion makes the prior needed to achieve credibility immediately clear and thus available to all for debate. By moving beyond the use of uninformative priors or priors based on expert elicitation, AnCred thus makes more of the inferential power of Bayesian methods, while increasing their transparency.

This article has examined the application of AnCred to basic inferential challenges, but there is clearly scope for further development, for example, involving different study designs and non-normal distributions. Using AnCred to re-examine implausible, "negative" and "discordant" claims based on NHST in the existing research literature may also prove illuminating.

AnCred is not a panacea for the inferential ills of the research community. It may, however, 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?

Appendix

What follows is a summary of the theory underpinning AnCred and its principal results; for full derivations, the reader is referred to Matthews (2018).

A1: Motivating Principles

AnCred seeks to provide researchers with a simple, intuitive and versatile means of assessing the credibility of effect sizes found by standard study designs. To achieve these goals, AnCred takes as input the familiar summary statistic of CIs, and extracts from them inferential content that "adds value" to the standard NHST-based concepts of statistical significance and nonsignificance. This is done through an inversion of Bayes' theorem, allowing the *deduction* of the prior needed for a given finding to constitute credible evidence for the existence of an effect, in the form of a posterior distribution that excludes no effect.

The required inversion is performed with the likelihood and the prior modelled as normal distributions, widely used to address standard inferential issues concerning means, proportions and ratios. They are also conjugate distributions, which in the case of AnCred leads to closed analytic results requiring no specialized computation. Mathematically, the inversion requires specification of the location and scale parameters of the prior distribution, and in AnCred this is achieved by applying the *Principle of Fair-Minded Challenge* (PFMC) to the two situations of statistically significant and nonsignificant findings; I now consider each in turn.

A2: The Case of Statistically Significant Findings

Here the PFMC implies challenging claims to have found a statistically significant nonzero effect on the basis of fair-minded *skepticism*. This is modelled by a prior distribution centered on no effect, but whose tails reflect the fair-minded concession that there is a finite probability that a nonzero effect exists. The resulting CPI is shown in Figure A1. For findings in the form of a statistically significant difference in a mean or proportion expressed as a CI of (L, U), the CPI has lower and upper bounds $(-\text{SL}, +\text{SL})$ where SL is the *skepticism limit* given by

$$\text{SL} = \frac{(U-L)^2}{4\sqrt{UL}}. \tag{A1}$$

A finding provides credible evidence of a nonzero effect if prior evidence exists for minimum effect sizes lying *outside* the skeptical CPI. For example, if the prior evidence for a positive effect is summarized (via, say, a meta-analysis) by a 95% CI of (L_o, U_o), then the finding is credible at the same level if $L_o > \text{SL}$. As (A1) shows, the width of the CPI is dictated by the evidential weight of the data, as indicated by the width and location of the CI summarizing the data. The stronger the evidence of a genuine effect, the tighter the CPI, putting skeptics of an effect under greater constraint. In colloquial terms, they must invoke more extreme levels of skepticism about effect sizes to "drag" the posterior distribution back to the no effect line.

The corresponding expression for statistically significant ratios can be found by applying the transformation $U \to \ln(U)$ etc., resulting in a CPI with bounds $(1/\text{SL}, \text{SL})$ where if U and L are expressed as ORs,

$$\text{SL} = \exp\left[\frac{\ln^2(U/L)}{4\sqrt{\ln(U)\ln(L)}}\right]. \tag{A2}$$

Note that the level of credibility reflected in the SL is set by the summary statistic: for example, if SL is calculated from a 90% CI, then (A2) is a 90% credibility level, etc.

As the CPI for skepticism is centered on no effect, it represents a conservative standard for assessing the credibility of a statistically significant result, reflecting the view that the most likely effect size is zero. As such, a finding may still be credible if only a proportion of the prior evidence lies inside the CPI; those wishing to demonstrate this can make their case via a conventional Bayesian analysis specific to the findings and prior evidence.

A3: The Case of Statistically Nonsignificant Findings

Here, the PFMC involves challenging a nonsignificant finding on the basis of fair-minded *advocacy* of the original substantive hypothesis. This is modelled by a prior distribution reflecting belief that despite the nonsignificance, a nonzero effect size still exists, albeit confined

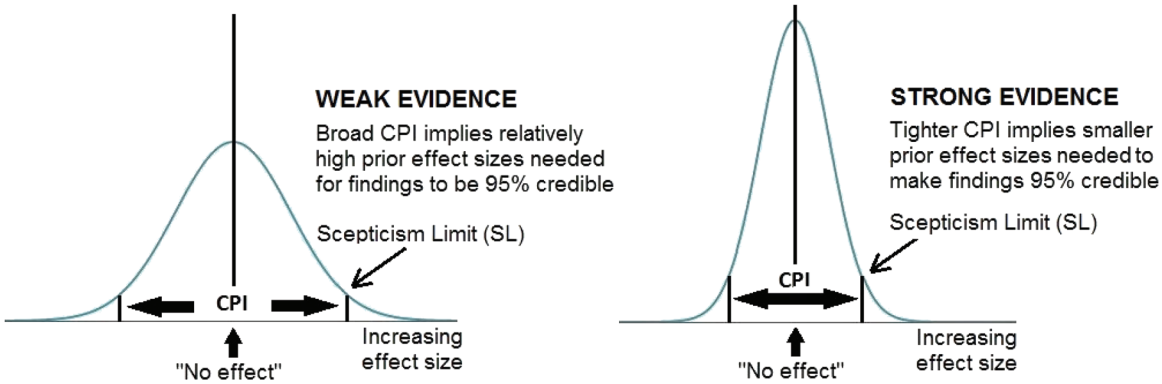


Figure A1. Under AnCred, statistically significant claims summarized by 95% CIs are subjected to fair-minded skeptical challenge represented by a prior distribution centered on no effect, whose 2.5% and 97.5% quantiles represent the critical prior interval. Weak evidence for the existence of an effect leads to relatively high skepticism limits (SLs), giving plenty of scope for skeptical challenge using prior knowledge. Strong evidence, in contrast, leads to tight CPIs, making skeptical challenge more demanding.

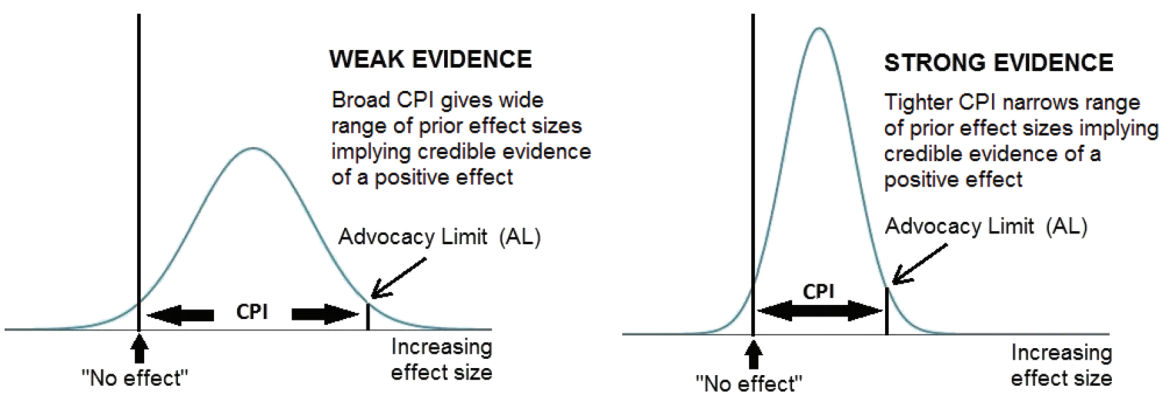


Figure A2. The advocacy prior distribution for assessing claims of statistical nonsignificance summarized by 95% CIs. For substantive hypotheses of a positive effect, the lower bound of the advocacy CPI is fixed at no effect, while the upper bound—the advocacy limit (AL)—is such that the CPI encompasses 95% of the prior distribution capable of making the finding credible evidence of a positive effect. Weak evidence against such an effect—that is a broad, nonsignificant 95% CI with a lower bound pointing to a small negative effect—leads to a relatively broad CPI, as such findings put little constraint on the prior evidence needed for credibility. Stronger evidence against the substantive hypothesis leads to a tighter CPI, thus constraining advocates to a narrower range of prior support.

to an interval extending from no effect to some finite bound. This reflects fair-mindedness in the sense that, despite the advocates' belief in *some* level of effect, they concede that its magnitude must have a finite limit. The resulting CPI is then the range of prior effect sizes capable of turning the nonsignificant finding into credible evidence for a nonzero effect. As such advocacy can be of effects either greater or less than the null, the CPI must reflect the sign as well as magnitude of the substantive hypothesis. For conciseness, I focus on the case where the substantive hypothesis implies differences in means exceeding zero and ratios exceeding unity (e.g., better test scores; higher survival rates); the results where these inequalities are reversed follow by symmetry. The CPI for nonsignificant differences in means and proportions is then (0, AL) where AL is the *advocacy limit* given by

$$\mathrm{AL} = \frac{-(U+L)}{2UL}(U-L)^2. \tag{A3}$$

The CPI for substantive hypotheses of ORs > 1 is (1, AL) where AL is given by

$$\mathrm{AL} = \exp\left[\frac{\ln(UL)\ln^2(U/L)}{2\ln(U)\ln(L)}\right] \tag{A4}$$

the level of credibility again being set by the summary statistic, with 95% CIs leading to ALs for 95% credibility etc. The width of the advocacy CPI again depends on the evidential weight of the data, such that the stronger the evidence *against* a positive effect, the narrower the CPI; see Figure A2.

If prior evidence points to effect sizes lying wholly within the advocacy CPI, the finding can be deemed credible evidence for a nonzero effect, despite its nonsignificance. For example, if the prior evidence for a positive effect is summarized by an OR 95% CI of (L_o, U_o) and both L_o and U_o lie within the relevant advocacy CPI—that is, $L_o \geq 1, U_o \leq \mathrm{AL}$—then the finding is credible evidence for a positive effect, despite its nonsignificance. On the other hand, the above expressions for AL show that a finding cannot provide credible evidence for the hypothesized effect under AnCred if the point value M lies on the other side of the null line from the hypothesized effect. Instead, it provides evidence for the opposite effect (e.g., decreased rather than increased test-scores), and this is credible if it meets the criteria set by AnCred, this time with the CPI transformed from (1, AL) $\rightarrow$ (AL, 1).

As with the skepticism CPI, AnCred imposes a conservative bound on the range of prior evidence needed to provide credible evidence of a nonzero effect, in this case because the Advocacy CPI has one bound fixed at no effect. If prior evidence exists for effect sizes substantially exceeding no effect, this allows the finding to retain credibility for values beyond AL. Again, those seeking to argue for this must make their case via a Bayesian analysis specific to the findings and prior evidence.

Figure A3 summarizes the use of AnCred for both statistically significant and nonsignificant findings, and shows the impact of evidence from earlier studies on findings with different weight of evidence.

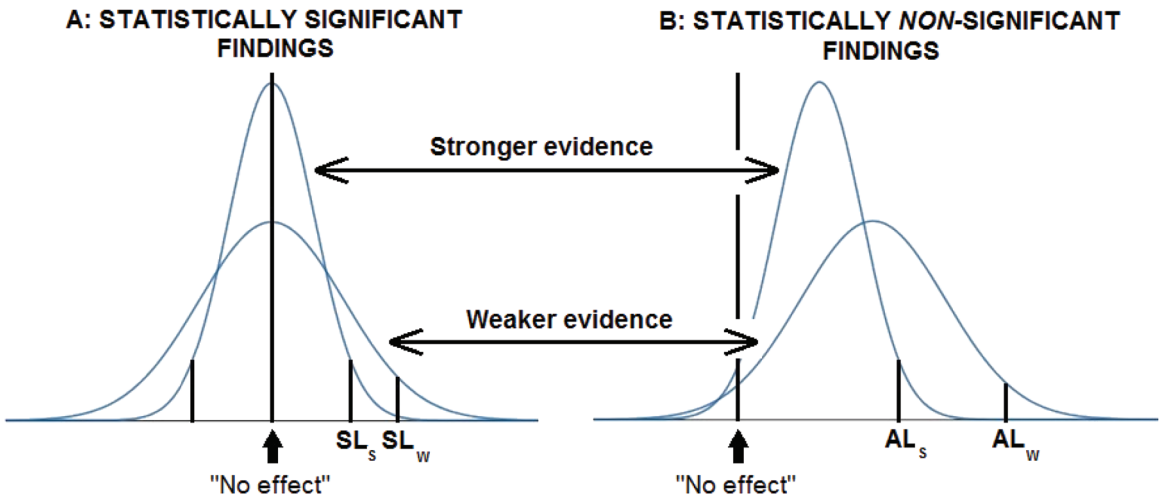


Figure A3. The effect of prior evidence summarized by the interval (L_o, U_o) on the credibility of findings from different studies. For statistically significant findings (A), if $L_o > \mathrm{SL}_w$ then the finding is credible evidence of a positive effect whether it came from the stronger or weaker study. If $\mathrm{SL}_s < L_o < \mathrm{SL}_w$, only the finding from the stronger study achieves credibility under AnCred. If $L_o < \mathrm{SL}_s$ then demonstration of the credibility of the finding requires a Bayesian analysis specific to the study and prior evidence. For statistically nonsignificant results (B), findings still provide credible evidence of a nonzero effect if both L_o and U_o lie between no effect and AL_s in the case of the stronger study, and AL_w for the weaker study. As before, the impact on credibility of prior evidence lying outside these ranges can be assessed using a specific Bayesian analysis.

A4: The Case of No Relevant Prior Evidence

In the case of entirely novel findings, there may be no obviously relevant prior evidence with which the CPI can be compared to establish credibility. In such cases, AnCred provides a framework leading to the assessment of *intrinsic credibility* (Matthews 2018, sect. 6). Here, the CPI calculated from the 95% CI is compared to the point value of the effect size, given by the point value of the CI. The application of AnCred then proceeds as usual, with the location of this point value relative to the CPI determining intrinsic credibility. An example is given in Example 2 of Section 3.

As shown in the above reference, the assessment can be performed by a simple rule of thumb: entirely novel findings possess intrinsic credibility if they have p-values less than around 0.01. It should be stressed, however, that this rule of thumb has no interpretation under NHST. Lack of intrinsic credibility means simply that the novel finding in question currently lacks the evidential weight to make its case without external support. Such support may emerge following publication of the finding, thus allowing a reassessment of the finding using AnCred.

Very recently, Held (2018) has developed an alternative approach to intrinsic credibility which casts intriguing new light on proposals to lower the threshold for significance of novel findings to $p \leq 0.005$ (Benjamin et al. 2018).

A5: Extension of AnCred to Other Inferential Challenges

This article has focused on the application of AnCred in the most widely used situation encountered in research: the assessment of evidence for nonzero effects expressed as differences in means, proportions, or as ORs. The basic principle of inverting Bayes' theorem and deducing the prior needed for posterior credibility at the 95% level can be applied to other inferential questions, such as the credibility of claims of worthwhile benefit. However, even with conjugate distributions the outcome typically lacks the analytic elegance made possible by the normal distribution, and the use of software packages is likely to be required for general acceptance by the research community. It is hoped that such developments will be encouraged by the potential benefits of using AnCred explored in this article.

Acknowledgments

I am indebted to Stephen Senn for prompting me to develop the ideas discussed here, to Ronald Wasserstein and David Spiegelhalter for their interest and encouragement, and to Denise Best, Doug Altman, Leo Held, Sander Greenland, Tom Louis, and Paul Parsons for valuable discussions. The final version benefited greatly from the comments of two anonymous referees and the editor and associate editor of *TAS*.

I dedicate this article to the memory of my friend and fellow member of the Summertown Soup Club, Doug Altman (1948–2018).

References

Al-Lamee, R., Thompson, D., Dehbi, H. M., Sen, S., Tang, K., Davies, J., Keeble, T., Mielewczik, M., Kaprielian, R., Malik, I. S., and Nijjer, S. S. (2018), "Percutaneous Coronary Intervention in Stable Angina (ORBITA): A Double-Blind, Randomised Controlled Trial," *The Lancet*, 391, 31–40. [207]

Altman, D. G., and Bland, M. J. (2004), "Confidence Intervals Illuminate Absence of Evidence," *BMJ*, 328, 1016–1017. [207]

Austin, P. C., Mamdani, M. M., Juurlink, D. N., and Hux, J. E. (2006), "Testing Multiple Statistical Hypotheses Resulted in Spurious Associations: A Study of Astrological Signs and Health," *Journal of Clinical Epidemiology*, 59, 964–969. [207]

Baker, M. (2016), "1,500 Scientists Lift the Lid on Reproducibility," *Nature*, 533, 452–454. [202]

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, 6–10. [208,211]

Bennett, C. M., Miller, M. B., and Wolford, G. L. (2009), "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument for Multiple Comparisons Correction," *Neuroimage*, 47, S125. [207]

Bracken, M. B. (2009), "Why Are So Many Epidemiology Associations Inflated or Wrong? Does Poorly Conducted Animal Research Suggest Implausible Hypotheses?," *Annals of Epidemiology*, 19, 220–224. [207]

Brown, D. L., and Redberg, R. F. (2017), "Last Nail in the Coffin for PCI in Stable Angina?," *The Lancet*, 391, 3–4. [207]

Carlin, B. P., and Louis, T. A. (1996), "Identifying Prior Distributions That Produce Specific Decisions, With Application to Monitoring Clinical Trials," in *Bayesian Analysis in Statistics and Econometrics: Essays in*

Honor of Arnold Zellner, eds. D. Berry, K. Chaloner, and J. Geweke, New York: Wiley, pp. 493–503. [203]

Colquhoun, D. (2018), "The False Positive Risk: A Proposal Concerning What to Do About *p*-Values," arXiv no. 1802.04888. [208]

Cumming, G. (2008), "Replication and *p* Intervals: *p* Values Predict the Future Only Vaguely, But Confidence Intervals Do Much Better," *Perspectives on Psychological Science*, 3, 286–300. [203]

Cumming, G., Williams, J., and Fidler, F. (2004), "Replication and Researchers' Understanding of Confidence Intervals and Standard Error Bars," *Understanding Statistics*, 3, 299–311. [206]

Devlin, H. (2017), "Exaggerations Threaten Public Trust in Science, Says Leading Statistician," *Guardian.com*, available at *www.tinyurl.com/Spiegelhalter-interview*. [207]

Eyding, D., Lelgemann, M., Grouven, U., Härter, M., Kromp, M., Kaiser, T., Kerekes, M.F., Gerken, M. and Wieseler, B. (2010), "Reboxetine for Acute Treatment of Major Depression: Systematic Review and Meta-Analysis of Published and Unpublished Placebo and Selective Serotonin Ruptake Inhibitor Controlled Trials." *BMJ*, 341, c4737. [206]

Fisher, R. A. (1929) "The Statistical Method in Psychical Research," *Proceedings of the Society for Psychical Research*, 39, 189–192. [208]

Gaist, D., Hallas, J., Friis, S., Hansen, S., and Sørensen, H. T. (2014), "Statin Use and Survival Following Glioblastoma Multiforme," *Cancer Epidemiology*, 38, 722–727. [206]

Gardner, M. J., and Altman, D. G. (1986), "Confidence Intervals Rather Than *P* Values: Estimation Rather Than Hypothesis Testing," *British Medical Journal*, 292, 746–750. [202]

Good, I. J. (1950), *Probability and the Weighing of Evidence*, London, UK: Griffin, pp. 35–36. [203]

Goodman, S. N. (1992), "A Comment on Replication, *P*-Values and Evidence," *Statistics in Medicine*, 11, 875–879. [206]

——— (2008), "A Dirty Dozen: Twelve *P*-Value Misconceptions," *Seminars in Hematology*, 45, 135–140. [205,207]

——— (2016a), "The Next Questions: Who, What, When, Where, and Why?," *Online Commentary to Wasserstein and Lazar (2016)*. [202,208]

——— (2016b), "Aligning Statistical and Scientific Reasoning," *Science*, 352, 1180–1181. [208]

Goodman, S. N., and Berlin, J. A. (1994), "The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results," *Annals of Internal Medicine*, 121, 200–206. [206]

GREAT Group (1992), "Feasibility, Safety, and Efficacy of Domiciliary Thrombolysis by General Practitioners: Grampian Region Early Anistreplase Trial," *British Medical Journal*, 305, 548–553. [207]

Greenland, S. (2011), "Null Misinterpretation in Statistical Testing and Its Impact on Health Risk Assessment," *Preventative Medicine*, 53, 225–228. [202]

——— (2017), "A Serious Misinterpretation of a Consistent Inverse Association of Statin Use With Glioma Across 3 Case-Control Studies," *European Journal of Epidemiology*, 32, 87–88. [206]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [203,205]

Held, L. (2013), "Reverse-Bayes Analysis of Two Common Misinterpretations of Significance Tests," *Clinical Trials*, 10, 236–242. [202]

——— (2018), "A New Argument for $p < 0.005$," arXiv no. 1803.10052. [211]

Hines, T. M. (1998), "Comprehensive Review of Biorhythm Theory," *Psychological Reports*, 83, 19–64. [207]

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E. J. (2014), "Robust Misinterpretation of Confidence Intervals," *Psychonomic Bulletin & Review*, 21, 1157–1164. [203]

Hoenig, J. M., and Heisey, D. M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55, 19–24. [206]

Horton, R. (2015), "Offline: What Is Medicine's 5 Sigma?," *Lancet*, 385, 1380. [208]

Hubbard, R. (2016a), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. Thousand Oaks, CA: Sage, pp. 209–213. [205]

——— (2016b), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. Thousand Oaks, CA: Sage, pp. 232–234. [208]

Ioannidis, J. P. A. (2013), "Implausible Results in Human Nutrition Research," *BMJ*, 347, I6698 [207]

Matthews, R. A. J. (2001a), "Why Should Clinicians Care about Bayesian Methods?," *Journal of Statistical Inference and Planning*, 94, 43–58. [202]

——— (2001b), "Methods for Assessing the Credibility of Clinical Trial Outcomes," *Drug Information Journal*, 35, 1469–1478. [202]

——— (2018) "Beyond 'Significance': Principles and Practice of the Analysis of Credibility," *Royal Society Open Science*, 5, 171047. [202,209,211]

Matthews, R. A. J., Wasserstein, R., and Spiegelhalter, D. (2017), "The ASA's *P*-Value Statement, One Year On," *Significance*, 14, 38–41. [202]

Morrison, L. J., Verbeek, P. R., McDonald, A. C., Sawadsky, B. V., and Cook, D. J. (2000), "Mortality and Prehospital Thrombolysis for Acute Myocardial Infarction: A Meta-Analysis," *Journal of the American Medical Association*, 283, 2686–2692. [207]

Nuzzo, R. (2014), "Scientific Method: Statistical Errors," *Nature*, 506, 150–152. [202]

Pocock, S. J., and Spiegelhalter, D. J. (1992), "Domiciliary Thrombolysis by General Practitioners," *British Medical Journal*, 305, 1015. [207]

Rothman, K. J. (1978), "A Show of Confidence," *New England Journal of Medicine*, 299, 1362–1363. [202]

Rothman, K. J., Greenland, S., and Lash, T. L. (Eds.) (2008), *Modern Epidemiology*, Philadelphia: Lippincott Williams & Wilkins, chapter 18. [202]

Sagan, C. (1980), *Broca's Brain: Reflections on the Romance of Science*, New York: Random House, p. 73, chapter 5. [207]

Schwab, A., Abrahamson, E., Starbuck, W. H., and Fidler, F. (2011), "Perspective: Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests," *Organization Science*, 22, 1105–1120. [208]

Seliger, C., Meier, C. R., Becker, C., Jick, S. S., Bogdahn, U., Hau, P., and Leitzmann, M. F. (2016), "Statin Use and Risk of Glioma: Population-Based Case–Control Analysis," *European Journal of Epidemiology*, 31, 947–952. [206]

Smaldino, P. E., and McElreath, R. (2016), "The Natural Selection of Bad Science," *Royal Society Open Science*, 3, 160384. [208]

Spiegelhalter, D. J. (2004), "Incorporating Bayesian Ideas Into Health-Care Evaluation," *Statistical Science*, 19, 156–174. [202,203]

Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester: Wiley & Sons, chapter 3. [202]

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., and Morey, R. D. (2015), "A Power Fallacy," *Behavior Research Methods*, 47, 913–917. [206]

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. [202]

The American Statistician

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

Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels

Mark Andrew Gannon, Carlos Alberto de Bragança Pereira & Adriano Polpo

To cite this article: Mark Andrew Gannon, Carlos Alberto de Bragança Pereira & Adriano Polpo (2019) Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels, The American Statistician, 73:sup1, 213-222, DOI: 10.1080/00031305.2018.1518268

To link to this article: https://doi.org/10.1080/00031305.2018.1518268

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 5172

View related articles

View Crossmark data

Citing articles: 6 View citing articles

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, 213–222: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518268

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels

Mark Andrew Gannon[a], Carlos Alberto de Bragança Pereira[a,b], and Adriano Polpo[c]

[a]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil; [b]Instituto de Matemática Aplicada, Universidade Federal de Mato Grosso do Sul, Campo Grande, Brazil; [c]Department of Statistics, Federal University of São Carlos

ABSTRACT

This article argues that researchers do not need to completely abandon the p-value, the best-known significance index, but should instead stop using significance levels that do not depend on sample sizes. A testing procedure is developed using a mixture of frequentist and Bayesian tools, with a significance level that is a function of sample size, obtained from a generalized form of the Neyman–Pearson Lemma that minimizes a linear combination of α, the probability of rejecting a true null hypothesis, and β, the probability of failing to reject a false null, instead of fixing α and minimizing β. The resulting hypothesis tests do not violate the Likelihood Principle and do not require any constraints on the dimensionalities of the sample space and parameter space. The procedure includes an ordering of the entire sample space and uses predictive probability (density) functions, allowing for testing of both simple and compound hypotheses. Accessible examples are presented to highlight specific characteristics of the new tests.

ARTICLE HISTORY
Received March 2018
Revised August 2018

KEYWORDS
Hardy–Weinberg equilibrium; Neyman–Pearson lemma; Predictive distribution; Significance test

1. Introduction

It has become clear that the tests performed by comparing a p-value to 0.05 are not adequate for science in the 21st Century. Classical p-values have multiple problems that can make the results of hypothesis tests difficult to understand, multiple kinds of "p-hacking" can be used to try to get a p-value below the "magic" number of 0.05, and as many researchers have discovered, tests using standard p-values just aren't useful when the sample size gets large, because they end up rejecting any hypothesis. Jeffreys's tests using Bayes factors with fixed cutoffs also tend to have problems with large samples. However, hypothesis testing is a useful tool and is now a more-than-familiar way of thinking about how to do experiments and a very broadly used way of understanding and reporting experimental results. What is a researcher to do?

The present article introduces one solution for researchers who recognize that the hypothesis tests of the 20th Century are inadequate, but don't want to have to change the way they think about experiments and the ways they interpret and report experimental results. Section 2 delves into some of the problems that arise with the most widely used hypothesis tests and why those problems occur. Section 3 presents a new kind of hypothesis test that avoids the problems described, but doesn't "throw out the baby with the bath water," retaining the useful concept of statistical significance and the same operational procedures as currently used tests, whether frequentist (Neyman–Pearson p-value tests) or Bayesian (Jeffreys's Bayes-factor tests). Section 4 presents examples of the new tests being used, to highlight some of the advantages of the new tests and show researchers operational details of how the new tests can be used. Final considerations are given in Section 5. The article is written for researchers who are interested in a more modern tool for hypothesis testing, but who are not necessarily statisticians. It therefore does not go into deep theoretical detail, focusing instead on issues of more direct relevance to would-be users of the new tests. However, references are provided for those interested in the details of the theory behind the tests.

2. Context and Motivation: What's Wrong With the Tests People Have Been Using for so Many Years?

The subject of hypothesis testing and some of the problems that arise has been the subject of vigorous debate for several decades. Because frequentist tests using p-values are the most widely used, the use of p-values has been the subject of the most and harshest criticism. The journal *Basic and Applied Social Psychology* even went so far as to prohibit the use of p-values in articles—see Trafimow and Marks (2015). The controversy over the use of p-values has been so great that the American Statistical Association issued an official statement on p-values: Wasserstein and Lazar (2016).

It is worth saying that hypothesis tests based on p-values are not the only tools subject to valid criticism. Hypothesis testing in general has been criticized, for example, in Cohen (1994), Levine et al. (2008), and Tukey (1969), and some of the alternative methods of hypothesis testing, such as the

CONTACT Mark Andrew Gannon mark@ime.usp.br Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo – SP 05508-090, Brazil.

Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/TAS.

© 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.

Bayes-factor hypothesis tests created by Jeffreys, have been the subject of specific criticisms like those of Gelman and Rubin (1995) and Weakliem (1999). Further, hypothesis tests are not the only statistical tools that can be criticized. Every statistical tool relies on a model of some kind, and ultimately seeks to do something that can't be done in an exact and perfect way: make inferences based on limited quantities of data. Even when there are many, many data available, the sample is always finite, and so the inference *must* be imperfect. As a result, there are valid criticisms of every statistical tool. Tests based on p-values get special attention because of the fact that they are so widely used, and their misuse has contributed greatly to the "reproducibility crisis" in science and medicine.

The way p-values are used in statistical tests is based on the work of Fisher and of Neyman and Pearson in the 1930s. Fisher produced "significance tests" in which a "null hypothesis" about a parameter in a statistical distribution is tested without any consideration of what the alternative or alternatives are. Neyman and Pearson created "hypothesis tests," in which a null hypothesis is tested against a specific alternative or alternatives. In both cases, the null hypothesis is usually something like "there is no effect."[1] When the null is rejected, the inference is that there is some interesting effect to be reported. In both cases, the null is rejected based on the comparison of a p-value to a "significance level." A p-value has multiple possible definitions. It is not always *defined* as a probability (see, for example, Schervish (2012)), but its *calculation* is always a probability calculation. Given an experimental result x_0, a p-value is calculated as the probability, if the null hypothesis were true, of an observation x that supports the null hypothesis as much or less than the actual experimental result x_0. When the p-value is sufficiently small, the researcher decides that the actual observation x_0 is sufficiently unlikely under the null hypothesis that he or she can conclude that the null is false.

How small is "sufficiently small"? This is where the concept of "statistical significance" comes into the story. The p-value is generally compared to some small number that ends up being the probability of incorrectly rejecting a null hypothesis that is true. In his original work on significance testing, Fisher mentioned a one-in-twenty chance (5%, or 0.05) of rejecting a correct null as a convenient cutoff for declaring a statistically significant result, but did not intend for this number to be used universally. He states in his 1956 article "Statistical Methods and Scientific Inference" that the significance level should be set according to the circumstances. In the seminal work on statistical hypothesis testing, Neyman and Pearson (1933), an attempt is made to explicitly control errors in finding the best kind of test to choose between a null hypothesis and one or more specific alternative hypotheses. Two kinds of possible errors are considered: rejecting a true null hypothesis, called "errors of the first kind" (here called"errors of type I" or "type-I errors") and accepting a false null hypothesis, an "error of the second type" (here, "type-II error"). Neyman and Pearson write "The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator," where "the balance" refers to the balance between the probabilities of the two types of error. Ironically enough, the very mathematical approach adopted by Neyman and Pearson makes it close to impossible for a researcher to determine "just how the balance should be struck." Neyman and Pearson fix the probability of a type-I error, which is denoted here and in many statistics texts as α, and then prove that the test that minimizes the probability of a type-II error, here and elsewhere denoted as β,[2] is a test based on comparing the ratio of the likelihoods[3] of the competing hypotheses to some cutoff value, with the cutoff chosen so that the probability of incorrectly rejecting a correct null hypothesis is α. As shown in Pericchi and Pereira (2016), adopting this approach can lead to imbalances so great that the probabilities of the two types of error can vary by orders of magnitude.

In addition to the imbalance between the error probabilities of the two types described above, fixing α can lead to an even more serious problem. Because p-values tend to decrease with increasing sample size, the easiest and most common form of "p-hacking" is to keep taking data until the p-value falls below 0.05 or whatever other level of significance is chosen and fixed. No matter what level is chosen, a large-enough sample will almost inevitably lead to rejection of *any* null hypothesis. In the words of Berger and Delampady (1987), "In real life, the null hypothesis will always be rejected if enough data are taken because there will be inevitably uncontrolled sources of bias." According to Pericchi and Pereira (2016), more information paradoxically ends up being "a bad thing" for Neyman–Pearson hypothesis testing (and Fisherian signficance testing). Fixing a lower value for declaring a statistically significant result, say, 0.005 instead of 0.05, as suggested recently by a group of 72 researchers, many of them quite prominent, and some of them highly respected statisticians (Benjamin et al. (2018)), will only postpone the problem to larger sample sizes. This is folly even at the dawn of the era of "Big Data," and not a good solution in general. More data will still be "a bad thing" for tests with a cutoff that is lower, but still fixed, in direct conflict with the commonsense idea that a result based on a larger sample should be more believable.

A definition of the p-value for a null hypothesis $\mathbf{H}$ and observation x_0 commonly used in statistics textbooks is represented in Pereira and Wechsler (1993) as follows[4]:

Definition 1 (PW2.2). The p-value is the probability, under $\mathbf{H}$, of the event composed by all sample points that are at least as *extreme* as x_0 is.

[1] Fisher's tests do not *require* the null to be a "nil," but tests are usually performed that way, even in Fisher's own work.

[2] Some readers may be familiar with the *power* of a test, the probability of correctly rejecting a false null hypothesis. The power of a test is given by $1 - \beta$.

[3] A likelihood function corresponding to a given probability (density) function is given by the same expression, but taken as a function of the parameter θ for a fixed value of the observation x. In general, a probability (density) function for any given value of the parameter θ, summed (integrated) over the entire space of all possible observations, yields 1, while the corresponding likelihood for a given observation x may not even be summable (integrable) over the entire space of all possible parameter values, much less yield the specific value 1.

[4] The definition numbers 2.2 and 2.1 from Pereira and Wechsler (1993) are maintained, even though they are presented in the opposite order in this article.

"[S]ample points" here refers to possible experimental observations. Note that this definition does not in any way take into account the alternative hypothesis **A**. One problem with this definition is that in some cases, a p-value defined this way can't even be calculated. For example, when the range of possible observations ("sample space") consists of separate intervals, or when it is multidimensional, it can be unclear what "extreme" means. Throughout this article, "p-value," with a lower-case "p," is used to refer to a quantity calculated using Definition 1. A contrasting definition of a quantity analogous to the small-p p-value was presented in Pereira and Wechsler (1993) and is reproduced in the next section, along with an explanation of its advantages over traditional p-values.

Another problem with commonly used hypothesis tests is that the results can be hard to interpret, and their correct interpretation depends on the intent of the experimenter. An issue that has been the subject of vigorous debate among statisticians for decades is the Likelihood Principle (LP). While it still has the status of a principle that one either accepts or does not[5], and there are greatly respected statisticians who do *not* accept the LP, it is still true that LP-compliant tests have intuitive appeal and can be easier to interpret. For example, Cornfield (1966), Lindley and Phillips (1976), and Berger and Wolpert (1988) show that while frequentist tests that are not LP-compliant can give conflicting results *for the same data,* depending on what "stopping rule" is chosen for an experiment, the results of LP-compliant tests do not depend on the stopping rule and permit a single, unique inference based on the data without needing to know the intent of the experimenters. This brings the added advantage of allowing the collected data to be used while permitting researchers to stop an experiment early. Ethics can demand an early end to an experiment, for example, in the case of a medical study in which it becomes very clear that the patients receiving a new treatment are recovering, while those in the control group are not, or when patients receiving a new treatment are suffering serious side effects. Non-LP-compliant tests would require the researchers either to carry out the experiments to their pre-planned end, possibly in conflict with medical ethics, or to throw away all the collected data upon being forced by ethical concerns to stop the experiments earlier than planned.

There are Bayesian alternatives to frequentist hypothesis tests, the tests based on Bayes factors created by Jeffreys in the 1930s (see Jeffreys (1935, 1939)) and reviewed 60 years later in Kass and Raftery (1995). They are not as widely used as frequentist tests, but they have gained some acceptance in certain areas of research and are sometimes held up as a better alternative. As noted earlier, no tool is perfect, and Bayes-factor tests have also been criticized by both Bayesian and frequentist statisticians. For the purposes of this article, it is worth noting that Bayes factors, once calculated, are also compared to fixed cutoffs in Jeffreys's tests. The calculation of Bayes factors is described in the next section. For now, it is enough to know that the Bayes factor $\mathrm{BF}_{\mathbf{HA}}$ is a measure of the evidence favoring **H** over **A**. Jeffreys proposed an initial table of evidence grades against a null hypothesis **H**, with cutoffs at half-integer powers of 10, and Kass and Raftery updated the table by reducing the number of grades of evidence, noting that the Bayes factor measuring evidence *against* a null hypothesis **H** in favor of an alternative **A**, $\mathrm{BF}_{\mathbf{AH}}$, is $1/\mathrm{BF}_{\mathbf{HA}}$, and compiling a table of cutoff values of $\mathrm{BF}_{\mathbf{AH}}$. No justification is given for the cutoffs, other than Kass and Raftery stating "From our experience, these categories seem to furnish appropriate guidelines." Jeffreys's Bayes-factor tests do not take experimental error into account, and the cutoffs, whether those proposed by Jeffreys or those proposed by Kass and Raftery, do not take sample size into account. As with frequentist hypothesis tests with fixed cutoffs, inconvenient behavior with large samples is to be expected. This manifests itself in multiple ways, including a tendency of Bayes factors to favor null hypotheses strongly for large samples, especially in cases of small effect sizes.

3. Solving Some of the Problems in Hypothesis Testing

Most readers of this article already knew before starting to read it that there are some problems with the kinds of hypothesis testing done in many fields of research. In the previous section, some of those problems have been described in a bit more detail than a news article can usually dedicate to the subject. So now what? What can a researcher do? In this section, one solution is presented, and some of its advantages over currently used hypothesis tests are described. The approach uses both frequentist and Bayesian methods and results in hypothesis tests that are operationally very similar to the commonly used tests that compare a p-value to a significance level α.

The authors of this article take the position taken over 60 years ago by both a Bayesian statistician, Lindley (1957), and a frequentist statistician, Bartlett (1957): a major part of the problems with p-value-based tests is in fixing a significance level that does not depend on the sample size. That is, the problem is not as much in the use of p-values themselves as in comparing p-values to fixed signficance levels. The solution to this issue is surprisingly simple and is rooted in the presentation of Neyman and Pearson's lemma in DeGroot's widely used textbook (DeGroot 1986), perhaps the greatest bridge between the frequentist and Bayesian "schools" of statistics. Instead of starting with a fixed α and determining the tests that minimize β as Neyman and Pearson did, DeGroot presented a generalized form of the Neyman–Pearson Lemma in which a linear combination of α and β is minimized. When α is then fixed, the result is the same as the one presented by Neyman and Pearson in 1933. In its full generality, the version presented by DeGroot has a major advantage: by minimizing a linear combination of α and β, it allows the probabilities of both types of error to vary, avoiding the kind of drastic imbalance between the probabilities of the

[5] A proof of the equivalence of the LP to the combination of two *much* less-controversial principles, the Conditionality Principle (CP) and Sufficiency Principle (SP), appears in Birnbaum (1962), and a modified version in Wechsler, Pereira, and Marques (2008), but the validity of this kind of proof has been questioned by, for example, Evans (2013) and Mayo (2014). Gandenberger (2015) presents a proof of the same equivalence designed to resist the kinds of attacks brought against Birnbaum's proof, but the controversy continues to rage. It is worth noting that even if both proofs were incorrect, that would not mean that the LP was *not* equivalent to the combination of the CP and SP. Further, even if that equivalence were not valid, that would not mean that the LP is not true. Even so, the LP is still not considered proved, which is why it is still just a principle that one chooses either to accept or not.

two types of errors described in Pericchi and Pereira (2016), because instead of α being fixed and β tending to decrease with increasing sample size, both error probabilities depend on the sample size. By controlling the ratio of the coefficients of α and β in the linear combination that is minimized, a researcher can actually determine "just how the balance should be struck," realizing the vision of Neyman and Pearson in a way Neyman-Pearson tests simply cannot. Cornfield (1966) suggested optimizing tests by minimizing a linear combination of α and β.

How should the coefficients a and b in the linear combination $a\alpha+b\beta$ be chosen? The coefficients a and b represent the relative seriousness of errors of the two types or, equivalently, relative prior preferences for the competing hypotheses. If, for example, $a > b$, that means type-I errors are considered more serious than type-II errors. That means incorrectly rejecting **H** in favor of **A** is considered more serious than incorrectly rejecting **A** in favor of **H**, which indicates a prior preference for **H**. Here is a concrete example: imagine a state in which there have been more cases of meningitis than usual, and where the governor is very budget-conscious. Take **H** to be the hypothesis that there is *not* a meningitis epidemic in the state, and **A** to be the competing hypothesis that there *is* an epidemic. The governor may consider the unnecessary spending from an incorrect rejection of **H** to be more serious than the consequences of not declaring an epidemic, or equivalently, favor hypothesis **H** over **A**, and so would set $a > b$. Decision theory allows for the underlying assumptions to be made even more explicit by going more deeply into the meaning of a and b. The details can be found in the section on "Bayes test procedures" in DeGroot (1986), but the important point here is that if the losses due to incorrect rejection of each hypothesis can be quantified, and the prior probability that **H** is true can be estimated, then a and b can be calculated from those numbers. It is worth mentioning that the absolute scale of a and b, and therefore the absolute scale of the losses from the two possible types of errors, do not matter; only the ratio of a and b affects the actual decision whether to reject a hypothesis.

The Neyman–Pearson Lemma and the extended version of it presented by DeGroot are proved for simple-vs.-simple hypotheses, that is, for comparing specific values of a parameter. For example, for a normal (Gaussian) distribution with mean θ and variance 1, $\mathcal{N}(\theta, 1)$, one might compare the hypotheses $\mathbf{H} : \theta = 0$ and $\mathbf{A} : \theta = 0.7$. For a given observation x, the ratio of likelihoods $\mathcal{L}(\theta = 0|x)/\mathcal{L}(\theta = 0.7|x)$ would be compared to a cutoff chosen so that the probability, under hypothesis **H**, that is, if x obeyed a $\mathcal{N}(0, 1)$ distribution, of an observation falling in the rejection region would be some fixed value α, like the commonly used 0.05 or the recently suggested 0.005. For certain types of distributions and certain types of hypotheses (see DeGroot (1986) or other statistics textbooks for details), the Neyman–Pearson Lemma can be extended to find the best tests for composite hypotheses, that is, hypotheses involving multiple values or continuous ranges of values of the parameter of interest. A Bayesian approach to extending beyond optimal simple-vs.-simple hypothesis tests offers a simple and obvious way to extend to tests where one or both of the hypotheses can be composite hypotheses, and where either or both hypotheses may be very complex.

As usual, consider a random vector x representing experimental results, with a probability (density) function[6] $f(x|\theta)$ having a parameter vector θ, with x and θ elements of real sample space $\mathfrak{X}$ and parameter space Θ, respectively, each space having some positive integer dimensionality. The competing hypotheses **H** and **A** must partition the parameter space, that is, divide it into nonoverlapping pieces $\Theta_{\mathbf{H}}$ and $\Theta_{\mathbf{A}}$ such that the hypotheses can be expressed as

$$\mathbf{H} : \theta \in \Theta_{\mathbf{H}} \text{ and } \mathbf{A} : \theta \in \Theta_{\mathbf{A}}. \tag{1}$$

As long as the two pieces make up the entire space ($\Theta = \Theta_{\mathbf{H}} \cup \Theta_{\mathbf{A}}$) and do not overlap ($\Theta_{\mathbf{H}} \cap \Theta_{\mathbf{A}} = \varnothing$), the hypotheses can be of any dimensionality and arbitrarily complex.

Define the binary parametric function $\lambda(\theta)$ as follows:

$$\lambda(\theta) = \begin{cases} 0 & \theta \in \Theta_{\mathbf{H}} \\ 1 & \theta \in \Theta_{\mathbf{A}}. \end{cases} \tag{2}$$

Because λ is a function of θ, one can write

$$f(x|\theta) = f(x|\theta, \lambda). \tag{3}$$

Now treat the original parameter θ as a "nuisance parameter" and remove it the Bayesian way: by taking averages of $f(x|\theta, \lambda)$, weighted by a prior $g(\theta)$, over the two pieces of the parameter space $\Theta_{\mathbf{H}}$ and $\Theta_{\mathbf{A}}$. The result is two predictive probability (density) functions

$$\begin{aligned} f_{\mathbf{H}}(x) &= f(x|\lambda = 0) \text{ and} \\ f_{\mathbf{A}}(x) &= f(x|\lambda = 1). \end{aligned} \tag{4}$$

Using the approach based on the generalized form of the Neyman–Pearson Lemma presented by DeGroot and previously used by Cornfield, as described earlier in this section, but now with the likelihoods averaged over $\Theta_{\mathbf{H}}$ and $\Theta_{\mathbf{A}}$ to produce $f_{\mathbf{H}}$ and $f_{\mathbf{A}}$, one obtains *averaged* error probabilities α and β that are optimal in the sense of the generalized Neyman–Pearson Lemma, and the optimal averaged α can be used as a significance level that depends strongly on the sample size.

As stated in the previous section, one problem with hypothesis tests comes from the use of definitions like Definition 1 to calculate p-values. A second definition that takes into account the alternative hypothesis **A**, unlike Definition 1, is presented in Pereira and Wechsler (1993). As in the previous definition, x_0 represents the observation and **H** is the null hypothesis.

Definition 2 (PW2.1). The P-value is the probability, under **H**, of the event composed by all sample points that favor **A** (against **H**) at least as much as x_0 does.

Note that the quantity defined here is a capital-P "P-value," to distinguish it from the small-p "p-values" defined by Definition 1. The P-value has the advantage that, unlike a small-p p-value, it can be calculated for arbitrarily complex hypotheses that lead to arbitrarily complex rejection regions (regions where $p < \alpha$, where the null hypothesis would be rejected if the experimental observations were to occur there). However, to do so,

[6] The notation "probability (density) function" is used to refer to a probability function for discrete sample spaces and a probability density function for continuous sample spaces.

it requires an ordering of the space of possible observations, the sample space, according to how much each possible observation favors one of the hypotheses over the other. The new approach, like Montoya-Delgado et al. (2001), uses the Bayes factor

$$\mathrm{BF}_{\mathbf{HA}}(x) = \frac{f_{\mathbf{H}}(x)}{f_{\mathbf{A}}(x)} \tag{5}$$

to order the sample space according to how much each point x favors **H** over **A**. Then, if the experiment yields a result x_0, the P-value is calculated as the probability, given probability (density) function $f_{\mathbf{H}}(x)$, of a point in the sample space favoring **A** as much as or more than x_0 does. That is, it is the sum or integral of the probability (density) function $f_{\mathbf{H}}(x)$ over the part of the sample space where the Bayes factor $\mathrm{BF}_{\mathbf{HA}}(x)$ is less than or equal to the Bayes factor calculated at x_0. Calling that part of the sample space Ψ, it is defined as

$$\Psi = \{x \in \mathfrak{X} : \mathrm{BF}_{\mathbf{HA}}(x) \leq \mathrm{BF}_{\mathbf{HA}}(x_0)\}, \tag{6}$$

and the P-value is

$$P = \begin{cases} \sum_{\Psi} f_{\mathbf{H}}(x), & \mathfrak{X} \text{ discrete} \\ \int_{\Psi} f_{\mathbf{H}}(x)dx, & \mathfrak{X} \text{ continuous.} \end{cases} \tag{7}$$

There is a one-to-one correspondence between the P-value used in this approach and the Bayes factors used to calculate them, so once a cutoff for P—a significance level α—is determined, a corresponding cutoff for Bayes factors at the same significance level can be determined. As a result, researchers who are more comfortable with Bayes-factor tests can continue to use them, but with a cutoff determined by this method, which takes experimental error into account and depends on the sample size, rather than with arbitrarily defined cutoffs like those of Jeffreys (1935) or Kass and Raftery (1995).

4. Examples

Here some specific examples of applications of the new tests are presented. An important point emphasized throughout is the way the significance level varies with sample size.

4.1. Bernoulli ("Yes–No") Examples

Though they are the simplest models, Bernoulli models, which are used for experiments with a "yes–no" result, like "heads-or-tails" coin flips, are very important. This is because many medical studies, among other kinds of experiments, fall into this category, with "yes" indicating, for example, a recovery from an illness or injury. The unknown parameter is the proportion or probability of success, and hypothesis tests are applied to that parameter in different ways, depending on the purpose of the study.

4.1.1. Comparing Two Proportions

A physician wants to show that the incorporation of a new technology in a treatment can produce better results than the conventional treatment. He plans a clinical trial with two arms: case and control, each with eight patients. The case arm receives the new treatment and the control arm receives the conventional one. The observed results in this example are that only one of the patients in the control arm responded positively, but in the case arm there were four positive outcomes.

The most common classical significance tests result in the following p-values: the Pearson χ^2 p-value is 0.106, changed to 0.281 with the Yates continuity correction applied, and Fisher's exact p-value is 0.282. Traditional analysts would conclude that there were no statistically significant differences between the two treatments, using any of the canonical significance levels. Note that these procedures were for testing a sharp hypothesis against a composite alternative – $\mathbf{H}: \ \theta_0 = \theta_1$ vs. $\mathbf{A}: \ \theta_0 \neq \theta_1$, comparing the proportions of success for the two treatments. Note that the hypothesis **H** is precise or "sharp," representing a line in the two-dimensional (θ_0, θ_1) parameter space. Here, the proposed P-value and the optimal significance level $\alpha(\delta^*)$ are calculated to choose one of the hypotheses using the new tests.

To be fair in our comparisons, we consider independent uniform (noninformative) prior distributions for θ_0 and θ_1. With these suppositions and the likelihoods being binomials with sample sizes $n = 8$, the predictive probability functions under the two hypotheses are

$$f_{\mathbf{H}}(x,y) = \frac{\binom{8}{x}\binom{8}{y}}{17\binom{16}{x+y}} \quad \text{and} \quad f_{\mathbf{A}}(x,y) = \frac{1}{81}, \tag{8}$$

where $(x,y) \in \{0,1,\ldots,8\} \times \{0,1,\ldots,8\}$ represent the possible observed values of the number of positive outcomes in the two arms of the study. Table 1 shows the Bayes factors for all possible results.

To obtain the P-value, define the set Ψ_{obs} of sample points (x,y) for which the Bayes factors are smaller than or equal to

Table 1. Bayes factor for all possible results in a clinical trial with two arms of size $n = 8$ each. Cells in **boldface** make up the region Ψ^*, and the observed Bayes factor is shown in ***boldface italics.*** See text.

	y									
x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Sum**
0	4.765	2.382	1.112	**0.476**	**0.183**	**0.061**	**0.017**	**0.003**	**4e-04**	**9**
1	2.382	2.541	1.906	1.173	***0.611***	**0.267**	**0.093**	**0.024**	**0.003**	**9**
2	1.112	1.906	2.052	1.710	1.166	**0.653**	**0.29**	**0.093**	**0.017**	**9**
3	**0.476**	1.173	1.710	1.866	1.633	1.161	**0.653**	**0.267**	**0.061**	**9**
4	**0.183**	***0.611***	1.166	1.633	1.814	1.633	1.166	**0.611**	**0.183**	**9**
5	**0.061**	**0.267**	**0.653**	1.161	1.633	1.866	1.710	1.173	**0.476**	**9**
6	**0.017**	**0.093**	**0.29**	**0.653**	1.166	1.710	2.052	1.906	1.112	**9**
7	**0.003**	**0.024**	**0.093**	**0.267**	**0.611**	1.173	1.906	2.541	2.382	**9**
8	**4e-04**	**0.003**	**0.017**	**0.061**	**0.183**	**0.476**	1.112	2.382	4.765	**9**
Sum	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**81**

the Bayes factor of the observed sample point, that is,

$$\Psi_{\text{obs}} = \left\{(x,y) \in \{0,1,\ldots,8\} \times \{0,1,\ldots,8\} : \ \text{BF} \le \text{BF}_{\text{obs}}\right\}, \tag{9}$$

and then the P-value is the sum of the prior predictive probabilities under $\mathbf{H}$ in Ψ_{obs} :

$$P = \sum_{(x,y)\in\Psi_{\text{obs}}} f_{\mathbf{H}}(x,y) = \sum_{(x,y)\in\Psi_{\text{obs}}} \frac{\binom{8}{x}\binom{8}{y}}{17\binom{16}{x+y}}. \tag{10}$$

Recalling the observed result of the clinical trial, (x,y) = (1,4), the observed Bayes factor is $\text{BF}_{\text{obs}} = 0.661$. Based on this, the P-value is $P = 0.0923$.

The test δ^* minimizes the linear combination $a\alpha(\delta) + b\beta(\delta)$. The Bayes factor is compared to the constant K, the ratio of the coefficients: $K = \frac{b}{a}$. Then, define the set $\Psi^* = \left\{(x,y) \in \{0,1,\ldots,8\} \times \{0,1,\ldots,8\} : \ \text{BF} \le K\right\}$, and the optimal averaged error probabilities from the generalized Neyman–Pearson Lemma are

$$\begin{aligned} \alpha(\delta^*) &= \sum_{(x,y)\in\Psi^*} f_{\mathbf{H}}(x,y) = \sum_{(x,y)\in\Psi^*} \frac{\binom{8}{x}\binom{8}{y}}{17\binom{16}{x+y}} \quad \text{and} \\ \beta(\delta^*) &= \sum_{(x,y)\notin\Psi^*} f_{\mathbf{A}}(x,y) = \sum_{(x,y)\notin\Psi^*} \frac{1}{81}. \end{aligned} \tag{11}$$

In DeGroot (1986), it is shown using decision theory that a linear combination $w_{\mathbf{H}}\pi\alpha(\delta) + w_{\mathbf{A}}(1-\pi)\beta(\delta)$ is minimized, where $w_{\mathbf{H}}$ is the expected loss from choosing to accept hypothesis $\mathbf{A}$ when $\mathbf{H}$ is true, and π is the prior probability that $\mathbf{H}$ is true. Taking the hypotheses to be equally likely *a priori*, $\pi = \frac{1}{2}$, and representing equal severity of type-I and type-II errors by taking $w_{\mathbf{H}} = w_{\mathbf{A}} = 1$, the result is $K = 1$. The set Ψ^* is identified by the cells with **boldface** numbers in Table 1. The observed Bayes factor is in ***boldface italics.*** The optimal significance level is $\alpha(\delta^*) = 0.1245$ and the optimal averaged type-II error probability is $\beta(\delta^*) = 0.4815$. The high type-II error probability is completely expected for small samples. Contrary to the classical results, the conclusion is now the most intuitive one: the null hypothesis is rejected because $P < \alpha(\delta^*)$. However, the rejection is only at the 12.45% level of significance. So what sample size would be necessary to obtain some better (lower) significance level, say 10%?

4.1.2. Comparing Two Proportions With Varying Sample Sizes

Consider first a clinical trial just like in the previous example, but now with arms of size $n = 20$. The observed result is $(x,y) = (4,10)$, that is, four patients had a positive result in the control arm, while 10 had a positive result in the arm receiving the new treatment. The same calculations done for the previous section are repeated, but with the appropriate expressions for $f_{\mathbf{H}}$ and $f_{\mathbf{A}}$ for a trial with two 20-patient arms. The observed Bayes factor in this case is $\text{BF}_{\text{obs}} = 0.415$, which leads to significance index $P = 0.02901$, optimal significance level $\alpha(\delta^*) = 0.0995$, and type-II error probability $\beta(\delta^*) = 0.3651$. The classical χ^2 p-value is $p = 0.0467$, indicating rejection of the null hypothesis at the canonical 5% significance level. The new test also rejects $\mathbf{H}$, because $P < \alpha(\delta^*)$.

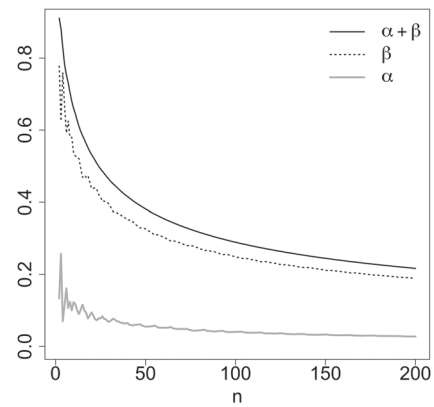


Figure 1. Optimal averaged type-I (solid gray line), type-II (dotted line), and total (solid black line) error probabilities as functions of the number of patients n in each arm of a two-arm medical study.

The same analysis can be done to calculate the optimal significance level and type-II error probability for any sample size. Figure 1 shows the optimal adaptive significance level α and optimal adaptive type-II error probability, plus the minimized linear combination $\alpha + \beta$, as functions of the size of each arm in a study.

Table 2 shows the optimal adaptive averaged error probabilities α and β for various arm sizes without the restriction of the two arms having equal size. For a given total sample size, an unbalanced sample can have higher probabilities of both type-I and type-II errors than a balanced sample. For example, the error probabilities for an unbalanced sample with $n_1 = 60$ and $n_2 = 20$ are larger than those for a balanced sample with $n_1 = n_2 = 40$, even though both experiments would have the same total sample size, $n_1 + n_2 = 80$. The effect of unbalanced samples can be as important as the effect of total sample size. For example, the error probabilities of an unbalanced sample with $n_1 = 60$ and $n_2 = 10$ are larger than those of a balanced sample with $n_1 = n_2 = 20$, even though the unbalanced sample has a total size of $n_1 + n_2 = 70$ and the balanced sample just 40.

4.1.3. Test for One Proportion and the Likelihood Principle

A common example in which the Likelihood Principle can be violated is the comparison of binomials to negative binomials in a coin-flipping experiment with a coin that may or may not have the expected 50–50 chance of coming up "heads," as described in Lindley and Phillips (1976).[7] For the same values of x, the number of successes in n independent coin flips, the two distributions produce different p-values, which can lead to different decisions at a given level of significance. That is, the inference can actually be different *for exactly the same data,* depending on the intent of the researcher before starting the experiment. If a result is nine heads and three tails, did the researcher start the experiment planning to flip the coin exactly 12 times? Was the intent to flip until the 9th occurrence of heads? Until the third occorrence of tails? Was it some other

[7] Lindley and Phillips actually consider throws of thumbtacks, called "drawing pins" in U.K. English, but the principle is identical.

Table 2. Optimal averaged error probabilities $\alpha(\delta^*)$ and $\beta(\delta^*)$ for comparison of two proportions for various arm sizes n_1 and n_2 in a two-arm medical study. Calculations were performed with $a = b$.

n_1	n_2	α	β	n_1	n_2	α	β	n_1	n_2	α	β
10	10	0.1639	0.4050	60	50	0.0626	0.2652	90	30	0.0707	0.2804
20	10	0.1318	0.3939	60	60	0.0591	0.2572	90	40	0.0648	0.2608
20	20	0.0995	0.3651	70	10	0.1130	0.3675	90	50	0.0575	0.2506
30	10	0.1159	0.3900	70	20	0.0865	0.3132	90	60	0.0550	0.2401
30	20	0.1045	0.3333	70	30	0.0727	0.2876	90	70	0.0529	0.2323
30	30	0.0997	0.3070	70	40	0.0645	0.2717	90	80	0.0493	0.2281
40	10	0.1250	0.3703	70	50	0.0603	0.2593	90	90	0.0468	0.2240
40	20	0.0868	0.3357	70	60	0.0575	0.2501	100	10	0.1111	0.3627
40	30	0.0850	0.3029	70	70	0.0539	0.2446	100	20	0.0818	0.3079
40	40	0.0706	0.2968	80	10	0.1130	0.3648	100	30	0.0684	0.2795
50	10	0.1126	0.3761	80	20	0.0834	0.3122	100	40	0.0617	0.2601
50	20	0.0883	0.3240	80	30	0.0704	0.2847	100	50	0.0559	0.2479
50	30	0.0767	0.2992	80	40	0.0634	0.2671	100	60	0.0538	0.2368
50	40	0.0718	0.2817	80	50	0.0603	0.2530	100	70	0.0512	0.2291
50	50	0.0667	0.2718	80	60	0.0553	0.2455	100	80	0.0483	0.2238
60	10	0.1097	0.3741	80	70	0.0531	0.2380	100	90	0.0467	0.2188
60	20	0.0860	0.3193	80	80	0.0508	0.2327	100	100	0.0449	0.2150
60	30	0.0765	0.2903	90	10	0.1131	0.3626				
60	40	0.0689	0.2747	90	20	0.0810	0.3114				

intent, but the experiment was interrupted? With frequentist hypothesis tests, this actually matters because they violate the Likelihood Principle.

In this example, the new tests are applied to show that they do *not* violate the Likelihood Principle. The reason the inference (decision to accept or reject a hypothesis about the parameter θ) ends up being the same for different models is that although the P-values for the two models are different from each other, the adaptive significance levels α for the two models are also different, and the decision about rejecting one hypothesis in favor of the other ends up being the same. Using different notation from the previous example, let the sample vector consist of the number of successes x and the number of failures y, and let the corresponding vector of probabilities be $(\theta, 1-\theta)$. Take $\mathbf{H}: \theta = \frac{1}{2}$ vs. $\mathbf{A}: \theta \neq \frac{1}{2}$, that is, a fair coin vs. an unbalanced coin, as the hypotheses to be tested. Taking a uniform prior for θ, taking the two hypotheses to be equally probable a priori ($\pi = \frac{1}{2}$), and considering the two types of error equally severe, the predictive probabilities for the tests are as follows.
For a binomial,

$$f_{\mathbf{H}}(x,y) = \binom{x+y}{x}\left(\frac{1}{2}\right)^{x+y} \quad \text{and} \quad f_{\mathbf{A}}(x,y) = \frac{1}{x+y+1}, \tag{12}$$

and for a negative binomial,

$$f_{\mathbf{H}}(x,y) = \binom{x+y-1}{x}\left(\frac{1}{2}\right)^{x+y}$$
$$\text{and} \quad f_{\mathbf{A}}(x,y) = \frac{y}{(x+y)(x+y+1)}. \tag{13}$$

The Bayes factors are equal for the two models, and since using the lemma will lead to comparing them to the same constant, the decisions about the hypothesis $\mathbf{H}: \theta = \frac{1}{2}$ end up being the same. The P-values and significance levels α are different for the two models, but the inference ends up being the same. Considering the observations $(x,y) = (3,10)$ and $(x,y) = (10,3)$ for a binomial, both samples yield the same results: $P = 0.02$, where the optimal error probabilities are $\alpha = 0.09$ and $\beta = 0.43$. For a negative binomial, the same observations produce different values of the significance index P, but the error probabilities are different. For the first (second) sample, one stops observing when the number of successes reaches 3 (reaches 10). For the first sample point, the P-value is 0.01, and the relevant error probabilities are $\alpha = 0.18$ and $\beta = 0.48$. For the second sample, $P = 0.01$, and the error probabilities are $\alpha = 0.12$ and $\beta = 0.33$. The decisions made for binomials are the same as those for negative binomials with the same (x,y). This behavior is much more general than this specific example, and in fact it is proven in Pereira et al. (2017) that the new tests are compliant with the Likelihood Principle for any discrete sample space.

4.1.4. An Important Note About "Yes–No" Experiments and the New Tests

The predictive densities under many common hypotheses in Bernoulli experiments can be calculated analytically for binomial and negative bionomial models when Beta priors are used. A few examples are $\mathbf{H}_1: \theta = \theta_0$; $\mathbf{H}_2: \theta \neq \theta_0$; $\mathbf{H}_3: \theta \leq \theta_0$; $\mathbf{H}_4: \theta > \theta_0$; $\mathbf{H}_5: \theta \in [\theta_1, \theta_2]$; and $\mathbf{H}_6: (\theta < \theta_1) \cup (\theta > \theta_2)$; etc. The predictive densities for several such sets of hypotheses are presented in Pereira et al. (2017).

4.2. Tests in Normal Distributions

Normal distributions are very widely used because the sample means of large random samples of random variables tend to act like normally distributed variables. This is the familiar tendency of large-sample distributions to look like "bell curves," described mathematically by the central limit theorem. Details can be found in DeGroot (1986) or other statistics textbooks. Because of the importance of normal distributions, two examples are presented here of how the tests can be used with normal distributions.

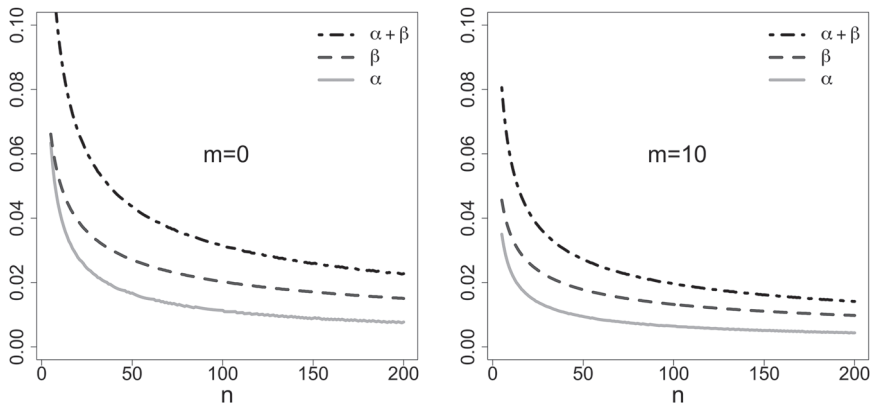


Figure 2. Type-I (solid gray lines), type-II (dashed black lines), and total (dot-dashed black lines) error probabilities as functions of sample size n for tests of $\mathtt{H}: \mu = 0$ on a normally distributed variable with variance 1 and unknown mean μ, $\mathcal{N}(\mu, 1)$, with priors for the mean $\mu \sim \mathcal{N}(m, 100)$ with $m = 0$ (left) and $m = 10$ (right).

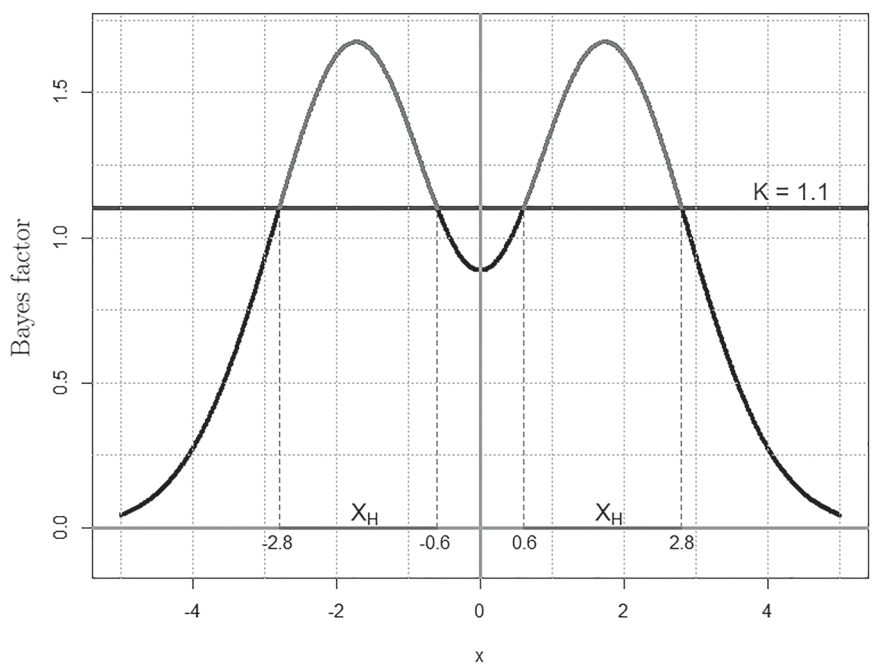


Figure 3. Bayes factor for $\mathcal{N}(0, 2)$ vs. Cauchy, arising from a test of a normal variance with hypotheses $\mathtt{H}: \sigma^2 = 2$ vs. $\mathtt{A}: \sigma^2 \neq 2$.

4.2.1. Test of a Mean in a Normal Distribution

Consider iid variables $X_1, \ldots, X_n \mid \mu$, that obey a normal distribution with mean μ and variance 1. Take as a prior for μ a normal distribution with mean m and variance v. Instead of working with the full n-dimensional sample space, the minimal sufficent statistic $\overline{X}_n$, the sample mean, can be used. The sampling distribution of $\overline{X}_n \mid \mu$ is a normal distribution with mean μ and variance $1/n$. For hypotheses $\mathbf{H}: \mu = 0$ and $\mathbf{A}: \mu \neq 0$, the predictive distributions are $\overline{X}_n \mid \mathbf{H} \sim \mathcal{N}(0, 1/n)$ and $\overline{X}_n \mid \mathbf{A} \sim \mathcal{N}(m, v + 1/n)$.

To calculate the probability of a type-I error, define the region

$$\mathfrak{X}_{\mathbf{A}} = \{\overline{x}_n : f_{\mathbf{H}}(\overline{x}_n)/f_{\mathbf{A}}(\overline{x}_n) < 1\}, \tag{14}$$

and evaluate $\alpha = \int_{\mathfrak{X}_{\mathbf{A}}} f_{\mathbf{H}}(\overline{x}_n) \mathrm{d}\overline{x}_n$. For the probability of a type-II error, define the region $\mathfrak{X}_{\mathbf{H}} = \{\overline{x}_n : f_{\mathbf{H}}(\overline{x}_n)/f_{\mathbf{A}}(\overline{x}_n) > 1\}$, and evaluate $\beta = \int_{\mathfrak{X}_{\mathbf{H}}} f_{\mathbf{A}}(\overline{x}_n) \mathrm{d}\overline{x}_n$. These error probabilities are plotted in Figure 2 for $v = 100$ and two values of the prior mean: $m = 0$ and $m = 10$. The plot with $m = -10$ would be identical to the plot with $m = 10$.

4.2.2. Test of a Variance in a Normal Distribution

This is an example used by Pereira and Wechsler (1993), showing that the region of the space of possible results where a test rejects the null hypothesis is not always the tails of the null distribution; it can be a union of disjoint intervals. In such cases, it can be impossible to calculate a classical p-value defined as in Definition 1, but the ordering of the entire sample space by Bayes factors allows for an unambiguous definition and calculation of the new index, a capital-P P-value in the sense of Definition 2.

Consider a normally distributed random variable X with mean zero and unknown variance σ^2. The hypotheses considered here are $\mathbf{H}: \sigma^2 = 2$ and $\mathbf{A}: \sigma^2 \neq 2$. A χ^2_1 (chi-squared with one degree of freedom) distribution is used as a prior density for σ^2. The predictive densities under hypotheses $\mathbf{H}$ and $\mathbf{A}$ are

$$f_{\mathbf{H}}(x) = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{x^2}{4}\right) \text{ and } f_{\mathbf{A}}(x) = \frac{1}{\pi\left(1 + x^2\right)}, \tag{15}$$

a normal density with mean zero and variance 2, and a Cauchy density, respectively. Figure 3 shows a plot of the Bayes factor,

using a value of 1.1 as a cutoff for a decision about the hypotheses. The sample points that do *not* favor **H** are in three separate regions: a central interval and the heavy tails of the Cauchy density. The set that favors **H** is the complement, made up of two intervals between the central interval and the tails:

$$X_{\mathbf{H}} = \{x | x \in (-2.8, -0.6) \cup (0.6, 2.8)\}. \tag{16}$$

The set that favors **A** over **H** includes, in addition to the tails, the central region $(-0.6, 0.6)$. Even with this or even more complex divisions of the sample space into regions that favor one hypothesis over another, the new method allows for calculation of a P-value.

4.3. Test of Hardy–Weinberg Equilibrium

"Hardy–Weinberg equilibrium" refers to the principle, proven by Hardy (1908) and Weinberg (1908), that allele and genotype frequencies in a population will remain constant from generation to generation, given certain assumptions about the absence of external evolutionary influences.

An individual's genotype is determined by a combination of alleles. If there are two possible alleles for some characteristic (say A and a), the possible genotypes are AA, Aa, and aa. Let x_1, x_2, x_3 be the observed frequencies of the genotypes AA, Aa, and aa, respectively, and $\theta_1, \theta_2, \theta_3$ the corresponding probabilities. Assuming a few premises, as described by Hartl and Clark (1989), the principle says that the allele probabilities in a population do not change from generation to generation. It is a fundamental principle for the Mendelian mating allelic model. If the probabilities of alleles are θ for allele A and $1-\theta$ for the allele a, the expected genotype probabilities are $\left((\theta^2, 2\theta(1-\theta), (1-\theta)^2\right), 0 \le \theta \le 1$.

The Hardy–Weinberg equilibrium hypothesis is

$$\mathbf{H}: \begin{cases} \theta_1 &= \theta^2, \\ \theta_2 &= 2\theta(1-\theta), \\ \theta_3 &= (1-\theta)^2, \end{cases} \tag{17}$$

Given n, using as a prior distribution for $\{\theta_1, \theta_2, \theta_3\}$ a Dirichlet$(1,1,1)$, that is, $f(\theta_1, \theta_2, \theta_3) = 2$, and using the multinomial probability

$$f(x_1, x_2, x_3 \mid \theta_1, \theta_2, \theta_3) = \frac{n!}{x_1! x_2! x_3!} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3} \tag{18}$$

for (x_1, x_2, x_3), the predictive under hypothesis **A** is

$$\begin{aligned} f_{\mathbf{A}}(x_1, x_2, x_3) &= \int_0^1 \int_0^{1-\theta_1} f(x_1, x_2, x_3 \mid \theta_1, \theta_2, 1-\theta_1-\theta_2) \\ &\quad \times f(\theta_1, \theta_2, 1-\theta_1-\theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2 \\ &= \int_0^1 \int_0^{1-\theta_1} \frac{n!}{x_1! x_2! x_3!} \theta_1^{x_1} \theta_2^{x_2} (1-\theta_1-\theta_2)^{x_3} \\ &\quad \times 2\, \mathrm{d}\theta_1 \mathrm{d}\theta_2 \\ &= \frac{2n!}{(n+2)!}. \end{aligned} \tag{19}$$

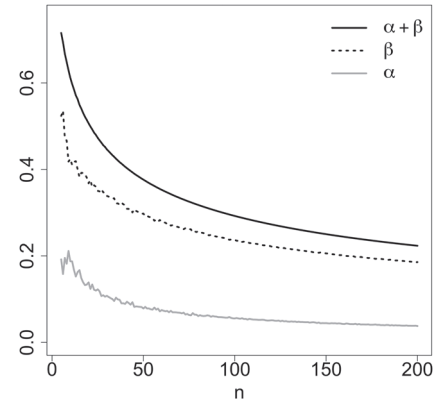


Figure 4. Type-I (solid gray line), type-II (dotted line), and total (solid black line) error probabilities as functions of the sample size n for the Hardy–Weinberg equilibrium hypothesis.

Under hypothesis **H**, the surface integral to normalize the values of the prior density in the set corresponding to the null hypothesis yields $f(\theta) = 1, 0 \le \theta \le 1$. Then,

$$\begin{aligned} f_{\mathbf{H}}(x_1, x_2, x_3) &= \int_0^1 f(x_1, x_2, x_3 \mid \theta) f(\theta) \mathrm{d}\theta \\ &= \int_0^1 \frac{n!}{x_1! x_2! x_3!} 2^{x_2} \theta^{2x_1+x_2} (1-\theta)^{2x_3+x_2} \mathrm{d}\theta \\ &= \frac{n! 2^{x_2} (2x_1+x_2)! (2x_3+x_2)!}{x_1! x_2! x_3! (2n+1)!}. \end{aligned} \tag{20}$$

The probability of a type-I error, α, is obtained from a sum of the predictive under **H** over all samples $\{x_1, x_2, x_3\}$ where $f_{\mathbf{H}}(x_1, x_2, x_3)/f_{\mathbf{A}}(x_1, x_2, x_3) < 1$. The probability of a type-II error, β, is obtained from a sum of the predictive under **A** over all samples $\{x_1, x_2, x_3\}$ where $f_{\mathbf{H}}(x_1, x_2, x_3)/f_{\mathbf{A}}(x_1, x_2, x_3) > 1$. These error probabilities are plotted in Figure 4.

5. Final Considerations

Using the hypothesis-testing procedure described here, the sample-size dependence of the optimal averaged α can be used to determine the best significance level for a given n, which can be relevant in studies where a limited number of trials can be done, or to determine the necessary sample size to achieve a desired α.

The procedure can be applied to other models and hypotheses, without restrictions on the dimensionality of the parameter space or the sample space. The sample space, regardless of its dimensionality, is ordered by a single real number, the Bayes factor. In all the examples used in this article, analytic (closed-form mathematical) solutions are available, but this is not the case for all distributions and hypotheses. For example, hypotheses about a normal distribution with unknown mean *and* variance involve calculations of significantly greater complexity and require some sort of analytic approximations or approximate methods for evaluating the necessary integrals, such as Monte Carlo methods. Note that the method requires the calculation of five integrals: two over the nuisance parameter(s) to find the predictive distributions $f_{\mathbf{H}}(x)$ and $f_{\mathbf{A}}(x)$; and three of predictive distributions over specific regions of the sample space to obtain α, β, and the capital-P P-value.

It is worth noting that the approach described here is compatible with guidelines in the ASA's statement on p-values (Wasserstein and Lazar (2016)). Specifically, because the signficance level depends on the sample size, and therefore is not the kind of predefined "bright-line" rule the ASA recommends avoiding, the approach is compatible with point 3, "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

Funding

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Finance Code 001. C.A. de B.P. thanks the *Conselho Nacional de Desenvolvimento Científico e Tecnológico*, CNPq, for support under grant number 308776/2014-3.

Acknowledgement

The authors thank Dr. Ronald L. Wasserstein for encouragement to write this article and Dr. Sergio Wechsler for illuminating conversations about the research described here and his invaluable contributions to research that led to the present work.

References

Bartlett, M. (1957), "A Comment on D. V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534. [215]

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., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [214]

Berger, J., and Wolpert, R. (1988), *The Likelihood Principle*, (2nd ed.), Haywood, CA: The Institute of Mathematical Statistics, available at *https://projecteuclid.org/euclid.lnms/1215466210* [215]

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–335. [214]

Birnbaum, A. (1962), "On the Foundations of Statistical Inference," *Journal of the American Statistical Association*, 57, 269–306. [215]

Cohen, J. (1994), "The Earth is round ($p < .05$)," *American Psychologist*, 49, 997–1003. [213]

Cornfield, J. (1966), "Sequential Trials, Sequential Analysis and the Likelihood Principle," *The American Statistician*, 20, 18–23. [215,216]

DeGroot, M. (1986), *Probability and Statistics*, Addison-Wesley Series in Statistics, Reading, MA, USA: Addison-Wesley Publishing Company. [215,216,218,219]

Evans, M. (2013), "What does the Proof of Birnbaum's Theorem Prove?" *Electronic Journalof Statistics*, 7, 2645–2655. [215]

Gandenberger, G. (2015), "A New Proof of the Likelihood Principle," *The British Journal for the Philosophy of Science*, 66, 475–503. [215]

Gelman, A. and Rubin, D. B. (1995), "Avoiding Model Selection in Bayesian Social Research," *Sociological Methodology*, 25, 165–173. [214]

Hardy, G. H. (1908), "Mendelian Proportions in a Mixed Population," *Science*, 28, 49–50. [221]

Hartl, D. L. and Clark, A. G. (1989), *Principles of Population Genetics*, (2nd ed.), Sunderland, Mass.: Sinauer Associates. [221]

Jeffreys, H. (1935), "Some Tests of Significance, Treated by the Theory of Probability," *Mathematical Proceedings of the Cambridge Philosophical Society*, 31, 203–222. [215,217]

——— (1939), *The Theory of Probability*, Oxford: The Clarendon Press. [215]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [215,217]

Levine, T. R., Weber, R., Hullett, C., Park, H. S., and Lindsey, L. L. M. (2008), "A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research," *Human Communication Research*, 34, 171–187. [213]

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192. [215]

Lindley, D. V., and Phillips, L. D. (1976), "Inference for a Bernoulli Process (a Bayesian View)," *The American Statistician*, 30, 112–119. [215,218]

Mayo, D. G. (2014), "On the Birnbaum Argument for the Strong Likelihood Principle," *Statistical Science*, 29, 227–239. [215]

Montoya-Delgado, L. E., Irony, T. Z., Pereira, C. A. B., and Whittle, M. R. (2001), "An Unconditional Exact Test for the Hardy-Weinberg Equilibrium Law: Sample-Space Ordering Using the Bayes Factor," *Genetics*, 158, 875–883. [217]

Neyman, J., and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337. [214]

Pereira, C. A. d. B., Nakano, E. Y., Fossaluza, V., Esteves, L. G., Gannon, M. A., and Polpo, A. (2017), "Hypothesis Tests for Bernoulli Experiments: Ordering the Sample Space by Bayes Factors and Using Adaptive Significance Levels for Decisions," *Entropy*, 19, 696. [219]

Pereira, C. A. d. B., and Wechsler, S. (1993), "On the Concept of p-value," *Revista Brasileira de Probabilidade e Estatística*, 7, 159–177. [214,215,216,220]

Pericchi, L., and Pereira, C. (2016), "Adaptative Significance Levels using Optimal Decision Rules: Balancing by Weighting the error Probabilities," *Brazilian Journal of . Probability and Statistics*, 30, 70–90. [214,216]

Schervish, M. (2012), *Theory of Statistics*, Springer Series in Statistics, New York: Springer. [214]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [213]

Tukey, J. W. (1969), "Analyzing Data: Sanctification or Detective Work?" *American Psychologist*, 24, 83–91. [213]

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. [213,222]

Weakliem, D. L. (1999), "A Critique of the Bayesian Information Criterion for Model Selection," *Sociological Methods & Research*, 27, 359–397. [214]

Wechsler, S., Pereira, C. A. d. B., and Marques F., P. C. (2008), "Birnbaum's Theorem Redux," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: Proceedings of the 28th International Workshop*, eds. de Souza Lauretto, M., de Bragança Pereira, C. A., and Stern, J. M., American Inst. of Physics, pp. 96–102. [215]

Weinberg, W. (1908), "Über den Nachweis der Vererbung beim Menschen," *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 64, 369–382. [221]

The American Statistician

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

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

Stanley Pogrow

To cite this article: Stanley Pogrow (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:sup1, 223-234, DOI: 10.1080/00031305.2018.1549101

To link to this article: https://doi.org/10.1080/00031305.2018.1549101

© 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: 8878

View related articles

View Crossmark data

Citing articles: 15 View citing articles

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, 223–234: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1549101

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

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

Stanley Pogrow[a,b]

[a]Equity, Leadership, and Instructional Technology, San Francisco State University, San Francisco, CA; [b]Emeritus Professor of Education, University of Arizona, Tucson, AZ

ABSTRACT

Relying on *effect size* as a measure of *practical significance* is turning out to be just as misleading as using *p*-values to determine the effectiveness of interventions for improving clinical practice in complex organizations such as schools. This article explains how effect sizes have misdirected practice in education and other disciplines. Even when effect size is incorporated into RCT research the recommendations of whether interventions are effective are misleading and generally useless to practitioners. As a result, a new criterion of *practical benefit* is recommended for evaluating research findings about the effectiveness of interventions in complex organizations where benchmarks of existing performance exist. Practical benefit exists when the unadjusted performance of an experimental group provides a noticeable advantage over an existing benchmark. Some basic principles for determining practical benefit are provided. Practical benefit is more intuitive and is expected to enable leaders to make more accurate assessments as to whether published research findings are likely to produce noticeable improvements in their organizations. In addition, practical benefit is used routinely as the research criterion for the alternative scientific methodology of improvement science that has an established track record of being a more efficient way to develop new interventions that improve practice dramatically than RCT research. Finally, the problems with practical significance suggest that the research community should seek different inferential methods for research designed to improve clinical performance in complex organizations, as compared to methods for testing theories and medicines.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
Applied research; Evidence-based decision-making; Improvement science; Interpreting effect sizes; Leadership-decision-making; Organizational effectiveness;

1. Introduction

The widely discussed problems created by relying on *p*-values to determine *statistical significance* have created a wave of advocacy for switching to, or incorporating, *effect sizes* as a measure of the *practical significance* of research findings (Kirk 1996; McCartney and Rosenthal 2000; Hojat and Xu 2004; Sullivan and Feinn 2012). Effect size has the advantages of (a) indicating the amount of difference between groups, and (b) not being affected by the size of the sample. However, I will demonstrate that in the field of education and other disciplines, effect size, typically Cohen's d, has been as problematic as *p*-values in producing misleading findings as to the effectiveness of interventions. In addition, emergent scholarship has documented widespread problems in other disciplines from relying on published effect size results to guide clinical practice.

As a result of the problems with both statistical and practical significance, I recommend switching to an alternative statistical criterion, *practical benefit,* for judging the validity and usefulness of applied research findings for improving the performance of complex organizations and for designing more effective interventions. While the focus of this article is on using practical benefit to inform education leadership decision-making, its use has implications for improving clinical practice in other disciplines. At the same time, the recommendations are not intended to inform research focused on generating or testing theory.

The fundamental question that has guided my work is:

> What kinds of quantitative evidence should leaders seek to feel reasonably confident that if they adopt a given practice based on this evidence that it is likely to improve their schools?

The following discussions will demonstrate why practical benefit, that relies on the actual performance of the experimental group, is a better criterion for answering this question about desired evidence and for supporting leadership decision-making and improving clinical practice in other disciplines, than either practical significance or statistical significance. However, before explaining the nature and advantages of practical benefit, it is necessary to first understand the context of educational research and its similarities and differences to research in other disciplines.

1.1. The Context of Education Research

Experimental education research was largely conducted in laboratory settings till the early 1990s. Criticism of the applicability

CONTACT Stanley Pogrow spogrow@sfsu.edu, stanpogrow@att.net San Francisco State University, San Francisco CA 94132; University of Arizona, Tucson, AZ 85721.

© 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.

of such research to real-world practice led to the major shift of conducting research to test the effectiveness of interventions in school settings. In time, pressure began to mount to increase the scientific rigor of such research by using RCT designs.

However, conducting gold-standard research in the complex, chaotic environment of schools is problematic. Placebos are usually not possible so the control groups tend to represent the existing practice which is usually not well described. Randomization is problematic for the following reasons:

- It is usually impossible to randomize at all the relevant levels of schools, grade levels, teachers, and students,
- Randomization tends not to equate all possible relevant student characteristics,
- Randomization invariably breaks down over time, as teachers leave, parents requests transfers for their children to another class or to a higher performing school, or students leave for other reasons. Student attrition alone over a 3-year period can be 50–70% high-poverty schools.

RCT research is also problematic because there are so many variables that can potentially impact learning that it is impossible to control for most of them. For example, there may be differences in the lead levels of the drinking water between schools, or different room temperatures, etc. Also, generalizing findings across different school districts and states is almost impossible because different tests are used and the contexts of the research vary. In addition, the cost of RCT experiments makes it unlikely that any study will be replicated. The appendix describes how all of the above problems played out in the most famous education experiment, Project STAR in Tennessee to determine the effects of class size reduction.

Given the difficulty of maintaining control of variables in the real world, methodologists have created advanced statistical methods that seek to simulate control of confounding variables or changing circumstances, but these are extremely complex and have their own errors.

The problems with conducting experiments in schools, or any complex organization, as compared to medicine are reflected in Glass's (2016) comparative analysis of meta-analyses in the two disciplines. Gene Glass, who named and pioneered the use of meta-analysis, found that while in medicine the effect sizes of experiments in a given meta-analysis tended to be large and stable across studies, in education meta-analyses the effect sizes are "relatively small" and their variation is "great" within the same meta-analysis. As a result, Glass argues that because of unstable small effect sizes meta-analysis has not been as useful for informing policy in education as in medicine. Indeed, it will be shown that the problem of basing recommendations for clinical practice on small effect sizes is also a problem in other disciplines.

The bright spot in education research has been the establishment of comprehensive databases at all levels. The federal National Assessment of Educational Progress (NAEP) periodically tests national performance by subject area, grade levels, student and district demographics. NAEP results are reported for the nation and for individual states. The federal National Center of Educational Statistics collects longitudinal data on a variety of educational outcomes and processes that are broken out in great detail in terms of student and school characteristics. Federal funding has also strengthened the data collection and statistical reporting of state departments of education on the performance of its schools and districts. In other words, education is now a field rich in benchmarks of current and historical performance.

Given (a) the existence of a wide array of benchmarks, (b) the difficulty of truly establishing causation in field-based research in complex organizations, (c) the desire of leaders for replicated noticeable improvements in research, a different statistical criterion, *practical benefit*, is recommended that compares the unadjusted actual performance of an experimental group to an existing benchmark. Such an approach for assessing research findings and the design of new interventions will make it easier for leaders to better determine whether a given intervention is likely to improve their schools to a noticeable extent. Practical benefit also offers the potential to use research more efficiently to discover better interventions, more quickly, and at a lower cost—than from RCT research.

2. How the Use of Effect Size Misdirects Practice in Education

Effect sizes provide the potential to be more relevant to improving clinical practice than *p*-values as leaders do care about magnitudes of benefits. However, that does not necessarily mean that researchers care about the same level and types of magnitudes that leaders and other clinicians do. What is clear is that education has institutionalized the use of effect sizes within RCT research as the key criterion for identifying effective practices with the establishment of the federal *What Works Clearinghouse* (WWC).

2.1. Has Increasing the Scientific Rigor for Identifying Effective Practices in Education Produced Valid Results?

The goal of the WWC is to bring the most rigorous scientific criteria to bear to inform educators as to what works. Rigorous science has come to be defined as the use of RCT with an effect size of at least 0.25. WWCs methodologies have become the standard for how the US Department of Education determines which research proposals to fund. In addition, the WWC list of practices it has certified as having strong evidence to support their use has become more than just a good housekeeping seal of approval. WWC certification also increases a program's chances of obtaining federal funds to disseminate its use. In addition, there is advocacy that the federal government require that low-performing schools use their federal funds only to implement practices approved by the WWC. So, the stakes for being certified as effective by WWC are quite high.

However, such institutionalization of methodology, combined with political pressures for schools to adopt certified interventions, will be very damaging if the interventions that have been certified are NOT actually effective. The result would be analogous to what would happen if the FDA approved an ineffectual or harmful drug.

Unfortunately, there is growing evidence that relying on *RCT*s and effect sizes are producing misleading findings that

conclude that interventions are effective when in fact they are not. As previously discussed, it is highly questionable whether the static and expensive nature of RCT research can capture the constantly shifting interactive dynamics and chaos of complex organizations such as schools, or that causation can ever be established in such an environment. Indeed, Ginsburg and Smith (2016) examined the evidence for all of the 18 math programs certified by the WWC as having evidence of effectiveness based on a total of 27 approved RCT studies. Ginsburg and Smith (2016, p. 44) found 12 potential threats to the usefulness of these studies and concluded "…none of the RCT's provides useful information for consumers wishing to make informed judgments about what mathematics curriculum to purchase." The questionable utility of RCT methods for identifying clinical practices that are effective in the real world does not just seem to be just a problem in education. Section 4 of this paper shows how a critique of relying on RCT to identify effective practices in another complex organization—hospitals—appeared in the Journal of the American Medical Association (Berwick 2008).

However, regardless of whether RCT or less rigorous methodologies are used, relying on effect sizes as the statistical criterion for determining the practical significance of findings is turning out to be highly misleading—as problematic as relying on *p*-values.

2.2. Problems With Using Effect Size to Determine the Effectiveness of Interventions

The biggest problem is that the magnitude of the effect size needed to claim that a finding has practical significance has been set way too low by the research community. Setting the desired magnitude of effect size so low enables researchers to claim that they found practical significance when there is no actual difference of real-world importance. The minimum standard for the effect size needed to claim practical significance is generally set at 0.2 based on Cohen's (1988) categories of small (0.2), medium (0.5), and large (0.8). However, researchers ignore that Cohen referred to an effect size of 0.2 as "difficult to detect." No leader should consider adopting an intervention based on such a result—Leaders seek *noticeable* improvement. Even worse, education researchers such as Borman, Grigg, and Hanselman (2016) are now seeking to reduce that cutoff to 0.1, or half of "difficult to detect," and some (Deke, Wei, and Kautz 2017) are even advocating for effect size as low as 0.03. Perhaps the silliest rationalization for the importance of a small effect size in the education research literature is the argument that an effect size of 0.18 is important because it is twice as good as other interventions tested which only produced an effect size of 0.09. *The correct conclusion is that neither is likely to produce any benefit in terms of real-world school improvement*.

Even the federal WWC over-estimates the real-world importance of small effect sizes. It is minimum effect size cutoff of 0.25 is simply too small given that Cohen (1988) estimated that effects become noticeable only at an effect size of 0.5.

Small effect sizes also appear to be a reason why so much of influential published experimental research cannot be replicated in psychology (Open Science Collaboration 2015), and oncology (Begley and Ellis 2012), and why laboratory research results often do not predict actual outcomes in psychiatry (Kraemer 2016). The best summary of the problem of exaggerating the importance of small effect sizes was the observation by Ioannidis (2005), co-founder of the Meta-Research Innovation Center at the Stanford University, that the smaller the effect sizes in any scientific research the less likely it is that the research findings are true. We seem to have traded *p*-hacking for effect size-hacking. In both cases, the importance of a small difference is exaggerated, and such exaggeration misdirects practice in education and other disciplines.

The second problem is that even where large effect sizes are reported they can mask the fact that the experimental students did terribly, and/or the actual benefits were so tiny as to not be of any practical benefit. There are generally two ways that effect sizes overstate results. The first way is that relative differences can ignore the key, most relevant, actual context information. The process of ignoring actual context is best illustrated is by the following scenario of a couple living in the upper Midwest during a particularly bad cold spell deciding where to vacation in January where the goal is to find a place where they can relax on a warm beach and get a tan:

Wife: I cannot wait for our vacation in January. Let's go somewhere warm.
Husband: Definitely.
Wife: Where should we go?
Husband: I just read that Greenland is warmer than Antarctica in January.
Wife: That sounds great.
Husband: Even better, due to climate warming Greenland will be warmer this year than last. Plus, it has 27,394 miles of coastline, so it will be no problem finding beaches.
Wife: That's great. It will be wonderful to go somewhere where we can leave our winter clothes behind.

Their decision is certainly evidence-based—but it is clearly a lousy decision. Why was this couple's *evidence-based decision* so bad? It was bad because they relied *solely* on relative data. They needed a key piece of absolute data about the context, which in this case was the actual temperature in Greenland in January. The actual temperature is −8°C with zero hours of sunshine. This couple is far more likely to die from hypothermia in Greenland in January than to get a tan. With the right actual data, the couple would arrive at the correct decision to reject vacationing in both Antarctica and Greenland as it would be warmer to stay at home or to seek other options.

The above example illustrates why you cannot make intelligent decisions relying *solely* on relative data that relate external outcomes to each other, no matter how compelling that evidence appears to be. You always need some absolute data. The key piece of absolute data needed to judge the quality of an education program is the answer to the obvious question: *How did the students in the experimental intervention actually perform?* For a program to be judged successful, we would expect students to do reasonably well on an absolute basis. Of course, reasonable people can debate as to what an expectation of students "doing well" is. The problem is that researchers are not being asked to provide this most basic information in published research and are able to get away with just reporting

the statistical or practical significance of the relative result. Not having to report actual outcomes leads to the following question: In what percentage of research claiming to have demonstrated the effectiveness of interventions for helping children born into poverty did the experimental students actually do well—or actually do terribly? We simply do not know! If it is the latter, education is wasting billions of dollars and decades in applying evidence-based practices that inadvertently maintain gross inequities.

The second way that reports of larger effect sizes often can mislead and hide the fact that the experimental students did terribly is the use of questionable adjustments to the *means*. The most typical adjustments in education are to use covaried *means*. However, covarying *means* creates several problems. First of all, since we are talking about small differences, it is easy to cherry pick samples in such a way so as to make it appear that the initial *mean* of the experimental group is lower on some measure in order to boost the final *mean* of the experimental group using an analysis of covariance and thereby produce a positive effect size. A second type of adjustment, normalization of *means* occurs if different tests are used. The relative results are then based on normalized covaried *means*. However, such results are abstractions with no real-world meaning that can be understood by practitioners and policy makers.

So, researchers have come up with a way to kill two birds with one stone. They convert the abstract numbers into ones with real world meaning by creating hypothetical extrapolations that make small differences in the normalized covaried *means* appear important. In education, the most common extrapolations are to convert adjusted differences between groups into (a) extra days of learning, or (b) advantages for the experimental group on a nationally normed test. However, these extrapolations are usually grossly invalid and are designed to make the relative results seem important.

For example, if an intervention in a high poverty school produces an effect size of 0.2, the researchers will note that this result is equivalent to increasing scores of the experimental students on a nationally normed test from the 50th to the 58th percentile. Such an increase indeed seems impressive! Of course, such extrapolation ignores that these schools are way below national norms which is why the improvement effort was conducted in the first place. Extrapolating results from distributions of low performing schools onto national norms is deceptive and invalid. In addition, since the actual results were usually not from a nationally normed test—if the experimental students had actually taken the test the results were extrapolated to, they might have scored at the 28th percentile as opposed to the 58th.

Such manipulated extrapolation is akin to telling our hapless vacationers that the difference between the temperatures in Antarctica and Greenland in January is equivalent to raising the temperature in Miami in January from 76° to 85°. Such a hypothetical extrapolation of temperature makes Greenland seem warm! At the end of the day the reality is that the average temperature in Greenland in January is not 80°—it is still −8.

As a result of the above problems, the statistical criterion of effect size appears to be as misleading for identifying effective practices as relying on *p*-values. The following examples show how manipulated effect size results have misdirected practice in education on a large scale.

2.3. Two Examples of how Effect Size Results in Influential Research Misdirected Clinical Practice in Education on a Large Scale

Example 1. The effectiveness of *Success for All* (SFA)

SFA became the leading reform for improving the performance of failing high poverty inner city elementary schools. It is an intensive reading intervention for grades k-5. SFA built its reputation on the basis of research published in the top research journals apparently showing it to be uniquely effective in improving the performance of high poverty urban schools. Between 1990 and 2008, there were eight articles in the prestigious journals of the American Educational Research Association (AERA) by the co-developer and/or researchers who were, or had been, associated with him that documented the success of *Success for All* in a series of urban districts (Slavin 1990; Madden et al. 1993; Ross et al. 1995; Borman and Hewes 2002; Borman et al. 2005, 2007). These publications were the tip of the iceberg of research articles demonstrating the success of the program.

SFA research took full advantage of the awesome power of modern statistical analysis. The state-of-the-art analyses established the superior relative average performance of the experimental SFA students over students in comparison schools. However, SFA research, amidst all the numbers in all the published articles in the highly ranked journals *never* (to my knowledge) revealed how the SFA students actually performed on the national measures used in the studies. It turns out that SFA students were actually doing terribly.

The most famous experiment with SFA occurred in high-poverty schools in Baltimore in the early 90s. SFA research findings claimed success based on large effect sizes. However, an independent reanalysis of the data by Venezky (1998) showed the students doing poorly. Pogrow (1998, 1999) elaborated Venezky's findings and showed how after five years in SFA the experimental students entered the sixth-grade reading almost 3 years below grade level. This deficit is actually an over-estimate of the experimental students' performance as Venezky reported that (a) the sample was limited to only stable students (who tend to score higher) and did not include the students who were in the program for less than 5 years, and (b) the special needs students disappeared from the post-test experimental sample. The latter, combined with the analysis of covariance, boosted the relative performance of the experimental group. In addition, SFA was one of the most expensive interventions available and more money was spent on the experimental schools, and the students spent substantially more time reading. If these variables two variables had been covaried, the comparison schools would most likely have been ended up with a higher adjusted *mean*.

The SFA students in the Baltimore experiment in fact did terribly. No educator would consider the actual results to be a success or even an acceptable result. The poor performance of the experimental group cited above was confirmed by the independent study that the district's research office conducted (Ruffini 1992). As a result, the district dropped the program.

Subsequent evidence has shown that SFA underperforms other available options. Dade County School District in Florida made a large commitment to SFA. Urdegar (2000) showed that SFA schools in Dade County were doing poorly and that similar

schools in the district using their own homegrown interventions were doing better at a lower cost. In addition, the only experiment that compared SFA to other "effective" reading programs while controlling for the amount of time spent reading, found that SFA was doing so poorly that most of the schools that had been randomly assigned to implement SFA dropped it after the first year for poor performance and refused to continue using it in the second year of the experiment (Burdumy et al. 2009).

Pogrow (2000a, 2000b, 2002) reported consistent failure across the U.S., with disillusioned schools dropping the program. Yet even as such contrary evidence appeared, and schools dropped the program, the government continued to provide even larger grants to disseminate it—and the SFA researchers continued to cite Baltimore as a "success" in their successful funding proposal for a new $50 million dissemination grant almost two decades later.

In other words, there was a dichotomy between the published *effect size* findings in the top journals and what was actually happening in practice. Under any common-sense metric that practitioners who the research was supposed to "inform" would use, the program was clearly failing. None of the peer-review panels of the research journals or the government panels reviewing proposals thought to ask the one question amidst all the data that were of most interest to practitioners and the public at large; that is, *How did the SFA students actually perform?* Even worse, despite the demonstration of failure over several decades, the federal WWC continues to list SFA as one of the programs with the strongest evidence of success. WWC's support of SFA is the equivalent of the federal government recommending that school leaders should vacation in Greenland during winter recess to get a tan. More importantly, WWC's support means that schools seeking to accelerate the achievement of their low-income students will continue to be pressured to adopt an intervention that has a long track record of failure.

Example 2. The CREDO study of the effectiveness of charter schools

The widely cited CREDO study at the Stanford University compared the effectiveness of charter schools to traditional public schools. The study concluded that charter schools produce an extra 14 days a year of learning a year for Black students as compared to traditional public schools. Fourteen extra days of learning a year seems to be a substantial advantage for charter schools. However, the actual *effect size* was .02, or a tenth of "difficult" to detect." How did CREDO (2013) extrapolate .02 to 14 days? CREDO (2013, p. 13) noted that their findings "…are only an estimate and should be used as a general guide rather than as empirical transformations." CREDO is essentially admitting that there is no real empirical basis for their published hypothetical extrapolation. Maul and McClelland (2013) noted that CREDO's conversion of *effect size* into days of learning was "insufficiently justified" and that there really was not a substantial difference between the performances of the two types of schools. My own thought experiment concluded that the tiny effect size of .02 was at best equivalent to an advantage of 2 hours a year of extra learning. Two hours is a tiny difference that is dwarfed by the errors involved in the analyses. In other words, the correct conclusion should have been that *there was no practical difference* between the two types of schools. Alas, advocates for charter schools cited the CREDO conclusion which helped spur increased support for such schools nationally, while unfairly damaging the reputation of traditional public schools.

2.4. Maintaining the Hype of the Importance of Small Effect Sizes

The majority of education researchers, practitioners, or journalists do not understand how to interpret the magnitude of effect sizes. In 2013, MDRC released its evaluation of SFA's effectiveness on its latest federal dissemination grant it had received several decades after the initial Baltimore experiments described earlier. The study (Quint et al. 2013) concluded that the results were promising. The findings were based on effect sizes of −0.01 and 0.18 in two different reading skills at the kindergarten level. This conclusion was parroted in an article in the widely read *Education Week* with the headline: "School Improvement Model Shows Promise in First i3 Evaluation (Sparks 2013)." The reporter then goes on:

> One of the biggest early bets in the U.S. Department of Education's Investing in Innovation program seems to be paying off: Success for All, a literacy-related, whole-school improvement model, shows signs of changing teaching practice and boosting students' early reading skills after a year in schools…

Aside from the question as to why you would consider results for a 25-year-old program promising, the fact that the early grades are the ones where it is easiest to show large *effect sizes*, the reality is that these results are terrible. A meta-analysis of the effects of intensive reading interventions in the earliest grades by Scammacca et al. (2007) found effect sizes at the ranging from 0.34 to 0.56 across five different reading skills. (A less direct comparison of meta-analysis effects of reading interventions by Hattie (2009) found effect sizes ranging from 0.60 to 0.67.) The correct interpretation of the effect sizes reported by MDRC is that these results are among the worst ever recorded for an intensive reading intervention at the early grades.

So, once again effect size findings are being put forth in a misleading fashion and parroted by the media. No one in the research community or media is raising obvious questions about such findings. Why not?

A key argument to accept small effect size results as a finding of practical significance is that rigorous experiments conducted in schools rarely finds large effect sizes. Researchers therefore argue that it is not fair to expect research to produce larger effect sizes. However, such logic makes no sense from the perspective of the leaders for whom the findings are intended. Why should leaders be expected, or possibly even mandated by the federal government, to adopt an intervention that is not likely to produce noticeable benefits simply because that is the best that researchers can do? Such logic around the interpretation of statistical criteria preserves a system that operates for the benefits of researchers with no evidence that it actually produces benefits for leaders and their schools. Of course, it may be that it is impossible to produce noticeable improvements in the many problems that continue to persist in education. However,

an alternative explanation for the ineffectual research results is that the educational research community is using the wrong scientific model of research for developing and testing interventions. (An alternative scientific method that other disciplines have used to produce major clinical improvements to seemingly intractable problems will be explored in Section 4 of this paper.) At the very least, sufficient evidence has been presented to conclude that the use of effect sizes has been just as misleading as *p-values* in both education and other disciplines. (The use of odds-ratios appears to be equally problematic in exaggerating the importance of outcomes in education and medicine, but that is a discussion for another paper.)

2.5. Where Do We Go From Here?

The magnitude of benefits from a new intervention, or what Ziliak and McClosky (2004, 2008) call "oomph," is clearly important. Leaders seek easy-to-understand and intuitive evidence that an intervention has produced "oomph" elsewhere and that such outcomes have been replicated in a variety of settings—ideally ones with a similar context to their own. Most of all they want to know how students in the intervention actually performed. However, RCT research and *effect size* reporting generally do not provide information on the actual (as opposed to relative) performance of the experimental students.

It is therefore critical to develop an alternative evidentiary criterion that (a) is more intuitive, (b) provides leaders with the type of information they seek, and (c) leaders can apply to their unique context. Such methods need to be less mathematically complex, and better able to distinguish which interventions are truly likely to be effective. As a result, I developed the alternative evidentiary criteria for informing leadership decision-making of practical benefit. These criteria are designed to provide leaders with an indication of whether adopting the findings and recommendations of specific studies are likely to benefit their particular schools in a clearly noticeable way.

3. Switching to Practical Benefit

The first time I presented the concept of practical benefit to a few statisticians over lunch they simply stared at me for several minutes without saying anything. When I broke the uncomfortable silence and asked them why they were not reacting, their response was that the idea was contrary to everything they had been taught, everything they believed in, and everything they taught their students. Anyway, here is that idea.

3.1. Determining the Practical Benefit of a Research Finding for Improving One's Schools

Practical benefit recognizes that the data that leaders most want to know is (a) how the experimental students actually did, and (b) whether the experimental students' actual results are noticeably better than *their students*' existing results. Knowing how an external control group did is of no interest to them—nor should it be. The relevant comparison is how the external experimental group (only) actually performed relative to their own students. The experimental students' performance is then compared to an existing benchmark of current performance in their own organization and possibly statewide or nationally.

In other words, when considering the likelihood that a published research finding will produce a noticeable benefit in one's own schools, practical benefit ignores how the external control group performed. That means that leaders can ignore all the statistical analyses trying to determine relative performance, and just focus on the unadjusted mean/median of the experimental group, and how it compares to their schools' performance. The only other relevant information in any study is the context in which the research was conducted. Every other stat or technical term in the research can be ignored. Using this method of focusing only on actual performance of the experimental group educators can peruse a stack of the most sophisticated quantitative research articles and determine in which, if any, the performance of the experimental group exceeds their existing benchmarks. Leaders can then focus on considering the interventions in those articles that report high actual performance.

A major argument against ignoring the control group results is that they are important for understanding the context of the research. That is probably true in highly controlled laboratory research. That is not true for research in complex organizations. The few variables that the study makes an effort to control are a minority of the many other interactive variables that exist. Therefore, while context is critical, a much better way to understand the context of the research is via a mixed-methods approach where there is a careful qualitative description. Alas, in most published quantitative research that I see there is little more than cursory discussion of context. (The consequences of ignoring context will be explored in the next section.)

But how does one measure "noticeable benefit" or "oomph?" Ziliak and McClosky (2004, p. 531) define oomph as "A big change, important for the science." I will paraphrase that and define practical benefit as a big change, important for the profession and/or individual leader. "Big" can be determined by human judgment. At the very least "oomph" is a clearly noticeable improvement or other benefit that does not require precise statistical criteria to discern.

While relying on human judgment to determine the effectiveness of an intervention is obviously imprecise and subject to its own biases, Ziliak and McClosky argue that creating the potential for big benefits is more important than being able to precisely predict the level of benefit. When a big improvement occurs, people can generally recognize it.

In addition, relying on human judgment to estimate whether the level of performance of an experimental group in a published research article is likely to produce noticeable benefits in one's schools is consistent with management theory which uses a goal setting process. Goal setting is also usually based on human judgment and aspiration.

Determining how likely the performance of the experimental group in a published study is likely to replicate in one's own schools also requires considering the context of the study. Examples of key context variables include (a) the types of students in the sample, (b) whether the duration of the intervention is a few days or an entire school year, and (c) whether the outcome measure is standardized. If there is a way to match up the results

for the sample or subsample in the studies with the characteristics of one's schools that increases the likelihood that the results will replicate. The final decision is whether resources and expertise needed to implement the intervention are available to the leader.

Is it possible to produce "oomph?" Yes! What does it look like? Here are two examples.

- The Carnegie Foundation established an initiative to try and improve the dismal record wherein half the students who enter community college fail to pass developmental math and therefore never earn any community college credits before giving up. This level of failure may be the biggest dropout rate in American education. Carnegie shifted the developmental sequence away from high school Algebra to statistics with its Statway program. The result was an increase in the percentage of students who passed developmental math and subsequently earned community college math credit from 15% after 2 years, to 50% after only one year.
- Studies of federal efforts to support students born into poverty (Title I students) consistently finds that students make gains in grades k-3, and then fall back after that. My own Higher Order Thinking Skills (HOTS) project provided general thinking development help in grades 4–8 instead of remediation. This program was adopted in approximately 2600 schools around the US and served 1/2 million students. Results consistently showed Title I students making three times the growth in reading comprehension and twice in math, and close to 15% of these low-performing students made honor roll in the first year of the program.[1]

In both cases the benefits relative to benchmarks are obviously substantial, and do not require calculations of p-values or effect sizes to justify their practical benefit.

In summation, determining the practical benefit of the findings of a published research study involves the following steps:

(a) Extract the average actual unadjusted performance of only the experimental group from the mass of data, and ignore all external relative results, adjusted results, and extrapolated results,
(b) Check the context of the research to determine how relevant the study is to your schools, and which findings are the most relevant—though leaders may choose to implement the study's approach even if their context is different if the benefits are substantial, much as my development work applied techniques from private schools to inner city schools, and
(c) Use human judgment to decide whether the actual results of the experimental group (and of the key subgroups) are sufficiently better than how your students are currently doing to warrant the time and money to invest in adopting the intervention.

3.2. Advantages of the Criterion of Practical Benefit

By focusing ONLY on how the experimental group did on an unadjusted basis and using human judgment to determine whether the experimental students did or did not do sufficiently better than your students, you bypass virtually all the complex statistics. In addition, the actual performance of the experimental group is the measure that leaders are most interested in knowing and most care about. There is no reason to use p-values or calculate effect sizes or any other external measure of relative differences such as odds-ratios. As a result, the use of practical benefit bypasses all the problems described above that have misdirected practice. There is reason to believe that the simple and intuitive measure of practical benefit will yield better decisions for improving clinical practice than results produced by traditional inferential analysis—for mission critical problems in complex organizations where benchmarks exist.

Practical benefit also makes the most complex quantitative research in the top journals accessible to practitioners. Practitioners can critically examine such research and reach their own conclusions as to its practical benefit by simply looking at the unadjusted *means* and standard deviations of the experimental group, and then possibly also medians—and then compare the results and context to the benchmark of how their schools are performing.

3.3. Limits on the Applicability of Practical Benefit

3.3.1. Limit 1—What if Research does not Report Unadjusted Means/Medians?

While I am not aware of journal policies in other fields, AERA does not require authors to report unadjusted results in its journals. To put it simply, you cannot rely on adjusted means/medians for decision-making purposes for the reasons already discussed. Therefore, you cannot really determine the practical benefit of the findings. The best you can do is look at the magnitude of effect sizes and seek studies that produce large ones and view that as an indication that the findings are worth looking into. How big should the effect size be?

Given that Hattie's (2009) meta-analysis of all meta-analyses in education found an overall effect size of 0.4, and Cohen (1988) concluded that effects become noticeable at 0.5, expecting an effect size of at least 0.45 seems to be a reasonable starting point for deciding that a research finding merits consideration. Expecting a minimum effect size of 0.45 is particularly true for research after the earliest grades and research conducted in schools over a period of at least several months with a widely used measure.[2]

For nonexperimental research, the recommended minimums for correlation coefficients to be considered as potentially important for leadership decision-making are $r > 0.39$, or $r < -0.39$. An r of 0.39 means that the variation in one variable is associated with 15% of the change in the other. A good argument can be made that this admittedly arbitrary minimum should be set higher. However, the ability to increase the predictability

[1] The HOTS program ran for 24 years and was discontinued 5 years ago This section discusses what was learnt as a researcher from the iterative development, evaluation, and large-scale dissemination of this program.

[2] Effect sizes tend to be higher for short-term laboratory research, research conducted at the earliest grades, and research with nonstandard measuring tools—particularly for measures developed by the researcher.

of another outcome by 15% is probably sufficient incentive for most leaders to consider that approach. For regression the suggested cutoff is whether the unique contribution of an actionable variable in a regression model increases R^2 by 15%.

There are three main problems with such cutoffs.

- These criteria will not prevent *p*- or effect size-hacking, however, the larger effect size minimum values might possibly make it more difficult to hide such behavior,
- Leaders may often NOT be able to find research in a particular area of interest that meet these criteria. Of course, knowing that research cannot help in a given situation is better than pursuing a course of action thinking that it has strong evidentiary support when it does not. Indeed, there is a long history of national fads in education that were supposedly evidence-based, and
- The biggest problem is that there are times when a small effect size can be important, depending on the context of the research. Sometimes a small effect size can mask a high potential clinical approach.

An example of the latter is the extensive research literature that consistently found *negative* effect sizes for the approach used in the previously mentioned HOTS program. *Fortunately, I was not aware of this research.* My ignorance enabled me to produce a major success in developing the thinking skills of students born into poverty that translated into major gains in reading and math (Pogrow 2005) by using a method that research conclusively showed did not work. How is this possible?

One of the most agreed-upon findings in the psychology research literature is that if you want to increase students' problem-solving ability in a content area such as math, you should provide more and better practice in solving thoughtful math problems—a *thinking in content* approach. A thinking in content approach is universally considered superior to a *general thinking development* approach. The latter tries to develop students' general thinking skills, such as metacognition, that might simultaneously improve problem solving in a variety of content areas. As a result, thinking in content is the basis of all reform movements in education that seek to increase the problem-solving ability of disadvantaged students in math, social studies, science, and reading comprehension. Of course, my "ignorant" approach was to improve the academic performance of children born into poverty in grades 4–8 solely by developing their general thinking over a year or two without providing any additional help in content specific skills.

However, when I shared the major gains we were seeing with my psychology friends they invariably told me that the approach could not possibly be working given that research consistently found negative effects from general thinking interventions. As a result, I decided to look into the apparent conflict between the large-scale results we were seeing and what research had universally concluded.

The extensive body of research on the effectiveness of thinking in content methods was well done and had indeed consistently found that general thinking did not work (Willingham 2007). However, the clear disconnect with the published research findings and the success we were producing resulted from the fact that the contexts in the published research had nothing to do with the context I was working in. The initial experiments were carried out with graduate students in math and physics. Clearly, students would not be at such an advanced level if they did not already possess a high level of general thinking ability. It is also clear that further skill development at such a high level of learning requires that students learn to solve ever more advanced math and physics problems—that is, thinking in content. What does the methods used with graduate students have to do with a fifth grader in Harlem reading 3 years below grade level? Nothing! The more recent research supporting the thinking in content approach with students in grades 4–8 was conducted in largely high-income schools, and the experiments ran from 1 to 3 days. Clearly, if you want students to learn how to solve a specific problem in 1 and 2 days they will clearly do better when they are taught the specific method for solving that problem as opposed to providing general thinking activities. As a result, no one had ever tried to provide intensive, daily, general thinking activities to low-income students—which in my ignorance I did successfully.

The point of my experience with the HOTS program is that depending on the context of the research, a small or negative effect size can mask a high potential intervention in a different context. This example also illustrates that it is always important to thoughtfully examine the context of research instead of just relying on some outcome number or numerical cutoff score.

3.3.2. Limit 2—Macro-Level Decision making at the Federal Level

The problems of establishing appropriate statistical criteria to measure effectiveness get more complicated when government seeks to impose improvement goals on all schools within an accountability framework. How much improvement should the government require? How should it be measured? The various efforts to date to set standards for improvement have generally not been very successful.[3]

A better approach is for the federal education department to support a better scientific method for developing interventions that produce replicable "oomph." Several successful applications, Statway and HOTS, were highlighted which met the criterion of practical benefit at scale and "oomph."

4. Practical Benefit and Improvement Science

While education was trying to do its best to emulate medical research as a sign of its growing professionalization, unbeknownst to it there were elements of medical practice that did not use RCT research. While gold-standard experiments are the universal method for testing drugs, there are branches of healthcare and medical practice that do not rely on experiments—

[3] The first time the federal government tried to set improvement standards it required all states to set improvement targets. Most states set the targets embarrassingly low, and this provision was dropped. The No Child Left Behind law passed in 2002 used a complex algorithm to establish annual improvement goals for each school in the nation to bring all students up to grade level in reading and math within 10 years. There were strong accountability penalties for schools that did not meet their annual progress goal. However, this improvement goal was too ambitious and made a mess. In 2015 Congress reversed course and eliminated federally mandated progress goals.

particularly for solving time-critical problems in complex organizations such as hospitals. A notable example of applying alternative research methods were successful efforts to improve the quality of health care in hospitals. The inadequacy of RCT research for improving hospital care is highlighted in the following quote by Berwick (2008) in the *Journal of the American Medical Association*:

> Changes in the current approach to evidence in health care would help accelerate the improvement of systems of practice…Educators and medical journals will have to recognize that, by itself, the usual… experimental paradigm is not up to this task. It is possible to rely on other methods without sacrificing rigor. Many assessment techniques developed in engineering and used in quality improvement…have more power to inform about mechanisms and contexts than do RCTs, as do ethnography, anthropology, and other qualitative methods. For these specific applications, these methods are not compromises in learning how to improve; they are superior.

Berwick's methods came to be known as improvement science and is based on the rapid prototyping of approaches to improve clinical outcomes, and the ones that demonstrate initial practical benefit are iterated to try and adapt to a wide variety of contexts. The goal is to produce clearly noticeable improvements against existing benchmarks (i.e., practical benefit) and collaboratively share the results of iterations in different contexts.

The methods of improvement science are on the surface a somewhat haphazard and imprecise method to design innovation. However, improvement science can produce dramatic, large-scale, clinical improvements. Gawande (2007) notes that the medical field of obstetrics has the best record of increasing the number of lives saved and that such improvement was produced without conducting formal experiments.

Elements of improvement science are now even being extended to the testing of new medicines. Kolata (2015) described how *precision medicine*, the newest federal initiative seeking to discover new and more powerful new cancer drugs, is moving away from conducting experiments in favor of seeking large benefits from the rapid testing of many different compounds for specific cancers. Instead of comparing the results to a formal control group, the results are compared to the known benchmark of the response rates of patients to currently available treatments. Kolata (2015, p. 2 of download) notes that "unlike previous efforts that looked for small differences between a new treatment and an older one…researchers are gambling on finding huge effects." Scientists are finding a patient response rate to the new drugs of 50–60% as compared to existing treatments that give a response rate of only 10–20%.

These alternative methods of scientific discovery that seek large improvements over existing benchmarks are essentially using practical benefit for demonstrating evidence of effectiveness. Improvement science networks are now springing up in medicine and other fields to try and solve heretofore-intractable problems using the criterion of practical benefit.

Improvement science is equally applicable to solving problems in education. The Carnegie Foundation has been taking the lead in working with school districts on how to use improvement science methods, and in creating networks and conferences to share results. Indeed, the dramatic results of the Statway and HOTS programs described earlier could not have been developed through rigorous experiments. There are simply too many variables and parameters to manipulate. The only way to find some near optimal combination of implementation parameters is to embrace the reality that there is tremendous variation in school contexts and learn from the beginning how early prototypes are working or not working in various contexts—and then make needed adjustments on the fly. Such adjustments cannot be made without researchers becoming intimately knowledgeable about the many facets of the different contexts that the intervention needs to be able to adapt to. As previously noted, such a focus on context has been noticeably lacking in more gold-standard forms of research, and such absence has led to misunderstanding of the applicability of research findings.

The methods of improvement science are in contrast to RCT research. RCT methodology pretends that it can control just a few of the interactive factors and that the findings will therefore generalize because of some elements of randomness and a significant p-value or nondetectable effect size. Such research has little actual understanding of the variations of contexts that exist and their potential impact on external validity. Nor will federal efforts to scale interventions that meet WWC standards to other contexts work because the standards are inadequate for reasons discussed earlier and the initial results probably had no "oomph."

Improvement science, on the other hand, relies on quick, flexible adaptation using practical benefit to determine which iterations of the intervention are working and which are not. If it is not initially effective it is critical to quickly figure out why not and make needed adjustments. Such flexibility and adaptability are critical because developing an intervention that provides practical benefit at scale requires developing a set of very precise parameters for all the key design and implementation factors. Improvement science eschews seeking causation to produce replicable, consistent improvement across contexts.[4]

Of course, it is not easy to produce such outcomes. Kerwin and Thornton (2018) describe how producing an effective intervention requires determining the near optimal mix of design parameters, and how minor changes in any parameter can substantially reduce effectiveness. They show how an early literacy intervention in northern Uganda found major gains in reading and writing. However, when relatively small programmatic changes were made the intervention was no longer effective. Kerwin and Thornton (2018, p.33) further note that:

> Evidence on the sensitivity of program results to implementation details is scarce…
> Thus even successful educational interventions implemented as a pilot may be completely uninformative to the results of a scaled-up version of the program: it is hard to know whether a seemingly small change can cause a large difference in a program's impacts, and there are innumerable such changes that can and will occur…Programs designed to

[4] Producing consistent improvements across contexts minimizes Ziliak and McClosky's notion of loss, that is, reduces the probability that adopting the intervention will do damage, as part of maximizing "oomph."

exploit complementarities—rather than to isolate the effects of individual inputs—are likely to be more effective at improving learning.

This is one of the few studies that captures how precise the values of the many different parameters need to be set if the intervention is to be effective at scale which is consistent with my development experience in developing the HOTS intervention. For example, we quickly discovered through informal trials wherein we varied the intensity of the program that providing the service less than 4 days a week produced little benefit.

There is now sufficient experience with successful application of the principles of practical benefit and improvement science producing major clinical improvements in practice to ask the following empirical question: Are these newer scientific methods more likely to develop and validate interventions with major clinical benefits at scale for complex organizations such as schools and hospitals as compared to the current RCT methods? The answer appears to be "yes," and this has major implications for changing how applied research is conducted, taught, and funded in education—and clinical research in other fields.

5. Conclusions and Recommendations

The prevailing rigorous methodology and statistical criteria used by research journals and government agencies in education research has routinely overstated the actual effectiveness of interventions. Relying on effect sizes to determine the practical significance of interventions in education has been as problematic as using *p*-values to determine statistical significance. Examples have been provided wherein effect size results have misdirected practice in education on a large scale, and evidence has been provided that the same thing has happened in other disciplines where influential published research has not been replicable and/or has misdirected clinical practice partly because of relying on small effect sizes.

In addition, the problem may not just be the statistical criteria that are used, but also the dominant scientific paradigm of relying on RCT research to identify effective interventions that can scale across complex organizations. There is no evidence to date that schools that adopt interventions recommended by the What Works Clearinghouse (WWC) actually improve, and a comprehensive study of its RCT based recommendations on its certified math programs found that the results did not provide useful information to practitioners.

As a result, it is recommended that education switch to a simpler measure of practical benefit that (a) measures the likely benefit of an intervention in simpler, actual outcomes for the experimental group (only), and (b) presents results in the types of measures that leaders can easily understand and that they value as key to their own improvement efforts. Practical benefit also has potential for improving the validity of research to improve clinical practice in other disciplines. It also appears that the alternative scientific approach of improvement science, that relies on looking for patterns of practical benefit in fast prototyping of iterative trials across contexts, has generated major improvements in previously intractable problems of practice in a variety of disciplines, including medical practice and education. Improvement science provides the potential to be a more appropriate methodology for developing interventions that can lead to large-scale improvements in schools than current reliance on gold-standard methods. Several examples of large-scale improvement in education using improvement science and practical benefit were provided. The key question is whether improvement science will be relegated to something that individual practitioners choose to use or whether they will be incorporated as a mainstream methodology for research and development.

Shifting applied education research to incorporate practical benefit and improvement science requires a variety of changes in the nature of knowledge production and editorial policy at the major research journals, as well as in the teaching of quantitative methods to researchers and leaders.

5.1. Recommended Changes in Knowledge Production

5.1.1. American Statistical Association

It may be important to create a dual track of recommendations for reforming inferential statistics—one track for basic research and another for applied research geared toward providing guidance for organizational improvement. Recommendations for reforming applied research should incorporate the principles of practical benefit across the disciplines.

5.1.2. Education

- Research journals need to require that applied research reveal the unadjusted means/medians of results for all groups and subgroups, and require review panels to take such data into account
- Research journals and the US Department of Education need to establish alternative standards for publishing studies and funding the research and dissemination of projects based on the demonstration of consistent practical benefit at scale, and
- Funding needs to be provided to establish the kinds of networks for sharing the results from iterative trials in different contexts similar to what exists in obstetrics.
- The WWC should stop certifying programs as effective, and federal legislation should stop mandating the use of evidence-based programs, until new criteria for identifying effective programs are developed that incorporate practical benefit.

Appendix

A1. Overview of the Project star Experiment

Project STAR was undertaken in 75–79 schools in Tennessee from 1985 to 1989 at a cost of approximately $12 million. Each year in the experimental period, 6000–7000 students in grades K-3 participated in the experiment, for a total of 12,000 students during the entire period.

As students initially entered STAR schools, they were randomly assigned to small classes with 13–17 students, regular classes with 22–25 students without teacher aides, and regular classes with 22–25 students with teacher aides.

Sohn (2010, 2015) identified the following problems with the research:

- Schools volunteered to participate, thus were not randomly selected,

- Kindergarten was not mandatory so the children who enrolled had unique characteristics,
- Students constantly left and entered STAR schools, so (a) the overall attrition rate was almost 50%, and class sizes became distorted from the original design with some small classes ending up with more students than the other categories,
- In the absence of pretest scores, it is not clear whether randomization controlled for initial ability, and randomization did not produce equal levels of low-income students (as measured by free and reduced lunch eligibility) across the treatments,
- Students switched from one type of class to another,
- While teachers were randomly assigned, some received training which may have affected outcomes.

References

Begley, C. G., and Ellis, L. M. (2012), "Drug Development: Raise Standards for Preclinical Cancer Research," *Nature*, 483(7391), 531–533. [225]

Berwick, D.M. (2008), "The Science of Improvement," *JAMA*, 299(10), 1182–1184. [225,231]

Borman, G. D., and Hewes, G. M. (2002), "The Long-term Effects and Cost-Effectiveness of Success for All," *Educational Evaluation and Policy Analysis*, 24, 243–266. [226]

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2005), "The National Randomized Field Trial of Success for All: Second-year Outcomes," *American Educational Research Journal*, 42, 673–696. 226

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2007), "Final Reading Outcomes of the National Randomized Field Trial of Success for All," *American Educational Research Journal*, 44, 701–731. 226

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2005), "Success for All: First-year Results From the National Randomized Field Trial," *Educational Evaluation and Policy Analysis*, 27, 1–22.

Borman, G. D., Grigg, J., and Hanselman, P. (2016), "An Effort to Close Achievement Gaps at Scale Through Self-affirmation," *Educational Evaluation and Policy Analysis*, 38, 21–42. [225]

Burdumy, J. S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., and Douglas, A. (2009, June 8), "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students," Mathematica Inc., Presentation at *The Institute of Education Sciences* research conference, available at *http://www.mathematica-mpr.com/~/media/publications/pdfs/education/ies_readcomp_james-burdumy0609.pdf*. [227]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum. [225]

CREDO (2013), National Charter School Study, Center for Research on Education Outcomes, Stanford, CA: Stanford University, available at *http://credo.stanford.edu/documents/NCSS%202013%20Final%20Draft.pdf*. [227]

Deke, J., Wei, T, and Kautz, T. (2017), *Asymdystopia: The Threat of Small Biases in Evaluation of Education Interventions That Need to be Powered to Detect Small Impacts*, Washington, DC: Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance. [225]

Gawande, A. (2007), *Better: A Surgeon's Notes on Performance*, New York: Metropolitan Books. [231]

Ginsburg, A., and Smith, M. S. (2016), *Do Randomized Control Trials Meet the "Gold Standard"? A Study of the Usefulness of RCTs in the What Works Clearinghouse*, Washington, DC: American Enterprise Institute. [225]

Glass, G. V. (2016), "One Hundred Years of Research: Prudent Aspirations," *Educational Researcher*, 45, 69–72. [224]

Hattie, J. A. C. (2009), *Visible Learning: A Synthesis of 800+ Meta-analyses on Achievement*, Abingdon, UK: Routledge. [227,229]

Hojat, M. and Xu, G. (2004), "A Visitor's Guide to Effect Sizes – Statistical Significance Versus Practical (clinical) Importance of Research Findings," *Advances in Health Sciences Education Theory and Practice* 9, 241–249. [223]

Ioannidis, J. P. (2005), "Why Most Published Research Findings are False," *PLoS Med*, 2, available at *http://journals.plos.org/plosmedicine/article/authors?id=10.1371%2Fjournal.pmed.0020124*. [225]

Kerwin, J. and Thornton, R. (2018), "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures," Paper presented at RISE Annual Conference, Oxford, UK, June 21–22, available at *https://www.riseprogramme.org/sites/www.riseprogramme.org/files/inline-files/Thornton.pdf*. [231]

Kirk, R. E. (1996), "Practical Significance: A Concept Whose Time has Come," *Educational and psychological measurement*, 56, 746–759. [223]

Kraemer, H. C. (2016), "Messages for Clinicians: Moderators and Mediators of Treatment Ooutcome in Randomized Clinical Trials," *American Journal of Psychiatry*, 173(7), 672–679. [225]

Kolata, G. (2015), "A Faster Way to try Many Drugs on Many Cancers," *New York Times*, available at *http://www.nytimes.com/2015/02/26/health/fast-track-attacks-on-cancer-accelerate-hopes.html?_r=0*. [231]

Lipsey, M. W., Puzio, K, Yun, C, Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., and Busick, M. D. (2012), *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms*, U.S. Department Of Education. Institute of Education Sciences, available at *https://ies.ed.gov/ncser/pubs/20133000/*.

Madden, N. A., Slavin, R. E., Karweit, N. L., Dolan, L. J., and Wasik, B. A. (1993), "Success for All: Longitudinal Effects of a Restructuring Program for Inner-City Elementary Schools," *American Educational Research Journal*, 30, 123–148. [226]

Maul, A., and McClelland, A. (2013), *Review of the National Charter School Study*, National Education Policy Center. Boulder, CO: University of Colorado. [227]

McCartney, K., and Rosenthal, R. (2000), "Effect Size, Practical Importance, and Social Policy for Children," *Child Development*, 71, 173–180. [223]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science.," *Science*, 349(6251), aac4716-1–aac4716-8. [225]

Plsek, P. E. (1999), "Quality Improvement Methods in Clinical Medicine," *Pediatrics*, 203 –214.

Pogrow, S. (1998), "What is an Exemplary Program and why Should Anyone Care? A Reaction to Slavin and Klein," *Educational Researcher*, 27, 22–29. [226]

——— (1999), "Rejoinder: Consistent Large Gains and High Levels of Achievement are the Best Measures of Program Quality: Author Responds to Slavin," *Educational Researcher*, 28, 24–26, 31. [226]

——— (2000a), "The Unsubstantiated 'Success' of *Success for All*. Implications for Policy, Practice, and the Soul of the Profession," *Phi Delta Kappan*, 81, 596–600. [227]

——— (2000b), "*Success for All* Does not Produce Success for Students," *Phi Delta Kappan*, 82, 67–80. [227]

——— (2002), "*Success for All* is a Failure," *Phi Delta Kappan*, 83, 463–468. [227]

——— (2005), "HOTS Revisited: A Thinking Development Approach to Reducing the Learning Gap After Grade 3," *Phi Delta Kappan*, 64–75. [230]

Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J., and Zhu, P. 2013), *The Success for All Model of School Reform: Early Findings from the Investing in Innovation (i3) Scale-Up. MDRC, available at https://www.mdrc.org/sites/default/files/The_Success_for_All_Model_FR_0.pdf* [227]

Ross, S. M., Smith, L. J., Casey, J., and Slavin, R. E. (1995), Increasing the Academic Success of Disadvantaged Children: An Examination of Alternative Early Intervention Programs, *American Educational Research Journal*, 32, 773–800. [226]

Ruffini, S. et al. (1992), *Assessment of Success for All* [Unpublished research study] Baltimore, MD: Baltimore City Public Schools. [226]

Scammacca, N., Vaughn, S., Roberts, G., Wanzek, J., and Torgesen, J. K. (2007), *Extensive Reading Interventions in Grades k– 3: From Research to Practice*, Portsmouth, NH: RMC Research Corporation, Center on Instruction. [227]

Sohn, K. (2010), "A Skeptic's Guide to Project STAR," *KEDI Journal of Educational Policy*, 7, 257–272. [232]

——— (2015), "Nonrobustness of the Carryover Effects of Small Classes in Project STAR. *Teachers College Record*, 117, 1–26. [232]

Slavin, R. E., Madden, N. A., Karweit, N. L., Livermon, B. J., and Dolan, L. (1990), "Success for All: First-year Outcomes of a Comprehensive

Plan for Reforming Urban Education," *American Educational Research Journal*, 27, 255–278. [226]

Sparks, S. D. (2013, October 30), "School Improvement Model Shows Promise in First i3 Evaluation," *Education Week* (online), available at *http://www.edweek.org/ew/articles/2013/10/30/11successforall.h33.html?qs=sparks+AND+%22Success+for+All%22*. [227]

Sullivan, G. M., and Feinn, R. (2012), "Using Effect Size—or why the P value is not Enough," *Journal of graduate medical education*, 4, 279–282. [223]

Urdegar, S. (2000), *Evaluation of the Success for All Program: 1998-1999,* [Unpublished study]. Dade County: MD, Office of Evaluation and Research, Miami-Dade County Public Schools. [226]

Venezky, R. L. (1998), "An Alternative Perspective on Success for All," in *Advances in Educational Policy*, ed. K. Wong, Vol. 4, Greenwich, CT: JAI Press, pp. 145–165. [226]

Willingham, D. T. (2007). "Critical Thinking: Why is it so Hard to Teach," *American Educator*, Summer, 31, 8–19. [230]

Ziliak, S. T., and McClosky, D. N. (2004), "Size Matters: The Standard Error of Regressions in the American Economic Review," *The Journal of Socio-Economics*, 33, 527–546. [228]

——— (2008), *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [228]

The American Statistician

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

Abandon Statistical Significance

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett

To cite this article: Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett (2019) Abandon Statistical Significance, The American Statistician, 73:sup1, 235-245, DOI: 10.1080/00031305.2018.1527253

To link to this article: https://doi.org/10.1080/00031305.2018.1527253

© 2019 The Authors. Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 37656

View related articles

View Crossmark data

Citing articles: 213 View citing articles

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, 235–245: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1527253

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Abandon Statistical Significance

Blakeley B. McShane[a], David Gal[b], Andrew Gelman[c], Christian Robert[d], and Jennifer L. Tackett[e]

[a]Department of Marketing, Kellogg School of Management, Northwestern University, Evanston, IL; [b]Department of Managerial Studies, College of Business Administration, University of Illinois at Chicago, Chicago, IL; [c]Department of Statistics and Department of Political Science, Columbia University, New York, NY; [d]Centre de Recherche en Mathématiques de la Décision (CEREMADE), Université Paris-Dauphine, Paris, France; [e]Department of Psychology, Northwestern University, Evanston, IL

ABSTRACT

We discuss problems the null hypothesis significance testing (NHST) paradigm poses for replication and more broadly in the biomedical and social sciences as well as how these problems remain unresolved by proposals involving modified *p*-value thresholds, confidence intervals, and Bayes factors. We then discuss our own proposal, which is to abandon statistical significance. We recommend dropping the NHST paradigm—and the *p*-value thresholds intrinsic to it—as the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences. Specifically, we propose that the *p*-value be demoted from its threshold screening role and instead, treated continuously, be considered along with currently subordinate factors (e.g., related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain) as just one among many pieces of evidence. We have no desire to "ban" *p*-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the currently subordinate factors. We also argue that it seldom makes sense to calibrate evidence as a function of *p*-values or other purely statistical measures. We offer recommendations for how our proposal can be implemented in the scientific publication process as well as in statistical decision making more broadly.

ARTICLE HISTORY
Received October 2017
Revised September 2018

KEYWORDS
Null hypothesis significance testing; *p*-Value; Replication; Sociology of science; Statistical significance

1. The Status Quo and Two Alternatives

The biomedical and social sciences are facing a widespread crisis with published findings failing to replicate at an alarming rate. Often, such failures to replicate are associated with claims of huge effects from subtle, sometimes even preposterous, interventions. Further, the primary evidence adduced for these claims is one or more comparisons that are anointed "statistically significant"—typically defined as comparisons with *p*-values less than the conventional 0.05 threshold relative to the sharp point null hypothesis of zero effect and zero systematic error.

Indeed, the *status quo* is that $p < 0.05$ is deemed as strong evidence in favor of a scientific theory and is required not only for a result to be published but even for it to be taken seriously. Specifically, statistical significance serves as a lexicographic decision rule whereby any result is first required to have a *p*-value that attains the 0.05 threshold and only then is consideration—often scant—given to such factors as 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 (for want of a better term, we hereafter refer to these collectively as the *currently subordinate factors*).

Traditionally, the $p < 0.05$ rule has been considered a safeguard against noise-chasing and thus a guarantor of replicability. However, in recent years, a series of well-publicized examples (e.g., Carney, Cuddy, and Yap 2010; Bem 2011) coupled with theoretical work has made it clear that statistical significance can easily be obtained from pure noise. Consequently, low replication rates are to be expected given existing scientific practices (Ioannidis 2005; Smaldino and McElreath 2016), and calls for reform, which are not new (see, e.g., Meehl 1978), have become insistent.

One proposal, suggested by Benjamin and 71 coauthors including distinguished scholars from a wide variety of fields, is to *redefine* statistical significance, "to change the default *p*-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005" (Benjamin et al. 2018). While, as they note, "changing the *p*-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance," we believe this "quick fix," this "dam to contain the flood" in the words of a prominent member of the 72 (Resnick 2017), is insufficient to overcome current difficulties with replication. Instead, we believe it opportune to proceed immediately with other measures, perhaps more radical and more difficult but also more principled and more permanent.

CONTACT Blakeley B. McShane b-mcshane@kellogg.northwestern.edu Marketing Department, Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208.

© 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.

In particular, we propose to *abandon* statistical significance, to drop the null hypothesis significance testing (NHST) paradigm—and the p-value thresholds intrinsic to it—as the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences. Specifically, rather than allowing statistical significance as determined by $p < 0.05$ (or some other threshold whether based on p-values, confidence intervals, Bayes factors, or some other purely statistical measure) to serve as a lexicographic decision rule in scientific publication and statistical decision making more broadly, we propose that the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with the currently subordinate factors as just one among many pieces of evidence.

To be clear, we have no desire to "ban" p-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the currently subordinate factors. We also argue that it seldom makes sense to calibrate evidence as a function of p-values or other purely statistical measures.

In the remainder of this article, we discuss general problems with NHST that motivate our proposal to abandon statistical significance and that remain unresolved by the Benjamin et al. (2018) proposal. We then discuss problems specific to the Benjamin et al. (2018) proposal. We next offer recommendations for how, in practice, the p-value can be demoted from its threshold screening role and instead be considered as just one among many pieces of evidence in the scientific publication process as well as in statistical decision making more broadly. We conclude with a brief discussion.

2. Problems General to Null Hypothesis Significance Testing

2.1. Preface

As noted, the NHST paradigm upon which the status quo and the Benjamin et al. (2018) proposal rest is the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences (see, e.g., Morrison and Henkel 1970; Sawyer and Peter 1983; Gigerenzer 1987; McCloskey and Ziliak 1996; Gill 1999; Anderson, Burnham, and Thompson 2000; Gigerenzer 2004; Hubbard 2004). Despite this, it has been roundly criticized both inside and outside of statistics over the decades (see, e.g., Rozenboom 1960; Bakan 1966; Meehl 1978; Serlin and Lapsley 1993; Cohen 1994; McCloskey and Ziliak 1996; Schmidt 1996; Hunter 1997; Gill 1999; Gigerenzer 2004; Gigerenzer, Krauss, and Vitouch 2004; Briggs 2016; McShane and Gal 2016, 2017). Indeed, the breadth of literature on this topic across time and fields makes a complete review intractable. Consequently, we focus on what we view as among the most important criticisms of NHST for the biomedical and social sciences.

2.2. Implausible Null Hypothesis

In the biomedical and social sciences, effects are typically small and vary considerably across people and contexts. In addition, measurements are often highly variable and only indirectly related to underlying constructs of interest; thus, even when sample sizes are large, the possibilities of systematic bias and variation can result in the equivalent of small or unrepresentative samples. Consequently, estimates from any single study—the typical fundamental unit of analysis—are themselves generally noisy.

In addition, the null hypothesis employed in the overwhelming majority of applications is the sharp point null hypothesis of zero effect—that is, no difference among two or more treatments or groups—and zero systematic error—which encompasses both the adequacy of the statistical model used to compute the p-value (e.g., in terms of functional form and distributional assumptions) as well as any and all forms of systematic or nonsampling error which vary by field but include measurement error; problems with reliability and validity; biased samples; nonrandom treatment assignment; missingness; nonresponse; failure of double-blinding; noncompliance; and confounding.

The combination of these features of the biomedical and social sciences and this sharp point null hypothesis of zero effect and zero systematic error is highly problematic. Specifically, because effects are generally small and variable, the assumption of zero effect is false. Further, even were the assumption of zero effect true for some phenomenon, the effect under consideration in any study designed to examine this phenomenon would not be zero because measurements are generally noisy and systematically so. Consequently, the sharp point null hypothesis of zero effect and zero systematic error employed in the overwhelming majority of applications is implausible (Berkson 1938; Edwards, Lindman, and Savage 1963; Bakan 1966; Meehl 1990; Tukey 1991; Cohen 1994; Gelman et al. 2014; McShane and Böckenholt 2014; Gelman 2015) and thus uninteresting.

These problems are exacerbated under a lexicographic decision rule for publication as per the status quo and the Benjamin et al. (2018) proposal. Specifically, because noisy estimates that attain statistical significance are upwardly biased in magnitude (potentially to a large degree) and often of the wrong sign (Gelman and Carlin 2014), a lexicographic decision rule results in a tarnished literature. In addition, because many smaller, less resource-intensive, noisier studies are more likely to yield (or can be made more likely to yield; Simmons, Nelson, and Simonsohn 2011) one or more statistically significant results than fewer larger, more resource-intensive, better studies, a lexicographic decision rule at least indirectly encourages the former over the latter. These issues are compounded when researchers engage in multiple comparisons—whether actual or potential (i.e., the "garden of forking paths"; Gelman and Loken 2014).

In sum, various features of the biomedical and social sciences—for example, small and variable effects, systematic error, noisy measurements, a lexicographic decision rule for publication, and research practices—make NHST and in particular the sharp point null hypothesis of zero effect and zero systematic error particularly poorly suited for these domains.

2.3. Categorization of Evidence

NHST is associated with a number of problems related to the dichotomization of evidence into the different categories "statistically significant" and "not statistically significant" (or,

sometimes, trichotomization with "marginally significant" as an intermediate category) depending upon where the p-value stands relative to certain conventional thresholds. Indeed, one well-known criticism of the NHST paradigm is that the conventional 0.05 threshold—or for that matter any other one—is entirely arbitrary (Fisher 1926; Yule and Kendall 1950; Cramer 1955; Cochran 1976; Cowles and Davis 1982).

A related line of criticism suggests that the problem is with having a threshold in the first place: the dichotomization (or trichotomization) of evidence into different categories of statistical significance itself has "no ontological basis" (Rosnow and Rosenthal 1989). Specifically, Rosnow and Rosenthal (1989) note that "from an ontological viewpoint, there is no sharp line between a 'significant' and a 'nonsignificant' difference; significance in statistics...varies continuously between extremes" and thus advocate "view[ing] the strength of evidence for or against the null as a fairly continuous function of the magnitude of p."

While we agree treating the p-value continuously rather than in a thresholded manner constitutes an improvement, we go further and argue that it seldom makes sense to calibrate evidence as a function of the p-value. We hold this for at least three reasons. First, and in our view the most important, it seldom makes sense because the p-value is, in the overwhelming majority of applications, defined relative to the generally implausible and uninteresting sharp point null hypothesis of zero effect and zero systematic error. Second, because it is a poor measure of the evidence for or against a statistical hypothesis (Edwards, Lindman, and Savage 1963; Berger and Sellke 1987; Cohen 1994; Hubbard and Lindsay 2008). Third, because it tests the hypothesis that one or more model parameters equal the tested values—but only given all other model assumptions. These other assumptions—in particular, zero systematic error—seldom hold (or are at least far from given) in the biomedical and social sciences. Consequently, "a small p-value only signals that there may be a problem with at least one assumption, without saying which one. Asymmetrically, a large p-value only means that this particular test did not detect a problem—perhaps because there is none, or because the test is insensitive to the problems, or because biases and random errors largely canceled each other out" (Greenland 2017). We note similar considerations hold for other purely statistical measures.

2.4. Erroneous Scientific Reasoning

The NHST paradigm and the p-value thresholds intrinsic to it are not only problematic in and of themselves but also they routinely result in erroneous scientific reasoning. For example, researchers typically take the rejection of the sharp point null hypothesis of zero effect and zero systematic error as positive or even definitive evidence in favor of some preferred alternative hypothesis—a logical fallacy. In addition, they often make scientific conclusions largely if not entirely based on whether or not a p-value crosses the 0.05 threshold instead of taking a more holistic view of the evidence that includes the consideration of the currently subordinate factors. Further, they often confuse statistical significance and practical importance (see, e.g., Freeman 1993). Finally, they often incorrectly believe a result with a p-value below 0.05 is evidence that a relationship is causal (Holman et al. 2001).

In addition, because the assignment of evidence to different categories (e.g., statistically significant and not statistically significant) is a strong inducement to the conclusion that the items thusly assigned are categorically different, NHST encourages researchers to engage in dichotomous thinking, that is, to interpret evidence dichotomously rather than continuously. Specifically, researchers interpret evidence that reaches the conventional threshold for statistical significance as a demonstration of a difference, and, in contrast, they interpret evidence that fails to reach this threshold as a demonstration of no difference.

An example of erroneous reasoning resulting from dichotomous thinking is provided by Gelman and Stern (2006) who show that applied researchers often fail to appreciate that "the difference between 'significant' and 'not significant' is not itself statistically significant." Additional examples are provided by McShane and Gal (2016) who show that researchers across a wide variety of fields including medicine, epidemiology, cognitive science, psychology, and economics (i) interpret p-values dichotomously rather than continuously, focusing solely on whether or not the p-value is below 0.05 rather than the magnitude of the p-value; (ii) fixate on p-values even when they are irrelevant, for example, when asked about descriptive statistics; and (iii) ignore other evidence, for example, the magnitude of treatment differences. McShane and Gal (2017) show that even statisticians are susceptible to these errors.

2.5. Misinterpretation of the p-Value

A final criticism against the NHST paradigm pertains to common misinterpretations of the p-value. While formally defined as the probability of observing data as extreme or more extreme than that actually observed assuming the null hypothesis is true, the p-value has often been misinterpreted as, *inter alia*, (i) the probability that the null hypothesis is true, (ii) one minus the probability that the alternative hypothesis is true, or (iii) one minus the probability of replication. For example, Gigerenzer (2004) reports an example of research conducted on psychology professors, lecturers, teaching assistants, and students. Subjects were given the result of a simple t-test of two independent means ($t = 2.7$, $df = 18$, $p = 0.01$) and were asked six true or false questions based on the result and designed to test common misinterpretations of the p-value. All six of the statements were false and, despite the fact that the study materials noted "several or none of the statements may be correct," (i) none of the 45 students, (ii) only four of the 39 professors and lectures who did not teach statistics, and (iii) only six of the 30 professors and lectures who did teach statistics marked all as false (members of each group marked an average of 3.5, 4.0, and 4.1 statements, respectively, as false). For related results, see Oakes (1986); Cohen (1994); Haller and Krauss (2002); Gigerenzer (2018).

3. Problems Specific to the Benjamin et al. (2018) Proposal

Beyond concerns about the NHST paradigm upon which the status quo and the Benjamin et al. (2018) proposal rest, there are additional problems specific to the latter proposal. First,

Benjamin et al. (2018) propose the 0.005 threshold because it (i) "corresponds to Bayes factors between approximately 14 and 26" in favor of the alternative hypothesis and (ii) "would reduce the false positive rate to levels we judge to be reasonable." However, little to no justification is provided for either of these choices of levels.

Second, Benjamin et al. (2018) "restrict [their] recommendation to claims of discovery of new effects" which is problematic for at least two reasons. First, the proposed policy is rendered entirely impractical because they fail to define what constitutes a new effect; this is especially so in domains where research is believed to be incremental and cumulative. Second, the proposed policy would lead to incoherence when applied to replication—the very issue their proposal is meant to address. In particular, the order in which two independent studies of a common phenomenon are conducted ought to be irrelevant but is not under the Benjamin et al. (2018) proposal. Specifically, given one study with $p < 0.005$ and another with $p \in (0.005, 0.05)$, it would matter crucially which study was conducted first (and thus was "new") under the definition of replication employed in practice (i.e., a subsequent study is considered to successfully replicate a prior study if either both fail to attain statistical significance or both attain statistical significance and are directionally consistent): the second (replication) study would be deemed a success under the Benjamin et al. (2018) proposal if the first study was the $p < 0.005$ study but a failure otherwise.

Third, the fact that uncorrected multiple comparisons—both actual and potential—are the norm in applied research strictly speaking invalidates all p-values outside those from studies with preregistered protocols and data analysis procedures. This concern is acknowledged by Benjamin et al. (2018). Nonetheless, what goes unacknowledged is that even with preregistration, p-values can be invalidated if the underlying model that generated the p-value is misspecified in an important manner.

Fourth, the mathematical justification underlying the Benjamin et al. (2018) proposal has come under no small amount of criticism. Specifically, the uniformly most powerful Bayesian tests (UMPBTs) that underlie the proposal were introduced and defended by Johnson (2013b) in parallel with his call in Johnson (2013a)—and now repeated in Benjamin et al. (2018)—to use 0.005 as the new threshold. We see a number of concerns with UMPBTs that we discuss in Appendix A. Perhaps most relevant for the biomedical and social sciences, the UMPBT approach is deeply entrenched in the century-old Neyman–Pearson formalism of binary decisions and 0–1 loss functions which does not in general map, even in an approximate way, to processes of scientific learning or costs and benefits. Consequently, the logic underlying the proposal to move to a lower p-value threshold avoids firmly confronting the nature of the issue: any such rule implicitly expresses 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 (Gelman and Robert 2014) which in turn depend on the problem at hand and which vary tremendously across studies and domains.

More speculatively, we are not convinced the more stringent 0.005 threshold for statistical significance would be helpful. In the short term, it could reduce the flow of low quality work that is currently polluting even top journals. In the medium term, it could motivate researchers to perform higher-quality work that is more likely to crack the 0.005 barrier. On the other hand, it could lead to even more overconfidence in results that do get published as well as a concomitant greater exaggeration of the effect sizes associated with such results. It could also lead to the discounting of important findings that happen not to reach the more stringent threshold. In sum, we have no idea whether implementation of the proposed 0.005 threshold would improve or degrade the state of science as we can envision both positive and negative outcomes resulting from it. Ultimately, while this question may be interesting if difficult to answer, we view it as outside our purview because we believe that thresholds whether based on p-values or other purely statistical measures are a bad idea.

Perhaps curiously, we do not necessarily expect that Benjamin et al. (2018) would disagree with our criticism that their proposal is insufficient to overcome current difficulties with replication (or perhaps even with our own proposal to abandon statistical significance). After all, they "restrict [their] recommendation to claims of discovery of new effects" and recognize that "the choice of any particular threshold is arbitrary" and "should depend on the prior odds that the null hypothesis is true, the number of hypotheses tested, the study design, the relative cost of Type I versus Type II errors, and other factors that vary by research topic." Indeed, "many of [the authors] agree that there are better approaches to statistical analyses than null hypothesis significance testing."

4. Abandoning Statistical Significance

4.1. Summation and Recommendations

What can be done? Statistics is hard, especially when effects are small and variable and measurements are noisy. There are no quick fixes. Proposals such as changing the default p-value threshold for statistical significance, employing confidence intervals with a focus on whether or not they contain zero, or employing Bayes factors along with conventional classifications for evaluating the strength of evidence suffer from the same or similar issues as the current use of p-values with the 0.05 threshold. In particular, each implicitly or explicitly categorizes evidence based on thresholds relative to the generally implausible and uninteresting sharp point null hypothesis of zero effect and zero systematic error. Further, each is a purely statistical measure that fails to take a more holistic view of the evidence that includes the consideration of the currently subordinate factors, that is, 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.

In brief, each is a form of statistical alchemy that falsely promises to transmute randomness into certainty, an "uncertainty laundering" (Gelman 2016) that begins with data and concludes with 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. A critical first step forward is to begin accepting uncertainty and embracing variation in effects (Carlin 2016; Gelman 2016) and recognizing that we can learn much (indeed,

more) about the world by forsaking the false promise of certainty offered by such dichotomization.

Toward this end, we offer recommendations for how, in practice, the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with the currently subordinate factors as just one among many pieces of evidence. First, we recommend authors use the currently subordinate factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters; we also recommend they analyze and report all of their data and relevant results. Second, we recommend editors and reviewers explicitly evaluate papers with regard to not only purely statistical measures but also the currently subordinate factors.

As a highly interdisciplinary research team with representation from statistics, political science, psychology, and consumer behavior, we are acutely aware that the implementation of our broad recommendations will and ought to vary tremendously across—and even within—domains. Further, we are not so supercilious to believe that we, by ourselves, are capable of providing concrete and specific guidance on the application of these recommendations across all or perhaps even in any of these domains. Indeed, we do not believe a "template" for our recommendations is possible or desirable. In fact, such a template could even be dangerous in that it might result in a rote and recipe-like application of our recommendations that would not be entirely dissimilar to, even if perhaps less harmful than, the current practice of rote and recipe-like application of NHST. To those who might argue that, without such a template, our recommendations are unrealistic or unlikely to be adopted in practice, we reiterate that statistics is hard and a formulaic approach to statistics is a principal cause of the current replication crisis. It is for these reasons we advocate this more radical and more difficult but also more principled and more permanent approach. Nonetheless, we below suggest some broad principles that show how our recommendations might be applied. We also provide a case study in Appendix B.

4.2. For Authors

We recommend authors use the currently subordinate factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters; we also recommend they analyze and report all of their data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold.

One specific operationalization of the first part of our recommendation might be to include in their manuscripts a section that directly addresses how each of the currently subordinate factors motivated their various decisions regarding data collection, statistical analysis, interpretation of results, and writing in the context of the totality of the data and results. Such a section could, for example, discuss study design in the context of subject-matter knowledge and expectations of effect sizes as discussed by Gelman and Carlin (2014). It could also discuss the plausibility of the mechanism by (i) formalizing the hypothesized mechanism for the effect in question and expounding on the various components of it, (ii) clarifying which components were measured and analyzed in the study, and (iii) discussing aspects of the results that support as well as those that undermine the hypothesized mechanism.

One might think that the second part of our recommendation —to analyze and report all of the data and relevant results—is such a fundamental principle of science that it need hardly be mentioned. However, this is not the case! As discussed above, the status quo in scientific publication is a lexicographic decision rule whereby $p < 0.05$ is virtually always required for a result to be published and, while there are some exceptions, standard practice is to focus on such results and to not report all relevant findings.

Given the current state of practice, authors may not have a sense for how they might go about this. Rather than attempt to provide broad guidance, we direct the reader to illustrations in clinical psychology (Tackett et al. 2014), epidemiology (Gelman and Auerbach 2016a,b), political science (Trangucci et al. 2018), program evaluation (Mitchell et al. 2018), and social psychology and consumer behavior (McShane and Böckenholt 2017) as well as our case study in Appendix B.

4.3. For Editors and Reviewers

We recommend editors and reviewers explicitly evaluate papers with regard to not only purely statistical measures but also the currently subordinate factors; this should be far superior to the status quo, namely giving consideration—often scant—to the currently subordinate factors only once some p-value or other statistical threshold has been reached.

One specific operationalization of this recommendation might be to incorporate consideration of these factors into various stages of the review process. For example, editors could require reviewers to provide quantitative evaluations of each factor—including domain-specific factors determined by the editor—as well as an overall quantitative evaluation of the strength of evidence as a supplement to the current open-ended, qualitative evaluations. These could then be weighted by the editors' publicly disclosed (or even reviewers' own) importance rating of each factor. Additionally, editors could discuss and address the evaluation and importance of each factor in decision letters, thereby providing a more holistic view of the evidence.

One might object here and call our position naive: do not editors and reviewers require some bright-line threshold to decide whether the data supporting a claim is far enough from pure noise to support publication? Do not statistical thresholds provide objective standards for what constitutes evidence, and does this not in turn provide a valuable brake on the subjectivity and personal biases of editors and reviewers? Against these, we would argue that even were such a threshold needed, it would not make sense to set it based on the p-value given that it seldom makes sense to calibrate evidence as a function of this statistic and given that the costs and benefits of publishing noisy results varies by field. Additionally, the p-value is not a purely objective standard: different model specifications and statistical tests for the same data and null hypothesis yield different p-values; to complicate matters further, many subjective decisions regarding data protocols and analysis procedures such as coding and exclusion are required in practice and these often strongly impact the p-value ultimately reported. Finally, we fail to see why such a threshold screening rule is needed: editors and reviewers

already make publication decisions one at a time based on qualitative factors, and this could continue to happen if the p-value were demoted from its threshold screening rule to just one among many pieces of evidence. Indeed, no single number—whether it be a p-value, Bayes factor, or some other statistical or nonstatistical measure—is capable of eliminating subjectivity and personal biases.

Instead, we believe it is entirely acceptable to publish a paper featuring a result with, say, a p-value of 0.2 or a 90% confidence interval that includes zero provided it is relevant to a theoretical or applied question of interest and the interpretation is sufficiently accurate. It should also be possible to publish a result with, say, a p-value of 0.001 without this being taken to imply the truth of some favored alternative hypothesis.

The p-value is relevant to the question of how easily a result could be explained by a particular null model, but there is no reason this should be the crucial factor in publication. A result can be consistent with a null model but still be relevant to science or policy debates, and a result can reject a null model without offering anything of scientific interest or policy relevance.

In sum, editors and reviewers can and should feel free to accept papers and present readers with the relevant evidence. We would much rather see a paper that, for example, states that there is weak evidence for an interesting finding but that existing data remain consistent with null effects than for the publication process to screen out such findings or encourage authors to cheat to obtain statistical significance.

4.4. Abandoning Statistical Significance Outside Scientific Publishing

While our focus has been on statistical significance thresholds in scientific publication, similar issues arise in other areas of statistical decision making, including, for example, neuroimaging where researchers use voxelwise NHSTs to decide which results to report or take seriously; medicine where regulatory agencies such as the Food and Drug Administration use NHSTs to decide whether or not to approve new drugs; policy analysis where nongovernmental and other organizations use NHSTs to determine whether interventions are beneficial or not; and business where managers use NHSTs to make binary decisions via A/B tests. In addition, thresholds arise not just around scientific publication but also within research projects, for example, when researchers use NHSTs to decide which avenues to pursue further based on preliminary findings.

While considerations around taking a more holistic view of the evidence and consequences of decisions are rather different across each of these settings and different from those in scientific publication, we nonetheless believe our proposal to demote the p-value from its threshold screening role and emphasize the currently subordinate factors applies in these settings. For example, in neuroimaging, the voxelwise NHST approach misses the point in that there are typically no true zeros and changes are generally happening at all brain locations at all times. Plotting images of estimates and uncertainties makes sense to us, but we see no advantage in using a threshold.

For regulatory, policy, and business decisions, cost-benefit calculations seem clearly superior to acontextual statistical thresholds. Specifically, and as noted, such thresholds 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.

That said, we acknowledge that thresholds—of a nonstatistical variety—may sometimes be useful in these settings. For example, consider a firm contemplating sending a costly offer to customers. Suppose the firm has a customer-level model of the revenue expected in response to the offer. In this setting, it could make sense for the firm to send the offer only to customers that yield an expected profit greater than some threshold, say, zero.

Even in pure research scenarios where there is no obvious cost-benefit calculation—for example, a comparison of the underlying mechanisms, as opposed to the efficacy, of two drugs used to treat some disease—we see no value in p-value or other statistical thresholds. Instead, we would like to see researchers simply report results: estimates, standard errors, confidence intervals, etc., with statistically inconclusive results being relevant for motivating future research.

While we see the intuitive appeal of using p-value or other statistical thresholds as a screening device to decide what avenues (e.g., ideas, drugs, or genes) to pursue further, this approach fundamentally does not make efficient use of data: there is in general no connection between a p-value—a probability based on a particular null model—and either the potential gains from pursuing a potential research lead or the predictive probability that the lead in question will ultimately be successful. Instead, to the extent that decisions do need to be made about which lines of research to pursue further, we recommend making such decisions using a model of the distribution of effect sizes and variation, thus working directly with hypotheses of interest rather than reasoning indirectly from a null model.

We would also like to see—when possible in these and other settings—more precise individual-level measurements, a greater use of within-person or longitudinal designs, and increased consideration of models that use informative priors, that feature varying treatment effects, and that are multilevel or meta-analytic in nature (Gelman 2015, 2017; McShane and Böckenholt 2017, 2018).

4.5. Getting From Here to There

How do we get from here—NHST, deterministic summaries, overconfidence in results, and statistical analysis focused on reporting just some of the data—to there—statistical analysis and interpretation of results that accepts uncertainty and embraces variation and that features full reporting of results rather than focusing on whatever happens to exceed some statistical threshold?

We have offered the recommendations that we believe will serve researchers best. However, we recognize that research takes place within an institutional structure that often encourages behavior that is counter to these recommendations. Researchers respond to the expectations of funding agencies in study design and editors and reviewers in writing. Conversely, funding agencies must choose among the submissions they receive and editors can only publish papers that are submitted to them. A careful research proposal that openly grapples

with uncertainty may unfortunately lose out in the funding competition to a more traditional proposal that blithely promises 80% power based on selected and overestimated effect sizes. Similarly, a paper that presents all the data without making inappropriate claims of certainty may not get published in a journal that also receives submissions in which statistically significant results are presented at face value.

These institutional problems are difficult and we do not propose solutions to them. We imagine improvement will come in fits and starts, in several parallel tracks, all of which we and others have tried to contribute to in our applied and methodological research: improved statistical methods that move beyond NHST and include multilevel modeling, machine learning, statistical graphics, and other tools for analyzing and visualizing large amounts of data; applied examples using these improved methods, thereby demonstrating that it is possible to perform successful statistical analyses without aiming for deterministic results; theoretical work on the statistical effects of selection based on statistical significance and other decision criteria; and criticism of published work with gross overestimates of effect sizes or inappropriate claims of certainty. While we recognize change will likely require institutional reform including major modifications of current practices of funding agencies and editors and reviewers, we are also optimistic that some combination of recognition of error and awareness of alternatives can also motivate change.

5. Discussion

In this article, we have proposed to abandon statistical significance and offered recommendations for how this can be implemented in the scientific publication process as well as in statistical decision making more broadly. We reiterate that we have no desire to "ban" p-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the currently subordinate factors.

While our proposal to abandon statistical significance may seem on the surface quite radical, at least one aspect of it—to treat p-values or other purely statistical measures continuously rather than in a thresholded manner—most certainly is not. Indeed, this was advocated by Fisher himself (Fisher 1956; Greenland and Poole 2013) as well as by other early and eminent statisticians including Pearson (Hurlbert and Lombardi 2009), Cox (Cox 1977, 1982), and Lehmann (Lehmann 1993; Senn 2001). It has also been advocated outside of statistics over the decades (see, e.g., Boring 1919; Eysenck 1960; Skipper, Guenther, and Nass 1967) and recently (see, e.g., Drummond 2015; Lemoine et al. 2016; Amrhein, Korner-Nievergelt, and Roth 2017; Greenland 2017; Amrhein and Greenland 2018). Finally, it is fully consistent with the recent American Statistical Association (ASA) Statement on Statistical Significance and p-values ("Principle 3: Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold;" Wasserstein and Lazar 2016). In sum, this aspect of our proposal is part of a long literature both inside and outside of statistics over the decades that stands in direct opposition to the threshold-based status quo and the Benjamin et al. (2018) proposal.

Where our proposal might move beyond this literature is in three ways. First, we suggest that p-values or other purely statistical measures, thresholded or not, should not take priority over the currently subordinate factors (that said, others too have emphasized this including the recent ASA Statement which advises that "researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" and cautions that "no single index should substitute for scientific reasoning;" Wasserstein and Lazar 2016). Second, as discussed above, while we believe treating the p-value continuously rather than in a thresholded manner constitutes an improvement, we go further and argue that it seldom makes sense to calibrate evidence as a function of the p-value or other purely statistical measures. Third, we offer recommendations for authors as well as editors and reviewers for how our proposal to abandon statistical significance can be implemented in the scientific publication process as well as in statistical decision making more broadly.

Our recommendations will not themselves resolve the replication crisis, but we believe they will have the salutary effect of pushing researchers away from the pursuit of irrelevant statistical targets and toward understanding of theory, mechanism, and measurement. We also hope they will push them to move beyond the paradigm of routine "discovery," and binary statements about there being "an effect" or "no effect," to one of continuous and inevitably flawed learning that is accepting of uncertainty and variation.

Appendix A. Problems With Uniformly Most Powerful Bayesian Tests

The mathematical justification underlying the Benjamin et al. (2018) proposal has come under no small amount of criticism. Specifically, the UMPBTs that underlie the proposal were introduced and defended by Johnson (2013b) in parallel with his call in Johnson (2013a)—and now repeated in Benjamin et al. (2018)—to use 0.005 as the new threshold. We see a number of concerns with UMPBTs.

First, and perhaps most relevant for the biomedical and social sciences, the UMPBT approach is deeply entrenched in the century-old Neyman–Pearson formalism of binary decisions and 0–1 loss functions. As Pericchi, Pereira, and Pérez (2014) note, "the essence of the problem of classical testing of significance lies in its goal of minimizing Type II error (false negative) for a fixed Type I error (false positive)." While this formalism allows for mathematical optimization under some restricted collection of distributions and testing problems, it is quite rudimentary from a decision theoretical point of view, even to the extent of failing most purposes of conducting a sharp point null hypothesis test.

More specifically, the 0–1 loss function implicit in the NHST paradigm does not in general map, even in an approximate way, to processes of scientific learning or costs and benefits. Consequently, the logic underlying the proposal to move to a lower p-value threshold avoids firmly confronting the nature of the issue: any such rule implicitly expresses 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 (Gelman and Robert 2014) which in turn depend on the problem at hand and which vary tremendously across studies and domains. Instead, the UMPBT is based on a minimax prior that does not correspond to any distribution of effect sizes but

rather represents a worst case scenario under a set of mathematical assumptions.

Second, there is no reason for non-Bayesians' to adopt UMPBTs when they can instead rely on the standard Neyman–Pearson approach to uniformly most powerful (non-Bayesian) tests.

Third, defining the dependence of the procedure over a threshold (γ in the notation of Johnson (2013b)) replicates the fundamental difficulty with the century-old Fisherian answer to hypothesis testing. To further seek a full agreement with the classical rejection region as advocated by Johnson (2013b) is to simply negate the appeal of a truly Bayesian approach to this issue; moreover, this agreement is impossible to achieve for realistic statistical models.

Fourth, the construction of a UMPBT relies on the assumption of a "true" prior, which can be criticized in a vast majority of cases and which in any case moves one away from the Bayesian paradigm: with a single and "true" prior, the Bayesian model becomes an errors-in-variables model.

Fifth, the argument to maximize a probability for the Bayes factor to exceed a certain threshold also moves one away from the Bayesian paradigm because: (i) it ignores the motives for running the NHST and the subsequent steps taken in decision making or inference; (ii) it further negates any prior modeling of the alternative hypothesis aimed at separating the parameter space into regions of different (prior) likelihood; (iii) it does not condition upon the actual observations but instead integrates over the observation space and hence may fall afoul of the likelihood principle; (iv) it posits a single and fixed threshold γ for rejecting the null when there is no reason for γ not to depend on the observed data, as also argued above; (v) the maximization step eliminates the role of the prior distribution, as also argued above; (vi) in the rare one-dimensional settings where the maximization step can be conducted in closed form, the solution is a distribution with finite support; (vii) in the event the null hypothesis is rejected, the uniformly most powerful prior (or alt-prior) corresponding to the alternative cannot be used as such in subsequent inference but must instead be replaced with a regular prior over the whole parameter space—a strong violation of Bayesian coherence.

Sixth, speaking more generally, the concept of uniformly most powerful priors (and tests) does not easily extend to multivariate settings and even less to realistic cases that involve complex null hypotheses that contain nuisance parameters. The first solution proposed in Johnson (2013b), to integrate out the nuisance parameters in the null hypothesis using a specific prior distribution, falls short of solving the issue of "objective Bayesian tests." The second solution, namely to replace the unknown nuisance parameters with standard estimates, stands even farther from a Bayesian perspective.

Indeed, the Bayes factor itself is a consequence of the rudimentary Neyman–Pearson formalism, which as such caters to the issue of statistical significance. A discussion of the difficulties with this from a Bayesian perspective is provided in Kamary et al. (2014), with a proposal of setting the hypothesis problem as one of mixture estimation.

Seventh, Johnson (2013b) contains very little support for the asymptotic relevance of the approach, beyond the limiting normal distribution of the uniformly most powerful log Bayes factor and the convergence of the support to the "true" value of the parameter.

In closing, we note that many of our criticisms of the Johnson (2013b) approach relate to the fact that it falls short of being truly Bayesian. However, we do not mean to say that hypothesis testing must be done in a Bayesian manner. Rather, we emphasize this because, to the extent that the Johnson (2013b) approach loses its Bayesian connection, it also loses a Bayesian justification for the 0.005 rule. Consequently, 0.005 becomes just another arbitrary threshold, justified by some implicit tradeoff between false positives and negatives which we think does not make sense in any absolute and acontextual way.

Appendix B. Case Study

In the context of a hypothetical case study on the effects of sodium on blood pressure, we discuss how authors as well as editors and reviewers might follow our recommendation to demote the *p*-value from its threshold screening role and instead treat it continuously along with the currently subordinate factors—related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain—as just one among many pieces of evidence.

We recommend authors use the currently subordinate factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters. In this example, the authors might consider *related prior evidence* that indicates the importance of blood pressure as a marker for healthy arteries, suggests the role of sodium in hemodynamics, and so forth. This evidence might also reveal a *plausible mechanism*, namely to excrete excess sodium the body must increase blood pressure.

In terms of *study design and data quality*, the authors might consider various possibilities for data collection. How should they recruit subjects? Should they randomize them to a low-sodium versus high-sodium diet? Or should they track them longitudinally, say via routine annual checkups over the course of years? Or is such data already available from some prior study? When and how often should sodium and blood pressure be measured? And how? The authors might measure sodium through a dietary recall questionnaire (noisy), through asking participants to maintain a food diary (somewhat less noisy), or through collection of urine to measure urinary sodium excretion (precise but restricted to a limited time point). Likewise, for blood pressure, they might rely on measurements conducted by someone convenient like friends or family members of the subjects who likely do not possess formal clinical training (noisy) or by paid clinicians instructed on the proper protocol for blood pressure measurement (precise but expensive).

Suppose the authors hypothesize a positive association between sodium consumption and high blood pressure. For the moment, let us assume that—while eschewing the NHST paradigm and the *p*-value thresholds intrinsic to it—the authors nonetheless perform a statistical analysis that results in a *p*-value. Further, let us assume they obtain a *p*-value of 0.001. How should this impact their interpretation of results, writing, and statistical decision making more broadly? Certainly, they have gained support for their hypothesis. However, can they conclude sodium is associated with—or even causes—high blood pressure as they would under the NHST paradigm?

Well, it would depend on the context and limitations of the *study design and data quality*. For example, supposing the study took place in Japan, perhaps the association exists in the Japanese subject population studied but does not in European populations whether because of some genetic differences between the two populations or because of some dietary differences (e.g., dietary sodium levels are much higher among Japanese so the association might not hold in levels typical among Europeans).

In terms of a causal interpretation, this would depend on *related prior evidence*, *plausibility of mechanism*, and *study design and data quality*. If prior studies show consistent and strong associations between sodium consumption and blood pressure, if evidence from physiological studies and animal models are consistent with an effect of sodium consumption on blood pressure, or if sodium levels were randomized, this increases the support for a causal role of sodium in increasing blood pressure.

Given, say, that the causal interpretation holds and holds broadly, the authors could then consider clinical significance, that is, *real world costs and benefits*. This depends not at all on a *p*-value but on the estimates of the magnitude of the effect—not only on blood pressure

but also on downstream outcomes such as cardiovascular disease—as well as the uncertainty in them. It also depends on the costs of potential interventions such as lower sodium diets and drugs. They could also discuss *novelty of finding* in light of all of the above.

Now, let us assume they had instead obtained a p-value of 0.2. Can they conclude sodium is not associated with high blood pressure as they would under the NHST paradigm? Again, this would depend on all the factors discussed above. For example, perhaps the association does not exist in the Japanese population but does in European ones and so on.

There are two key points in this. First, results need not first have a p-value or some other purely statistical measure that attains some threshold before consideration is given to the currently subordinate factors. Instead, and as illustrated above, statistical measures should be considered along with the currently subordinate factors as just one among many pieces of evidence and should not take priority thereby yielding a more holistic view of the evidence. Second, statistical measures should be treated continuously in this more holistic view of the evidence. Specifically, a lower p-value constitutes continuously stronger evidence—and this holds regardless of the level of the p-value. Further, this continuously stronger evidence can be balanced along with the strengths and weaknesses of the currently subordinate factors in assessing the level of support for a hypothesis.

Of course, we believe not only that the authors' statistical analysis should not be restricted to the NHST paradigm and the p-value thresholds intrinsic to it but also that it need not—and often should not—even result in a p-value (i.e., because it seldom makes sense to calibrate evidence as a function of the p-value). As noted, we recommend authors report all of their data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold. In this context, this might involve modeling the association between sodium and blood pressure as a function of additional health and dietary variables, demographic variables, and geography using a multilevel model. Such a model would not yield one single p-value thereby encouraging dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect." Instead, it would yield many estimates that vary based on, for example, health and dietary variables, demographic variables, and geography, as well as the uncertainty in these estimates. Indeed, by accepting uncertainty and embracing variation in effects, the authors would uncover and present a much richer and more nuanced story about the association between sodium and blood pressure.

Turning to editors and reviewers, we recommend they explicitly evaluate papers with regard to not only purely statistical measures but also the currently subordinate factors. How might this work? We envision it would be rather similar to the above but in reverse. Specifically, editors and reviewers evaluating the authors' paper on sodium and blood pressure would systemically assess, and possibly even indicate the weight they assign to, each of the following: How does the paper fit in with and build upon *related prior evidence*? Is the *mechanism plausible*? Are the *study design and data quality* sufficient to justify the conclusions? What are the implications in terms of *real world costs and benefits*? How *novel* are the findings? And, of course, how appropriate are the statistical analyses employed and how strong is the statistical support, whether in the form of a p-value or some other measure, resulting from these analyses?

In this more holistic view of the evidence, statistical measures are just one among many pieces of evidence considered by editors and reviewers and do not take priority. Of course, this does not mean that they cannot or will not strongly impact or alter their evaluation decisions. For example, in the context of the authors' paper on sodium and blood pressure, strong statistical support, whether in the form of a low p-value or otherwise, for a finding that sodium consumption is associated with low blood pressure—the direction opposite of that indicated by prior evidence—in the context of a high quality study design featuring large samples and precise measurements might be deemed more novel and worthy of publication than if the statistical support had been weaker or if the finding was in the same direction as that indicated by prior evidence.

In sum, authors as well as reviewers and editors need not use statistical significance as a lexicographic decision rule. Results need not first have a p-value or some other purely statistical measure that attains some threshold before consideration is given to the currently subordinate factors. Instead, treated continuously, statistical measures should be considered along with the currently subordinate factors as just one among many pieces of evidence and should not take priority thereby yielding a more holistic view of the evidence.

Acknowledgment

We thank the National Science Foundation, the Institute for Education Sciences, and the Office of Naval Research for partial support of Andrew Gelman's work.

References

Amrhein, V., and Greenland, S. (2018), "Remove, Rather Than Redefine, Statistical Significance," *Nature Human Behaviour*, 2, 4. [241]

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). "The Earth is Flat ($p > 0.05$): Significance Thresholds and the Crisis of Unreplicable Research," *PeerJ*, 5, e3544. [241]

Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000), "Null Hypothesis Testing: Problems, Prevalence, and an Alternative," *Journal of Wildlife Management*, 64, 912–923. [236]

Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66(6), 423–437. [236]

Bem, D. J. (2011), "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect," *Journal of Personality and Social Psychology*, 100, 407–425. [235]

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, 6–10. [235,236,237,238,241]

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconciliability of p Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [237]

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of the American Statistical Association*, 33, 526–536. [236]

Boring, E. G. (1919), "Mathematical vs. Scientific Significance," *Psychological Bulletin*, 16, 335–338. [241]

Briggs, W. M. (2016), *Uncertainty: The Soul of Modeling, Probability and Statistics*, New York: Springer. [236]

Carlin, J. B. (2016), "Is Reform Possible Without a Paradigm Shift?" *The American Statistician*, 901, 10 (supplemental material to the ASA statement on p-values and statistical significance). [238]

Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010), "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance," *Psychological Science*, 21, 1363–1368. [235]

Cochran, W. G. (1976), "Early Development of Techniques in Comparative Experimentation," in *On the History of Statistics and Probability*, New York: Dekker. [237]

Cohen, J. (1994), "The Earth is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [236,237]

Cowles, M., and Davis, C. (1982), " On the Origins of the .05 Level of Significance," *American Psychologist*, 44, 1276–1284. [237]

Cox, D. R. (1977), "The Role of Significance Tests," *Scandinavian Journal of Statistics*, 4, 49–70. [241]

——— (1982), "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325–331. [241]

Cramer, H. (1955), *The Elements of Probability Theory*, New York: Wiley. [237]

Drummond, G. (2015), "Most of the Time, *P* Is an Unreliable Marker, So We Need No Exact Cut-Off," *British Journal of Anaesthesia*, 116, 894–894. [241]

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193. [236,237]

Eysenck, H. J. (1960), "The Concept of Statistical Significance and the Controversy About One-Tailed Tests," *Psychological Review*, 67, 269. [241]

Fisher, R. A. (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture*, 33, 503–513. [237]

——— (1956), *Statistical Methods and Scientific Inference*, New York: Hafner Publishing Co. [241]

Freeman, P. R. (1993), "The Role of *p*-Values in Analysing Trial Results," *Statistics in Medicine*, 12, 1443–1452. [237]

Gelman, A. (2015), "The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective," *Journal of Management*, 41, 632–643. [236,240]

——— (2016), "The Problems With *p*-Values Are Not Just With *p*-Values," *The American Statistician*, 70, 10 (supplemental material to the ASA statement on *p*-values and statistical significance). [238]

——— (2017), "The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to do About It," *Personality and Social Psychology Bulletin*, 44, 16–23. [240]

Gelman, A., and Auerbach, J. (2016a), "Age-Aggregation Bias in Mortality Trends," *Proceedings of the National Academy of Sciences of the United States of America* , 113, E816–E817. [239]

——— (2016b), "Mortality Trends by Race/Ethnicity, Sex, Age and State," Technical Report, Columbia University. [239]

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations Assessing Type s (Sign) and Type m (magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [236,239]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014), *Bayesian Data Analysis* (3rd ed.), Boca Raton, FL: Chapman and Hall/CRC. [236]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [236]

Gelman, A., and Robert, C. P. (2014), "Revised Evidence for Statistical Standards," *Proceedings of the National Academy of Sciences of the United States of America*, 111, E1933–E1933. [238,241]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [237]

Gigerenzer, G. (1987). *The Probabilistic Revolution. Vol. II: Ideas in the Sciences* (Vol. II), Cambridge, MA: MIT Press. [236]

——— (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [236,237]

——— (2018), "Statistical Rituals: The Replication Delusion and How We Got There," *Advances in Methods and Practices in Psychological Science*, 1, 198–218. [237]

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004), "The Null Ritual: What You Always Wanted to Know About Null Hypothesis Testing But Were Afraid to Ask," in *Handbook on Quantitative Methods in the Social Sciences*, Thousand Oaks, CA: Sage Publications, Inc., pp. 389–406. [236]

Gill, J. (1999), "The Insignificance of Null Hypothesis Significance Testing," *Political Research Quarterly*, 52, 647–674. [236]

Greenland, S. (2017), "Invited Commentary: The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–646. [237,241]

Greenland, S., and Poole, C. (2013), "Living With Statistics in Observational Research," *Epidemiology*, 24, 73–78. [241]

Haller, H., and Krauss, S. (2002), "Misinterpretations of Significance: a Problem Students Share With Their Teachers?," *Methods of Psychological Research*, 7, 1–20, *http://www.mpr-online.de*. [237]

Holman, C. J., Arnold-Reed, D. E., de Klerk, N., McComb, C., and English, D. R. (2001), "A Psychometric Experiment in Causal Inference to Estimate Evidential Weights Used by Epidemiologists," *Epidemiology*, 12, 246–255. [237]

Hubbard, R. (2004), "Alphabet Soup: Blurring the Distinctions Between p's and α's in Psychological Research," *Theory and Psychology*, 14, 295–327. [236]

Hubbard, R., and Lindsay, R. M. (2008), "Why *p* Values Are Not a Useful Measure of Evidence in Statistical Significance Testing," *Theory and Psychology*, 18, 69–88. [237]

Hunter, J. E. (1997), "Needed: A Ban on the Significance Test," *Psychological Science*, 8, 3–7. [236]

Hurlbert, S. H., and Lombardi, C. M. (2009), "Final Collapse of the Neyman–Pearson Decision Theoretic Framework and Rise of the Neofisherian," *Annales Zoologici Fennici*, 46, 311–349. [241]

Ioannidis, J. P. A. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124. [235]

Johnson, V. E. (2013a), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19313–19317. [238,241]

——— (2013b), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741. [238,241,242]

Kamary, K., Mengersen, K., Robert, C., and Rousseau, J. (2014), " Testing Hypotheses as a Mixture Estimation Model," Technical Report, *https://arxiv.org/pdf/1214.4436.pdf*. [242]

Lehmann, E. L. (1993), *Testing Statistical Hypotheses*, New York: Chapman and Hall. [241]

Lemoine, N. P., Hoffman, A., Felton, A. J., Baur, L., Chaves, F., Gray, J., Yu, Q., and Smith, M. D. (2016), "Underappreciated Problems of Low Replication in Ecological Field Studies," *Ecology*, 97, 2554–2561. [241]

McCloskey, D. N., and Ziliak, S. (1996), "The Standard Error of Regression," *Journal of Economic Literature*, 34, 97–114. [236]

McShane, B. B., and Böckenholt, U. (2014), "You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic," *Perspectives on Psychological Science*, 9, 612–625. [236]

——— (2017), "Single Paper Meta-Analysis: Benefits for Study Summary, Theory-Testing, and Replicability," *Journal of Consumer Research*, 43, 1048–1063. [239,240]

——— (2018), "Multilevel Multivariate Meta-Analysis With Application to Choice Overload," *Psychometrika*, 83, 255–271. [240]

McShane, B. 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. [236,237]

——— (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112, 885–895. [236,237]

Meehl, P. E. (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," *Journal of Counseling and Clinical Psychology*, 46, 806–834. [235,236]

——— (1990), "Why Summaries of Research on Psychological Theories Are Often uninterpretable," *Psychological Reports*, 66, 195–244. [236]

Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U. K., Harris, M. W., Sachs, S. E., Stuart, E. A., Feller, A., and Makela, S. (2018), "The Millennium Villages Project: A Retrospective, Observational, Endline Evaluation," *The Lancet*, 6, e500–e513. [239]

Morrison, D. E., and Henkel, R. E. (1970), *The Significance Test Controversy*, Chicago: Aldine. [236]

Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioral Sciences*, New York: Wiley. [237]

Pericchi, L., Pereira, C. A., and Pérez, M.-E. (2014), "Adaptive Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences of the United States of America*, 111, E1935–E1935. [241]

Resnick, B. (2017), "What a Nerdy Debate About *p*-Values Shows About Science—And How to Fix It," Technical Report. [235]

Rosnow, R. L., and Rosenthal, R. (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44, 1276–1284. [237]

Rozenboom, W. W. (1960), "The Fallacy of the Null Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428. [236]

Sawyer, A. G., and Peter, J. P. (1983), "The Significance of Statistical Significance Tests in Marketing Research," *Journal of Marketing Research*, 20, 122–133. [236]

Schmidt, F. L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers," *Psychological Methods*, 1, 115–129. [236]

Senn, S. S. (2001), "Two Cheers for p-Values?," *Journal of Epidemiology and Biostatistics*, 6, 193–204. [241]

Serlin, R. C., and Lapsley, D. K. (1993), "Rational Appraisal Psychological Research and the Good Enough Principle," in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale, NJ: Lawrence Erlbaum Associates. [236]

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), "False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [236]

Skipper, J. K., Guenther, A. L., and Nass, G. (1967), "The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science," *The American Sociologist*, 5, 16–18. [241]

Smaldino, P. E., and McElreath, R. (2016), "The Natural Selection of Bad Science," Technical Report, *https://arxiv.org/pdf/1605.09511v1.pdf*. [235]

Tackett, J. L., Kushner, S. C., Herzhoff, K., Smack, A. J., and Reardon, K. W. (2014), "Viewing Relational Aggression Through Multiple Lenses: Temperament, Personality, and Personality Pathology," *Development and Psychopathology*, 26, 863–877. [239]

Trangucci, R., Ali, I., Gelman, A., and Rivers, D. (2018), "Voting Patterns in 2016: Exploration Using Multilevel Regression and Poststratification (MRP) on Pre-Election Polls," arXiv preprint arXiv:1802.00842. [239]

Tukey, J. W. (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100–116. [236]

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. [241]

Yule, G. U., and Kendall, M. G. (1950), *An Introduction to the Theory of Statistics* (14th ed.), London: Griffin. [237]

The American Statistician

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

Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

Christopher Tong

To cite this article: Christopher Tong (2019) Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science, The American Statistician, 73:sup1, 246-261, DOI: 10.1080/00031305.2018.1518264

To link to this article: https://doi.org/10.1080/00031305.2018.1518264

© 2019 The Authors. Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 21561

View related articles

View Crossmark data

Citing articles: 22 View citing articles

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, 246–261: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518264

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

Christopher Tong

United States Department of Agriculture, Center for Veterinary Biologics, Ames, IA

ABSTRACT

Scientific research of all kinds should be guided by statistical thinking: in the design and conduct of the study, in the disciplined exploration and enlightened display of the data, and to avoid statistical pitfalls in the interpretation of the results. However, formal, probability-based statistical inference should play no role in most scientific research, which is inherently exploratory, requiring flexible methods of analysis that inherently risk overfitting. The nature of exploratory work is that data are used to help guide model choice, and under these circumstances, uncertainty cannot be precisely quantified, because of the inevitable model selection bias that results. To be valid, statistical inference should be restricted to situations where the study design and analysis plan are specified prior to data collection. Exploratory data analysis provides the flexibility needed for most other situations, including statistical methods that are regularized, robust, or nonparametric. Of course, no individual statistical analysis should be considered sufficient to establish scientific validity: research requires many sets of data along many lines of evidence, with a watchfulness for systematic error. Replicating and predicting findings in new data and new settings is a stronger way of validating claims than blessing results from an isolated study with statistical inferences.

ARTICLE HISTORY
Received February 2018
Revised June 2018

KEYWORDS
Many sets of data; Model uncertainty; Optimism principle; Researcher degrees of freedom; Shoe leather; Triangulation

1. Introduction

Statistical inferences are claims made using probability models of data generating processes, intended to characterize unknown features of the population(s) or process(es) from which data are thought to be sampled. Examples include estimates of parameters such as the population mean (often attended by confidence intervals), hypothesis test results (such as p-values), and posterior probabilities. Such methods are often intended to quantify (and thus to tame) uncertainty, and to evaluate the plausibility of chance as an explanation of a data pattern. The widespread use of statistical inference methods in scientific research has recently been scrutinized and questioned, for reasons outlined in the "ASA Statement on Statistical Significance and P-values" (Wasserstein and Lazar 2016). The *ASA Symposium on Statistical Inference* (Bethesda, Maryland, USA; October 11–13, 2017) was expected to "lead to a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical science—and science itself—into a new age," according to the call for articles for this special issue of *The American Statistician*. This article is intended to offer one such rethinking. The core of our perspective is indeed the essential link between statistical science and science itself.

Much of the recent discussion of statistical inference focuses on null hypothesis testing, p-values, and even the very notion of statistical significance. Unease with these commonly used procedures and concepts has a lengthy history, eloquently discussed by others. We do not add to this dimension of the conversation; it suffices to cite the ASA Statement itself (Wasserstein and Lazar 2016) and two recent commentaries (Argamon 2017; McShane et al. 2017). Among these criticisms, McShane and Gelman (2017) succinctly stated that null hypothesis testing "was supposed to protect researchers from over-interpreting noisy data. Now it has the opposite effect." The *ASA Symposium on Statistical Inference* was intended to move the conversation to what should be done instead.

However, as a prelude to a discussion of remedies, we will argue here, as Andrew Gelman (2016) did, that nearly *all* forms of statistical inference share serious deficiencies, except in cases when the study protocol and statistical model are fully prespecified. We shall argue that these issues stem largely from the *Optimism Principle* (Picard and Cook 1984) that is an inevitable byproduct of the necessarily flexible data analysis and modeling work that attends most scientific research. Moreover, we contend that the well-established distinction between exploratory and confirmatory objectives provides a framework for understanding the proper roles of flexible versus prespecified statistical analyses. Unfortunately, we think that in much of the current use of inferential methods in science, except in specialized fields such as human clinical trials, this distinction is absent. This absence has enabled the widespread dissemination of biased statistical inferences and encouraged a *Cult of the Isolated Study* (Nelder 1986) that short-circuits the iterative nature of research. Statistical inference should not be used to avoid wearing "shoe leather" (Freedman 1991), a metaphor for the hard work of gathering more and better data, discovering and dealing with systematic sources of error, and building a scientific argument along many lines of evidence. The statistical contribution to science must focus on data production, data description and

CONTACT Christopher Tong ctong@mailaps.org United States Department of Agriculture, Center for Veterinary Biologics, P. O. Box 844, Ames, IA 50010.
© 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.

exploration, and statistical thinking, rather than statistical inference. None of the concepts we discuss are new, but collectively this perspective argues for abandoning much of conventional statistical practice and teaching. The perspective outlined here has little in common with many of the proposals discussed at the *ASA Symposium*, although a few have specific points of commonality.

2. Statistical Inference and the Optimism Principle

A stereotypical framework for statistical inference in a simple univariate setting is as follows (Fisher 1922). We are presented with a dataset, and we have reason to believe (or are prepared to assume) that the data were generated as an independent, identically distributed (IID) sample from an underlying population. This population may be characterized by a probability model, and our goal is to make inferences about parameters characterizing this unknown model. Such inferences will be attended by uncertainty, since we do not have access to the population, but only IID samples drawn from it. Examples of statistical inferences include estimates of parameters or other properties of the population, confidence intervals (confidence sets, credible sets) for such estimates, and hypothesis tests about such parameters or properties. More generally, the probability model may include both fixed and random components, and the latter may have variance-covariance structures more complex than the IID assumption. There are several contending schools of thought about how such inferences should be made, of which the frequentist and Bayesian are the most widely known (see *Appendix* A.1). What all these ideologies have in common is that *they express uncertainty through a probability claim*, such as a confidence level, Type 1 error, p-value, posterior probability, likelihood function, and so on. Our focus on statistical inference in this article will mainly be on these probabilistic expressions of uncertainty. This is distinct from both the common language and general scientific uses of the term "inference."

On the first page of the first chapter of their recent book, *Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science*, Efron and Hastie (2016, p. 3) presented the sample mean and its standard error. They then stated that

> It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy.

This principle is indeed typical of statistical teaching and practice. However, when it comes to *prediction*, Efron and Hastie (2016, p. 227) instead advocated a procedure common in machine learning, namely, *data-splitting*. The available data are randomly split into a training set and a test set. The test set is hidden in a vault, and the training set is used to train/fit the predictive algorithm or model. Once that algorithm/model is developed, the test set is unveiled and run through the algorithm to obtain an "honest" assessment of its performance on data it has never seen. (We return to data splitting and more sophisticated approaches in Section 5.) All model building activities, such as variable selection, must be confined to the training set (e.g., Ambroise and McLachlan 2002; Reunanen 2003). The use of data-splitting helps to diagnose *overfitting*, the tendency of a predictive algorithm to optimize its performance on the training data, at the expense of its ability to perform well on data *outside* the training set. Generalizability, to data from the future or to subjects not sampled, for instance, is usually of greater ultimate interest to the user than training data performance.

Statistical inferences also can suffer from this overfitting problem when "the same data that supplies an estimate" are used "to assess its accuracy." In fact, Efron and Hastie (2016) alluded to this in chap. 20 of their book, titled "Inference after model selection." There they wrote that in the past, such inferences were "typically done ignoring the model selection process" but that "Electronic computation has opened the door to a more honest analysis of estimation accuracy, one that takes account of the variability induced by data-based model selection" (ibid, p. 394). They conceded that while "Currently, there is no overarching theory for inference after model-selection" they instead provide "promising analyses of individual situations" (ibid, pp. 395).

Let us unpack these ideas further. Harrell (2015, p. ix) observed that

> Using the data to guide the data analysis is almost as dangerous as not doing so.

This seems like an oxymoron, but there is wisdom here. Statistical model building is usually a multi-step, interactive process because the model is not completely prespecified prior to data collection. Consequently, the model building/model criticism/model selection process may include the following steps that depend on the data at hand, are often partly subjective, and are ideally informed by knowledge of earlier studies and/or existing scientific theory.

- Screening the data for unusual, extreme, incomplete, or inconsistent data records, for possible adjustment or removal;
- Consideration of rescaling or transforming some variables (e.g., logarithm? standardize?) or even a change of coordinates (e.g., principle components analysis);
- Variable selection;
- Decisions to keep or remove interaction terms, or higher-order polynomials;
- Graphs and tables of data to identify patterns, expected or otherwise, for possible inferential reporting;
- Graphical examination of residuals from fitted models, to assess goodness of fit, and applying remedial measures if needed; and
- Methodology selection for fitting the model (e.g., should a robust regression be used instead of least squares?).

Each of these activities offers a chance to improve the fit of the model by repeated comparison with the data—and therefore a chance to *overfit* the data. Simmons, Nelson, and Simonsohn (2011) called these opportunities *researcher degrees of freedom*, and when abused to fish for publishable p-values, *p-hacking*. Wicherts et al. (2016) cataloged 34 kinds of researcher degrees of freedom spanning design, conduct, analysis, and reporting of psychological studies; they state that their list is "in no way exhaustive."

The resulting inferences from the final model tend to be biased, with uncertainties underestimated, and statistical sig-

nificance overestimated, a phenomenon dubbed the *Optimism Principle* by Picard and Cook (1984). They quoted Mosteller and Tukey (1977, p. 37):

> Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimizing process that chose it from among many possible procedures will have made the greatest use possible of any and all idiosyncrasies of those particular data . . . As a result, the procedure will likely work better for these data than for almost any other data that will arise in practice.

Nearly seven decades ago Koopmans (1949), who was later a 1975 Nobel laureate in economics, alluded to this problem, and it has been studied extensively by statisticians and econometricians since the early 1980s. A classic result was given by Freedman (1983), who showed that a commonly taught and practiced procedure of screening available predictor variables for inclusion in a regression model, using a statistical test, can result in the inclusion of bogus variables with high statistical significance. He showed this in the extreme case where all the candidate predictor variables consist of Gaussian noise, unrelated to the response variable (which is also Gaussian noise). In other words, it is possible to obtain a seemingly informative linear model, with decent R^2 and several statistically significant predictor variables, from data that is utter nonsense. This finding was later dubbed "Freedman's paradox" (Raftery, Madigan, and Hoeting 1993).

Chatfield (1995) used the term *model selection bias* to describe the distorted inferences that result when using the same data that determines the form of the final model to also produce inferences from that model. Parameter estimates themselves may be biased (e.g., Hjorth 1989; Berk, Brown, and Zhao 2010); their uncertainties (standard errors and confidence intervals) underestimated; and prediction intervals will be too narrow. Thus, the highly precise claims of statistical inference tend to be misleading if taken at face value; indeed, Gelman (2016) used the term "uncertainty laundering" to describe this behavior.

Closely related is the term *model uncertainty*, which reflects the fact that a statistical model can rarely be prespecified (as known) without alteration in light of the data, with the important exception of statistical models used in Phase III clinical trials (discussed further in Section 3). "It is indeed strange that we often admit uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true" Chatfield (1995) says. Moreover, this source of uncertainty, which is invisible to conventional statistical inference, can be the largest component of uncertainty. We refer readers who wish to pursue the sizable literature on model uncertainty and the Optimism Principle to two influential review/discussion articles, Draper (1995) and Chatfield (1995), and the more recent discussions by Berk, Brown, and Zhao (2010), Gelman and Loken (2014), and Holmes (2018).

While these problems are often attributed to misuses of statistical methodology, Gelman and Loken (2014) noted that even *without* a conscious effort to abuse statistical inference (e.g., testing multiple hypotheses but only reporting the significant ones), it is still possible for a null effect to appear highly significant:

> Given a particular data set, it can seem entirely appropriate to look at the data and construct reasonable rules for data exclusion, coding, and analysis that can lead to statistical significance. In such a case, researchers need to perform only one test, but that test is conditional on the data …with the same effect as if they had deliberately fished for those results.

The choice of what test to carry out, even if only one such test is made, is data-dependent. Had a different dataset obtained, a different choice could have been made. "This error carries particular risks in the context of small effect sizes, small sample sizes, large measurement error, and high variation," they wrote. However, such issues also exist in "big data" problems, such as functional imaging in neuroscience, where damaging reuse of the data is known as "double dipping" (Kriegeskorte et al. 2009). Gelman and Loken (2014) used the term "the garden of forking paths" to describe the many potential data-dependent choices a researcher can make during data analysis, resulting in "a sort of invisible multiplicity: data-dependent analysis choices that did not appear to be degrees of freedom because the researchers analyze only one data set at a time."

Gelman and Loken (2014) did *not* decry the researcher's ability to refine hypotheses in light of the data, for on the contrary, such activity is actually "good scientific practice," they said. The only trouble is that the use of statistical inference to guard against being fooled by randomness just does not work when a necessarily flexible data analysis procedure is pursued. Gelman and Loken (2014) "did not want demands of statistical purity to strait-jacket our science," and we completely agree. Making the distinction between exploratory and confirmatory objectives helps us understand where either flexibility or statistical purity are called for in study design and analysis.

3. Exploratory and Confirmatory Objectives in Scientific Research

The obvious way to avoid the difficulties of overfitting and produce valid statistical inferences is to completely prespecify the study design and statistical analysis plan prior to the start of data collection. This can only be done once a great deal is known about the scientific problem at hand. Hence, most scientific research occurs long before it is possible to entertain such rigid prespecification.

Tukey (e.g., 1969, 1977) made a distinction between exploratory and confirmatory analyses. We contend that most scientific research is *exploratory* in nature: the 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. In particular, inferences from such models provide no ability to evaluate "statistical significance"—the implausibility of chance as an explanation of the data. Formal statistical inferences are only valid and appropriate for *confirmatory* analyses, where rigid prespecification of design and analysis methods can be made,

and which constitute only the latter stages of the iterative learning process that characterizes most scientific research (Box 1976, 1999). As medical researchers Mogil and Macleod (2017) put it, "most preclinical research articles describe a long chain of experiments, all incrementally building support for the same hypothesis."

In George Box's discussion of Draper (1995), he made an observation that provides the motivating philosophy of this article:

> Statistics has no reason for existence except as a catalyst for scientific enquiry in which only the last stage, when all the creative work has already been done, is concerned with a final fixed model and a rigorous test of conclusions. The main part of such an investigation involves an inductive-deductive iteration with input coming from the subject-matter specialist at every stage. This requires a continuously developing model in which the identity of the measured responses, the factors considered, the structure of the mathematical model, the number and nature of its parameters and even the objective of the study change. With its present access to enormous computer power and provocative and thought-provoking graphical display, modern statistics could make enormous contributions to this—the main body of scientific endeavour. But most of the time it does not.

There is, of course, one arena of science where the exploratory/confirmatory distinction is clearly made, and attitudes toward statistical inferences are sound: the phased experimentation of medical clinical trials. The epistemological framework of exploratory and confirmatory objectives articulated in the ICH guidelines, E8 and E9 (International Conference on Harmonisation 1997, 1998), seems to be unique in scientific research. This framework helps to separate therapeutic exploratory (typically Phase II) with therapeutic confirmatory (typically Phase III) objectives (see *Appendix* A.2). The latter are intended to inform licensing decisions, and imply comprehensive prespecification of (and adherence to) the study protocol, standard operating procedures, and statistical analysis plan, including writing the analysis software code, prior to collecting any data. Exploratory objectives may be pursued in earlier phase trials, where data-driven choices may influence trial conduct and analysis in a flexible way. However ICH E9 states, "Such trials cannot be the basis of the formal proof of efficacy, though they may contribute to the total body of relevant evidence." Although statistical inferences are reported in these earlier phase trials, their cognitive status is very different: they are considered much less definitive than inferences from later phase trials, and licensing decisions are *not* typically made solely on their basis. Moreover, even for Phase III trials, usually at least *two* are required for a New Drug Application. As Piantadosi (2017) noted, "Medicine is a conservative science and behavior usually does not change on the basis of one study."

Nonetheless, the ICH guidance acknowledges that a clinical trial can have both exploratory and confirmatory objectives. In a confirmatory trial, once the prespecified analysis of the primary endpoint is complete, a large, rich, and expensive dataset typically remains at hand. It seems responsible to use that dataset for further exploratory work—to generate hypotheses for further testing in later experiments. A *subgroup analysis*, for example, may seek to identify a subpopulation (perhaps defined by the presence of a particular biomarker at elevated levels) for which efficacy is particularly pronounced (or suppressed). Statistical inferences are often reported, but like Phase II inferences, they only serve as fodder for proposing and designing future confirmatory trials, rather than for defining the current label indication. A succinct perspective on such inferences is given by Sir Richard Peto, often quoted (e.g., Freedman 1998) as saying "you should always do subgroup analysis and never believe the results."

This shift in attitude about the cognitive status of statistical inferences in exploratory versus confirmatory analyses is wholly absent from much of the rest of scientific endeavor. Typically, flexible data analysis methods are used to generate biased statistical inferences that are taken at face value, and used to justify publication decisions. However, in an interview (Shell 2016), the then-Editor of *Science*, and now President of the U.S. National Academy of Science, Marcia McNutt, stated:

> At *Science*, the paradigm is changing. We're talking about asking authors, 'Is this hypothesis testing or exploratory?' An exploratory study explores new questions rather than tests an existing hypothesis. But scientists have felt that they had to disguise an exploratory study as hypothesis testing and that is totally dishonest. I have no problem with true exploratory science. That is what I did most of my career. But it is important that scientists call it as such and not try to pass it off as something else. If the result is important and exciting, we want to publish exploratory studies, but at the same time make clear that they are generally statistically underpowered, and need to be reproduced.

This call for reproduction is reminiscent of the phased clinical trials framework, where multiple studies are carried out, in series and sometimes in parallel, to build an evidence base for a licensing decision of a medical product. Piantadosi (2017) reminded us that "Readers of clinical trials tend to protect themselves by reserving final judgment until findings have been verified independently or assimilated with other knowledge." A single set of data can rarely give definitive results.

4. From the Cult of the Isolated Study to Triangulation

The treatment of statistical inferences from exploratory research as if they were confirmatory enables what Nelder (1986) called *The Cult of the Isolated Study*, so that

> The effects claimed may never be checked by painstaking reproduction of the study elsewhere, and when this absence of checking is combined with the possibility that the original results would not have been reported unless the effects could be presented as significant, the result is a procedure which hardly deserves the attribute 'scientific.'

The use of iterative experimentation and analysis, described by Box above, is obviously more labor intensive, time-consuming, and expensive than worshipping at the Cult of the Isolated Study, but it is more likely to converge to reproducible results. However, we must depart from Tukey's (1977) view that "exploratory and confirmatory can—and should—proceed side by side," and

Gelman's (2003) comment that they "can both be applied at various stages of the analysis." Instead, we agree with Andrew Ehrenberg's (1990) hope for the future of statistics: *Many Sets of Data* (MSOD), which "seems the only way in which we can produce results that are generalizable, lawlike, and predictable—which in fact hold for many different sets of data." Chatfield (1995) added that "The (over?) emphasis on analyzing single sets of data permeates the statistical literature and is a serious disease of statistical teaching." As Freedman (1991) also noted:

> Generally, replication and prediction of new results provide a harsher and more useful validating regime than statistical testing of many models on one data set. Fewer assumptions are needed, there is less chance of artifact, more kinds of variation can be explored, and alternative explanations can be ruled out.

That said, simple replication is usually not sufficient. Many scientific theories have implications that can be tested in multiple ways, as illustrated by John Snow's work on the 1854 cholera outbreak in London, a major step in developing the germ theory of disease. Freedman (1991) wrote of this episode that "The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data." Moreover "He made steady progress from shrewd observation through case studies to analysis of ecological data. In the end, he found and analyzed a natural experiment." There were many sets of data of different kinds, but no statistical inferences were involved. A second example is the discovery of the link between smoking and lung cancer, built on many epidemiological studies of varying designs, again discussed by Freedman (1999). He wrote that because of the doubtfulness of the usual statistical model assumption that patients in a case–control study constitute a random sample,

> Scientifically, the strength of the case against smoking rests not so much on the P-values, but more on the size of the effect, on its coherence and on extensive replication both with the original research design and with many other designs. Replication guards against chance capitalization and, at least to some extent, against confounding—if there is some variation in study design.

In addition he wrote, "Great care was taken to exclude alternative explanations for the findings. Even so, the argument depends on a complex interplay among many lines of evidence." A third example is the invention of powered flight by the Wright Brothers, discussed by Box (1999). The Wright Brothers conducted a lengthy series of experiments, using wind tunnels, kites, gliders, and finally powered aircraft. No formal statistical modeling or inference was used. Such examples support the assertions made by William Feller (1969):

> The aim of basic research is not to produce statistically valid results but to study new phenomena. An evaluation of experimental findings depends on many factors, such as compatibility with other results, predictions to which it leads and so on—such evidence can rarely be evaluated statistically.

Munafo and Davey Smith (2018) define *triangulation* as "the strategic use of multiple approaches to address one question. Each approach has its own unrelated assumptions, strengths and weaknesses. Results that agree across different methodologies are less likely to be artifacts."

A particular weakness of the Isolated Study is that systematic errors may contaminate an entire study but remain hidden if no further research is done. An example is the notorious report by the OPERA collaboration of a detection of faster-than-light neutrinos, with a claimed statistical significance of six sigmas. The team even reproduced this finding using their own apparatus. However, the collaboration later discovered an error due to an improperly connected fiber optic cable. Correcting for this error eliminated the superluminal claim (the OPERA collaboration 2012). Several other examples from physics and astronomy, results with unambiguous statistical significance that "crumbled to dust" due to systematic errors, are given by Seife (2000).

Lithgow, Driscoll, and Phillips (2017) began an instructive tale with initially nonreproducible results on worm lifetimes among different labs, an example of systematic lab-to-lab differences swamping any claims of statistical significance. Three different labs, to achieve reproducible results among themselves, examined and standardized a host of experimental sources of variation, such as

- Lighting and temperature in the labs.
- Amount of heat emitted by microscopes in different labs.
- Stirring versus rocking in the cell isolation procedure.
- Procedure for picking up worms to place them in a new agar dish. (Gentler technicians added a day to worm lifetime!)
- Positioning of flasks in autoclave runs.
- Defining the worm lifetime: starting from the laying of an egg, or from the hatching of an egg?

Lithgow, Driscoll, and Phillips (2017) concluded that such factors had to be addressed more vigorously at the study design stage to improve reproducibility. More commonly the influence of such technical artifacts is discovered post hoc, for instance, through comparison of datasets from different labs, if they are discovered at all. (See also the related discussion of batch effects in Section 7.4.)

Regrettably, statisticians have continued attempts to salvage the validity of statistical inference from the Isolated Study and the "final model." This makes such systematic errors an inherent risk of the Isolated Study paradigm (see Youden 1972; Bailey 2018). Before we discuss positive ideas on how to proceed, we first examine (and dispense with) a few examples of such salvaging attempts.

5. Technical Solutions and Their Deficiencies

Many have tried to address the concerns with the Optimism Principle using a set of methods that we call "technical," which here means attempts to use mathematics and computation to salvage valid statistical inference in the Isolated Study. A general framework for all of them is that the data represent, at least approximately, a representative sample of the population of interest, an assumption that itself calls for validation in many scientific contexts—validation that can only be sought with MSOD.

The most widely known class of such methods is based on adjusting for multiple inferences. These range from the simple Bonferroni inequality to the modern methods of false discovery rate and false coverage rate (e.g., Dickhaus 2014). However, in a case study of human microbiome data discussed by Holmes (2018), "if one counts the number of possible analyses on the same data—allowing for the choice of up to nine outliers, different transformations of the data, choice from 40 different possible distances, and five different ordination methods—the result is more than 200 million possibilities. No multiple hypothesis correction can protect the user." Moreover, these methods do not account for other researcher degrees of freedom, data-dependent choices that influence model selection, which Gelman and Loken (2014) called "invisible multiplicity."

A second class of methods incorporates resistance to overfitting into the statistical modeling process, often through an optimization procedure that penalizes model complexity, an approach sometimes called *regularization*. A signature example is the *lasso* (Tibshirani 1996). While such methods can certainly mitigate overfitting, it remains unclear how badly mis-calibrated the resulting statistical inferences remain. Similar arguments could be made for *robust* statistical methods (designed to be less sensitive to some model assumptions) and *nonparametric* methods (designed to minimize model assumptions). Such methods must still fail to guarantee reliable statistical inference, because they cannot eliminate model uncertainty and systematic error, though they are safer to use than conventional statistical methodology in exploratory data analysis, which we discuss in Section 7.4.

Among other proposed remedies are *data splitting*, to which we alluded earlier in the context of prediction, and various versions of *cross-validation* (see also Stone 1974; Picard and Cook 1984; Faraway 2016). These methods are also widely used in machine learning (e.g., Hastie, Tibshirani, and Friedman 2009). Data splitting can also be considered for more traditional statistical inferences like hypothesis testing (e.g., Cox 1975; Dahl et al. 2008). (Cross-validation is discussed in *Appendix* A.3. Another approach, using bootstrap methodology, is discussed briefly in *Appendix* A.4.) Unfortunately, such procedures (or their variants) are still vulnerable to the Optimism Principle, because random splitting implies that "left-out" samples are similar to the "left-in" samples (Gunter and Tong 2017). Put another way, if "a dataset from model A happens to have features which suggest model B, then the resampled data are also likely to indicate model B rather than the true model A" (Chatfield 1995). Obtaining "more than one set of data, whenever possible, is a potentially more convincing way of overcoming model uncertainty and is needed anyway to determine the range of conditions under which a model is valid" (Chatfield 1995).

Another widely advocated category of technical solutions is *model averaging*, which comes in both Bayesian (Hoeting et al. 1999; Fragoso, Bertoli, and Louzada 2018) and frequentist (Hjort and Claeskens 2003) flavors. This approach acknowledges the futility of assuming a single "true" model; instead many models are fit to one (isolated) dataset, and their outputs are averaged in some fashion. All entertained models must be known and specified; the Bayesian version also requires their priors. This approach doubles down on the notion of reusing a single set of data, and inherits all the hazards of the Cult of the Isolated Study.

As "post model selection inference" is currently an active area of statistical research, there are still other different approaches, some of which are reviewed by Holmes (2018). A common theme of such methods is that an initial full model needs to be prespecified. Once this is done, inferences from a selected final model can be made that account for variable selection, if that model is a submodel of the initially specified one. The work of Berk et al. (2013), Lee et al. (2016), and the Bayesian methods discussed by Efron and Hastie (2016), seem to share this common feature. Such proposals do not capture the full range of "invisible multiplicity" and model uncertainty that we discussed earlier, which cannot be boiled down to just variable selection. Taylor and Tibshirani (2015) conceded as much: "The challenge of correcting for the effects of selection is a complex one, because the selective decisions can occur at many different stages in the analysis process." Like the others cited here, they focus on "more limited problems." Another theme is conditional inference, conditioning on either the selection itself (Lee et al. 2016) or on the training data (Leeb 2009). Such conditioning severely limits the interpretation and generalizability of statistical inferences in the setting of exploratory research and many sets of data.

Only through the iterative learning process, using multiple lines of evidence and many sets of data, can systematic error be discovered, and model refinement be continually guided by new data. Moreover, a retreat into mathematical and computational "remedies" can distract us from interactively confronting the natural phenomena under investigation by acquiring more and better data, under an increasingly wider range of conditions, and with constantly improving experimental methodology. For example, Galileo's late 16th century "Leaning Tower of Pisa" experiment, showing that bodies of different mass and composition fall at the same rate, has been tested under various scenarios for over four centuries, including by Apollo 15 astronaut David Scott (Allen 1972) and, most recently, on board the MICROSCOPE satellite (Touboul et al. 2017; see further references therein for other historical experiments). Incidentally, this example also suggests that our use of the term "confirmatory" is a verbal shorthand, not meant to imply definitive confirmation. All scientific findings are tentative. A "confirmatory" study and analysis are simply those designed so that previously established models may be fit and prespecified effects may be estimated with greater rigor than in a flexible setting. (Experience with late-stage clinical trials shows that even the knowledge they generate must still be considered tentative, e.g., Gauch 2009.)

6. More Thoughtful Solutions

A second set of solutions considers the larger framework for scientific research, not just statistical inference methodology.

One strategy requires *preregistering* both the research hypotheses to be tested and the statistical analysis plan prior to data collection, much as in a late-stage clinical trial (e.g., Nosek et al. 2018). If the temptation to adjust the analysis in light of the data can be resisted, this approach inherently avoids the flexible, data-dependent analysis choices that can

lead to overfitting, regardless of the style of inference being practiced (e.g., frequentist or Bayesian). Indeed, Nosek et al. (2018) argued that such preregistration helps clarify the distinction between hypothesis-generating and hypothesis-testing activities. Hypothesis-generating activities are not ruled out, but are intended for future confirmatory hypothesis-testing, as in the clinical trials framework.

While such an approach is commendable, the simple fact is that much—indeed most—scientific research cannot fit this paradigm (e.g., Scott 2013; Goldin-Meadow 2016). As we have emphasized, most scientific research is (and should be) highly exploratory: this means that not enough can be known a priori to prespecify a fully formed statistical analysis plan. Moreover, as Gelman and Loken (2014) said, we "do not want demands of statistical purity to strait-jacket our science" by limiting the ability to explore modeling alternatives. If science is to be evidence-driven, research must be open to what the data tell us, rather than rigidly committed to a prespecified analysis plan based on prior expectations. As Feller (1969) wrote, "No statistics should stand in the way of an experimenter keeping his eyes open, his mind flexible, and on the lookout for surprises."

A variation on this theme is *preregistered replication*, where a *replication* study, rather than the original study, is subject to strict preregistration (e.g., Gelman 2015). 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 converges to a single working hypothesis, it can *then* be subjected to a preregistered, randomized, and blinded, appropriately powered confirmatory experiment, carried out by another laboratory, in which valid statistical inference may be made. (Unlike with preregistered replication, the confirmatory study here need not be a literal replication of one of the earlier exploratory studies.) The key is that publication of the exploratory experiments would require that they be accompanied by the confirmatory study in the same manuscript. Mogil and Macleod (2017) explained how this publication model may change the incentives for research for the better. The proposal is appealing, but there are still many challenges and situations (such as observational data studies) that might not fit it well (Gunter and Tong 2017). For example, when a confirmatory study fails to go according to plan, troubleshooting should be prioritized over statistical validity. When this happens, the study simply reverts to becoming another in the series of exploratory studies, rather than confirmatory, as originally intended.

In summary, then, we have argued that the paradigms and assumptions of statistical inference do not fit the inherently exploratory nature of science, and therefore should rarely be applied. We next examine the many positive contributions of statistical methodology.

7. Enabling Good Science

7.1. A Taxonomy of Statistical Activity

We adapt a taxonomy of statistical activity that has previously been used, in different forms, by Cox (1957) and Moore (1992):

- Data production. The planning and execution of a study (either observational or experimental).
- Descriptive and exploratory analysis. Study the data at hand.
- Generalization. Make claims about the world beyond the data at hand.

Data production includes the experimental design, sampling plan, measurement and data collection procedures, and the operational conduct of the study. Descriptive and exploratory analysis includes summary statistics, statistical graphics and tables, and disciplined data exploration. Generalization includes both prediction and statistical inferences. Much of current statistical teaching disproportionately focuses on statistical inference, just one piece of this taxonomy, one that we have argued is the least appropriate for most scientific research. As Freedman (1995) lamented:

> I wish we could learn to look at the data more directly, without the fictional models and priors. On the same wish list: We should stop pretending to fix bad designs and inadequate measurements by modeling.

However, statistical modeling can still contribute to all three sectors of the taxonomy. For example, statistical models help us understand the advantage of factorial designs over one-factor-at-a-time designs; succinctly characterize potential patterns observed in the dataset, such as linear or curved relationships among variables; and produce statistical inferences and predictions for confirmatory analyses. Also contributing to all three sectors is the amorphous notion of *statistical thinking*, which we try to partially characterize in Section 7.5.

The first step of statistical thinking is to understand the *objective* of the study, its *context*, and its *constraints*, so that planning for study design and analysis can be fit for purpose. In this respect, the use of statistical inference as a universal mechanism for scientific validity must be replaced by mainly noninferential statistical methods that are discipline- and problem-specific (Gigerenzer and Marewski 2015). Despite this, the thoughts below may yet be of broad general interest to data analysts in many disciplines. Only the first two items in the taxonomy are discussed here, since as we argue above, generalization best emerges from the iterative learning process described by Box (Section 3). (Statistically literate readers may skip Section 7.2, which is intended for nonspecialist readers.)

7.2. Data Production

Feller (1969) pronounced that "The purpose of statistics in laboratories should be to save labor, time, and expense by efficient experimental designs" rather than null hypothesis significance testing. Some useful basic principles of study design include *concurrent control*, *replication*, *randomization*, *blinding*, and *blocking*. Each of these is not always mandatory (or possible), but they should be carefully considered. In a comparative study, for instance, a group of subjects exposed to an experimental intervention is compared to a control group. Attributing causality of outcomes to the intervention can then be made only if the two groups were treated the same in other respects. Use of a *concurrent control* group helps compensate for certain sources of

systematic errors that may occur, for instance, if the two groups are widely separated in time and/or space. *Random allocation* of subjects to treatment groups helps "even out" other systematic differences between groups that may not be known or measured. *Blinding* of subjects, and others involved in the study, to their assigned treatment reduces the likelihood of unconscious bias. Together, these mechanisms also account for the placebo effect, when present. In many observational studies, *random sampling* (a guarantee that each subject in the target population has an equal chance of being selected) helps to reduce sampling bias in the results. Finally, *randomization of treatment and measurement order* helps to average out the corresponding systematic effects, such as learning effects in a study involving radiologists' interpretation of medical images. Together, these concepts are aimed at bias suppression, including bias due to unknown or unobserved sources, that contribute to misleading conclusions. The impact of such simple procedures can be substantial: Couzin-Frankel (2013) observed that only about 1/3 of mouse studies of stroke therapeutics report randomization or blinding, but studies reporting neither one "gave substantially and significantly higher estimates of how good these drugs were," according to interviewee Malcolm Macleod. In one case the same drug had twice the effectiveness in a study *without* randomization as in one that *did* randomize.

Adequate *replication* allows us to get a handle on experimental variation, such as variation between and within subjects, and variation due to the measurement process. The notion of an *experimental unit* helps us understand the replication structure of the experiment. The experimental unit is the smallest subset of experimental materials or subjects that can be randomly allocated to separate treatments. For instance, if experiments are done with mice in a cage, and treatments are administered through a common food tray in each cage, then the experimental unit is the cage, not the individual mice (which are usually the units of *observation* or *measurement*). Treatments can only be administered at the cage level, not the mouse level. Such a design is often necessary since mice are social, and ideally should live with their litter-mates, not in single-animal cages. (Statistical models can accommodate such nuances during the analysis.) In many such designs, several units of replication may be identified, and the identification of *one* of these as the experimental unit for analysis purposes is an important and sometimes controversial issue impacting scientific interpretability. As another example, a cluster-randomized trial of a new pedagogical method could have schools as units of allocation, classes (within schools) as units of intervention, and students (within classes) as units of measurement (Murray 1998). In biology, we often make the distinction between *biological* replicates (samples from different people or animals) and *technical* replicates (replicate measurements on the same sample); often both terms must be enriched depending on the context.

The concept of *blocking* recognizes that experimental units or runs are not all equally alike. There are subgroups of units or runs whose data are more similar to each other than with those of other subgroups of units or runs. Examples include plots of land in close proximity, in an agronomy study involving multiple fields; litter-mates in animal studies involving multiple litters of animals; experimental runs made on the same day versus those from other days; and clinical trial subjects enrolled close to each other in time and at the same clinical site. These known sources of variation can be accounted for in both the study design and analysis, to improve the quality of information obtained from the experiment. Randomization should still be applied *within* these subgroups ("blocks") when they have been identified. As the classic text, Box et al. (2005), advised: "Block what you can and randomize what you cannot." (In observational studies, *stratified sampling* plays an analogous role to blocking.)

Confounding occurs when the treatment groups differ systematically in some way besides the intervention. Randomization, blinding, and blocking help to eliminate some sources of confounding. When prognostic covariates are known for enrolled subjects, methods for allocating subjects to treatments that seek approximate balance on such covariates can further reduce the risk of confounding (e.g., Lock Morgan and Rubin 2012; Kallus 2018).

More advanced notions of experimental design include matched-pairs designs, factorial designs, split-plotting, crossover trials, and many other concepts that we have not the space to discuss here, but may be appropriate in different situations. Beyond these standard topics of study design, other aspects of data production must be considered. Measurement processes must be well-defined, standardized, and validated, with quality assurance procedures in place. Data acquisition and storage systems should have appropriate resolution and reliability. (We once worked with an instrument that allowed the user to retrieve stored time series data with a choice of time-resolution. Upon investigation, we found that the system was artificially interpolating data, and reporting values not actually measured, if the user chose a high resolution.) Again, many issues will be discipline- and problem-specific. In biological research for instance, authentication of cell lines and validation of antibodies can address major sources of systematic error (see Harris 2017).

Earlier we described "researcher degrees of freedom" that are characteristic of flexible data analysis. Other researcher degrees of freedom can affect study design and execution. An instructive example for the latter is the decision to terminate data collection. Except in clinical trials, where this decision is tightly regulated and accounted for in the subsequent analysis (e.g., Chow and Chang 2012), many researchers have no formal termination rule, stopping when funding is exhausted, lab priorities shift, apparent statistical significance is achieved (or becomes clearly hopeless), or for some other arbitrary reason, often involving unblinded interim looks at the data. *Any* formal or informal rule for terminating data collection, whether an explicit calculation of statistical inference is used or not, has inherent risks of chasing a false positive for too long, or burying a false negative too soon. In the exploratory setting, probability claims about such risks (e.g., "power analysis") are just as invalid as the statistical inferences that obtain during and after data collection.

Principles of study design are a major positive contribution of statistical methodology dating from the time of Fisher (1926), and we feel that much of the reproducibility crisis in science (e.g., Harris 2017) can be addressed using this sector of statistical activity, quite apart from the inference issues that have dominated statisticians' conversations about reproducible research. Many general topics are covered in standard texts on experimental design (e.g., Winer, Brown, and Michels 1991; Kuehl 2000; Box et al. 2005; Montgomery 2017) and sampling

(e.g., Cochran 1977; Thompson 2012) though again each discipline has its own specific issues of data production to deal with. Coleman and Gunter (2014) provided a brief introduction to some of the main ideas of experimental design, and emphasized how varying multiple variables simultaneously can be used profitably, contrary to the conventional wisdom of varying one variable at a time, while keeping all others fixed. Despite the evident utility of all these ideas—and their relevance to many of the problems of nonreproducibility—many scientists seem largely unaware of or confused by them. We therefore think this is fertile ground for improving statistical education for scientists and should be an important component of such efforts.

7.3. Data Description

Data description and exploration are related activities that focus on the data at hand, without attempting to make inferences to the world beyond. Consider first data description. Many data analytical problems do not lend themselves to statistical inference at all: when there is no sense in which the data are even approximately a sample from a population; or when a probability model (conveying some notion of randomness) is irrelevant (Mallows and Walley 1980) or even misleading, as in Taleb's (2007) "ludic fallacy." However almost all data analysis requires at least some data reduction or description, which may include statistical summaries, tables, and visualizations. Moses (1992) warned us that

> Good statistical description is demanding and challenging work: it requires sound conceptualization, and demands insightfully organizing the data, and effectively communicating the results; not one of those tasks is easy. To mistakenly treat description as 'routine' is almost surely to botch the job.

One of the best developed branches of data description is statistical graphics and data visualization, the subject of many books (e.g., Cleveland 1993, 1994; Robbins 2013) and software packages. Scientists could use help in this arena. Perhaps the most egregious yet ubiquitous graphical method found in life science articles is a set of barplots featuring one-sided error bars (sometimes called "skyscraper" or "dynamite" plots), used to represent univariate data from different comparison groups. Such plots are poor representations of the data (Pikounis 2001; Koyama 2011). Krzywinski and Altman (2014) advocated using *boxplots* (Tukey 1977) as a less distorting alternative. Superimposing a representation of the actual data points on a boxplot mitigates some of its deficiencies, such as its inability to indicate bimodality.

Mallows (1983) provided an interesting perspective on a Theory of Description. He noted that "A good descriptive technique should be appropriate for its purpose; effective as a mode of communication, accurate, complete, and resistant." By *effective as a mode of communication*, he included the notions of "familiarity, relative to the target audience, simplicity, and honesty." By *accurate*, he meant "a measure of the closeness to which the description approximates the data," and proposed a discrepancy measure. By *complete*, he meant "the degree to which there is an absence of structure in the residual or undescribed variation in the data." To elaborate, statistical models may be deployed to examine patterns in the data. Deviations between the observed data and the model's fitted values are called *residuals*, and can themselves be examined for patterns suggesting systematic effects not fully captured in the fitted model. By *resistant*, Mallows (1983) meant that methods should minimize "sensitivity …to small changes in the data—either small perturbations in all the data, or arbitrarily large perturbations in a small part of the data." Resistance to distortion by extreme values, potential outliers, is a common example. In this respect the *median* is often used in place of the mean, and the *median absolute deviation* in place of the standard deviation. For curve fitting, *least absolute error* is more resistant than *least squared error*. Mallows acknowledged that his criteria are sometimes in competition with each other. Median absolute deviations are not widely familiar outside of statistics, so using them prioritizes resistance over effectiveness in communication. Nonetheless, his criteria provide a useful guide when designing a data description for a given audience.

Though we might not quantify *uncertainty* using probability statements, we can attempt to convey the observed *variability* of the data at hand, while acknowledging that it does not fully capture uncertainty. For univariate data, quantities such as Tukey's (1977) five-number summary (the minimum, first quartile, median, third quartile, and maximum) or other empirical quantiles partially describe the distribution of the data, as does a graph of the empirical distribution function. In addition, resistant measures of dispersion, such as the median absolute deviation, Gini's mean difference (David 1968), and others (e.g., Rousseeuw and Croux 1993) are available to characterize observed variability of the data at hand. However, the use of such data summaries is not free of assumptions (e.g., unimodality, in some cases symmetry), so they are descriptive only in relation to these assumptions, not in an absolute sense. They cannot serve as measures of uncertainty due to the selection bias inherent in making such assumptions.

The relationship among many variables is less easily characterized by such simple methods; in such cases, "a statistical model is often the best descriptive tool, even when it's not used for inference" (Harrell 2018). For example, bivariate data can be examined with scatterplots and simple curve fits, such as the *loess smoother*, a type of nonparametric regression that incorporates resistance to atypical values (Cleveland 1979). The loess fit is determined by a user-selected parameter called the *span*, and varying this parameter controls the tradeoff between variability captured by the curve fit and variability left over to the residuals. These residuals may be graphed, and when devoid of clear patterns, univariate measures of dispersion such as those just mentioned could be applied to quantify their dispersion from the fitted curve. Of course, the user's choice of the span parameter illustrates that model selection bias remains a hazard. Any quantification of observed variability only describes the relationship between the model and the data at hand. (See Appendix A.5 for a discussion of standard errors as descriptive statistics.)

The use of variability as a benchmark for measuring effect size is often emphasized by statisticians, but is rarely the only benchmark of interest when assessing practical/clinical significance. There are even cases where "variability is so small (or the

data are so abundant) that variability is not the central issue" (Mallows 1998). Two examples (depending on the context, as always) include the power spectrum in time series (as noted by Tukey 1962) and the mean kinetic temperature in applied thermodynamics (see Tong and Lock 2015); examples from materials science are given by Wenmackers and Vanpouke (2012). The selection of appropriate benchmark(s) and the subsequent design of meaningful descriptive statistics are highly problem-dependent and require both a grasp of the subject matter and creative thinking.

7.4. Disciplined Data Exploration

Effective data *exploration* goes beyond data *description*. According to Tukey (1973), exploratory analysis of the data is not "just descriptive statistics," but rather an "actively incisive rather than passively descriptive" activity, "with a real emphasis on the discovery of the unexpected." Later, Tukey (1977) added, "Exploratory data analysis is detective work—numerical detective work—or counting detective work—or graphical detective work." The founding text is Tukey (1977), which was followed by a series known as the *Statistician's Guide to Exploratory Data Analysis* (Hoaglin, Mosteller, and Tukey 1983, 1985, 1991). Pearson (2011) provided a contemporary perspective on exploratory data analysis.

An example of how exploratory analysis may be essential for scientific inquiry is in the detection of and adjustment for *batch effects*. Leek et al. (2010) defined batch effects as "sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study." Their examples include data obtained under varying experimental processing conditions, perhaps from multiple laboratories using different protocols. They found that

> For example, multiple laboratory comparisons of microarray experiments have shown strong laboratory-specific effects. In addition, in nearly every gene expression study, large variations are associated with the processing date, and in microarray studies focusing on copy number variation, large effects are associated with DNA preparation groups. The processing group and date are therefore commonly used to account for batch effects. However, in a typical experiment these are probably only surrogates for other sources of variation, such as ozone levels, laboratory temperatures and reagent quality. Unfortunately, many possible sources of batch effects are not recorded, and data analysts are left with just processing group and date as surrogates.

Leek et al. (2010) also observed that "In gene expression studies, the greatest source of differential expression is nearly always across batches rather than across biological groups, which can lead to confusing or incorrect biological conclusions owing to the influence of technical artefacts." While known batch effects can be accounted for at the study design stage, unknown batch effects may well remain. Statistical methods for the detection of batch effects rely on (noninferential) exploratory analysis using multivariate statistical methods, such as principal components analysis and hierarchical clustering. Adjusting the data for such effects often makes (noninferential) use of linear statistical models. Of course, the comparison of many sets of data, from different laboratories and varying experimental protocols, further helps elucidate batch effects.

Diaconis (1985) warned that undisciplined exploratory analysis of a dataset lends itself to finding spurious patterns as well as real ones, as we discussed earlier. If such patterns are accepted as gospel without considering that they *may* have arisen by chance, he considers it *magical thinking*, which he defines as "our inclination to seek and interpret connections and events around us, together with our disinclination to revise belief after further observation." (See also the related discussion of Grolemund and Wickham 2014.) Statistical methods mentioned earlier that resist overfitting (regularized methods, such as the lasso), are less sensitive to some model assumptions (robust estimators, e.g., Maronna, Martin, and Yohai 2006), or make few such assumptions (nonparametric estimators, e.g., Wasserman 2006) help guard against being fooled by randomness, though their associated probabilistic claims about uncertainty remain unreliable. (A new class of statistical methods with the property of *differential privacy* may also be worthy of further consideration; see Dwork et al. 2016.)

7.5. Statistical Thinking

Mallows (1998) reviewed several definitions of statistical thinking, and provided his own:

> Statistical thinking concerns the relation of quantitative data to a real-world problem, often in the presence of variability and uncertainty. It attempts to make precise and explicit what the data has to say about the problem of interest.

Statistical thinking begins with a relentless focus on *fitness for purpose* (paraphrasing Tukey 1962: seeking approximate answers to the right questions, not exact answers to the wrong ones), sound attitudes about data production and its pitfalls, and good habits of data display and disciplined data exploration. An incomplete list of other attributes of statistical thinking includes the following.

- Evidence, often in the form of data, matters, "for without data, everyone is an expert" (Snee 1986). As Frederick Mosteller said, "It is easy to lie with statistics, but a whole lot easier without them" (quoted in Holmes 2018).
- All observations are variable; all conclusions are uncertain. (However, as discussed above, variability is not always the aspect of the data that should most be emphasized.)
- A holistic view of the data. For instance, we study measurement *processes*, not just measurement *instruments*, which comprise only one component of the whole process of measurement, which also involves personnel, reagents, environmental conditions, etc. Another example is the *Intention to Treat* (ITT) principle in clinical trials, whereby a patient's data are analyzed according to the random assignment of the patient to treatment group, rather than whether the patient actually complied with the protocol (e.g., Piantadosi 2017). This reflects interest in treatment *policy*, including the occurrence of realistic phenomena such as errors and noncompliance.

- Statistical significance is *not* a measure of practical, clinical, or scientific significance.

Statisticians have a useful language and a mathematical apparatus for decomposing sources of variation, both systematic and random. These include the basic bias-variance decomposition, the discussion of replication structure in Section 7.2, and so on, up to complex hierarchical models with both nested and crossed factors and nontrivial correlation structures among variables. Unfortunately, in many cases such statistical models may be difficult to fit, due to lack of data or unrealistic model assumptions. Any inferences from such models are subject to model selection biases and the Optimism Principle. Nonetheless, the paradigm represented by such models is a guide to precise quantitative thinking characteristic of the statistical approach to data analysis. Statisticians are also taught to be sensitive to sources of bias and variability, as well as confounding, issues nicely illustrated in the discussion of batch effects given by Leek et al. (2010), an article in which statistical inference plays almost no role.

Statistical thinking also involves a keen awareness of the pitfalls of data analysis and its interpretation, including:

- The correlation versus causation fallacy.
- The distinction between interpolation and extrapolation.
- The distinction between experimental and observational data.
- Regression to the mean.
- Simpson's paradox, and the ecological fallacy.
- The curse of dimensionality.

We have only outlined some of the elements of statistical thinking, and many readers may have more to add.

8. Discussion

To summarize the argument of this article:

- Most scientific research is exploratory, not confirmatory. The research process is iterative, requiring many sets of data and many lines of evidence (Freedman's "shoe leather" and triangulation). Avoid the Cult of the Isolated Study.
- Attention to statistical issues in the design and execution of the study should be the primary concern. Remain on guard for systematic error.
- 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.
- Formal statistical inference may only be used in a confirmatory setting where the study design and statistical analysis plan are specified prior to data collection, and adhered to during and after it. That is the only setting where we may rely on the Efron and Hastie (2016) principle of using the same data that produces an estimate to assess its precision; in any other setting, statistical inferences are undermined by the Optimism Principle.
- Statistical analysis of exploratory research data should rely only on descriptive methods (summary statistics, tables, and graphics) and disciplined data exploration, which often involves statistical modeling. The latter is often enabled by statistical methods that are regularized, robust, and/or nonparametric, which are safer to use than conventional methodology, but still do not fully eliminate the Optimism Principle. Bear in mind the Harrell (2015) maxim, "Using the data to guide the data analysis is almost as dangerous as not doing so" and Diaconis' (1985) warning about magical thinking.
- The framework outlined by Mogil and Macleod (2017) is an example of an approach to research and publication that seems consistent with the exploratory/confirmatory distinction outlined here.
- Exploratory analysis of data from a confirmatory study (subsequent to the completion of the preplanned analysis) should also be entertained, for the purpose of hypothesis *generation* rather than hypothesis *testing*. Subgroup analysis for late-stage clinical trials is a signature example.
- Statistical thinking provides a framework for critical thinking that can benefit every stage of the research program.

A counterargument to our position is that inferential statistics (p-values, confidence intervals, Bayes factors, and so on) could still be used, but considered as just elaborate descriptive statistics, without inferential implications (e.g., Berry 2016; Lew 2016). We do not find this a compelling way to salvage the machinery of statistical inference. 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. The danger is that both cultural inertia and the seductive appearance of quantified uncertainty may continue to incentivize the inappropriate reporting of statistical inferences.

A second counterargument is that, as George Box (1999) reminded us, "All models are wrong, but some are useful." Statistical inferences may be biased per the Optimism Principle, but they are reasonably approximate (it might be claimed), and paraphrasing John Tukey (1962), we are concerned with *approximate* answers to the right questions, not *exact* answers to the wrong ones. This line of thinking also fails to be compelling, because we cannot safely estimate how large such approximation errors can be. In a single study, hypothesis test selection can lead to underestimation of error by an order of magnitude (e.g., Huber 1985). Recall also Freedman's Paradox, a "final model" with bogus significant variables. Of course unknown systematic error is completely unaccounted for. Many sets of data and triangulation are more reliable ways to explore the approximation error in our (tentative) conclusions.

One reviewer of this article characterized our view as "the proposed solution for imperfect variance estimation is no variance estimation," and then asked "Is no quantification of uncertainty truly better than imperfect quantification?" We think many readers will share this question. In Section 7.3, we made a distinction between characterizing observed variability and quantifying uncertainty. In the exploratory/learning phases of research, it is way too early to pretend to be able to quantify uncertainty. However, we *can* describe how variable the data at hand are, relative to one or more models, using well-chosen descriptive statistics and graphs, though this variability should

not be conflated with uncertainty. Uncertainty also includes model uncertainty, due to model selection bias (the Optimism Principle) and the potential for systematic error, which both require many sets of data to fully evaluate. There is no scientifically sound way to quantify uncertainty from a single set of data, in isolation from other sets of data comprising an exploratory/learning process.

Tukey (1962) once suggested that "it might be well if statisticians looked to see how data was actually analyzed by many sorts of people." In fact, some statisticians have shown just such an interest, for example, Box (1999) and Freedman (1991, 1999), as discussed earlier. For the most part, however, statistics teaching and practice ignores the fact that most of the great historical discoveries in science, engineering, and medicine were made without the crutch of statistical inference (Gigerenzer and Marewski 2015). Kepler's laws of planetary motion, the periodic table of the elements, the germ theory of disease, plate tectonics, the molecular structure of DNA, and the quantization of energy are other examples where data and modeling, without statistical inferences, were crucial for discovery. Financial accounting, a data profession much older and larger than statistics, often depends on estimation in an uncertain setting, but does not report any precise uncertainty quantification in the balance sheet. This brings to mind an observation made about certain research in materials science: "Even if the studies had reported an error value, the trustworthiness of the result would not depend on that value alone" (Wenmackers and Vanpouke 2012).

Nevertheless, we do think that the discipline of statistics has much to offer science. By focusing on the methods that are of broadest usefulness to science, engineering, and medicine—rather than on an obsession with statistical inference and the Cult of the Isolated Study—statistical thinking has much to contribute to scientific work. By emphasizing principles of data production, data description, enlightened data display, disciplined data exploration, and exposing statistical pitfalls in interpretation, there is much that statisticians can do to ensure that statistics is "a catalyst to iterative scientific learning" (Box 1999). *Statistical inference enables bad science; statistical thinking enables good science.*

Appendix: Further Elaborations

A.1. Frequentist and Bayesian Schools of Inference

There are several schools of thought on how statistical inference should be carried out (e.g., Barnett 1999). For instance, the survey by Geisser (2006) focuses on the four most prominent: frequentist, Bayesian, likelihoodist, and fiducial. The first two of these are by far the most often used in practice, and we describe them briefly here. Bear in mind that within each school of thought, there is further variation of both perspective and methodology.

The frequentist approach is based on the interpretation of probability as a fixed long-run frequency, imagined as resulting from replication of the data generating process ad infinitum. This requires defining a fixed probability space a priori, which in turn requires knowing the intentions of the investigator (Berger and Berry 1988), a setting wholly inconsistent with flexible data analysis (Gunter and Tong 2017). Frequentist inference is only conceivable for prespecified confirmatory analyses.

The Bayesian approach views probability as a subjective degree of belief. A Bayesian analysis begins by specifying prior probability distributions for the parameters in a stochastic model, then updating these probabilities in light of the data (using Bayes' theorem, and a model-dependent quantity known as the *likelihood function*). It represents a computationally intensive set of procedures attempting to capture the intuition that prior knowledge should be combined with current data to make current knowledge. This sounds deceptively consistent with the iterative learning process advocated in this article. However, it is not. The placement of prior distributions on model parameters is a highly artificial way of embedding prior knowledge; few scientists formulate their beliefs in terms of real-valued (i.e., infinitely precise) probability functions. The more natural (and more important) way that prior knowledge is incorporated (for both frequentists and Bayesians) is in the choice of study design (including what variables to measure, and how to measure them) and the class of models the analyst is prepared to entertain (Harrell 2018). Harrell (2018) noted that for the Bayesian "the model choice does not 'wear off' nearly as much as the prior does as the sample size gets large," so this choice is more impactful than the choice of priors. However, this model choice remains subject to researcher degrees of freedom and the Optimism Principle.

In the iterative learning process described by Box (Section 3), there is no fixed model whose parameters can be repeatedly updated, because the model itself changes as the research program evolves. In the most extreme cases, such as alleged scientific revolutions (Kuhn 1970), the incommensurate paradigms cannot transfer information to each other through prior probabilities. Bayesian inference cannot guide an entire research program, but like frequentist inference, may be of value within the confines of a confirmatory analysis.

Harrell (2018) observed that even in confirmatory analyses, model uncertainty remains, because we never know if the model is misspecified. He argued that Bayesian inference can accommodate some degree of model uncertainty in this setting: "If the model contains a parameter for everything we know we don't know (e.g., a parameter for the ratio of variances in a two-sample t-test), the resulting posterior distribution for the parameter of interest will be flatter, credible intervals wider, and confidence intervals wider. This makes them more likely to lead to the correct interpretation, and makes the result more likely to be reproducible." This approach may be considered by those whose response to model uncertainty is to double down on probability modeling. Others might instead back off, and rely instead on our suggestions for exploratory analysis: regularization, robustness, and nonparametricness. The differences in interpretation among the various styles of statistical inference are another factor to consider.

A.2. Phased Clinical Trials

The framework of phased clinical trials is an exemplar of scientific epistemology. The ICH E8 states:

> The cardinal logic behind serially conducted studies of a medicinal product is that the results of prior studies should influence the plan of later studies. Emerging data will frequently prompt a modification of the development strategy.
> […]
> Drug development is ideally a logical, step-wise procedure in which information from small early studies is used to support and plan larger, more definitive studies. To develop new drugs efficiently, it is essential to identify characteristics of the investigational medicine in the early stages of development and to plan an appropriate development based on this profile.

> Initial trials provide an early evaluation of short-term safety and tolerability and can provide pharmacodynamic and pharmacokinetic information needed to choose a suitable dosage range and administration schedule for initial exploratory therapeutic trials. Later confirmatory studies are generally larger and longer and include a more diverse patient population. …Throughout development, new data may suggest the need for additional studies that are typically part of an earlier phase.
>
> To oversimplify, the typical stages of clinical trials are as follows:
>
> - Phase I. This represents the first administration of the drug in humans, usually to a small number of healthy volunteers. The goal is to evaluate tolerability of the dose range, observe potential side effects, and evaluate clinical pharmacokinetics, such as drug absorption, distribution, metabolism, and excretion (ADME).
> - Phase II. These studies are typically therapeutic exploratory studies. Patients are enrolled under "relatively narrow criteria" to obtain data on drug efficacy, safety, and dose response. Flexibility in design and analysis may be entertained.
> - Phase III. These studies are typically therapeutic confirmatory studies. A large number of patients, from a wider population than Phase II, are enrolled to confirm drug efficacy and safety. Very tight specification of study protocols, standard operating procedures, and statistical analysis plans should be made prior to data collection, and adhered to during and after it. There is rightful indignation when the prespecified analysis is altered for publication purposes (see *http://compare-trials.org/*).

For a more detailed exposition, consult the guidance documents ICH E8 and E9 (International Conference on Harmonisation 1997, 1998) and textbooks such as Piantadosi (2017). (See also Scheiner 1997, for another perspective.)

The intellectual heritage of phased clinical trials is not statistical, but statisticians have helped sharpen the reasoning this framework embodies. Carpenter (2010, chap. 4) traced the origins of phased clinical trials to pharmacologists at the U.S. Food and Drug Administration (FDA), such as A.J. Lehman and O. Garth Fitzhugh, and clinical oncologists at the National Cancer Institute (NCI), who made the distinction between Phases I and II in the 1950s. The notion of Phase III was formulated by FDA physicians, attorneys, and other officials, around the time that Congress passed the 1962 Kefauver-Harris amendment to the 1938 Federal Food, Drug, and Cosmetic Act. Phases I, II, and III first appear in formal regulatory guidance in the FDA's draft Investigational New Drug rules of January 1963. Carpenter writes of the latter, "The authorship of phased experimentation appears to have been collaborative, with [Julius] Hauser and [Frances] Kelsey assuming crucial leadership roles." Of particular interest to statisticians, the drafting of the ICH E9 guidance in the 1990s is described briefly by Lewis (1999).

A.3. Cross-Validation

Hastie, Tibshirani, and Friedman (2009, p. 241) wrote that "Probably the simplest and most widely used method for estimating prediction error is cross-validation." As an example, we consider five-fold cross-validation (5CV), which proceeds as follows.

1. Randomly partition the dataset into five roughly equal sized subsets.
2. For each of the five subsets separately, perform model fitting/criticism/selection on an aggregation of the other four subsets, and then use the left-out fifth to evaluate performance of the model, as in data splitting. The model fitting procedure should be coded to run automatically (without human intervention) so that it can be called in each iteration. (Of course, such an automated procedure could not reflect the human judgment usually applied in actual model building.)
3. After this has been done five times, each data record has one predicted value, from a model built on 4/5 of the dataset, that can be compared with the actual value. Prediction performance can be then be evaluated by comparing this whole set of predicted values with the actual values. The resulting performance estimate does not correspond to any particular one of the five models that was trained, but to the automated modeling *procedure* that was coded.

A.4. Bootstrap Methodology

Bootstrap methodology is a signature achievement of statistical inference in the late 20th century, and it illustrates the maxim (disputed here) that the same data used to provide an estimate can be used to assess its precision (Efron 1982). However, our interest in the bootstrap here is in its use to examine model selection bias in various ways (e.g., Efron 1983; Gong 1986; Dijkstra 1988; Efron 2014). A bootstrap resampling procedure for assessing prediction performance could operate as follows. Generate multiple bootstrap samples, which are samples with replacement from the original dataset. On each of these, fit a statistical model, then assess the model's performance on the original dataset. (Like cross-validation, this requires coding an automated model building procedure that can be called iteratively.) An overall estimate of performance is obtained by averaging over the results from all bootstrap samples. An enhanced approach is to use the bootstrap to evaluate the expected bias due to overfitting. Harrell (2015), sec. 5.3, provided a more thorough discussion than we can give here. However, the criticisms we raised earlier about data splitting and cross-validation mostly extend to such bootstrap methods as well. As Harrell (2018) observed, all these methods attend to *internal* not *external* validation, and it is external validation that matters most to the user. However, Harrell (2018) argued that among the *internal* validation methods, the bootstrap is least wasteful of the data.

A.5. Standard Errors as Descriptive Statistics

For univariate data, the standard error is often used as an error bar for the mean. In regression modeling, regression coefficients are sometimes reported with standard errors to characterize their precision.

As Motulsky (2014) observed, the standard error is *not* a measure of variability, since it shrinks with sample size regardless of how variable the data are. Rather, the standard error is a measure of *precision* for a point estimate under certain probability models. From this perspective, the standard error is associated with inferential probability statements derived from a statistical model, and such statements are often invalid due to the Optimism Principle. For this reason, we advise avoiding the standard error even as a descriptive statistic. Regardless, often the variability of the data themselves (think standard deviation, not standard error) can be a practically or clinically meaningful benchmark. (For the cognoscenti, Cohen's D may be more meaningful than a t-statistic, though neither properly characterizes uncertainty.)

A.6. Recommended Reading

Even in a manuscript of this length, we have not been able to rehearse the compelling arguments made by earlier writers with full justice. It would be pointless to do so when the original writers have been so eloquent. Readers interested in specific topics are invited to track down the literature cited in sections of this article that interest them. More generally, if we had to recommend just three articles that capture

the spirit of the overall approach outlined here, they would be (in chronological order) Freedman (1991), Gelman and Loken (2014), and Mogil and Macleod (2017).

Acknowledgments

This article was written on the author's personal time, not as part of his official duties. Many of the ideas in this article were developed jointly with Bert Gunter, to whom the author also owe thanks for considerable improvements to the manuscript. The author is grateful to several readers for valuable discussions and critique, particularly Prof. Frank Harrell and the anonymous reviewers, and for feedback from audiences when he presented some of these ideas in Reno (October 2017) and Indianapolis (May 2018). The author takes the sole responsibility for the final content. The views expressed do not necessarily represent the policies, views, or opinions of the authors's employer.

References

Allen, J. P. (1972), "Summary of Scientific Results," Apollo 15 Preliminary Science Report NASA SP-289, NASA Manned Spacecraft Center, Washington, DC. [251]

Ambroise, C., and McLachlan, G. J. (2002), "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proceedings of the National Academy of Sciences*, 99, 6562–6566. [247]

Argamon, S. E. (2017), "Don't Strengthen Statistical Significance—Abolish It," *American Scientist*, Macroscope blog [online], available at *https://www.americanscientist.org/blog/macroscope/dont-strengthen-statistical-significance-abolish-it* [246]

Bailey, D. (2018), "Why Outliers Are Good for Science," *Significance*, 15, 14–19. [250]

Barnett, V. (1999), *Comparative Statistical Inference* (3rd ed.), Chichester, UK: Wiley. [257]

Berger, J. O., and Berry, D. A. (1988), "Statistical Analysis and the Illusion of Objectivity," *American Scientist*, 76, 159–165. [257]

Berk, R., Brown, L., and Zhao, L. (2010), "Statistical Inference After Model Selection," *Journal of Quantitative Criminology*, 26, 217–236. [248]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *Annals of Statistics*, 41, 802–837. [251]

Berry, D. (2016), "P-values Are Not What They're Cracked Up To Be," Online supplement to Wasserstein & Lazar (2016). [256]

Box, G. E. P. (1976), "Science and Statistics," *Journal of the American Statistical Association*, 71, 791–799. [249]

——— (1999), "Statistics as a Catalyst to Learning by Scientific Method Part II—A Discussion," *Journal of Quality Technology*, 31, 16–29. [249,250,256,257]

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd ed.), Hoboken, NJ: Wiley. [253]

Carpenter, D. (2010), *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*, Princeton, NJ: Princeton University Press. [258]

Chatfield, C. (1995), "Model Uncertainty, Data Mining and Statistical Inference" (with discussion), *Journal of the Royal Statistical Society*, Series A, 158, 419–466. [248,250,251]

Chow, S.-C., and Chang, M. (2012), *Adaptive Design Methods in Clinical Trials* (2nd ed.), Boca Raton, FL: CRC Press. [253]

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836. [254]

——— (1993), *Visualizing Data*, Summit, NJ: Hobart Press. [254]

——— (1994), *The Elements of Graphing Data* (2nd ed.), Summit, NJ: Hobart Press. [254]

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley. [254]

Coleman, D., and Gunter, B. (2014), *A DOE Handbook: A Simple Approach to Basic Statistical Design of Experiments*, Seattle, WA: Amazon CreateSpace. [254]

Couzin-Frankel, J. (2013), "When Mice Mislead," *Science*, 342, 922–925. [253]

Cox, D. R. (1975), "A Note on Data-Splitting for the Evaluation of Significance Levels," *Biometrika*, 62, 441–444. [251]

Cox, G. M. (1957), "Statistical Frontiers," *Journal of the American Statistical Association*, 52, 1–12. [252]

Dahl, F. A., Grotle, M., Benth, J. J. S., and Natvig, B. (2008), "Data Splitting as a Countermeasure Against Hypothesis Fishing: With a Case Study of Predictors for Low Back Pain," *European Journal of Epidemiology*, 23, 237–242. [251]

David, H. A. (1968), "Gini's Mean Difference Rediscovered," *Biometrika*, 55, 573–575. [254]

Diaconis, P. (1985), "Theories of Data Analysis: From Magical Thinking Through Classical Statistics," in *Exploring Data Tables, Trends, and Shapes*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley, pp. 1–36. [255,256]

Dickhaus, T. (2014), *Simultaneous Statistical Inference*, Berlin: Springer. [251]

Dijkstra, T. K. (ed.) (1988), *On Model Uncertainty and its Statistical Implications* (*Lecture Notes in Economics and Mathematical Systems*, Vol. 307), Berlin: Springer. [258]

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Series B, 57, 45–97. [248,249]

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2016), "Preserving Statistical Validity in Adaptive Data Analysis," [online], available at *https://arxiv.org/abs/1411.2664v3* [255]

Efron, B. (1982), *The Jacknife, the Bootstrap and Other Resampling Plans*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [258]

——— (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331. [258]

——— (2014), "Estimation and Accuracy After Model Selection," *Journal of the American Statistical Association*, 109, 991–1007. [258]

Efron, B., and Hastie, T. (2016), *Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science*, New York: Cambridge University Press. [247,251,256]

Ehrenberg, A. S. C. (1990), "A Hope for the Future of Statistics: MSOD," *The American Statistician*, 44, 195–196. [250]

Faraway, J. (2016), "Does Data Splitting Improve Prediction?" *Statistics and Computing*, 26, 49–60. [251]

Feller, W. (1969), "Are Life Scientists Overawed by Statistics? (Too Much Faith in Statistics)," *Scientific Research*, 4, 24–29. [250,252]

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London A*, 222, 309–368. [247]

——— (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. [253]

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018), "Bayesian Model Averaging: A Systematic Review and Conceptual Classification," *International Statistical Review*, 86, 1–28. [251]

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155. [248]

——— (1991), "Statistical Models and Shoe Leather," *Sociological Methodology*, 21, 291–313. [246,250,257,259]

——— (1995), "Issues in the Foundations of Statistics: Probability and Statistical Models," *Foundations of Science*, 1, 19–39. [252]

——— (1998), "Oasis or Mirage?," *Chance*, 21, 59–61. [249]

——— (1999), "From Association to Causation: Some Remarks on the History of Statistics," *Statistical Science*, 14, 243–258. [250,257]

Gauch, R. R. (2009), *It's Great! Oops, No It Isn't: Why Clinical Research Can't Guarantee the Right Medical Answers*, Berlin: Springer. [251]

Geisser, S. (2006), *Modes of Parametric Statistical Inference*, Hoboken, NJ: Wiley. [257]

Gelman, A. (2003), "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing," *International Statistical Review*, 71, 369–382. [250]

——— (2015), "Statistics and Research Integrity," *European Science Editing*, 41, 13–14. [252]

——— (2016), "The Problems with P-values Are Not Just With P-values," online supplement to Wasserstein & Lazar (2016). [246,248]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [248,251,252,259]

Gigerenzer, G., and Marewski, J. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440. [252,257]

Goldin-Meadow, S. (2016), "Why Preregistration Makes Me Nervous," *Association for Psychological Science Observer*, [online], available at *https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous* [252]

Gong, G. (1986), "Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression," *Journal of the American Statistical Association*, 81, 108–113. [258]

Grolemund, G., and Wickham, H. (2014), "A Cognitive Interpretation of Data Analysis" (with discussion), *International Statistical Review*, 82, 184–213. [255]

Gunter, B., and Tong, C. (2017), "What Are the Odds!? The 'Airport Fallacy' and Statistical Inference," *Significance*, 14, 38–41. [251,252,257]

Harrell, F. E., Jr. (2015), *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.), New York: Springer. [247,256,258]

——— (2018), "Improving Research Through Safer Learning from Data," Statistical Thinking blog [online], available at *http://www.fharrell.com/post/improve-research/* [254,257,258]

Harris, R. (2017), *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*, New York: Basic Books. [253]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), New York: Springer. [251,258]

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [251]

Hjorth, U. (1989), "On Model Selection in the Computer Age," *Journal of Statistical Planning and Inference*, 23, 101–115. [248]

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (eds.) (1983), *Understanding Robust and Exploratory Data Analysis*, New York: Wiley. [255]

——— (1985), *Exploring Data Tables, Trends, and Shapes*, New York: Wiley. [255]

——— (1991), *Fundamentals of Exploratory Analysis of Variance*, New York: Wiley. [255]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [251]

Holmes, S. (2018), "Statistical Proof? The Problem of Irreproducibility," *Bulletin of the American Mathematical Society*, 55, 31–55. [248,251,255]

Huber, P. J. (1985), "Data Analysis: In Search of an Identity," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (Vol. 1), Wadsworth, pp. 65–78. [256]

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1997), *ICH Harmonised Tripartite Guideline: General Considerations for Clinical Trials*, E8. [249,258]

——— (1998), *ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials*, E9. [249,258]

Kallus, N. (2018), "Optimal *A Priori* Balance in the Design of Controlled Experiments," *Journal of the Royal Statistical Society*, Series B, 80, 85–112. [253]

Koopmans, T. C. (1949), "Identification Problems in Economic Model Construction," *Econometrica*, 17, 125–144. [248]

Koyama, T. (2011), "Dynamite Plots," [online], available at *http://biostat.mc.vanderbilt.edu/wiki/Main/DynamitePlots* [254]

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I., (2009), "Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping," *Nature Neuroscience*, 12, 535–540. [248]

Krzywinski, M., and Altman, N. (2014), "Visualizing Samples with Box Plots," *Nature Methods*, 11, 119–120. [254]

Kuehl, R. O. (2000), *Design of Experiments: Statistical Principles of Research Design and Analysis* (2nd ed.), Pacific Grove, CA: Duxbury. [253]

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions* (2nd ed.), Chicago, IL: University of Chicago Press. [257]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-Selection Inference, with Application to the Lasso," *Annals of Statistics*, 44, 907–927. [251]

Leeb, H. (2009), "Conditional Predictive Inference Post Model Selection," *Annals of Statistics*, 37, 2838–2876. [251]

Leek, J. T., Scharpf, R. B., Corrada Bravo, H., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010), "Tackling the Widespread and Critical Impact of Batch Effects in High-throughput Data," *Nature Reviews Genetics*, 11, 733–739. [255,256]

Lew, M. (2016), "Three Inferential Questions, Two Types of P-value," Online supplement to Wasserstein & Lazar (2016). [256]

Lewis, J. A. (1999), "Statistical Principles for Clinical Trials (ICH E9) An Introductory Note on an International Guideline," *Statistics in Medicine*, 18, 1903–1904. [258]

Lithgow, G. J., Driscoll, M., and Phillips, P. (2017), "A Long Journey to Reproducible Results," *Nature*, 548, 387–388, available at *https://www.nature.com/news/a-long-journey-to-reproducible-results-1.22478?WT.mc_id=FBK_NatureNews&sf108251523=1* [250]

Lock Morgan, K., and Rubin, D. B. (2012), "Rerandomization to Improve Covariate Balance in Experiments," *Annals of Statistics*, 40, 1263–1282. [253]

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2017), "Abandon Statistical Significance," [online], available at *https://arxiv.org/abs/1709.07588*. [246]

McShane, B. B., and Gelman, A. (2017), "Abandon Statistical Significance," *Nature*, 551, 558, available at *https://www.nature.com/articles/d41586-017-07522-z* [246]

Mallows, C. L. (1983), "Data Description," in *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and C.-F. Wu, New York: Academic Press, pp. 135–151. [254]

——— (1998), "The Zeroth Problem," *The American Statistician*, 52, 1–9. [255]

Mallows, C. L., and Walley, P. (1980), "A Theory of Data Analysis?" in *Proceedings of the Business and Economics Statistics Section*, American Statistical Association, pp. 8–14. [254]

Maronna, R., Martin, D., and Yohai, V. (2006), *Robust Statistics: Theory and Methods*, Hoboken, NJ: Wiley. [255]

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* [249,252,256,259]

Montgomery, D. C. (2017), *Design and Analysis of Experiments* (9th ed.), Hoboken, NJ: Wiley. [253]

Moore, D. S. (1992), "What is Statistics?" in *Perspectives on Contemporary Statistics*, eds. D. C. Hoaglin, and D. S. Moore, Washington, DC: Mathematical Association of America, pp. 1–17. [252]

Moses, L. E. (1992), "The Reasoning of Statistical Inference," in *Perspectives on Contemporary Statistics*, eds. D. C. Hoaglin, and D. S. Moore, Washington, DC: Mathematical Association of America, pp. 107–122. [254]

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley. [248]

Motulsky, H. J. (2014), "Common Misconceptions About Data Analysis and Statistics," *Journal of Pharmacology and Experimental Therapeutics*, 351, 200–205. [258]

Munafo, M. R., and Davey Smith, G. (2018), "Robust Research Needs Many Lines of Evidence," *Nature*, [online], available at *https://www.nature.com/articles/d41586-018-01023-3* [250]

Murray, D. M. (1998), *Design and Analysis of Group-Randomised Trials* (*Monographs in Epidemiology and Biostatistics*, Vol. 27), New York: Oxford University Press. [253]

Nelder, J. A. (1986), "Statistics, Science and Technology," *Journal of the Royal Statistical Society*, Series A, 149, 109–121. [246,249]

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018), "The Preregistration Revolution," *Proceedings of the National Academy of Sciences*, 115, 2600–2606. [251,252]

Pearson, R. K. (2011), *Exploring Data in Engineering, the Sciences, and Medicine*, New York: Oxford University Press. [255]

Piantadosi, S. (2017), *Clinical Trials: A Methodologic Perspective* (3rd ed.), Hoboken, NJ: Wiley. [249,255,258]

Picard, R. R., and Cook, R. D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583. [246,248,251]

Pikounis, B. (2001), "One-Factor Comparative Studies," in *Applied Statistics in the Pharmaceutical Industry*, eds. S. P. Millard, and A. Krause, New York: Springer, pp. 17–40. [254]

Raftery, A., Madigan, D., and Hoeting, J. (1993), "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," Technical Report No. 262, University of Washington (Seattle), Department of Statistics. [248]

Reunanen, J. (2003), "Overfitting in Making Comparisons Between Variable Selection Methods," *Journal of Machine Learning Research*, 3, 1371–1382. [247]

Robbins, N. B. (2013), *Creating More Effective Graphs*, Wayne, NJ: Chart House. [254]

Rousseeuw, P. J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283. [254]

Scheiner, L. B. (1997), "Learning Versus Confirming in Clinical Trials," *Clinical Pharmacology and Therapeutics*, 61, 275–291. [258]

Scott, S. (2013), "Pre-Registration Would Put Science in Chains," *Times Higher Education Supplement*, [online], available at *https://www.timeshighereducation.com/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article* [252]

Seife, C. (2000), "CERN's Gamble Shows Perils, Rewards of Playing the Odds," *Science*, 289, 2260–2262. [250]

Shell, E. R. (2016), "Hurdling Obstacles: Meet Marcia McNutt, Scientist, Administrator, Editor, and Now National Academy of Sciences President," *Science*, 353, 116–119. [249]

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [247]

Snee, R. D. (1986), "In Pursuit of Total Quality," *Quality Progress*, 20, 25–31. [255]

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–147. [251]

Taleb, N. N. (2007), *The Black Swan: The Impact of the Highly Improbable*, New York: Random House. [254]

Taylor, J. E., and Tibshirani, R. (2015), "Statistical Learning and Selective Inference," *Proceedings of the National Academy of Sciences*, 112, 7629–7634. [251]

The OPERA Collaboration (2012), "Measurement of the Neutrino Velocity with the OPERA Detector in the CNGS Beam," *Journal of High Energy Physics* [online], available at *https://doi.org/10.1007/JHEP10(2012)093* [250]

Thompson, S. K. (2012), *Sampling* (3rd ed.), Hoboken, NJ: Wiley. [254]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [251]

Tong, C., and Lock, A. (2015), "A Computational Procedure for Mean Kinetic Temperature Using Unequally Spaced Data," in *Proceedings of the Biopharmaceutical Section*, American Statistical Association, pp. 2065–2070. [255]

Touboul, P., Metris, G., Rodrigues, M., Andre, Y., Baghi, Q., Berge, J., Boulanger, D., Bremer, S., Carle, P., Chhun, R., et al. (2017), "MICROSCOPE Mission: First Results of a Space Test of the Equivalence Principle," *Physical Review Letters*, 119, 231101. [251]

Tukey, J. W. (1962), "The Future of Data Analysis," *Annals of Statistics*, 33, 1–67. [255,256,257]

——— (1969), "Analyzing Data: Sanctification or Detective Work?," *American Psychologist*, 24, 83–91. [248]

——— (1973), "Exploratory Data Analysis as Part of a Larger Whole," in *Proceedings of the Eighteenth Conference on the Design of Experiments in Army Research Development and Testing* (Part I), Durham, NC: Army Research Office, pp. 1–10. [255]

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley. [248,249,254,255]

Wasserman, L. (2006), *All of Nonparametric Statistics*, New York: Springer. [255]

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. [246]

Wenmackers, S., and Vanpoucke, D. E. P. (2012), "Models and Simulations in Material Science: Two Cases Without Error Bars," *Statistica Neerlandica*, 66, 339–355. [255,257]

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., and van Assen, M. A. L. M. (2016), "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking," *Frontiers in Psychology*, 7, 1832. [247]

Winer, B. J., Brown, D. R., and Michels, K. M. (1991), *Statistical Principles in Experimental Design* (3rd ed.), Boston, MA: McGraw-Hill. [253]

Youden, W. J. (1972), "Enduring Values," *Technometrics*, 14, 1–11. [250]

The American Statistician

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

Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication

Valentin Amrhein, David Trafimow & Sander Greenland

To cite this article: Valentin Amrhein, David Trafimow & Sander Greenland (2019) Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication, The American Statistician, 73:sup1, 262-270, DOI: 10.1080/00031305.2018.1543137

To link to this article: https://doi.org/10.1080/00031305.2018.1543137

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 42053

View related articles

View Crossmark data

Citing articles: 90 View citing articles

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, 262–270: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543137

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication

Valentin Amrhein[a], David Trafimow[b], and Sander Greenland[c]

[a]Zoological Institute, University of Basel, Basel, Switzerland; [b]Department of Psychology, New Mexico State University, Las Cruces, NM; [c]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA

Abstract

Statistical inference often fails to replicate. One reason is that many results may be selected for drawing inference because some threshold of a statistic like the *P*-value was crossed, leading to biased reported effect sizes. Nonetheless, considerable non-replication is to be expected even without selective reporting, and generalizations from single studies are rarely if ever warranted. 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 or desires. A general perception of a "replication crisis" may thus reflect failure to recognize that statistical tests not only test hypotheses, but countless assumptions and the entire environment in which research takes place. Because of all the uncertain and unknown assumptions that underpin statistical inferences, we should treat inferential statistics as highly unstable local descriptions of relations between assumptions and data, rather than as providing generalizable inferences about hypotheses or models. And that means we should treat statistical results as being much more incomplete and uncertain than is currently the norm. Acknowledging this uncertainty could help reduce the allure of selective reporting: Since a small *P*-value could be large in a replication study, and a large *P*-value could be small, there is simply no need to selectively report studies based on statistical results. Rather than focusing our study reports on uncertain conclusions, we should thus focus on describing accurately how the study was conducted, what problems occurred, what data were obtained, what analysis methods were used and why, and what output those methods produced.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
Auxiliary hypotheses; Confidence interval; Hypothesis test; *P*-value; Posterior probability; Replication; Selective reporting; Significance test; Statistical model; Unreplicable research

The "crisis of unreplicable research" is not only about alleged replication failures. It is also about perceived nonreplication of scientific results being interpreted as a sign of bad science (Baker 2016). Yes, there is an epidemic of misinterpretation of statistics and what amounts to scientific misconduct, even though it is common practice (such as selectively reporting studies that "worked" or that were "significant"; Martinson, Anderson, and de Vries 2005; John, Loewenstein, and Prelec 2012). But all results are uncertain and highly variable, even those from the most rigorous studies.

Because a small *P*-value could result from random variation alone, Fisher (1937) wrote that "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon." And Boring (1919) said a century ago, "scientific generalization is a broader question than mathematical description." Yet today we still indoctrinate students with methods that claim to produce scientific generalizations from mathematical descriptions of isolated studies. Naturally, such generalizations will often fail to agree with those from other studies—and thus statistical inference will fail to replicate. Because our current academic reward system is built on single publications (usually reporting the results of one or a few similar studies), it should come as no surprise that many conflicting generalizations are published, and hence that a high proportion of generalizations must be wrong.

A core problem 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.

In the following, we argue that the crisis of unreplicable research is mainly a crisis of overconfidence in statistical results. We recommend that we should use, communicate, and teach inferential statistical methods as describing logical relations between assumptions and data (as detailed in the Appendix), rather than as providing generalizable inferences about universal populations.

1. Inferences Are Not About Hypotheses

A statistical model is a set of assumptions, and thus a compound hypothesis, about how the data could have been generated. The

CONTACT Valentin Amrhein v.amrhein@unibas.ch Zoological Institute, University of Basel, Basel 4501, Switzerland.
© 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.

model matches reality to the degree that assumptions are met, starting from the assumptions that we measured what we think we measured and that measurement errors were either absent or adequately accounted for. Such model assumptions are part of what are called "auxiliary hypotheses" (Popper 1968), "auxiliary theories" (Meehl 1990), or "currently subordinate factors" (McShane et al. 2019).

Thus, statistical models imply countless assumptions about the underlying reality. A null hypothesis such as "the means of these two populations do not differ" is an explicit assumption. Further assumptions that are often explicitly addressed in research reports are that sampling was random or that residuals are independent and identically distributed. Other assumptions may not even be recognized or mentioned in research reports, such as that there was no selection of particular results for presentation, or that the population from which we drew our random sample is equivalent to the population we have targeted for inference. Whether it is assumptions that are reviewed by inspecting residuals, or further assumptions that link statistics to reality, the validity of statistical inferences depends on the entire set of assumptions.

For example, we should think of a *P*-value[1] as referring not only to the hypothesis it claims to test, such as a null hypothesis. A *P*-value refers to the entire model including other usually explicit assumptions like randomization of treatment and linearity of effects, plus usually implicit procedural assumptions such as that the equipment for taking measurements was in perfect working order (Greenland et al. 2016; Greenland 2017; Amrhein 2018). Whether recognized or not, these assumptions underpin the usual inferences from a test (Greenland 2019b). A small *P*-value is the net result of some combination of random variation and violations of model assumptions, but does not indicate which (if any) assumption is violated.

Yes, a small *P*-value may arise because the null hypothesis is false. But it can also mean that some mathematical aspect of the model was not correctly specified, that sampling was not a hundred percent random, that we accidentally switched the names of some factor levels, that we unintentionally, or intentionally, selected analyses that led to a small *P*-value (downward "P-hacking"), that we did not measure what we think we measured, or that a cable in our measuring device was loose (Amrhein 2018). And a large *P*-value may arise from mistakes and procedural errors, such as selecting analyses that led to a large *P*-value (upward P-hacking), or using a measurement so noisy that the relation of the measured construct to anything else is hopelessly obscured.

2. Replication Studies Have a False-Negative Problem

Even the best single studies will be imperfect. In addition to random variation, their results will usually vary from replication to replication because of varying assumption violations, whether recognized or hidden, and thus the observed effect sizes can easily differ across settings. Consider the replication project by the Open Science Collaboration (2015): Of 97 psychological studies with "significant" results ($p \leq 0.05$) out of 100 that were subject to replication, 35 had $p \leq 0.05$ in the replication. This is much less than would have been expected if all original null hypotheses were false and only random variation caused differences between the original and replication—under these circumstances, with an average power of 92% in the replications, 89 of the 97 replicates were expected to have $p \leq 0.05$. One explanation by the authors is that in the original studies, results were selected for reporting based on having $p \leq 0.05$, which led to inflated effect sizes (see next section) that could not be replicated.

The reports of such replication projects are often misinterpreted as showing that the original "significant" studies were mostly or entirely false positives. To see the error in such interpretations, consider that the Open Science Collaboration (2015) observed 97-35 = 62 replications with $p > 0.05$ for which the original study had $p \leq 0.05$, which is 64% of the 97 replications. This emphatically does *not* mean that 64% of the 97 original null hypotheses with "significant" *P*-values were correct, or that 62 of the 97 "significant" results were false positives. Why not? If as suggested (and indicated by the replications) the original reported effect sizes were largely inflated due to selective reporting based on $p \leq 0.05$, then the actual effect sizes (both in the original and replication studies) could easily be too small to provide high power for replication (Camerer et al. 2018).

Suppose, for example, that among the 97 "significant" original studies, 70% of the null hypotheses were false, for a total of 0.70(97) = 68 false nulls, and the average power for the effects in these non-null cases was 50% in the replication studies (and not 92% as calculated based on the probably inflated original effect sizes). Then the expected number of "true positive" replications (those with $p \leq 0.05$ for a false null) would be 0.50(68) = 34, while the number of "false positive" replications (those with $p \leq 0.05$ for a true null) would be 0.05(97-68) = 1.45, resulting in a total of about 35 out of 97 replications having $p \leq 0.05$, as observed. That means that, given selective reporting in the original studies, the observed 64% of the 97 replication attempts with $p > 0.05$ could have been expected even if only 97-68 = 29 or 30% of the null hypotheses were correct!

Thus, with selective reporting in the original studies, it may be unsurprising to get "nonsignificant" results in about two-thirds of replications. And even with no selective reporting and only random variation present, replication studies remain subject to what may be severe *false-negative* errors. Consequently, "non-significant" replications (those with $p > 0.05$) do not reliably flag original studies as being false positives. For example, with a statistical power of 80%, two studies will be traditionally judged as "conflicting," meaning that one is "significant" and the other is not, in one third of the cases (Greenland et al. 2016; Amrhein, Korner-Nievergelt, and Roth 2017). This means that, unless statistical power of the replication is nearly 100%, interpretations of replication attempts must allow for false-negative errors as well as false-positive errors, and that "significance" and "non-significance" cannot be used to reliably judge success or failure of replication (Goodman 1992; Senn 2001, 2002). As importantly, valid interpretation must take into account any background research (prior information) on the

[1] We focus on *P*-values and confidence intervals not because they are better or worse than other methods, but because they are probably the most often used, and most often misused, inferential statistics. Throughout the text, *P* is the random variable and *p* is its observed value (realization) in a given sample.

hypothesis under study, especially if there are conflicts between the original and replication studies.

Another often overlooked point about hacking and replication failures is that both random and selection artefacts will arise even if we ban all tests and focus instead on estimates. Some degree of variation, and hence nonreplication, is the norm across honestly reported studies, even when all assumptions are met, and estimates can be selected for reporting based directly on their size rather than their *P*-value. Add other assumption violations to these problems and we will find that error rates are usually far from the nominal rates given in the report—for example, a 95% confidence interval cannot realistically be claimed to have as much as 95% coverage of the true effect when study imperfections exist. Consequently, having confidence in generalizations from single studies means having overconfidence in most cases. Inference that could be called trustworthy would require merging information from multiple studies and lines of evidence.

3. Overconfidence Triggers Selection Bias

Unfortunately, even a combination of studies does not guarantee that inferences will be valid. As noted above, published results tend to be biased, for example because they may be selected from unpublished results based on some statistical criterion. Such bad yet common scientific practice introduces bias by accumulating statistical inferences that go into a certain direction, typically emphasizing results that cross some *P*-value threshold (Amrhein, Korner-Nievergelt, and Roth 2017; Locascio 2017).

We suspect that a major driver of result-selection bias is overconfidence in statistical inference. For decades, scientists were taught to judge which results are trustworthy and which are not, and which results are thus worth being published or not, based on statistics obtained from single studies. Statistics was misused as an automated scientific decision machine, both for statements about hypotheses and for selection of studies for publication. And this made interpretation, publication, and public advertising much easier, because everybody assumed that statistical inferences based on *P*-value thresholds or other rigid criteria would be "reliable," in the sense that a replication would probably meet the same thresholds or criteria again. So if researchers expect that a small *P*-value or a short interval estimate indicate "reliable" results, while all other results are "unreliable," they may be "prepared to ignore all results which fail to reach this standard" (Fisher 1937, p. 16; one of many published pleas by various authors encouraging selective reporting).

But any selection criterion will introduce bias. If there is a tendency to publish results because the estimates are yellow, because interval estimates are short, and because *P*-values are small or the point estimates are far from null, then the published literature will become biased toward yellow results with underestimated variances and overestimated effect sizes. The latter effect is the "winner's curse," or inflation of effect sizes, that is reflected in the findings of the Open Science Collaboration (2015): the average effect size in the original studies was about twice as large as in the replication studies that reported all results and thus did not suffer from selection bias.

Even if authors report all study outcomes, but then select what to discuss and to highlight based on *P*-value thresholds or other aids to judgment, their conclusions and what is reported in subsequent news and reviews will be biased (Amrhein, Korner-Nievergelt, and Roth 2017). 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.

One way to reduce selective reporting and attention is to maintain that all results are uncertain. If we obtain a small *P*-value, a large effect estimate, or a narrow interval estimate—or even all three—we should not be confident about textbook inferences from these results. In one of the next replications, p will be large, the effect estimate will be small, or the interval estimate wide, and thus the textbook inferences will shift dramatically due to random variation or to assumptions we have not modeled. Because of this uncertainty, there is simply no need to selectively report studies based on statistical results.

We should thus "move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016) and focus on describing accurately how the study was conducted, what problems occurred (e.g., nonresponse of some subjects, missing data), and what analysis methods were used, with detailed data tabulation and graphs, and complete reporting of results. The advent of online supplements and preprint servers eliminate the common excuse that space limitations prevent reporting such detail.

4. Don't Blame the *P*-value

A clear sign that overconfidence has ruled the era of hypothesis testing is that many people still are surprised by the "dance of the *P*-values" (Cumming 2014), that is, by the way a valid *P*-value bounces around its range even in the largest of samples. This variability means that $p < 0.05$ is no guarantee for $p < 0.05$ in a replication (Goodman 1992; Senn 2001, 2002; Gelman and Stern 2006); after all, if the (null) hypothesis tested is correct and experimental conditions are ideal, the *P*-value will vary uniformly between 0 and 1. And even if our alternative hypothesis is correct, the *P*-value in the next sample will typically differ widely from our current sample: "The fickle *P*-value generates irreproducible results" (Halsey et al. 2015), at least if reproducibility is defined by whether p is above or below a threshold and the power is not very high.

But the *P*-value itself is not supposed to be "reliable" in the sense of staying put (Greenland 2019a). Its fickleness indicates variation in the data from sample to sample. If sample averages vary among samples, then *P*-values will vary as well, because they are calculated from sample averages. And we don't usually take a single sample average and announce it to be the truth. But if instead of simply reporting the *P*-value, we engage in "dichotomania" (Greenland 2017) and use it to decide which hypothesis is wrong and which is right, such scientifically destructive behavior is our fault, even if socially encouraged; it is not the fault of the *P*-value.

Further, if we overlook the sensitivity of *P*-values to possible violations of background assumptions, by assuming that *P*-values are only about deciding whether to reject "null hypotheses," we are privileging what may be a scientifically irrelevant hypothesis and are engaging in "nullism," a compulsion to test

only one hypothesis among many of importance. But again, such bad behavior is our fault, even if socially encouraged. We could instead provide P-values for relevant alternatives, and arguably *should* do so if we compute any P-value at all (Poole 1987a, 1987b). And if we interpret a small null P-value as providing support for some alternative hypothesis (which currently seems to be a standard interpretation) without testing the alternative as well, this too is our fault, not the fault of the P-value.

5. Ban Statistical Tests?

It may help to ban some practices, at least temporarily and in specific contexts. Despite the general acceptance of alcohol[2] in relaxed settings and possible beneficial effects from light use, we ban its drinking before or during driving, and recognize that its use easily becomes abuse. Reaching for statistical tests to force out "inferences" (whether traditional "$p \leq \alpha$" testing or substitutes like tests using Bayes-factor criteria) is, like drinking alcohol, a culturally ingrained habit. Statistical testing (like alcohol) often gives the wrong impression that complex decisions can be oversimplified without negative consequences, for example, by making decisions solely because p was above or below some cutoff like 0.05. And many researchers are addicted to such oversimplification. These addictions are worth breaking.

At the very least, partial or temporary bans are one way to force researchers to learn how to analyze data in alternative ways (Trafimow and Marks 2015). Hopefully, thinking about advantages and disadvantages of alternatives will lead to more sober interpretation of statistics. One concern, however, is that complete prohibition could lead to misuse and abuse of other methods, such as Bayesian techniques—which have an additional source of non-replicability insofar as what are acceptable priors can vary dramatically across research groups.

6. Long Live No King

Fixed-cutoff (α-level) hypothesis testing has been king for over 80 years. We propose not to banish the king—a fixed-cutoff decision procedure may be useful, for example, in industrial or laboratory quality control, or "sampling tests laid down in commercial specifications" (Neyman and Pearson 1933), in which automated decisions to stop a production line or to recalibrate equipment may be necessary. For scientific inference, however, we hope that dichotomania from which such procedures suffer can be cured by abandoning them in favor of data description and direct presentation of precise P-values—including P-values for alternative hypotheses (Poole 1987a, 1987b; Cohen 1994; Ziliak and McCloskey 2008; Hurlbert and Lombardi 2009; Amrhein, Korner-Nievergelt, and Roth 2017; Greenland 2017; Amrhein and Greenland 2018; Trafimow et al. 2018; Greenland 2019a, 2019b; McShane et al. 2019). Parallel criticisms and remedies apply to tests or decisions (such as whether to report or highlight results) based on Bayes factors, posterior odds, or any other statistical criterion.

Yes, sometimes we need to make decisions in science, for example, whether to further pursue a study or not. For such a decision, we will usually weigh scientific and personal costs and benefits of our decision, applying informed personal judgment (Gigerenzer 1993). But when it comes to weigh evidence against, or in favor of, a scientific hypothesis, statistical tests cannot suffice, and may even be destructive if degraded into a binary form as in reporting tests as significant/nonsignificant, or in basing conclusions on whether the null value was included in or excluded from an interval. This is especially true when (as almost always) these results are sensitive to doubtful assumptions, such as absence of measurement-error dependencies. And even in the unlikely case that all model assumptions are met, we would still need to consider costs and benefits, as well as the published and unpublished literature, to judge a scientific hypothesis as being largely correct (subject to further evidence to the contrary). We hope that classical hypothesis testing will be retired quickly from research reporting, so that regicide is not necessary.

7. Empire of Diversity

But what comes next? There are countless possibilities. The most common proposal is to replace hypothesis tests with interval estimates. While doing so is helpful for sophisticated researchers, it has not reduced what we see as the core psychological problems—which is unsurprising, because the classical confidence interval is nothing more than a summary of dichotomized hypothesis tests. Consider that a 95% confidence interval encompasses a range of hypotheses (effect sizes) that have a P-value exceeding 0.05. Instead of talking about hypothetical coverage of the true value by such intervals, which will fail under various assumption violations, we can think of the confidence interval as a "compatibility interval" (Greenland 2019a, 2019b), showing effect sizes most compatible with the data according to their P-values, under the model used to compute the interval. Likewise, we can think of a posterior probability interval, or Bayesian "credible interval," as a compatibility interval showing effect sizes most compatible with the data, under the model *and prior distribution* used to compute the interval (Greenland 2019a). Again, whether such intervals include or exclude zero should play no role in their interpretation, because even with only random variation the intervals from different studies can easily be very different (Cumming 2014).

With additional (and inevitable) nonrandom variation, the true effect size will frequently be outside the interval. In reality, it will not happen that every assumption is met, nor will we be aware of every assumption. Stating that our data "support" any value in the compatibility interval (e.g., a zero effect), or that, upon unlimited replication, the intervals would cover the true value at some rate, or that the interval "measures uncertainty" by indicating the range of *possible* effect sizes (as opposed to *compatible* effect sizes, given the model), makes the compatibility interval into an *overconfidence* interval.

To avoid the dichotomization retained by interval estimates, one could report a measure of refutational evidence, such as a traditional P-value in a continuous fashion (as recommended by classic texts on testing such as Lehmann 1986, p. 71), reporting

[2] We prefer this analogy above comparing P-values or hypothesis tests with guns, as we have heard or read sometimes.

an observed P-value as a measure of the degree of compatibility between the hypothesis or model it tests and the data (Greenland 2019a). Better still, we could report the Shannon information or S-value (surprisal) of the test, $-\log_2(p)$, which is a measure of the evidence against the model supplied by the test, expressed in units of *bits* (binary digits) of information. Among its advantages, the S-value is unbounded above and thus difficult to misinterpret as a hypothesis probability (Greenland 2017, 2019a, 2019b). Considering that the 95% confidence interval is the range in which the parameter values have $p > 0.05$, the values in the interval are those for which the S-value s is less than $-\log_2(0.05) = 4.3$. This means that, under the model used to construct the interval (e.g., a regression model), the values in a 95% confidence interval have only about 4 bits or less information against them; that is, very little information indeed (4 bits is the same as the evidence against "fairness" of coin tosses versus loading for heads provided by obtaining 4 heads in a row).

If we want to compare the relative support for different parameter values or models, we could use likelihood ratios or Bayesian methods. But we should not lapse back into dichotomous thinking by using some P-value threshold, or by making binary inferences based on confidence intervals or Bayes factors. And we should not expect posterior probabilities to solve the problems, especially since they rely on the same often questionable assumptions about the data-generating model that both hypothesis tests and confidence intervals use.

The empire of "statistical significance" has its roots in the 19th century writings of Edgeworth (1885) and reached full dominance with the spread of cutoffs for testing, formalized by Jerzy Neyman and Egon Pearson as Type-I error rates. Like the political empires of their period, such hypothesis testing for scientific (as opposed to mechanical) inference is a relic of a bygone era, whose destructive effects reverberate to this day. We hope this era is over. As for what comes next, there is no substitute for accepting methodologic diversity (Good 1957; Cox 1978; Box 1980; Barnard 1996; Little 2006; Senn 2011; Efron and Hastie 2016; Crane 2017), with careful assessment of uncertainty as the core motivation for statistical practice (e.g., by discussing the effect sizes compatible with the data, given the model, as outlined above).

8. The Replacement for Hypothesis Tests

We "don't look for a magic alternative to NHST [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist" (Cohen 1994). And if it existed, we would probably not recommend it for scientific inference. What needs to change is not necessarily the statistical methods we use, but how we select our results for interpretation and publication, and what conclusions we draw. Why would we want a mechanical decision procedure for single studies, if not for selecting results for publication or interpretation? As we described above, every selection criterion would introduce bias. Therefore, we join others who have advised that we should, to the extent feasible:

(a) Target results for publication and interpretation *before* data are collected, that is, state our hypotheses and predictions in a defined protocol or a binding research proposal.
(b) Before analyzing data (and preferably before collecting them), make an analysis plan (i.e., a pre-analysis protocol), setting out how data will be analyzed; and, in the publication, show what results the protocol produced before displaying the results of any analyses deviating from the predefined protocol.
(c) Emphasize and interpret our estimates rather than tests, explicitly discussing both the lower and upper limits of our interval estimates.
(d) When reporting statistics, give their precise values rather than mere inequalities; for example, if we are reporting a P-value and it is 0.03, report "$p = 0.03$," not "$p < 0.05$."[3]
(e) Not use the words "significant" or "confidence" to describe scientific results, as they imply an inappropriate level of certainty based on an arbitrary criterion, and have produced far too much confusion between statistical, scientific, and policy meanings.
(f) Acknowledge that our statistical results describe relations between assumptions and the data *in our study*, and that scientific generalization from a single study is unwarranted.
(g) Openly and fully report our detailed methods, materials, procedures, data, and analysis scripts.

As an example, consider a study by Brown et al. (2017), who reported that "in utero serotonergic antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child," based on an estimated hazard-rate ratio (HR) of 1.61 (a 61% rate increase in the exposed relative to the unexposed) and a 95% confidence interval of 0.997 to 2.59. As is often the case, the authors misused the confidence interval as a hypothesis test, and they claim to have demonstrated no association because the lower limit of the interval was slightly below no association (which corresponds to a hazard-rate ratio of HR = 1), ignoring that the upper limit exceeded 2.50. A more correct summary of the results would have been: "Our estimate of the hazard-rate ratio was 1.61, and thus exposure could be associated with autism; however, possible hazard-rate ratios that are highly compatible with our data, given our model, ranged from 0.997 (essentially no association) to 2.59 (a relatively strong association)." If applicable, this could then be followed by a discussion of why the authors seem to think the exposure effect might be negligible despite the association, and how strong they judge their evidence not only based on the width of an interval estimate, but also in view of possible shortcomings of their study, of their prior knowledge about other studies on autism, and of possible costs of their interpretation for the health of the patients.

Had the authors found an interval of 1.003 to 2.59 rather than 0.997 to 2.59, the reporting should have been the same. Even with an interval of 0.900 to 2.59, the description of the results should be largely similar—the point estimate would still be a HR well above 1, indicating a possible positive association. What would need to change with the latter interval is the description that not only relatively large positive, but also small negative associations would fall within the interval.

[3] An exception: If a P-value is below the limit of numerical accuracy of the data, an inequality would be called for, but the precision would be context dependent: e.g., $p < 10^{-8}$ is typical in genomics, $p < 0.001$ is common in medicine and ecology.

9. Anything Goes?

So what do we conclude from a study like Brown et al. (2017)? If we would interpret an interval of 1.003 to 2.59 and of 0.997 to 2.59 in the same way, does that mean that the floodgates of "anything goes" are wide open? Yes, the floodgates should be open—for reporting our results. 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. If after doing the study, it appears the measure did not make sense or the methods were faulty, at least other researchers can learn that lesson without repeating the error—provided our report contains enough detail to allow such critical evaluation.

However, the floodgates should be closed for drawing conclusions from virtually any single study. For example, because they found a confidence interval that barely included the null value, Brown et al. (2017) reported conflict with previously observed associations that were nearly the same size (hazard ratios around 1.7) but had confidence intervals that did not include the null value. We think that, at most, a conclusion could be drawn that the new study was largely consistent with previous studies, but that the null value was also compatible with their data, given the model.

In view of all the unmodeled uncertainties, it would be good to plan and publish single studies with the goal of their easy entry into meta-analyses. In the words of Trafimow et al. (2018): "It is desirable to obtain precise estimates in those studies, but a more important goal is to eliminate publication bias by including wide confidence intervals and small effects in the literature, without which the cumulative evidence will be distorted."

10. Abandon Statistical Inference?

We do not suggest to completely abandon inference from our data to a larger population (although the title of a preprint of this paper was "Abandon statistical inference"; Amrhein, Trafimow, and Greenland 2018). But we say this inference must be scientific rather than statistical, even if we use inferential statistics. Because all statistical methods require subjective choices (Gelman and Hennig 2017), there is no objective decision machine for automated scientific inference; it must be we who make the inference, and claims about a larger population will always be uncertain.

So when can we be confident that we know something? This is the topic of the vast domains epistemology, scientific inference, and philosophy of science, and thus far beyond the scope of the present paper (and its authors). Nonetheless, a successful theory is one that survives decades of scrutiny. If every study claims to provide decisive results (whether from inferential statistics or narrative impressions—or a confusion of the two), there will be ever more replication failures, which in turn will further undermine public confidence in science. We thus believe that decision makers must act based on cumulative knowledge—which means they should preferably not rely solely on single studies or even single lines of research (even if such contributions may determine a decision when all other evidence appears ambiguous or unreliable).

11. If We Are Researchers …

…and we obtained a large P-value for the null hypothesis or an interval estimate that includes a null effect, our interval will show that the null hypothesis is only one of many different hypotheses that are compatible with the data (Rothman, Greenland, and Lash 2008; Greenland et al. 2016). Unlike what Brown et al. (2017) suggested with their hazard-rate ratios, we cannot claim our statistics indicate there is no effect, because even if the data remain consistent with a zero effect, they remain consistent with many other effects as well. A "proof of the null hypothesis" such as "the earth is flat ($p > 0.05$)" is, therefore, not possible (Greenland 2011; Amrhein, Korner-Nievergelt, and Roth 2017). And we should remember there are lots of additional hypotheses outside the interval estimate that will also be compatible with our data, due to methodologic limitations that we have not modeled.

Thus, we should not be overconfident about our "weak evidence" from a large P-value or an interval estimate that includes the null. Almost never will we have found absolutely no effect. Let us free our "negative" results by allowing them to be positive contributions to knowledge (Amrhein, Korner-Nievergelt, and Roth 2017). This means that, unless the interval estimate is entirely within an explicit interval of "practical equivalence" to the null,[4] we should first and foremost ban the following statements from our thinking, from our papers, and from our talks: "there was no effect," "there was no difference," "there was no interaction," or "we deleted this term from the model because it had no influence."

Such statements will usually be wrong even if our point estimate is exactly null (and thus $p = 1$), because our interval estimate will usually show there are many important non-null effects that are highly compatible with our data. This means that an outcome of a study can only be "negative" in a falsificationist sense of finding little incompatibility between the data and the predictions of a model that includes a hypothesis of no effect. A P-value of 1 for the test of the null only means our data are perfectly compatible with our model (including the null hypothesis); but "perfectly compatible" with one hypothesis, or model, does not mean that all other hypotheses, or models, are refuted. Indeed, typical interval estimates will reveal a large number of non-null hypotheses that we would call highly compatible with our data (e.g., because their P-values exceed 0.05), given our model is correct.

Thus, most studies with large P-values or interval estimates that include the null should be considered "positive" in the sense that they leave open the possibility of important effects, even if they also leave open the possibility of no effect. The best we can do is describe the values covered by our interval estimates—and if those values have qualitatively different practical consequences, we should admit that our current set of data could not settle the matter even if we knew that all the auxiliary assumptions were correct.

[4] In terms of α-level equivalence testing (Wellek 2010; Lakens et al. 2018), for example, this means that all effects with symmetric two-sided $p > 2\alpha$ are inside the interval of equivalence.

Conversely, if we believe we have "strong evidence" because our *P*-value is small and our point estimate is large, or because our interval estimate is not near the null, we are placing too much faith in our inferential statistics. Keep in mind that these statistics do not "measure" uncertainty. At best, the interval estimate may give a rough idea of uncertainty, given that all the assumptions used to create it are correct. And even then, we should remember the "dance of the confidence intervals" (Cumming 2014) shows a valid interval will bounce around from sample to sample due to random variation.

Because every inferential statistic (including *P*-values, interval estimates, and posterior probabilities) is derived from the multiple implicit as well as explicit assumptions that compose the model, we should treat these statistics as descriptions of the relation of the model to the data rather than as statements about the correctness of the model. This discipline may be more difficult for Bayesians to accept, since Bayesian methods do produce hypothetical probabilities of models by using assumptions hidden in restrictions on the models considered. Regardless, it is hard enough to describe the known assumptions about our model. We should not draw inference and generalize based on assumptions we cannot be certain about or we do not even think about.

For example, a *P*-value is merely the probability of one particular test statistic being as or more extreme than observed in our particular study, given that the model it is computed from is correct. No inferential meaning need be attached to that. For the next set of data, the *P*-value will be different. A small *P*-value is just a warning signal that the current model could have a problem, so we should check our model assumptions (including an assumption such as "the means of these two populations do not differ," i.e., our tested hypothesis). And this assumption checking does not only mean inspecting residuals, but also checking the extent of deviations of our study from a perfect randomized experiment or random survey, whether from failures of protocol, measurement, equipment, or any of the innumerable details that real research must confront (Greenland 2017; Stark and Saltelli 2018).

Science includes learning about assumption violations, then addressing those violations and improving the performance of our models about reality. Statistics can help by formalizing parts of the models, and by assisting in careful assessment of uncertainty. We should thus communicate our limited conclusions about our data, not our generalized inferences about some ill-defined universal population. And decisions to communicate and interpret a result should not be based on *P*-values, nor on any other statistic. Presentations that start with analysis plans that were formulated before the analysis (pre-analysis protocols) can help strengthen both the validity and credibility of our inferences. The reported description of our results will be a good description if it is complete and honest. If we think we did a good study, we should thus be modest about our conclusions, but be proud about our painfully honest and thorough description and discussion of our methods and of our data.

And if we are working as journal editors, we should be proud about our exhaustive methods sections and consider "results blind evaluation" of manuscripts (Locascio 2017), that is, basing our 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.

12. If We Are Science Writers and Journalists …

…we should continue writing about isolated experiments and replications. Single studies are the life blood of science. If we think we found a good study, or a bad study, we may report it. But let us try not to be impressed by what researchers say is surprising about their study—surprising results are often products of data dredging or random error, and are thus less reproducible (Open Science Collaboration 2015). So surprising results will often not point to general scientific discoveries, although they may still be valuable because they lead to new insights about study problems and violations of assumptions.

Then too, we should not overemphasize what researchers say was unsurprising, since that may largely reflect their conformity to group expectations rather than what the data would actually show under close scrutiny. Indeed, we might consider not asking researchers about surprising or unsurprising results, but instead ask which results appeared most boring because they were shown several times before and thus seem to be trustworthy. More generally, we should not fall for overconfident claims by researchers or by other science writers or journalists. Rather, we should try to uncover overconfident claims and the bad incentives that lead to those claims.

Clear signs of overconfidence are formulations like "we proved" or "we disproved" or "we rejected" a hypothesis, or "we have shown" or "demonstrated" a relation exists or is explained in some manner. So are "there was *no* effect/*no* association/*no* difference" (which almost always would be an impossible proof of the null hypothesis), and "our study confirms/validates/invalidates/refutes previous results" (because a single study has nothing definitive, it can only add one further data point to the larger picture; at most it can be "consistent/inconsistent with previous results"). If we find any of those or related phrases, we should question the interpretations being offered in the paper and search for arguments provided by the authors. If the main argument for a conclusion is that the results were "significant" or "not significant," this does not automatically mean that the study is bad. But it does flag the paper as likely providing an unreliable interpretation of the reported results.

The hard truth is that journalists cannot decide whether a result from a single study can be generalized—and the same is usually true for the authors of the study, and for editors and external experts. An important role for statistics in research is the summary and accumulation of information. 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.

Acknowledgments

The authors thank three referees, the associate editor, and the editor Allen Schirm for exceptionally helpful critical comments. For additional

comments, we thank Natalie Jeanneret, Fränzi Korner-Nievergelt, Timothy L. Lash, Lilla Lovász, Michel Montagner, Freya Pappert, and Tobias Roth.

Appendix: A Descriptive View of *P*-values and Posterior Probabilities

We here provide a more technical explanation of why *inferential* statistics like *P*-values and posterior probabilities can be interpreted as being *descriptive* of logical relations between assumptions and the observed data.

Consider a data-generating model M we wish to test and a test statistic (function of the data) T that has been selected as a summary measure of the absolute deviation of the data from the model predictions, with an observed value for T of t.[5] Here, M is the assumption set or constraints used to derive the distribution of T. To say that this test gave back $p = 0.04$ is to say that under the model M, the probability that $T \geq t$ is 0.04, or symbolically that $\Pr(T \geq t|M) = 0.04$. This Fisherian[6] *P*-value of 0.04 is thus a *logical* inference from the model M and the observation that $T = t$, the final deduction "$p = 0.04$" from a derivation that begins with the premises "the data were generated from M" and "those data produced $T = t$." Thus the model and data are in a logical relation in which their conjunction implies a probability p for the tail event $T \geq t$.

Without further elements (such as an α-level cutoff) this observed *P*-value p implies absolutely no decision, inference, bet, or behavior. What people make of $p = 0.04$ in these or other practical terms requires additional contextual detail such as a loss function, acceptable error rates, or whatever else the analysis team can bring to bear (although usually it is just social conventions like $\alpha = 0.05$ and program defaults that determine what gets claimed).

The same comments apply to posterior probabilities. Now, however, "the model" (the set of assumptions used for deducing posterior probabilities) must include or entail a prior distribution $\Pr(M)$ over some restricted model family for generating a probability $\Pr(x|M)$ for the observed data x so that the full model provides a formula for $\Pr(x, M) = \Pr(x|M)\Pr(M)$ (Box 1980). Again the deduction of $\Pr(M|x)$ follows by conditioning on the observed data x (or equivalently on a sufficient statistic t, using $T = t$ as a premise); that is, the posterior $\Pr(M|x)$ becomes a deduction from the observed data x and the full model $\Pr(x,M)$. Nonetheless, because one cannot construct a prior distribution over every conceivable model, this deduction is limited by the assumption that the model family used can approximate (within statistical error) the true data-generating process. In contrast, Fisherian *P*-values can be used to test this assumption before using it in the Bayesian deduction (Bayarri and Berger 2000; Robins et al. 2000).

References

Amrhein, V. (2018), "Inferential Statistics is Not Inferential," sci five, University of Basel, available at *http://bit.ly/notinfer*. [263]

Amrhein, V., and Greenland, S. (2018), "Remove, Rather Than Redefine, Statistical Significance," *Nature Human Behaviour*, 2, 4. [265]

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The Earth is Flat ($p > 0.05$): Significance Thresholds and the Crisis of Unreplicable Research," *PeerJ*, 5, e3544. [263,264,265,267]

Amrhein, V., Trafimow, D., and Greenland, S. (2018), "Abandon Statistical Inference," *PeerJ Preprints*, 6, e26857v1. [267]

Baker, M. (2016), "Is There a Reproducibility Crisis?" *Nature*, 533, 452–454. [262]

Barnard, G. A. (1996), "Fragments of a Statistical Autobiography," *Student*, 1, 257–268. [266]

Bayarri, M. J., and Berger, J. O. (2000), "P-values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142. [269]

Boring, E. G. (1919), "Mathematical vs. Scientific Significance," *Psychological Bulletin*, 16, 335–338. [262]

Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modeling and Robustness," *Journal of the Royal Statistical Society*, Series A, 143, 383–430. [266,269]

Brown, H. K., Ray, J. G., Wilton, A. S., Lunsky, Y., Gomes, T., and Vigod, S. N. (2017), "Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children," *JAMA: Journal of the American Medical Association*, 317, 1544–1552. [266,267]

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T. Z., Chen, Y. L., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E. J., and Wu, H. (2018), "Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* Between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644. [263]

Cohen, J. (1994), "The Earth Is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [265,266]

Cox, D. R. (1978), "Foundations of Statistical Inference: The Case for Eclecticism," *Australian Journal of Statistics*, 20, 43–59. [266]

Crane, H. (2017), "Why 'Redefining Statistical Significance' Will Not Improve Reproducibility and Could Make the Replication Crisis Worse," available at *https://arxiv.org/abs/1711.07801*. [266]

Cumming, G. (2014), "The New Statistics: Why and How," *Psychological Science*, 25, 7–29. [264,265,268]

Edgeworth, F. Y. (1885), "Methods of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume, 181–217. [266]

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, New York: Cambridge University Press. [266]

Fisher, R. A. (1937), *The Design of Experiments* (2nd ed.), Edinburgh: Oliver and Boyd. [262,264]

Gelman, A. (2016), "The Problems with P-values are not Just with P-values," *The American Statistician*, Supplemental Material to the ASA Statement on P-values and Statistical Significance. [264]

Gelman, A., and Hennig, C. (2017), "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society*, Series A, 180, 967–1033. [267]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' is Not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [264]

Gigerenzer, G. (1993), "The Superego, the Ego, and the ID in Statistical Reasoning," in *A Handbook for Data Analysis in the Behavioral Sciences*, eds G. Keren and C. Lewis, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 311–339. [265]

Good, I. J. (1957), "Some Logic and History of Hypothesis Testing," in *Philosophical Foundations of Economics*, ed. J. C. Pitt, Dordrecht, Holland: D. Reidel, pp. 149–174. (Reprinted as Ch. 14 in Good, I. J. (1983), *Good Thinking*, 129–148, Minneapolis, MN: University of Minnesota Press). [266]

Goodman, S. N. (1992), "A Comment on Replication, P-values and Evidence," *Statistics in Medicine*, 11, 875–879. [263,264]

Greenland, S. (2011), "Null Misinterpretation in Statistical Testing and Its Impact on Health Risk Assessment," *Preventive Medicine*, 53, 225–228. [267]

——— (2017), "Invited Commentary: The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. [263,264,265,266,268]

——— (2019a), "Valid P-values Behave Exactly as They Should: Some Misleading Criticisms of P-values and Their Resolution with S-values," *The American Statistician*, this issue. [264,265,266]

[5] There are of course many technical details we must gloss over such as optimization of T to detect a scientifically relevant type of deviation, and how to proceed when as usual M does not fully specify the distribution of T; see Bayarri and Berger (2000) for further discussion. For simplicity we here assume discrete data and model spaces.

[6] That is, a tail probability of an observed statistic. This device actually precedes Fisher, for example in the goodness-of-fit tests of Karl Pearson.

——— (2019b), "The Unconditional Information in *P*—values, and Its Refutational Interpretation via *S*-values," submitted. [263,265,266]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. C., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician*, 70, online supplement 1 at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf*; reprinted in *European Journal of Epidemiology*, 31, 337–350. [263,267]

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015), "The Fickle *P*-value Generates Irreproducible Results," *Nature Methods*, 12, 179–185. [264]

Hurlbert, S. H. and Lombardi, C. M. (2009), "Final Collapse of the Neyman-Pearson Decision Theoretic Framework and Rise of the neo Fisherian. *Annales Zoologici Fennici*, 46, 311–349. [265]

John, L. K., Loewenstein, G., and Prelec, D. (2012), "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling," *Psychological Science*, 23, 524–532. [262]

Lakens, D., Scheel, A. M., and Isager, P. M. (2018), "Equivalence Testing for Psychological Research: A Tutorial," *Advances in Methods and Practices in Psychological Science*, 1, 259–269. [267]

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Springer. [265]

Little, R. J. (2006), "Calibrated Bayes: A Bayes/Frequentist Roadmap," *The American Statistician*, 60, 213–223. [266]

Locascio, J. (2017), "Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39, 239–246. [264,268]

Martinson, B. C., Anderson, M. S., and de Vries, R. (2005), "Scientists Behaving Badly," *Nature*, 435, 737–738. [262]

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019), "Abandon Statistical Significance," *The American Statistician*, this issue. [263,265]

Meehl, P. E. (1990), "Why Summaries of Research on Psychological Theories Are Often Uninterpretable," *Psychological Reports*, 66, 195–244. [263]

Neyman, J., and Pearson, E. S. (1933), "The Testing of Statistical Hypotheses in Relation to Probabilities a priori," *Mathematical Proceedings of the Cambridge Philosophical Society*, 29, 492–510. [265]

Open Science Collaboration. (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [263,264,268]

Poole, C. (1987a), "Beyond the Confidence Interval," *American Journal of Public Health*, 77, 195–199. [265]

——— (1987b), "Confidence Intervals Exclude Nothing," *American Journal of Public Health*, 77, 492–493. [265]

Popper, K. R. (1968), *The Logic of Scientific Discovery* (2nd English ed.), London: Routledge. [263]

Robins, J. M., van der Vaart, A., and Ventura, V. (2000), "Asymptotic Distribution of P-values in Composite Null Models," *Journal of the American Statistical Association*, 95, 1143–1156. [269]

Rothman, K., Greenland, S., and Lash, T. L. (2008), *Modern Epidemiology* (3rd ed., Ch. 10), Philadelphia, PA: Lippincott Williams & Wilkins. [267]

Senn, S. J. (2001), "Two Cheers for P-values?" *Journal of Epidemiology and Biostatistics*, 6, 193–204. [263,264]

——— (2002), "'Letter to the Editor' Re: Goodman 1992," *Statistics in Medicine*, 21, 2437–2444. [263,264]

——— (2011), "You May Believe You Are a Bayesian But You Are Probably Wrong," *Rational Markets and Morals*, 2, 48–66. [266]

Stark, P. B., and Saltelli, A. (2018), "Cargo-Cult Statistics and Scientific Crisis," *Significance*, 15, 40–43. [268]

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C., Beh, E. J., Bilgiç, Y., Bono, R., Bradley, M. T., Briggs, W. M., Cepeda-Freyre, H. A., Chaigneau, S. E., Ciocca, D. R., Carlos Correa, J., Cousineau, D., de Boer, M. R., Dhar, S. S., Dolgov, I., Gómez-Benito, J., Grendar, M., Grice, J., Guerrero-Gimenez, M. E., Gutiérrez, A., Huedo-Medina, T. B., Jaffe, K., Janyan, A., Karimnezhad, A., Korner-Nievergelt, F., Kosugi, K., Lachmair, M., Ledesma, R., Limongi, R., Liuzza, M. T., Lombardo, R., Marks, M., Meinlschmidt, G., Nalborczyk, L., Nguyen, H. T., Ospina, R., Perezgonzalez, J. D., Pfister, R., Rahona, J. J., Rodríguez-Medina, D. A., Romão, X., Ruiz-Fernández, S., Suarez, I., Tegethoff, M., Tejo, M., van de Schoot, R., Vankov, I., Velasco-Forero, S., Wang, T., Yamada, Y., Zoppino, F. C., and Marmolejo-Ramos, F. (2018), "Manipulating the Alpha Level Cannot Cure Significance Testing," *Frontiers in Psychology*, 9, 699. [265,267]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [265]

Wellek, S. (2010), *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd ed.), New York: Chapman & Hall. [267]

Ziliak, S. T., and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, MI: University of Michigan Press. [265]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known

Robert J. Calin-Jageman & Geoff Cumming

To cite this article: Robert J. Calin-Jageman & Geoff Cumming (2019) The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known, The American Statistician, 73:sup1, 271-280, DOI: 10.1080/00031305.2018.1518266

To link to this article: https://doi.org/10.1080/00031305.2018.1518266

© 2019 The Authors. Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 11230

View related articles

View Crossmark data

Citing articles: 32 View citing articles

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, 271–280: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518266

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known

Robert J. Calin-Jageman[a] and Geoff Cumming[b]

[a]Psychology, Dominican University, River Forest, IL; [b]Psychology, La Trobe University, Melbourne VIC, Australia

Abstract

The "New Statistics" emphasizes effect sizes, confidence intervals, meta-analysis, and the use of Open Science practices. We present three specific ways in which a New Statistics approach can help improve scientific practice: by reducing overconfidence in small samples, by reducing confirmation bias, and by fostering more cautious judgments of consistency. We illustrate these points through consideration of the literature on oxytocin and human trust, a research area that typifies some of the endemic problems that arise with poor statistical practice.

ARTICLE HISTORY
Received March 2018
Revised August 2018

KEYWORDS
Estimation; Confidence intervals; Meta-Analysis; Open science; The New Statistics

Scientists try to discern general laws from limited data. A key tool in this effort is inferential statistics—the techniques that help leverage the information obtained from a sample to make reasonable conclusions about the world at large. Surprisingly, many scientists are not using this critical tool well. In some fields, inferential statistics are frequently misunderstood and misused, leading to erroneous conclusions, unreplicable research, and tremendous waste.

Solving this problem is difficult. The statistical issues faced by scientists are complex and domain-specific, so even though there are some clearly wrong practices being used, there is no one "right" way that can cure all statistical ills. Still, we can and must do more to develop practices that will less frequently lead to error.

In this spirit, we advocate use of a "New Statistics"—a simple and fairly generalizable set of principles for making inferences from data that we think can have a positive impact across many disciplines. Specifically, we think that scientists analyzing data should:

- Ask quantitative questions and give quantitative answers.
- Countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error (there are multiple ways to do this, including frequentist confidence intervals and Bayesian credible intervals).
- Seek replication, and use quantitative methods to synthesize across data sets as a matter of course (there are multiple ways to do this, including meta-analysis and the use of informed priors for Bayesian analyses).
- Use Open Science practices to enhance the trustworthiness of research results.

There is nothing new about these recommendations (e.g. Maxwell et al. 2008; Smithson 2002); we have dubbed this a "New Statistics" approach because widespread adoption would be novel for many fields of inquiry. Another good name is the "Estimation Approach".

Currently, many scientists approach inference very differently (Chavalarias et al. 2016); they are trained in what is called null-hypothesis significance testing (NHST). In this tradition, complex studies are summarized primarily by a series of *p*-values, and these are then used to make categorical judgments (e.g. "This drug works" or "This drug doesn't work"), usually with little to no attention to the magnitude of the effect (to what extent does the drug work?) or to uncertainty in generalizing from the sample (how misleading might these results be?).

In this article, we illustrate ways in which the New Statistics approach is better than the dominant NHST tradition. The fault, we argue, is not in *p*-values, but in ourselves[1]. Human cognition is subject to many biases and shortcomings. We believe the NHST approach exacerbates some of these failings, making it more likely that researchers will make overconfident and biased conclusions from data. The New Statistics approach, on the other hand, focuses the researcher's attention on uncertainty to help push back against these tendencies; we believe this helps foster more cautious and even-handed conclusions. In this paper, we illustrate three specific ways in which we think the New Statistics approach can be beneficial: by reducing overconfidence in small samples, by reducing confirmation bias, and by fostering more cautious judgments of consistency. We then consider some common objections to this approach and some of the important cautions in moving forward with reforms toward better inference.

[1] Actually, there are problems with *p*-values themselves, too. But our focus in this paper is specifically on our own cognitive foibles and how these can be mitigated by a New Statistics approach.

CONTACT Robert J Calin-Jageman rcalinjageman@dom.edu Dominican University, 7900 West Division, River Forest, IL 60305.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.
© 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.

To help illustrate the benefits of the New Statistics approach, we review literature that has examined the effect of the neurohormone oxytocin on human trust. Like many other areas of preclinical research, there has been a growing awareness that the literature on oxytocin and human trust may be unreliable. We choose to focus on this literature because it provides a particularly clear case of problems that are endemic across many research domains that depend on NHST.

1. The New Statistics Approach Can Help Reduce Overconfidence in Small Samples

Samples can provide useful information about the world at large. Mathematically speaking, we know that larger samples provide more certain information, whereas smaller samples provide less certain information (assuming all else is equal). In practice, though, humans often show undue confidence in conclusions drawn from small samples (e.g. Tversky and Kahneman 1971; this has been ironically dubbed the "law" of small numbers, but a better name is the fallacy of hasty generalization). The New Statistics seeks to mitigate this problem by making uncertainty salient. The use of p-values, on the other hand, seems to exacerbate overconfidence in small samples.

As an example, consider a study by Kosfeld et al. (2005) examining the influence of the neurohormone oxytocin on social behavior in humans. In this experiment, 58 participants were assigned to receive an intranasal dose of either oxytocin or placebo. Next, all participants played an economic trust game. Kosfeld et al. (2005) used the NHST approach to analyze the data, reporting that oxytocin had a statistically significant effect on trust (Mann–Whitney $U = 301, p = 0.029$ one-sided, Figure 1A). From this (and the fact that oxytocin did not seem to influence performance on a control task) the researchers made a categorical and unequivocal claim: "Oxytocin increases trust in humans". This claim was perceived by other scientists as well-founded: Kosfeld et al. (2005) was published in *Nature* and has now been cited over 3000 times (according to Google Scholar).

The New Statistics approach offers a very different lens from which to interpret the same data (Figure 1B). Under this approach, we would ask the *quantitative* question: By how much does oxytocin increase trust in humans? The answer given would not be a p value but an effect-size estimate: median trust behavior in the oxytocin group is estimated to be 12.5% higher than in the placebo group, a notable increase (technical details on our re-analysis of this data are presented in the last section of this article). This effect size should not stand on its own—we also need to countenance the uncertainty inherent in generalizing from a sample of this size. One popular way to express uncertainty is to report a confidence interval; in this case the 90% confidence interval is [0.0002%, 37.5%]. This means that if the real effect was very large (37.5% increase in median trust) the data obtained would not be especially unusual or remarkable. On the other hand, the data would also be unsurprising if the true effect is vanishingly small, as small as 2 ten-thousandths of a percent. Based on this, we should consider the experiment uninformative; the study is intriguing but it does *not* demonstrate that oxytocin has a meaningful impact on trust.

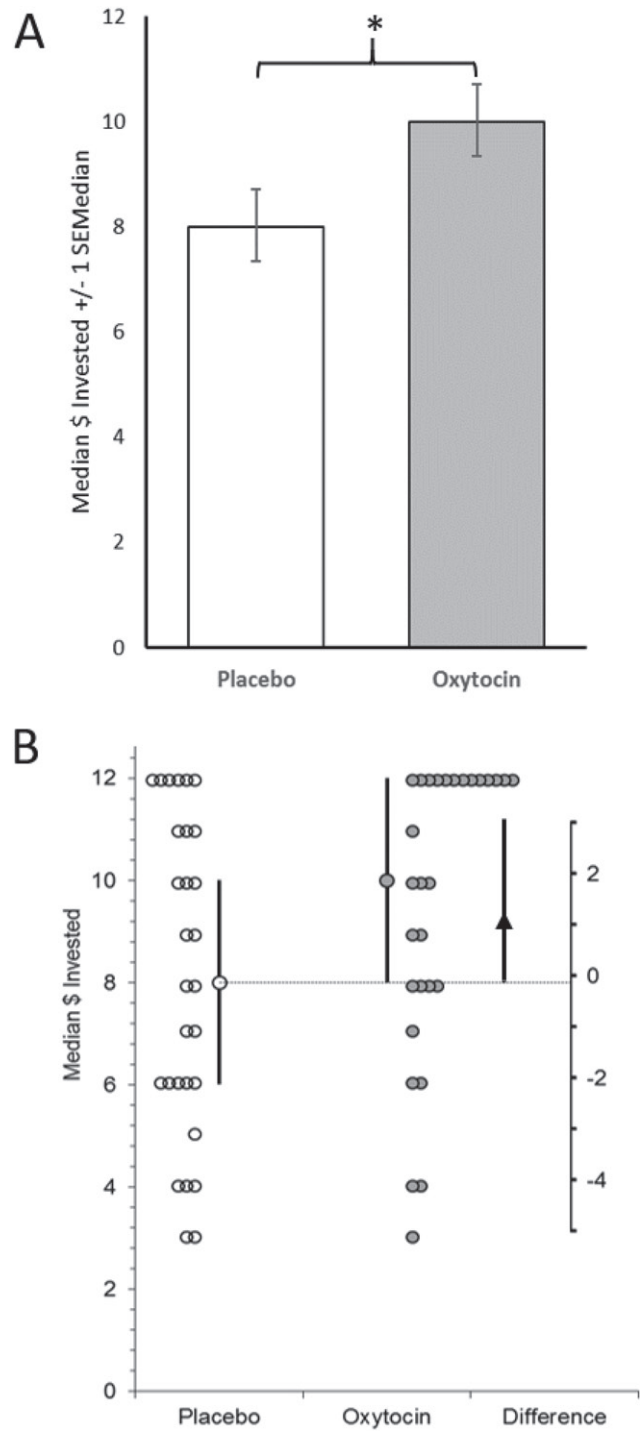


Figure 1. Two ways of looking at the same data. This figure compares the NHST approach (A) and the New Statistics approach to visualizing the same data (B). The data is from Kosfeld et al. (2005) on the effect of intranasal oxytocin on dollars invested in a trust game. In A, a bar graph is used to show median trust and standard error for each group. The * indicates a statistically significant difference in a one-tailed test ($p = .029$). In B, all the individual data is shown (circles). Each circle with an error bar represents the group median along with the 95% CI for the median. The plot emphasizes the effect size, which is the difference between the two groups (marked by a triangle, which is an increase of $1 in median trust). The error bar represents the uncertainty about that estimate; it is the 90% CI of the difference, which is [0.00001, 2.99]. The confidence interval is not symmetrical around the point estimate. See the last section of the paper for technical details on how this data was summarized.

With the benefit of hindsight, we know that this cautious conclusion is probably accurate: only a very small quantity of intranasally delivered oxytocin is now thought to reach the

brain, so it may not be possible for this route of administration to reliably influence trust (reviewed in Leng and Ludwig 2016a; b; though see the commentary by Quintana and Woolley 2016). Consistent with this, attempts to replicate the findings of Kosfeld et al. (2005) have so far been unsuccessful (summarized in Nave et al. 2015; see also a critical commentary by Quintana 2018).

In this example, the New Statistics leads to a very different (and more judicious) conclusion than the NHST approach. What is striking is that the confidence interval reported here is based on the exact same data, models, assumptions, and standards of evidence as the *p*-value originally reported. Although both approaches *ought* to lead to a very tentative conclusion, the *p*-value approach does not make the uncertainty in the sample salient and the tendency towards hasty generalization is left unchecked. The New Statistics emphasizes quantifying and interpreting uncertainty; this should help researchers better calibrate their conclusions to the evidence.

The example provided here is not unusual. There is diversity across different fields of inquiry, but many research communities rely primarily on samples that are far too small relative to the research claims that are made (e.g., Button et al. 2013; Szucs and Ioannidis 2017). In oxytocin research, for example, Kosfeld et al. (2005) is entirely typical. A recent meta-analysis of studies manipulating oxytocin in humans found a median sample-size of just 25 participants/group (Walum et al. 2016). The authors of the meta-analysis concluded that the published research literature on intranasal oxytocin is generally unreliable.

2. The New Statistics Approach Can Help Reduce Confirmation Bias

Confirmation bias is the selective appraisal of evidence: evidence supporting one's position is accepted uncritically while contrary evidence is discounted or even ignored. Humans have a strong tendency toward confirmation bias; the scientific method is a cognitive prosthetic designed to mitigate this bias.

Although scientists agree that evidence must be weighed impartially, few seem to recognize that the NHST approach institutionalizes confirmation bias. The NHST approach dichotomizes research results into "statistically significant" and "not statistically significant". That would be fine if both types of results were weighed equally, but they are not. Researchers are taught that a lack of statistical significance is inconclusive—it *could* mean their hypothesis is wrong, but it could also mean the experiment was not conducted properly or that not enough data were collected. Statistical significance, on the other hand, is treated as conclusive—worries about sample size or procedural error might go right out the door. This "heads I win, tails you lose" approach means that negative results are routinely discarded into the "file drawer" rather than being published and weighed alongside more favorable results. The file drawer problem has been well-documented but unsolved for almost 60 years (Sterling 1959; Sterling et al. 1995).

The New Statistics approach can help decrease confirmation bias in the published literature. Under the New Statistics approach, results are not dichotomized into "significant" or "nonsignificant"—there are simply estimates and uncertainty. Of course, every research result needs to be evaluated carefully. There are factors that can cause underestimation (e.g. insufficient manipulation, careless responding, noisy measurement, etc.). But there are just as many factors that can cause overestimation (e.g. experimenter bias, failure of random assignment, differential dropout, etc.). Thus, the New Statistics approach emphasizes applying the same epistemic standards regardless of the result. Moreover, the New Statistics emphasizes synthesis, especially through meta-analysis, to help ensure that all evidence is weighed. Although meta-analysis cannot magically open the file drawer, it can often indicate when the data available for synthesis is biased, helping to sound the alarm that a hypothesis is not being fairly evaluated.

The impact of confirmation bias has been acutely illustrated in the literature on oxytocin and human social behavior. We now know that oxytocin delivered via the nose could be inert in terms of brain function (Leng and Ludwig 2016b). And yet there is a large published literature showing statistically significant and large effects of intranasal oxytocin on human behavior: it is reported to increase self-confidence, boost trust, facilitate social interactions, increase altruism, and much more (reviewed in Walum et al. 2016). How could an inactive treatment be "proven" effective? Some researchers hold out hope that intransal oxytocin may be active, perhaps via the peripheral nervous system or by reaching the brain at important but unmeasurable levels (Neumann and Landgraf 2012; Quintana and Woolley 2016). A recent meta-analysis, however, offers a much simpler explanation: confirmation bias. Specifically, it was found that the published literature shows "excess significance": more studies are statistically significant than plausible given the small sample sizes used (Walum et al. 2016). This means that many "negative" findings have been either suppressed or inappropriately coaxed under the threshold for statistical significance. Consistent with this possibility, one brave lab has opened its file drawer for inspection (Lane et al. 2016), reporting that the five positive results it has published represent just 39% of the 13 different tests the lab had conducted. The lab had tried to publish the eight less favorable results, but these were "rejected time and time again" (p. 38). Having only positive results reach the public distorts research findings: although the lab's public output seemed to show a strong effect of oxytocin, meta-analysis across all of its work indicates oxytocin had little to no effect on the trust behaviors that had been measured.

At this stage, there is a voluminous published literature on oxytocin and human trust, but it is unclear if any very reliable knowledge has been generated. Moreover, some harm has been done. Based on what seemed like an unequivocal research literature, a number of clinical trials have been launched examining whether intranasal oxytocin could enhance social function in children with autism. These trials have not yet yielded strong evidence of a benefit (though see Yatawara et al. 2016). Unfortunately, almost 40% of the 261 children so far treated with oxytocin have suffered an adverse event (DeMayo et al. 2017; compared with only 12% of the 170 children assigned a placebo). Thankfully, most (93) of these adverse events were mild; but six were moderate, and three severe.

Could the New Statistics and Open Science have helped prevent the oxytocin debacle? We think so. First, focusing on effect sizes might have made it clear that some reported effects of oxytocin on trust are too large to be credible. Second, making

uncertainty salient would have indicated the acute need for direct replication, something which has been unacceptably rare in this literature and in many other fields of study (Makel et al. 2012). Third, dispelling the false dichotomy between "positive" and "negative" results might have made it more likely for unfavorable results to be published and considered. Fourth, Open Science practices might have improved transparency and completeness in reporting. Finally, meta-analysis has already played a critical role in bringing the problems in the oxytocin literature to light.

The file drawer problem is anathema to good science. It is a disgrace that this problem has persisted for so long since its initial recognition. If the New Statistics approach can even slightly improve the situation, it will be well worth adoption.

3. The New Statistics Approach Can Foster More Cautious Judgments of Consistency

Researchers frequently judge the consistency of results across different conditions or contexts. For example, a researcher may test the effects of a drug on both memory and anxiety and then try to judge if the drug influences both measures similarly, or if there is a selective effect that is specific to just memory or just anxiety. Judgments of consistency are important tools in discerning mechanisms (why an effect occurs) and boundary conditions (in what contexts an effect will occur). Accurate judgments of consistency are essential to fruitful science.

The correct method for judging the consistency of two results is to make a direct statistical comparison that factors in the uncertainty in both data sets. In the NHST approach, this means examining the "interaction" to see if it is statistically significant. In the New Statistics approach, this means estimating and interpreting the "difference of the difference" to quantify the degree to which the two sets of results differ and the uncertainty in generalizing from the samples.

If that last paragraph made your head spin, you are not alone: comparing sets of results is a difficult topic in inferential statistics. Unsurprisingly, many competent researchers avoid the complex (but correct) solution in favor of heuristic approaches that are simpler, but sometimes wrong. There are heuristic approaches for both the NHST approach and the New Statistics approach. Both can lead to error, but in very different ways: the NHST approach sets a standard that is too low; the New Statistics sets a standard that is too high.

Under the NHST approach, researchers often take the shortcut of judging consistency through the comparison of statistical significance status: two results are judged to be consistent if their statistical significance matches (both significant or both not significant), or inconsistent if one is statistically significant and the other is not. For example, Kosfeld et al. (2005) noted that oxytocin had a statistically significant effect on a game of trust, but did not have a statistically significant effect in a similar game that involved risk but not trust (the other player was not a human but an algorithm that selected responses at random). Because one effect was statistically significant but the other was not, the results were judged to be inconsistent and it was concluded that oxytocin has a selective effect on trust (Figure 2).

This approach is frequently found in the published literature (e.g. Nieuwenhuis et al. 2011) but it is too generous by far. Specifically, "selective" effects can be obtained by comparing a statistically significant finding against another set of results drawn from an inadequate sample. For example, Kosfeld et al. (2005) recruited only 61 participants for the study of how oxytocin influences behavior in the nontrust game. That's not a fair test: even if oxytocin affects behavior in both games similarly, the second study is more likely than not to find nonsignificant results due to inadequate sample size. When this heuristic approach is replaced by a formal test for consistency of results, the outcome is not statistically significant ($p = 0.23$). Thus, the short-cut used by Kosfeld et al. (2005) led to a claim that oxytocin selectively influences trust, but the correct analysis does not provide adequate support for this conclusion. This is not an uncommon situation—many researchers use this shortcut approach and many fields of inquiry rely on inadequate samples. This combination will often lead to spurious claims of specificity[2].

With the New Statistics approach, researchers can also take a shortcut by simply comparing the confidence intervals for both sets of results (Figure 3B): the results are judged inconsistent only if the intervals have no overlap. This approach is not valid; results can be statistically inconsistent even with a small degree of overlap between confidence intervals (see Greenland et al. 2016 for a detailed explanation). Thus, the natural inclination to use a "no overlap" rule is actually overly cautious, in many cases equivalent to requiring $p < 0.01$ to judge results inconsistent (see Cumming and Finch 2005 for an explanation and important caveats).

It may seem a bit odd that we are highlighting here a way in which the New Statistics fosters invalid research conclusions. But we think this reveals several important points. First, there are complex interactions between human cognition and statistical inference. Even the same statistical models can evoke different reactions when summarized in different ways. Second, the New Statistics is not a panacea—it may protect against some cognitive foibles but foster others. Finally, there is much more work to be done in the study of *statistical cognition*—in the way both experts and the public respond to different ways of presenting inferential statistics. We think the New Statistics approach is often the best way of summarizing results because it makes uncertainty salient, but continued empirical work can help work out what types of statistical summaries best support good judgments in different contexts.

4. Open Science Practices Are Needed, Too

Predictive power is one of the ultimate signs that science is working well. Because of this, data analysis is usually divided into two domains: exploratory, where a researcher is discovering trends but must be extremely tentative about conclusions, and planned,

[2] Comparing significance levels can also mask real differences. For example, suppose both Drug A and Drug B have statistically significant benefits on feelings of depression. Many researchers would therefore judge the drugs to have consistent effects. In reality, though, the drugs might differ substantially in *how much* they influence feelings of depression. Being cognizant of a difference in magnitude can have important implications for making good inferences.

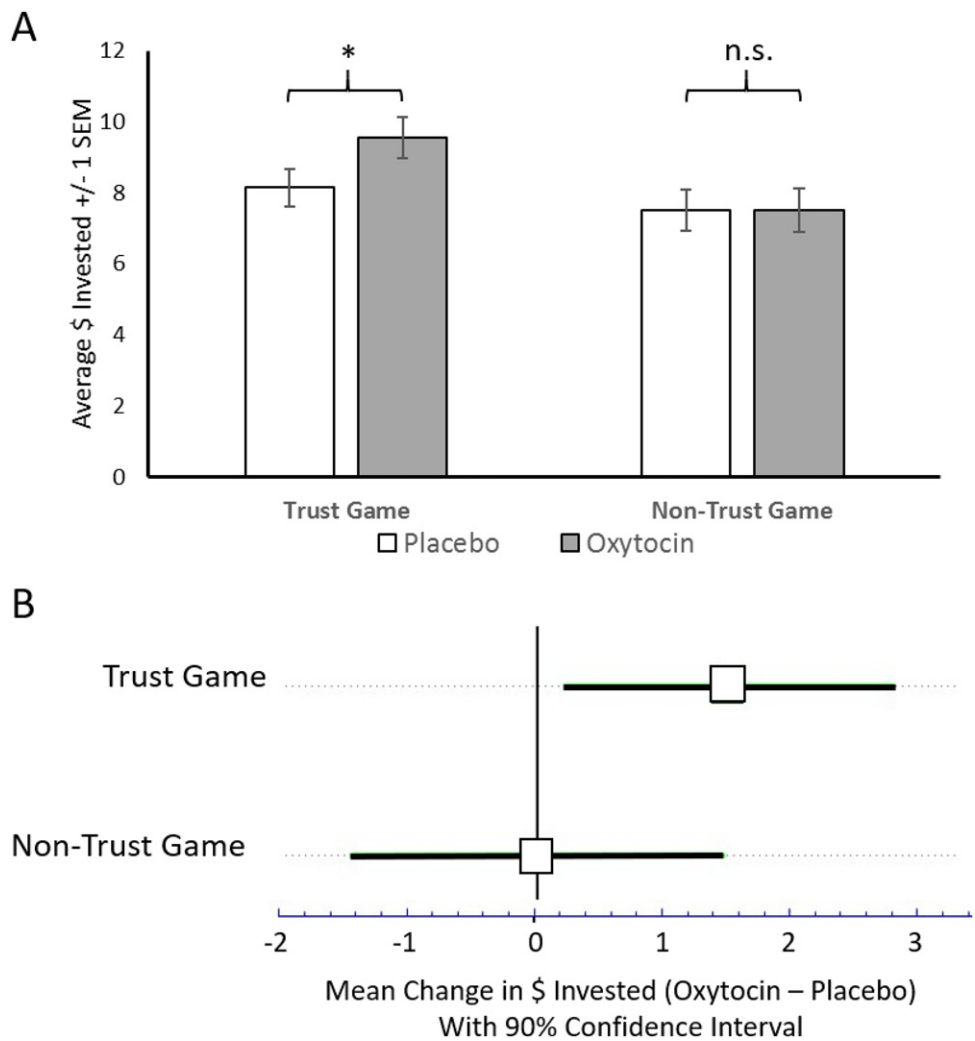


Figure 2. Two heuristic approaches to judging consistency. This figure compares the NHST approach (A) and the New Statistics approach (B) to making heuristic judgements of consistency. The data is from Kosfeld et al. (2005) on the effect of intranasal oxytocin on dollars invested in a trust game and nontrust game that involved only risk. In A, a bar graph is used to show average trust with standard error. The * indicates a statistically significant difference in a one-tailed test ($p = 0.04$) for the trust game whereas the n.s. indicates a nonsignificant test in the nontrust game. On this basis, many researchers erroneously judge the results to be inconsistent. However, a direct test for the interaction of oxytocin and game type is not statistically significant ($p = 0.20$ for a standard ANOVA test for interaction). In B we focus on the effect of oxytocin in each context, plotting the difference in mean trust between the oxytocin and the placebo groups, along with the 90% CIs. The strong overlap between the CIs suggests, correctly, that this is not enough data to judge the results inconsistent. In this figure means are compared for ease of analysis. Kosfeld et al. (2005) actually used nonparametric tests and focused on comparing medians, but this also indicates a nonsignificant interaction between drug and task ($p = 0.23$). The last section of this article has technical details on how we re-analyzed the data from Kosefeld et al. (2005).

where a researcher seeks to confirm specific predictions. While this sounds simple in principle, 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 (e.g., O'Boyle et al. 2014). Moreover, researchers can be tempted to use flexible definitions of what counts as confirmation.

The literature on oxytocin and human social behavior again offers a striking example of this problem. One research group recently defended the quality of the research in this field by pointing out that they had found a beneficial effect of oxytocin on social-cognition in patients with schizophrenia and that this "particular effect" had been independently replicated (Quintana and Woolley 2016). That seems like striking evidence. Here, though, is a more complete description of these results provided by Leng and Ludwig (2016b):

- In the first study (Woolley et al. 2014), patients treated with oxytocin or placebo were given two different tests of social cognition. There were no overall effects. Instead, there was an effect on one of the subelements of the second test.
- In the second study (Davis et al. 2014), three tests of social cognition were administered. None showed significant effects of oxytocin. However, a marginally significant effect was found when combining one subelement of the third part of the second test with the results from one of the other tests.

Underneath the claim of consistently replicated results, then, there is very generous standard for what it means for "particular effect" to be replicated.

Changing statistical approaches will not help with these types of issues. Any approach can be used to selectively report positive evidence, to dress up exploratory research as planned, and/or to allow undue flexibility in what counts as confirmation of a

hypothesis. Fortunately, Open Science practices can help ameliorate these problems. One practice that is quickly gaining popularity is public preregistration (Mellor and Nosek 2018). This is where a researcher publicly posts full details of their planned research, including their predictions and the specific planned analyses that will be used to check these predictions. Preregistration helps foster more thoughtful initial planning of the research and provides a publicly verifiable firewall between exploratory and planned analysis. Even better, some journals are now implementing preregistered review (e.g., Bernard 2017), where peer review occurs *prior* to data collection as well as after the study is complete. This enables peer reviewers to help researchers identify and fix flaws in their approach *when it still might be helpful*. In addition, preregistered review commits the journal to an "in principle" acceptance of the paper regardless of results. This helps focus the journal evaluation process on selecting good research questions being answered with reliable methods. Most importantly, preregistered review offers a fighting chance to finally solve the file drawer problem[3].

The Open Science movement is still young and evolving. We expect that any new practice adopted may bring concomitant problems and could also open up new opportunities for the system to be gamed. But it would be foolish to let the best be the enemy of the good. Increasing transparency in the research process, especially as it relates to exploratory vs. planned analysis, can have a strong positive impact on science.

5. Some Unpersuasive Counterarguments

There is tremendous resistance to abandoning *p* values and the NHST approach:

- Some have argued that there is there is no need to change because there are no systematic or unusual problems in current scientific practice (e.g. Gilbert et al. 2016). According to this point of view replicability in science is high, and any areas of low replicability are not representative of the whole. Statistically speaking, we know that this sunny outlook is implausible. The oxytocin and trust literature went astray due to insufficient sample sizes, publication bias, and post-hoc reasoning. These problems are not unique to neurohormone research, rather they are endemic across many of the behavioral and life sciences (Button et al. 2013; Dumas-Mallet et al. 2017; Szucs and Ioannidis 2017).
- Some have argued against change due to optimism, arguing that if we simply taught and used the NHST approach correctly all would be fine. We do not believe that the cognitive biases which *p*-values exacerbate can be trained away. Moreover, those with the highest levels of statistical training still regularly interpret *p*-values in invalid ways (McShane and Gal 2015). Vulcans would probably use *p*-values perfectly; mere humans should seek safer alternatives.
- Others have argued against change due to cynicism. For example, Nelson, Simmons, and Simonsohn (2018) recently argued that *p*-values are no more problematic than any other approach, because all approaches can be used mindlessly. We agree that all approaches can be misused. Indeed, researchers can make rote reports of confidence intervals without actually countenancing uncertainty (Fidler et al. 2004). Still, the argument by Nelson et al. (2018) is like claiming that all cars are equally safe just because any car can crash. What matters is how likely one is to go wrong with the NHST approach compared to the alternatives. Further study is required, but we think the New Statistics approach can help researchers more regularly make good judgments from data.
- Another frequent concern is that scientists need to make clear Yes/No decisions (e.g. Does this drug work? Is this project worth funding?). No problem! Focusing on effect sizes and uncertainty does not preclude making decisions—in fact, it makes it easier because one can easily test a result against any desired standard of evidence. For example, suppose you know that a drug improves outcomes by 10% with a 95% confidence interval from 2% up to 18%. If the standard of evidence required is at least a 1% increase in outcomes, the drug would be considered suitable (because a 1% increase is not within the range of the confidence interval). For a standard of evidence of at least a 5% increase, though, the drug would be considered not yet proven to be suitable (because although 5% is plausible, so is the drug providing only a 2, 3, or 4% benefit). Thus, when focusing on effect sizes and uncertainty over *p* values one can (if absolutely necessary) make clear-cut Yes/No decision. It is important to note, though, that in practice Yes/No decisions should not be reduced to a simplistic ritual. Decision-making should be informed not only by the best available statistical evidence but also costs and benefits, the alternatives available, relevant values or norms, and more. This is especially important when it comes to publication decisions—these should be guided by the *quality* of the study, not the data that has been obtained.
- Some have argued that we can retain *p*-values but interpret them quantitatively, as a measure of strength of evidence against the null. This is a step forward from the dichotomous interpretation of *p*-values, but it still fails to highlight practical significance. Moreover, it seems that in practice even experts fall into the trap of thinking of *p*-values in a Yes/No fashion (McShane and Gal 2015).
- Finally, others have sought to rescue the NHST approach by tightening standards, advocating for a redefined and more stringent threshold for declaring statistical significance (Benjamin et al. 2018). This pours gasoline onto the fire. This "solution" doubles down on the arbitrary dichotomy between "significant" and "not significant" and will exacerbate publication bias by pushing more results into the file drawer.

Not all resist abandoning *p*-values; many are ready to move on to other approaches. It has been difficult, though, to rally reformers around a common set of principles. Notably, some have criticized the New Statistics approach for not going far enough, arguing confidence intervals have too many of the foundational flaws that plague the NHST approach. Traditional confidence intervals do have weaknesses, but this is not a reason to reject a New Statistics approach. There are many ways to quantify and express uncertainty, including Bayesian credible

[3] There is much more to the Open Science movement (e.g., sharing data and code, being fully transparent in reporting). We are brief here because others have discussed the benefits of Open Science with more detail and authority (e.g., Nosek et al. 2015, 2018).

intervals (Kruschke and Liddell 2017). Thus, one can make a clean break from frequentist statistics while still focusing on estimation, uncertainty, and meta-analysis. What does not make sense, to our mind, is to adopt Bayesian inference but then to use it in the same mindless decision-making framework that makes the NHST approach so dangerous. We suspect that, at present, confidence intervals are more likely to be understood and used by researchers steeped in NHST. The larger jump to Bayesian estimation could follow.

6. Some Persuasive Cautions

One important caution is that expressions of uncertainty are themselves uncertain. Confidence intervals and credible intervals are based on mathematical ideals; they express the sampling error one would expect in an otherwise perfect study. In practice, samples are often more misleading than the math would predict (e.g., Shirani-Mehr et al. 2018). Researchers should always keep in mind that expressions of uncertainty are a *best case* scenario. In particular, it is important not to put too much emphasis on the boundaries of a particular confidence interval or credible interval; these boundaries are arbitrary.

Another important caution is that inferential statistics are not the ultimate arbiter of scientific truth. McShane et al. (2017) have argued that it is important for researchers to refocus on what have lately been neglected factors in evaluating scientific results: the reliability and validity of the measurement, the quality of the design, interpretation of positive and negative controls, the plausibility of the mechanism, and the practical real-world implications. We whole-heartedly agree.

A final important caution is that the problems in science are not merely statistical. There are also structural, economic, and cultural issues that are increasingly problematic as science nears its 300th year of nearly continuous exponential growth. Inferential statistics is just one area in need of reform in order to maintain the unparalleled productivity of modern science.

7. Ways Forward

The inertia behind p-values is considerable, and 60+ years of incisive criticism has not yet dethroned NHST as the dominant approach to inference in many fields of science. Momentum, though, seems to finally be on the side of reform. Journals are increasingly revising publication guidelines to require authors to report effect sizes and confidence intervals (e.g., Eich 2014). The American Psychological Association has revised its publication manual (2010) to enjoin reporting *and interpreting* effect sizes and confidence intervals (reviewed in Fidler 2010). Open Science practices are becoming increasingly common across many branches of scientific inquiry (see, e.g., Giofrè et al. 2017).

In addition to internal reforms, we think an important catalyst for change will be from the outside, from the many stakeholders that fund, consume, and depend on scientific research. One of the persistent barriers for meaningful interactions with stakeholders has been 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. If stakeholders realized that instead of simplistic yes/no judgments and mysterious p-values scientists could speak to them in the simple language of effect size and uncertainty, they would rightfully demand that all results be presented this way. We strongly encourage this response. When engaging with research outputs, stakeholders should be unafraid to ask:

- How much does this help?
- How wrong might you be?
- What else is known that might influence your conclusions?

Leverage from outside forces would help internal reforms gain more traction and might finally help end the pernicious NHST tradition.

Although we are optimistic about the future, it is important to be clear-eyed about the daunting task ahead. The NHST approach is a complex ecosystem. It is embedded in textbooks, in the assessments that colleges use to evaluate their students, in software tools, and most importantly in the way several generations of scientists have been trained to think. Moving towards a New Statistics approach will thus require systemic change across the scientific landscape, not just updates to journal and publishing guidelines. There is lots of interesting and helpful work to be done; we hope many will take up the challenge.

8. Further Reading

For those interested in learning more about the New Statistics approach, we recommend:

- Video Tutorials on the New Statistics designed specifically for those already trained in the NHST approach: *http://www.psychologicalscience.org/members/new-statistics* (see also Cumming 2011).
- Our recent textbook for students just getting started with inferential statistics (Cumming and Calin-Jageman 2017). It emphasizes the New Statistics and Open Science practices.
- ESCI – a free set of Excel workbooks for calculating and visualizing effects sizes and confidence intervals for a number of common research designs. ESCI is available for download at: *https://thenewstatistics.com/itns/esci/*
- Works on Bayesian New Statistics, especially this short overview (Kruschke and Liddell 2017) and this excellent textbook (Kruschke 2014)
- Resources for sample-size planning to obtain accurate effect size estimates, especially the work of Ken Kelley and colleagues (e.g. Kelley 2007; Maxwell et al. 2008) and the useful web tools they have made available: *https://designingexperiments.com/*
- The excellent book by Kline (2004) that summarizes the history of the NHST approach, outlines its flaws, and provides detailed exposition and references on the calculation of frequentist confidence intervals. Smithson (2002) also provides a highly accessible introduction to confidence intervals.

9. Technical Details

Kosfeld et al. (2005) summarized their research on oxytocin and trust using the NHST approach. In Sections 2 and 3, we showed how this data would be summarized using the New

Statistics approach. Our goal was to stay as true to Kosfeld's approach as possible, using the same models and assumptions, but to summarize the data through effect sizes and uncertainty. It is important to note that regardless of the analytic strategy the data from Kosfeld et al. (2005) is highly uncertain; it is compatible with a wide range of oxytocin effects including the possibility that effects are vanishingly small. The point of our re-presentation is simply to illustrate how differently the same results are judged when summarized with p values compared to with estimates of effect sizes and uncertainty. Data and analysis scripts are posted at *https://osf.io/54n9q/*.

9.1. Data Extraction

To recover the raw data from Kosfeld et al. (2005), we drew on their Figure 2, which presents histograms of the data from the trust and risk experiments. However, we found that the bar heights in Figure 1A sum to just 26 control participants even though the text of the manuscript reports that there were 29 control participants. Through trial and error we found that imputing three additional scores of 10 in the control group of the trust experiment yielded a dataset that reproduced all the statistics reported in the main text and Table 1. Notably, Figure 1A seems to be missing a bar for control group participants scoring a 10; it seems likely that this aspect of the figure was clipped or misprinted.

9.2. Original Research Strategy

Kosfeld et al. (2005) made inferences about the population by using the Mann–Whitney U test (also known as the Wilcoxon test). This tests the equality of the distribution functions for the two groups. Kosfeld et al. (2005) interpreted statistically significant results to indicate differences not only in group distributions but also in group medians (e.g. "These differences in the distribution of trust result in higher average and median trust levels for subjects given Oxytocin", p. 674). The interpretation of differing group medians requires the assumption that the distributions are symmetrical. For the critical test comparing the placebo and oxytocin in the trust experiment, a one-tailed test was reported ("$p = 0.025$ one sided"). All other tests reported were two-sided.

9.3. Comparing Oxytocin And Placebo Groups in The Trust Experiment

As Kosfeld primarily focused on median trust, we selected the difference between median trust as our measure of effect size. This was calculated using the Hodges–Lehmann estimator. Technically, this calculates the psuedomedian, but if one assumes the distributions are symmetrical (as Kosfeld et al., 2005 seem to have assumed), then the psuedomedian coincides with the median. For clarity, we elided this distinction in the main text and discussed the effect size simply as a difference in medians. Note, again, that other analytic approaches (e.g., estimating the difference between means with a t-test) would still lead to the same conclusions.

There are multiple ways to quantify uncertainty, including Bayesian credible intervals and frequentist confidence intervals. As Kosfeld et al. (2005) used a frequentist approach, we report frequentist confidence intervals for the difference between medians. This was again based on the Hodges–Lehmann estimator. Calculations were made in R using the Wilcox.test function which can also provide the estimated psuedomedian and its confidence interval. To match the stringency of the one-tailed test used by Kosefeld et al. (2005), we calculated and report 90% confidence intervals.

In the text, we express the difference between medians in % terms (e.g. a 12.5% increase in median trust). To obtain % changes, we divided the estimated change in location (1) by the median in the placebo group (8).

9.4. Estimated Power for the Non-trust Experiment

In Section 3 we estimate the power Kosfeld et al. (2005) had to detect an oxytocin effect in the nontrust experiment in which participants played a game that involved the same risk but not trust. For the trust experiment, the standardized mean difference in performance was $d = 0.47$. We estimated the power to detect this effect size in the non-trust experiment, which involved 61 total participants and was analyzed with a two-tailed test with an alpha of 0.05. For an independent samples t-test, power would be 0.45 if all assumptions were perfectly met. Kosfeld et al. (2005) actually used a nonparametric test. Parametric tests can lose power dramatically with violations of assumptions whereas non-parametric tests are more robust. However, our estimated power of 0.45 is for a t-test under ideal circumstances and thus is optimistic for the analysis strategy utilized by Kosfeld et al. (2005). Although the ideal way to estimate a priori power for the nontrust experiment could be debated, it seems unequivocal that the experiment did not have a sample-size sufficient to regularly detect the expected effect of oxytocin, and thus the negative results are unconvincing as a test for specificity of an oxytocin effect.

9.5. Drug × Task Interaction

Kosfeld et al. (2005) made a number of comparisons between groups in the trust and nontrust experiments. They did not, however, test the critical interaction between drug and task. To do this with the nonparametric approach used in the original study, we used the aligned rank transform test developed by Higgins and Tashtoush (1994), implemented as the aligned.rank.transform function from the ART package in R. This indicated a nonsignificant interaction, $p = 0.23$. An ANOVA test for an interaction also indicated a nonsignificant interaction between group and task, $p = 0.20$.

Acknowledgment

The authors thank Rebecca Pliske and Daniel Quintana for helpful comments on an earlier draft of this article.

Funding

This work was partially supported by NIMH Grant 1R15MH107892-01.

References

American Psychological Association (2010), *Publication Manual of the American Psychological Association*, Washington, DC: Author. [277]

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., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [276]

Bernard, C. (2017), "Editorial: A Message from the Editor-in-Chief," *Eneuro*, 4, ENEURO.0023-17.2017. [276]

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013), "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 14, 365–376. [273,276]

Chavalarias, D., Wallach, J. D., Li, A. H. T., and Ioannidis, J. P. A. (2016), "Evolution of Reporting P Values in the Biomedical Literature, 1990–2015," *JAMA*, 315, 1141. [271]

Cumming, G. (2011), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, New York: Routledge. [277]

Cumming, G., and Calin-Jageman, R. J. (2017), *Introduction to the New Statistics: Estimation, Open Science, and Beyond*, New York: Routledge. [277]

Cumming, G., and Finch, S. (2005), "Inference by Eye: Confidence Intervals and How to Read Pictures of Data," *The American Psychologist*, 60, 170–180. [274]

Davis, O. S. P., Band, G., Pirinen, M., Haworth, C. M. A., Meaburn, E. L., Kovas, Y., Harlaar, N., Docherty, S. J., Hanscombe, K. B., Trzaskowski, M., Curtis, C. J. C., Strange, A., Freeman, C., Bellenguez, C., Su, Z., Pearson, R., Vukcevic, D., Langford, C., Deloukas, P., Hunt, S., Gray, E., Dronov, S., Potter, S. C., Tashakkori-Ghanbaria, A., Edkins, S., Bumpstead, S. J., Blackwell, J. M., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A., Duncanson, A., Jankowski, J. A. Z., Markus, H. S., Mathew, C. G., Palmer, C. N. A., Rautanen, A., Sawcer, S. J., Trembath, R. C., Viswanathan, A. C., Wood, N. W., Barroso, I., Peltonen, L., Dale, P. S., Petrill, S. A., Schalkwyk, L. S., Craig, I. W., Lewis, C. M., Price, T. S., Donnelly, P., Plomin, R., and Spencer, C. C. A. (2014), "The Correlation Between Reading and Mathematics Ability at Age Twelve Has a Substantial Genetic Component," *Nature Communications*, 5, 4204. [275]

DeMayo, M. M., Song, Y. J. C., Hickie, I. B., and Guastella, A. J. (2017), "A Review of the Safety, Efficacy and Mechanisms of Delivery of Nasal Oxytocin in Children: Therapeutic Potential for Autism and Prader-Willi Syndrome, and Recommendations for Future Research," *Springer International Publishing*, 19, 391–410. [273]

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., and Munafò, M. R. (2017), "Low Statistical Power in Biomedical Science: A Review of Three Human Research Domains," *Royal Society Open Science*, 4, 160254. [276]

Eich, E. (2014), "Business Not as Usual," *Psychological Science*, 25, 3–6. [277]

Fidler, F. (2010), "The American Psychological Association Publication Manual: Implications for Statistics Education (6th ed.)," *Proceedings of ICOTS-8, Eighth International Conference on Teaching Statistics*, 8. [277]

Fidler, F., Thomason, N., Cumming, G., Finch, S., and Leeman, J. (2004), "Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think: Statistical Reform Lessons from Medicine," *Psychological Science*, 15, 119–126. [276]

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016), "Comment on 'Estimating the Reproducibility of Psychological Science,'" *Science*, 351, 1037–1037. [276]

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., and Tressoldi, P. (2017), "The influence of Journal Submission Guidelines on Authors' Reporting of Statistics and Use of Open Research Practices," (J. M. Wicherts, ed.), *PLOS ONE*, 12, e0175583. [277]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, P-values, Confidence Intervals, and Power?: A Guide to Misinterpretations," *The American Statistician*, 15, 1–31. [274]

Kelley, K. (2007), "Sample Size Planning for the Coefficient of Variation from the Accuracy in Parameter Estimation Approach," *Behavior Research Methods*, 39, 755–766. [277]

Kline, R. B. (2004), *Beyond Significance Testing*, Washington, DC: American Psychological Association. [277]

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005), "Oxytocin Increases Trust in Humans," *Nature*, 435, 673–676. [272,273,274,275,277,278]

Kruschke, J. K. (2014), *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). [277]

Kruschke, J. K., and Liddell, T. M. (2017), "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective," *Psychonomic Bulletin & Review*. [277]

Lane, A., Luminet, O., Nave, G., and Mikolajczak, M. (2016), "Is there a Publication Bias in Behavioural Intranasal Oxytocin Research on Humans? Opening the File Drawer of One Laboratory," *Journal of Neuroendocrinology*, 28. [273]

Leng, G., and Ludwig, M. (2016a), "Review Intranasal Oxytocin?: Myths and Delusions," *Biological Psychiatry*, 243–250. [273]

Leng, G., and Ludwig, M. (2016b), "Reply to: Intranasal Oxytocin Mechanisms Can Be Better Understood, but Its Effects on Social Cognition and Behavior are Not to Be Sniffed At," *Biological Psychiatry*, 79, e51–e52. [273,275]

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012), "Replications in Psychology Research: How Often Do They Really Occur?" *Perspectives on Psychological Science*, 7, 537–542. [274]

Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008), "Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation," *Annual Review of Psychology*, 59, 537–563. [271,277]

McShane, B. B., and Gal, D. (2015), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, mnsc.2015.2212. [276]

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2017), "Abandon Statistical Significance," arXiv preprint arXiv:1709.07588. [277]

Mellor, D. T., and Nosek, B. A. (2018), "Easy Preregistration will Benefit any Research," *Nature Human Behaviour*, 2, 98–98. [276]

Nave, G., Camerer, C., and McCullough, M. (2015), "Does Oxytocin Increase Trust in Humans? A Critical Review of Research," *Perspectives on Psychological Science*, 10, 772–789. [273]

Nelson, L. D., Simmons, J., and Simonsohn, U. (2018), "Psychology's Renaissance," *Annual Review of Psychology*, 511–534. [276]

Neumann, I. D., and Landgraf, R. (2012), "Balance of Brain Oxytocin and Vasopressin: Implications for Anxiety, Depression, and Social Behaviors," *Trends in Neurosciences*, 35, 649–659. [273]

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011), "Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance," *Nature Neuroscience*, 14, 1105–1107. [274]

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., and Yarkoni, T. (2015), "Promoting an Open Research Culture," *Science*, 348, 1422–1425. [276]

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018), "The Preregistration Revolution," *Proceedings of the National Academy of Sciences*, 115, 2600–2606. [276]

O'Boyle, E. H., Banks, G. C., and Gonzalez-Mule, E. (2014), "The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles," Journal of Management. [275]

Quintana, D. S. (2018), "Revisiting Non-Significant Effects of Intranasal Oxytocin Using Equivalence Testing," *Psychoneuroendocrinology*, 87, 127–130. [273]

Quintana, D. S., and Woolley, J. D. (2016), "Intranasal Oxytocin Mechanisms Can Be Better Understood, but Its Effects on Social Cognition and Behavior Are Not to Be Sniffed At," *Biological Psychiatry*, 79, e49–e50. [273,275]

Shirani-Mehr, H., Rothschild, D., Goel, S., and Gelman, A. (2018), "Disentangling Bias and Variance in Election Polls," *Journal of the American Statistical Association*, 113, 607–614 [277]

Smithson, M. (2002), *Confidence Intervals*, London: SAGE Publications. [271,277]

Sterling, T. D. (1959), "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa," *Journal of the American Statistical Association*, 54, 30–34. [273]

Sterling, T. D., Rosenbaum, W. L., and Weinkam, J. J. (1995), "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa," *The American Statistician*, 49, 108. [273]

Szucs, D., and Ioannidis, J. P. (2017), "Empirical Assessment of Published Effect Sizes and Power in The Recent Cognitive Neuroscience and Psychology Literature," *PLOS Biology*, 15, e2000797. [273,276]

Tversky, A., and Kahneman, D. (1971), "Belief in the Law of Small Numbers," *Psychological Bulletin*, 76, 105–110. [272]

Walum, H., Waldman, I. D., and Young, L. J. (2016), "Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies," *Biological Psychiatry*, 79, 251–257. [273]

Woolley, J. D., Chuang, B., Lam, O., Lai, W., O'Donovan, A., Rankin, K. P., Mathalon, D. H., and Vinogradov, S. (2014), "Oxytocin Administration Enhances Controlled Social Cognition in Patients with Schizophrenia," 47, 116–125. [275]

Yatawara, C. J., Einfeld, S. L., Hickie, I. B., Davenport, T. A., and Guastella, A. J. (2016), "The Effect of Oxytocin Nasal Spray on Social Interaction Deficits Observed in Young Children with Autism: A Randomized Clinical Crossover Trial," *Molecular Psychiatry*, 21, 1225–1231. [273]

The American Statistician

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

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

Stephen T. Ziliak

To cite this article: Stephen T. Ziliak (2019) How Large Are Your *G*-Values? Try Gosset's Guinnessometrics When a Little "*p*" Is Not Enough, The American Statistician, 73:sup1, 281-290, DOI: 10.1080/00031305.2018.1514325

To link to this article: https://doi.org/10.1080/00031305.2018.1514325

© 2019 The Authors. Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 6742

View related articles

View Crossmark data

Citing articles: 11 View citing articles

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, 281–290: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1514325

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

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

Stephen T. Ziliak

Roosevelt University, College of Arts and Sciences, Department of Economics, Chicago, IL; Newcastle Business School, University of Newcastle, NSW, Australia

Abstract

A crisis of validity has emerged from three related crises of science, that is, the crises of statistical significance and complete randomization, of replication, and of reproducibility. Guinnessometrics takes commonplace assumptions and methods of statistical science and stands them on their head, from little *p*-values to unstructured Big Data. Guinnessometrics focuses instead on the substantive significance which emerges from a small series of independent and economical yet balanced and repeated experiments. Originally developed and market-tested by William S. Gosset aka "Student" in his job as Head Experimental Brewer at the Guinness Brewery in Dublin, Gosset's economic and common sense approach to statistical inference and scientific method has been unwisely neglected. In many areas of science and life, the 10 principles of Guinnessometrics or G-values outlined here can help. Other things equal, the larger the G-values, the better the science and judgment.

By now a colleague, neighbor, or YouTube junkie has probably shown you one of those wacky psychology experiments in a video involving a gorilla, and testing the limits of human cognition. In one video, a person wearing a gorilla suit suddenly appears on the scene among humans, who are themselves engaged in some ordinary, mundane activity such as passing a basketball. The funny thing is, prankster researchers have discovered, when observers are asked to think about the mundane activity (such as by counting the number of observed passes of a basketball), the unexpected gorilla is frequently unseen (for discussion see Kahneman 2011). The gorilla is invisible. People don't see it.

ARTICLE HISTORY
Received May 2018
Revised August 2018

KEYWORDS
Balanced design; Economic approach to uncertainty; External validity; Fisher's *p*; Randomization; Reproducibility; Statistical significance; "Student's" *t*

1. Invisible Giant of Statistics Inside The Archives (And Out)

I am an economic statistician and historian who has studied the history of statistical significance and experimental design in what is considered by most people to be a foreign and ancient land—the archives, that is, historical archival libraries—for nearly 25 years. I've logged countless hours in some of the grandest and oldest reading rooms and their opposites in England, Ireland, and the United States. On and off since 2008, my research efforts have been especially focused in Dublin, Ireland, primarily in the archives of the Guinness Storehouse Museum (wicked, I know: nice work if you can get it).

I have found by comparing the archives with published literature a similar pattern in the history of statistics, an "unseeing" of a statistical giant whose methods, when heeded, could help out much of the economy, from Big Pharm and agri-business to the World Bank and higher education, more or less instantly. Statisticians and their clients have narrowed their focus on the wrong thing, the back and forth of statistical significance testing, "pass" or "no pass". We have become so routinized in our focus on bright-line significance levels such as $p < 0.05$ (an example of what the behavioral economist Daniel Kahneman calls "thinking fast") that we are often blinded to larger, perhaps more important things—such as estimation of the substantive meaning of our results (requiring "slower", deliberate thought).

In fact, I argue that this unseen giant of statistics holds the keys for unlocking our current crises in science: the crises of *replication*, of *reproducibility*, and of *statistical significance*; in sum, close study of this giant's methods can help to undo the *crisis of validity* currently haunting science and society, including business and law (Goodman 2002; Ziliak and McCloskey 2008; Greenland, et al. 2016; Wasserstein and Lazar, 2016; Pierson, Broman, et al. 2017).

2. How a Giant Statistician and Brewer Sidestepped the Crisis of Validity

Who is this invisible giant with a hero's cape? "Student" is the name which graces Student's *t*, the most popular test of statistical significance in the world, and the foundation from which Fisher's *p*-values were originally derived (Student 1908a, 1925; Fisher 1925; Ziliak 2008). Most statisticians have heard

CONTACT Stephen T. Ziliak sziliak@roosevelt.edu Roosevelt University, College of Arts and Sciences, Department of Economics, 430 S. Michigan Ave., Chicago, IL 60605.

© 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.

that "Student" is in reality the pen name of William Sealy Gosset (1876–1937), a brewer of Guinness beer.[1]

But even professional statisticians struggle to distinguish "Student's" actual methods from those of Fisher, erroneously conflating the ideas and opinions of the two men (Ziliak and McCloskey 2008, Chaps. 20–23). The much older and far more experienced "Student", who had been comparing random with balanced layouts in agricultural plots since 1905 was not, as some observers seem to think, just another "Fisher"-replicate donning a brewer's costume, a minor figure appearing on the scene to spice up a boring statistics lecture. Brewing and shipping unpasteurized beer for profit, taste, and quality assurance on the global scale is serious business—the economic element is fundamental, and a lot can be lost, "Student" knew and warned. Conflation of Student's methods with the admittedly hackneyed mathematical and antieconomic re-interpretation of them by Fisher (1925, 1933, 1935) and most other textbook authors is both a historical blind spot and scientific blunder with, as I have shown elsewhere, enormous practical and human consequences, many of them far from good.[2] I call this valuable if neglected approach to statistics, "Guinnessometrics".

3. G-Values: The 10 Principles of Guinnessometrics

Guinnessometrics is an experimental philosophy of inference and decision-making innovated and market-tested between 1904 and 1937 by William S. Gosset, a self-trained statistician who rose to Head Brewer of Guinness during the decades when Guinness was the largest brewery in the world (Ziliak 2008). At Guinness the scientific brewers, including Gosset, were allowed by the company to publish research so long as they did not mention (1) beer, (2) Guinness, or (3) their own surname. Ironically the hundred-million gallon a year brewery did not rely on statistical significance, Student's t, randomized trials, or Big Data. Closer to the opposite. Guinnessometrics takes a repeated small-sample economic approach to experimental statistics and decisions, in cooperation with agents up and down supply chains, all with real "skin in the game" (Taleb 2018). Gosset's *Guinnessometrics* inverts the usual matrix of statistical science and emerges with little or no need for a p-value or placebo-controlled randomized trial. And as any Guinness drinker can tell you, results of this science are not only repeatable and replicable, they are reproducible. Crisis averted.

The crisis of validity in the statistical sciences has been caused largely, though not entirely, by the following common yet erroneous practices undermining trust in data-based decisions:

- Attempted falsification of an assumed-to-be true null hypothesis without a loss function;
- Statistical significance testing at a bright line level (such as $p < 0.05$ or $t > 1.96$) independent of the substantive meaning of the result, coefficient, or model;
- Running a randomized, placebo controlled trial (RCT) assuming the independence of observations;
- Making decisions based on one, large-scale experiment, random or not ("One and done");
- Accepting or rejecting a hypothesis based on a single sample of convenience and arbitrary size (the majority of social science, and much of life science);
- Promoting what turns out to be irreproducible results;
- Investing prematurely in "Big Data";
- Assuming in statistical tests there is "no prior" subject-matter knowledge; and
- Making false binary and merely qualitative assessments based on the alleged bright line "significance" or "insignificance" of a result (important/unimportant, yes/no, exists/does not exist)

Guinnessometrics reverses these misleading yet widespread practices and replaces them with 10 principles or G-values, just as Gosset always said. In listicle form, the Ten Principles of Guinnessometrics are:

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 on Statistical Significance and P-values 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 "i.i.d." 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 (Ziliak 2014).

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 experiment. "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 min-

[1] Harold Hotelling (1930, p. 189), a vice president of the American Statistical Association and a teacher of many leading economists and statisticians, wrote: "I have heard guesses in this country, identifying 'Student' with Egon S. Pearson and the Prince of Wales."

[2] See also Student 1942; Pearson 1990; Ziliak 2014, 2010a, 2010b; Ziliak and Teather-Posadas 2016.

imizes 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/or 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, Chp. 22). And last but not least:

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" (the expected benefit and cost, for example) together with the cost of obtaining new material.

No one could plausibly claim that the 10 G-values are the end-all, be-all of statistical science. Only that for a great variety of business, medical, and scientific purposes, the Guinnessometric approach to data and decision-making answers far more questions, and far more satisfactorily, than the conventional, unrepeated observational study or placebo controlled RCT judged by the level of a p-value (compare Ziliak 2010a, 2010b; Senn 2010).

4. How to Get Large G-Values

Gosset's routine was to produce as many of the 10 G-values as possible, and to labor at them at maximum possible strength, so long as doing so does not subtract too much from one or more of the other G-values. We briefly illustrate below the value of each G-value. For example, throwing too many resources into G-value Number 7 by replicating *too much* (such as by spending $10 million on a 12th replication of the overrated "invisible gorilla" experiment made famous by YouTube) would encroach too negatively on G-value Number 6: the need to "Economize." Economizing in Gosset's sense means studying a series of small but independent and well-structured samples (rather than plunking down millions for Big, unstructured Data, today's fashion) to learn about regression input X or dependent variable Y. It should be said that the validity of G-values is completely general and does not depend on whether or not one fancies a beverage by Guinness. Thus:

5. G-10 Consider The Purpose Of The Inquiry, And Compare With Best Practice

Falsification of a null hypothesis is not, we have said, the "purpose" of a study or experiment helped along by statistical methods. Saving lives or money or malted barley, or advancing the health and wealth of schools and nations, is. In his magisterial *Theory of Probability* Harold Jeffreys (1961 [1961], p. 377) wrote in agreement with Gosset that the null test-and-p-value procedure advocated by the Fisher School "is merely something to set up like a coconut to stand until it is hit." "Hence the hypotheses made by "Student" are completely equivalent to mine; they have merely been introduced in a different order" (Jeffreys 1961, p. 380).

Presumably there is prior interest, otherwise, why test? Why invest in an experiment that is thought in advance to make no difference whatsoever? In their hearts, most investigators do not. But in their statistical science most are in Jeffreys's sense dedicated coconut swatters.

Whatever the purpose of the experiment, best practice research compares a novel treatment or variable with best practice and/or prevailing wisdom, not with an assumed-to-be-true null hypothesis or blank placebo. At the largest brewery in the world measured by annual output and sales, Gosset was determined, indeed he was incentivized, to mash the beer, not a low hanging coconut. At Guinness, he could not afford to spend his days taking whacks at easy to hit coconuts. At stake was nothing less than 100 million gallons of Guinness stout produced and sold annually by one of the most recognized brands in the world, first established by Arthur Guinness in 1759.

In an important letter of 1905, Gosset told Karl Pearson that one can "aim at" the odds of attaining some result by way of repeated experimentation. He told Pearson he decided that one cannot judge the "significance" of results—or decide a course of business action—without, in effect, employing some scale of human values capable of balancing the utility of expected gains against the disutility of losses (Gosset 1905). The point is fundamental:

> When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery" (1904)], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [in mathematics, such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours *the degree of certainty to be aimed at* must depend on the *pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment.*

Gosset upped the odds of farming and brewing success by designing and repeating small samples of stratified and balanced experiments, reducing what he called the "real error." The object or purpose of an experiment has little or nothing to do with the "significance" of a null hypothesis in and of itself, independent of some scale of values and range of action.

In a seminal *Biometrika* article "On Testing Varieties of Cereals" Gosset wrote (Student 1923):

> The object of testing varieties of cereals is to find out which will pay the farmer best. This may depend on quality, but in general it is an increase of yield which is profitable, and since yield is very variable from year to year and from farm to farm it is a difficult matter upon which to obtain conclusive evidence.

Citing his own research undertaken in cooperation with the Irish Department of Agriculture and Technical Instruction (IDATI), Gosset continued:

> Yet it is certain that very considerable improvements in yield have been made as the result of replacing the native cereals by improved varieties; as an example of this I may cite the case of Ireland, where varieties of barley have been introduced which were shown by experiment [since 1898] to have an average yield of 15 to 20% above those which they replaced. This represents, probably a gain to the country of not less than £ 250,000 per year. As the cost of experiments from the commencement to the present time [namely, about 25 years] cannot have reached £ 40,000 the money has been well spent.

In the same article Gosset told the assembled at the Royal Statistical Society—including Fisher, Neyman, and Egon Pearson—that the net gain to Ireland alone after 25 years of continuous experimentation using balanced (not RCT) designs was approximately:

Gain: + £ 250,000 per year × 25 years

Cost: - £ 40,000

That is, a ballpark net gain = £ 6.21 million on a £ 40,000 investment

In his last published paper, "Comparison of Random and Balanced Arrangements of Field Plots," Gosset said (Student 1938, p. 206):

I personally choose the method which is most likely to be profitable when designing the experiment rather than use Prof. Fisher's system of a posteriori choice* which has always seemed to me to savour rather too much of "heads I win, tails you lose."

6. G-9 Estimate The Stakes (Or Eat Them)

Thus, the purpose of an experiment puts something substantive at stake—a belief, an input, a technology, or policy variable. The purpose of the experiment is not to reject a null hypothesis; the purpose of a study is usually not even "statistical" in meaning. Science is substantive. Its meaning lies in magnitudes of bees, or beer, or basketball, of love, or family, or famous movie stars. Not in columns and rows of Student's t or Fisher's p.

The stakes of a study are normally speaking the substantive stakes of life—a gamble on a new pill or product or pilot study—with expected magnitudes of gains and losses expressed in the form of an expected loss function (Manski 2018, Press 2005), Jeffreys's minimum difference (Jeffreys 1961), a Ramsey bet (Ramsey 1926) or Gosset gamble (Gosset 1905) on a scale of values not captured by tables of probability alone. As Ramsey (1926, p. 51) wrote in "Truth and Probability":

We all agree that a man who did not make inductions would be unreasonable: the question is only what this means.

Like Gosset, Ramsey believed that probability "is a measurement of belief qua basis of action."[3] Quantities of substantive gain or loss are central to the Guinnessometric approach. The much-admired Frank Plumpton Ramsey (1926)—who by chance had attended the same school as Gosset (Winchester College) and studied mathematics with one of Gosset's friends (the mountain-climber, Robert Irving)—observed that the "expected value" of an experiment or a set of observations is not a reliable guide for action. (Most economists first learned the point from Savage, though Savage credits Ramsey.) Expected value is an unreliable metric for gauging human behavior, though many social and life scientists interpret regression models as if. As the stakes rise, or as the gamble is rephrased or otherwise altered rhetorically and thus psychologically, risk taking behavior changes, Ramsey perceived. And this fact has led economists and statisticians to favor "expected utility" over expected value (in statistics, see Lindley (1991) and Press (2005); also see Taleb (2018) for the weakness of utility under "ruin" probabilities). The expected value of something is the sum of all the possible outcomes (in nominal terms) weighted by their respective probabilities of occurrence.

It is easy to see Ramsey's point in a comparison of three different gambles subject to one random flip of a fair coin (illustrated by Frank 2017, pp. 179–180):

Gamble 1: If a coin flip shows "heads", win \$100; if tails, lose \$0.50

Gamble 2: If a coin flip shows "heads", win \$200; if tails, lose \$100

Gamble 3: If a coin flip shows "heads", win \$20,000; if tails, lose \$10,000.

Now Gamble 1 would be accepted by many. Assuming as convention does that the probability of heads and tails are equal (at 1/2 each) the expected value of the gamble is (1/2)(win \$100) + (1/2)(lose \$0.50) for an expected win of \$49.75. The worst case scenario is "lose 50 cents", which many would be willing to risk for an equal chance at winning \$100.

Gamble 2 will draw fewer contestants, though many will still accept the gamble. Gamble 3 has the highest expected value but

[3] *https://plato.stanford.edu/entries/probability-interpret/*

this bet would be rejected by many (though not all) even though the payoff structure is identical to Gamble 2 (both have a 2-to-1 win-loss ratio). The reason is that, although winning $20,000 could be lovely, losing $10,000 from a single coin flip could be tragic.

Suppose now that subjects are expected utility maximizers instead of expected value maximizers. Utility theory has problems and limitations, too, but it offers something more. Expected utility is a scale of values which accounts for, among other things, attitudes toward risk and the diminishing marginal utility of money (for example Savage 1954; Lindley 1991; Press 2005). Suppose you are a conventional risk-averse person with a strictly concave utility function, such as the square-root function, which is concave in money accumulation (EU $= \sqrt{}$ money).

Your initial amount of money (or wealth) is $10,000. What is the expected utility of each of the three gambles? Notice that the "value" ranking of the gambles is now *reversed*.

The expected utility of Gamble 1 is *highest* even though its expected value is lowest ($(1/2)(\sqrt{10{,}100}) + (1/2)(\sqrt{9999.50}) = 100.248$ utils) while the expected utility of Gamble 3 is *lowest* even though its expected value is by far the highest ($= 86.60$ utils). "The old established way of measuring a person's belief is to propose a bet, and see what are the lowest odds which he will accept. This method I regard as fundamentally sound" (Ramsey 1926, p. 34). "[B]ut it suffers from being insufficiently general, and from being necessarily inexact. It is inexact partly because of the diminishing marginal utility of money, partly because the person may have a special eagerness or reluctance to bet, because he either enjoys or dislikes excitement or for any other reason, e.g., to make a book" (Ramsey, pp. 34–35).

Yet from the point of view of a statistical scientist who is charged with practical estimation and interpretation in a particular context, Ramsey's approach is backward way around, and too abstract. What Gosset's expected loss function approach lacks in abstract shine, it gains in economic profit and plain common sense.

7. G-8 Study Correlated Data: Abba, Take a Chance on Me

Completely randomized studies, where treatment and control groups are determined as if by random coin flip, are now fashionable. Yet, statisticians have long known that stratification or blocking adds precision and efficiency to a study otherwise based on complete randomization. Gosset (Student 1911) used blocking or stratification long before the synonymous words existed in the statisticians' vocabulary. He called his balanced approach to field layouts, "ABBA," the closely arranged mirror pattern of the layout, treatments, and controls, the As and Bs being compared. Student (1923, p. 273) said:

The art of designing all experiments lies even more in arranging matters so that ρ [the correlation coefficient] is as large as possible than in reducing σ_x^2 and σ_y^2 [the variance].

The peculiar difficulties of the problem lie in the fact that the soil in which the experiments are carried out is nowhere really uniform; however little it may vary from eye to eye, it is found to vary not only from acre to acre but from yard to yard, and even from inch to inch. This variation is anything but random [Gosset himself noted], so the ordinary formulae for combining errors of observation which are based on randomness are even less applicable than usual.

As Deming (1938, p. 879), an admirer of Gosset, noted: "Stratification is equivalent to blocking in the design of an experiment." Box, Hunter, and Hunter (2005, p. 92) explain that "A block is a portion of the experimental material (the two shoes of one boy, two seeds in the same pot) that is expected to be more homogenous than the aggregate (the shoes of all the boys, all the seeds not in the same pot). By confining comparisons to those within blocks (boys, girls), greater precision is usually obtained because the differences associated between the blocks are eliminated."

Deming, who before turning to manufacturing did a long stint as researcher at the U.S. Department of Agriculture, agreed with Gosset's nonrandom point: random sampling and randomized experiments are at best preliminary steps to scientific study. Complete randomization has a purpose when the investigator knows little or nothing at all about strata or when the cost of being wrong is negligible. Said Deming (p. 879):

The primary aim of stratified sampling is to increase the amount of information per unit of cost. A further aim may be to obtain adequate information about certain strata of special interest. One way to carry out stratification is to rearrange the sampling units in the frame so as to separate them into classes, or strata, and then to draw sampling units from each class. The goal should be to make each stratum as homogeneous as possible, within limitations of time and cost.[4]

Likewise in his book, *Planning of Experiments*, David Cox (1958) recommends "completely randomized arrangement …[only] in experiments in which no reasonable grouping into blocks suggests itself"—that is, when ignorance prevails, or priors are flat.

Normally speaking, ignorance does not prevail, and real economic and statistical gains can be found by stratifying. Deming (1938) and Tippett (1952) simplified Student's (1911, 1923) proof that stratification (blocking) can reduce sample size requirements by 40% or more, holding variance constant.[5] And as Tippett noted, "At the worst"—assuming the rare case that calculated variance between strata is zero—"sampling in strata is no better than random sampling, but it is never worse."

8. G-7 Minimize "Real Error" With The 3 R'S: Represent, Replicate, Reproduce

Recently the ASA has adopted the following widely used definitions of reproducibility and replicability in a set of "Recommendations" designed to assuage the reproducibility "crisis" which

[4] Deming (1978, p. 879). Deming said he learned the technique from Neyman (1934). In the seminal article Neyman demonstrates the statistical and economic advantages of stratified sampling over random sampling (Neyman 1934, pp. 579-585). Neyman credits the idea of "purposive selection" to earlier writers, such as Bowley and Gini and Galvani.

[5] Deming (1978, p. 880-881), Tippett (1958, p. 356). In a Riesling vine-and-wine experiment, Meyers, Sacks, van Es, and Vanden Heuvel (2011) used blocking, balancing, and repetition (at $n = 3$ vineyards) to reduce sample size requirements by up to 60%.

is currently being observed throughout the sciences (Pierson, Broman, et al. 2017):

1. *Reproducibility*: A study is reproducible if you can take the original data and the computer code used to analyze the data and reproduce all of the numerical findings from the study. This may initially sound like a trivial task but experience has shown that it's not always easy to achieve this seemingly minimal standard.

For Gosset and Guinnessometrics, "reproducibility" is the ability to brew Guinness stout or ale (Smithwick's) in such a way as to taste and to otherwise behave the same, pint after pint, gallon after gallon, millions of times over worldwide.

2. *Replicability*: This is the act of repeating an entire study, independently of the original investigator without the use of original data (but generally using the same methods)

In 35 years of barley yield and quality trials, Gosset and Guinness commissioned experiments with Irish barley farmers scattered across the different barley growing regions of the country. On average there were 10 farmers running simultaneously the same or nearly identical experimental design on new barley (treatment) and one or more "old" barley. And for insurance each planted a replicate locally, thus there were approximately 20 replications in the average year.]

These definitions suit well the Guinnessometric practice of replication and reproducibility. Gosset explained the 3 R's of minimum real error in a letter of April 1937 to Egon S. Pearson, who was his close friend and editor of *Biometrika* (quoted in Pearson 1939, pp. 247–248):

Many thanks for yours of [April] 10th; I feel I'm wasting your time but as long as you ask questions you must expect to get answers …Now I was talking about Cooperative experiments and obviously the important thing in such is to have a low real error, not to have a "significant" result at a particular station. *The latter seems to me to be nearly valueless in itself.* Even when experiments are carried out only at a single station, if they are not mere five finger exercises, they will have to be part of a series in time …. But in fact experiments at a single station [that is, tests of statistical significance on a single set of data] *are* almost valueless; you can say "In heavy soils like Rabbitsbury potatoes cannot utilise potash manures", but when you are asked "What *are* heavy soils like Rabbitsbury?" you have to admit—until you have tried elsewhere—that what you mean is "At Rabbitsbury etc." And that, according to *X* may mean only "In the old cow field at Rabbitsbury". What you really want to find out is "In what soil and under what conditions of weather do potatoes utilise the addition of potash manures?"{PRIVATE}

To do that you must try it out at a *representative sample* of the farms of the country and correlate with the characters of the soil and weather. It may be that you have an easy problem, like our barleys which come out in much the same order whatever–in reason–you grow them or like Crowther's cotton which benefitted very appreciably from nitro-chalk in seven stations out of eight, but even then *what you really want is a low real error. You want to be able to say not only "We have significant evidence that if farmers in general do this they will make money by it"*, but also "we have found it so in nineteen cases out of twenty and we are finding out why it doesn't work in the twentieth". To do that you have to be as sure as possible which is *the 20th—your real error must be small* (emphasis added).

Representation, the third "R," we can define in a number of different ways but something like this: coverage and stratification of all systematic sources of fluctuation, whether the source be a temporal, a spatial, or other exogenous force (such as weather: rainy or not rainy, or soil quality: loamy or clay). These definitions work as a first approximation. It's not about random error: that's the main point. We are trying to minimize and control for the systematic errors, which are of larger importance and number than are the assumed-to-be "random sampling errors." These are the 3 R's of Guinnessometrics and G-value No. 7.

The reason for G-value Number 7, the 3 R's of Minimum Real Error is simple: the out-of-sample experience of life is difficult to predict, and for both systematic and random reasons. Unless you are a late night psychic broadcasting on cable television to well sedated individuals, you do not hold the crystal ball for ensuring external validity and economic profit should we decide to "scale up" from an unbalanced, un-stratified village-level experiment on eyeglasses or to nets to, let's say, a whole nation (the lack doesn't stop some from donning a purple scarf at The World Bank: Banerjee and Duflo 2011; Glewwe et al, 2012).

Yet most statistical studies in the social sciences, economics and psychology included, and many more in health and medicine are of the "one and done" variety: one RCT (which assumes independence) is conducted on a single sample; one approach is taken to regression modeling, followed by 100 tests of significance on the data using NHST and *p*-values on the single sample on offer. P-hack until you publish the paper (Ziliak and Teather-Posadas 2016 document the ethical side of this). This treating of single samples as if they are repeated samples is old news to older psychologists, many of whom remember the great Sterling (1959) survey showing that the probability of *replication* decreased with the level of statistical significance (the lower the *p* value, the less likelihood the study will be replicated).

From 1901 until the start of World War II, Guinness (and thus Gosset) invested heavily in the 3 R's of G-values: that is by repeating annually a small series of independent, representative, and balanced experiments (what sociologists call repeated, stratified-random samples). The positive results of the Guinness and Irish investment are nothing short of astonishing, with one of the byproducts being that Gosset invented or inspired half of the toolkit of modern statistics and experimental science.

9. G-6 Economize With "Less is More": Small Samples of Independent Experiments

Replication and reproduction does not always require millions of dollars in grant money, though at the NIH, NSF, and elsewhere it can.

Small samples are an economic and scientific choice, not a mathematical problem to be solved in abstract mathematical terms. For example, in a small-sample analysis, a brewer may wish to know with 10 to 1 or better odds how many samples of malt extract he needs to mix to be confident that the saccharine level of the beer stays within 0.5 degrees of the 133 degree standard he is targeting. The example is "Student's": brewing

over 100 million gallons of Guinness stout per annum, "Student" and Guinness stakeholders needed to know (Ziliak 2008, p. 206). "Real" errors in this context include uneven temperature changes, heterogeneous barley malt, and mismeasurement of saccharine levels—adding up to more error than is allegedly described by p or t.

Student rejected artificial rules about significance from the beginning of his inquiries at the Brewery—at least four years before he published the first table and small sample test of significance (Gosset 1904; Student 1908a). In November, 1904, Gosset—he would not be known as Student until three years later—discussed his first break-through on the economic meaning of statistical significance, in an internal report titled "The Application of the 'Law of Error' to the Work of the Brewery." The Apprentice Brewer said:

Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment. What the odds should be depends —

1. On the degree of accuracy which the nature of the experiment allows, and
2. On the importance of the issues at stake.

Comparing the level of saccharine content in a series of malt extracts which he and others' mixed in the Experimental Brewery with that found in malts being used in the Main Brewery, Gosset brought attention to a positive correlation he found between "the square root of the number of observations"— that is, the number of calculated differences in saccharine content between Experimental and Main Brewery malts—and the level of statistical significance. Other things equal, he said "the greater the number of observations of which means are taken [the larger the sample size of extract differences], the smaller the [probable or standard] error" of the estimates. "And the curve which represents their frequency of error," he showed in a graph and plot drawing, "becomes taller and narrower."

Prior to Gosset the relation between sample size and the level of statistical significance was rarely explored. For example, while looking at biometric samples with up to thousands of observations, Karl Pearson declared that a result departing by more than three standard deviations is "definitely significant." The normal tables assumed very large samples. Yet Gosset, self-trained in statistics, found by experiment that at such large samples nearly everything is statistically "significant"— though not, in Gosset's terms, economically or scientifically "important". And, likewise, Gosset found that a small number of observations can be profitable, though not statistically significant in Pearson's conventional sense. Regardless, Gosset did not have the luxury of large samples. One of his earliest experiments employed a sample size of $n = 2$, which helps to explain why in the original 1908 article Gosset calculated a z statistic for $n = 2$ (Student 1908a).

His 1904 article is worth exploring a bit further – especially for the econometrician and real-world firm that wants to earn more with less. Guinness malt was produced in Gosset's time primarily from Irish and English barley stock—Old Irish, Prentice, Plumage Archer, and Spratt Archer were effective varieties. Malt extract was measured by "degrees saccharine" per barrel of 168 pounds malt weight.

An extract in the neighborhood of 133 degrees saccharine gave the targeted level of alcohol for Guinness's beer. A much higher degree of saccharine would affect the stability and life of the beer, but it also increases alcohol content—which in turn increases the excise tax which Guinness owes to the British government, which—sad to say—ups the price of Dad's pint. If, on the other hand, the alcohol content comes in too low, if the degree of saccharine is insufficient, customers would riot, or switch to Beamish and Beck's. In Gosset's view, $+/-$ 0.5 degrees saccharine was a difference or error in malt extract which Guinness and its customers could swallow. "It might be maintained," he said, "that malt extract "should be [estimated] within 0.5 of the true result with a probability of 10 to 1." Gosset calculated the odds of observing the stipulated accuracy for small and then large numbers of extracts. He found that:

Odds in favour of smaller error than 0.5 [are with:]
2 observations 4:1
3 " 7:1
4 " 12:1
5 " 19:1
82 " practically infinite

Thus, Gosset concluded, "In order to get the accuracy we require [that is, 10 to 1 odds with 0.5 accuracy], we must, therefore, take the mean of [at least] four determinations." The Guinness Board cheered. The Apprentice Brewer found an economical way to assess the behavior of population parameters, using very small samples.

Small samples and their analysis originate from a fundamental economic cause: scarcity—the economic scarcity and expense of gaining new information about barley, malt, hops, and other beer inputs. Thus, one can say in general that Gosset took an economic approach to the logic of uncertainty, from the choice of sample size on up.

10. G-5 Keep Your Eyes on The Size Matters/How Much? Questions

We're all in search of that "Goldilocks" zone, not too high, not too low; not too hot, not too cold, et cetera. In 1995, some cancer epidemiologists made history (discussed by Ziliak and McCloskey 2008, pp. 184–186). The authors of 10 independent and randomized clinical trials involving thousands of patients in treatment and control groups had come to an agreement on an effect size. Consensus on a mere direction of effect—up or down, positive or negative—is rare enough in science. After four centuries of public assistance for the poor in the United States and Western Europe for example, economists, do not speak with one voice on the direction of effect on labor supply exerted by tax-financed income subsidies. Medicine is no different. Disagreement on the direction of effect—let alone the size of effect—is more rule than exception.

So the Prostate Cancer Trialists' Collaborative Group was understandably eager to publicize the agreement. Each of the 10 studies showed that a certain drug "flutamide" —for the treatment of prostate cancer—can increase the likelihood of patient survival by an average of 12% (the 95% confidence interval in the pooled data put an upper bound on flutamide-enhanced survival at about 20% [Rothman, Johnson, and Sugano 1999]).

Odds of 5 in 100 is not the best news to deliver to a prostate patient. But if castration followed by death is the next best alternative, a noninvasive 12-to-20% increase in survival sounds good.

But in 1998 the results of still another, 11th trial were published in the *New England Journal of Medicine* (Eisenberger et al. 1998, pp. 1036–1042). The authors of the new study found a similar size effect. But when the two-sided p value for their odds ratio came in at .14 they dismissed the efficacious drug, concluding "no clinically meaningful improvement" (pp. 1036, 1039). Kenneth Rothman, Eric Johnson, and David Sugano examined the individual and pooled results of the 11 separate studies, including the study conducted by Eisenberger et al..

> One might suspect that [Eisenbergers et al.'s] findings were at odds with the results from the previous ten trials, but that is not so. From 697 patients randomised to flutamide and 685 randomised to placebo, Eisenberger and colleagues found an OR of 0.87 (95% CI 0.70–1.10), a value nearly identical to that from the ten previous studies. Eisenberger's interpretation that flutamide is ineffective was based on absence of statistical significance. (Rothman, Johnson, and Sugano 1999, p. 1184)

Rothman and coauthors display the flutamide effect graphically in a manner consistent with a Gosset-Deming-and-Savage approach to visualization. Does the effect hit you between the eyes? Does it cause interocular trauma? Rothman and others pool data from the separate studies and plot the flutamide effect (measured by an odds ratio, or the negative of the survival probability in a hazard function) together with the p-value function. With the graphical approach, Rothman and his coauthors are able to show pictorially how the $p-$values vary with increasingly positive and increasingly negative large effects of flutamide on patient survival. And what they show is substantively significant:

Eisenberger's new data only reinforce the findings from the earlier studies that flutamide provides a small clinical benefit. Adding the latest data makes the p value function narrower, which is to say that the overall estimate is now more precise, and points even more clearly to a benefit of about 12% in the odds of surviving for patients receiving flutamide.

Rothman and others conclude: "the real lesson" from the latest study is "that one should eschew statistical significance testing and focus on the quantitative measurement of effects." That sounds right. Statistical significance is hurting people, indeed killing them. It is leaving their illnesses and a defective notion of significance "unexplained."

Still, a recent correspondent points out that although the relative risk is high, the measured relative difference is small, adding only a month or so extra life. (Other things equal, perhaps even a little life is better than none.) If you want to draw a line in your science, in any case, draw a line of minimally important effect or minimally important magnitude of a regression coefficient. Keep your eyes on the size.

11. G-4 Visualize

And visualize. Model uncertainty is not the same thing as parameter uncertainty. From farming to pharmaceuticals, we want to know what the entire distribution looks like from the point of view of oomph and precision, magnitudes of relationships, and attendant uncertainty. Not just the point mean or median, with a superscript of asterisks declaring "significant" or "highly" so. Remember Stephen Jay Gould's far-above-the-median experience with surviving stomach cancer, discussed in his essay "The Median Isn't the Message" (Gould 1985). Gould's doctor cited a median survival time from diagnosis of about 8 months; but the prolific scholar and writer looked at the graphs showing estimates. He noticed a thin but long right-hand tail, changed his lifestyle and eating habits, and lived and worked for another 22 years (rather more than 8 months)!

An illuminating study by Soyer and Hogarth (2011) tested the predictive ability of more than 200 econometricians using linear models. Prediction was most accurate when the experts were only given a theoretical regression line and scatter plot of data. Take away the plots and their ability to relate model error to levels of the dependent variable fell dramatically. For novice and seasoned alike, the several books by Tufte on the art and science of visualization are invaluable.

The variance can vary, error bounds vary, degree and direction of uncertainty vary. Pictures help us to see by how much. Like Gosset, we should pay more attention to the actual distribution of our data, not only to rejections of the normal or exponential distribution (which few deign to discover) but also simulations or other confirmations of the actual distribution.

12. G-3 Consider Posteriors And Priors Too ("It Pays To Go Bayes")

Be explicit about how prior information is or is not incorporated. Perhaps the most commonly used alternative to classical $t$and p is the Bayes factor (Carlin and Louis 2008; Press 2003). Gosset, I have mentioned, was a Bayesian who for reasons of efficiency, and in a world without a supercomputer, resorted frequently to frequentist methods (especially power: Student 1938). For discrete data and simple hypotheses, the Bayes factor represents the ratio between the probability assigned to the data under an alternative hypothesis and the null hypothesis (Johnson 2013). One big advantage of Bayesian analysis is that one can compute the probability of a hypothesis, given the evidence, whereas with the null hypothesis test of significance, measured by a p value, one can only speak to the probability of seeing data more extreme than have actually obtained, assuming the null hypothesis of "no difference" (or whatever) to be true. As the Bayesian Jeffreys noted (1961, p. 409):

Whether statisticians like it or not, their results are used to decide between hypotheses, and it is elementary that if p entails q, q does not necessarily entail p. We cannot get from "the data are unlikely given the hypothesis" to"the hypothesis is unlikely given the data" without some additional rule of thought. Those that reject inverse probability have to replace it by some circumlocution, which leaves it to the student to spot where the change of data has been slipped in [, in] the hope that it will not be noticed.

Jeffreys went on to explain that if one assigns prior odds between the alternative and null hypotheses, multiplication

of the Bayes factor by these prior odds yields the posterior odds between the hypotheses. From the posterior odds between hypotheses, scientists can compute the posterior probability that a null hypothesis is true (or in any case useful or persuasive) relative to an explicit alternative. Classical tests of significance, measured by t and p, cannot.[6] Johnson (2013) observes that in certain hypothesis tests the alternative hypothesis can be specified so that an equivalence between Bayes factors and p-values can be established. Technically speaking, Johnson and others have shown, in one parameter exponential family models in which a point null hypothesis has been specified on the model parameter, specifying the size of the test is equivalent to specifying a rejection threshold for the Bayes factor, provided that it is further assumed that the alternative hypothesis is specified so as to maximize the power of the test. The correspondence between Bayes factors and p-values in this setting is just one example of the false demarcation line between objective and subjective.

When an alternative hypothesis exists—and that's the usual situation of science: otherwise, why test?—Bayes factors can be easily reported. Bayes factors permit individual scientists and consumers to use prior information or the principle of insufficient reason together with new evidence to compute the posterior probability that a given hypothesis, H, is true (or to repeat, useful or persuasive) based on the prior probability that they assign to each hypothesis. After all—fortunately—we do not have to begin every new observation or experiment from *tabula rasa*; we know some stuff, but we want to know more stuff, however imperfectly. Bayes factors add that information into the calculation comparing the likelihood of alternative hypotheses. For example, Bayes factors provide a clear interpretation of the evidence contained in the data in favor of or against the null: a Bayes factor of 10 simply means that the data were 10 times more likely under the alternative hypothesis than they were under the null hypothesis. Better than mushy p's.

13. G-2 Cooperate Up, Down, And Across (Networks And Value Chains)

Cooperation was a hallmark of Irish agricultural development (economic historians will think of Horace Plunkett, John Bennett, William Gosset and others), the growth of the Guinness brewery, and—in one very large positive externality—the development of modern statistics. Cooperation includes the human side, such as the academic Karl Pearson inviting the unknown brewer, Mr. Gosset, to London for sabbatical in 1906–1907 to work and study at University College London Biometrics Lab. Despite a strong-man reputation which Guinness has held in some quarters, for trying to monopolize, the legendary brewery cooperated up and down supply chains, and sometimes to a remarkable extent. The 40-something years of repeated experiments on barley, already mentioned, is just one example: Guinness subsidized one-half (50%) of barley and malt investment by the Irish Department of Agriculture. There is something to be learned from that: remember the yield and net profit figures shared by Gosset above. Guinness sending scientific brewers (Gosset was hardly the only one) on sabbatical for postgraduate studies in statistics and chemistry and genetics is another example. Sharing seeds and technical know-how with the Carlsberg Brewery, in Denmark, is another (and Carlsberg reciprocated generously by offering a stellar brewing barley called "Prentice" for Guinness to try: it turned out to be the mother of the greatest barley in early and midcentury Europe). One could write a book on the importance of cooperation to the history and development of science and statistics. And finally:

[6] Lavine and Schervich (1999) caution that Bayes factors can sometimes lead to incoherence in the technical statistical sense of that term.

14. G-1 Answer The Brewer's Original Question ("How Should You Set The Odds?")

Finally, how should you set the odds? As a 28-year-old brewer and self-trained statistician told Karl Pearson way back in 1905: "the degree of certainty to be aimed at must depend on the *pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment*." Add to Gosset's pecuniary advantage "lives saved," "jobs gained," "racism abolished," "health crisis averted," and the like and we have the makings of a general approach.

Acknowledgment

Sincere thanks to the editors and reviewers, together with Roosevelt colleagues Gary Langer and Justin Shea, for commenting on a previous version of this paper. At the Guinness Storehouse Museum (Diageo) I have enjoyed for many years the benefit of assistance from archivist Eibhlin Colgan. The seeds of the present paper were sown during a 2017 sabbatical visit at Trinity College Dublin (TRISS Research Institute) and the University of Oxford, New College. Many thanks to Ronan Lyons (Trinity College) and to Warden Miles Young (New College) for making those visits possible. Any errors are my own.

References

Banerjee, A., and Duflo, E.. (2011), *Poor Economics*, New York: Public Affairs. [286]

Carlin, B., and Louis, T.. (2008), *Bayes and Empirical Bayes Methods for Empirical Analysis* (3rd rev. ed.), London: Chapman and Hall/CRC Press. [288]

De Finetti, B. (1971 [1976]), Comments on Savage's "On Rereading R. A. Fisher," *Annals of Statistics*, 4, 486–487.

Deming, W. E. (1938 [1943]), *Statistical Adjustment of Data*, New York: Dover. [285]

——— (1961), *Sample Design in Business Research*, New York: Wiley.

——— (1982), *Out of the Crisis*, Cambridge, MA: MIT Center for Advanced Engineering Study.

Eisenberger, M. A., et al. (1998), "Bilateral Orchiectomy with or Without Flutamide for Metastatic Protate Cancer," *New England Journal of Medicine*, 339, 1036–1042. [288]

Fisher, R. A. (1925 [1928]), *Statistical Methods for Research Workers*, Edinburgh, UK: Oliver and Boyd. [281,282]

——— (1933), "The Contributions of Rothamsted to the Development of the Science of Statistics," *Annual Report of Rothamsted Experimental Station*, 43–50. [282]

——— (1935), *The Design of Experiments*, Edinburgh, UK: Oliver & Boyd. [282]

Frank, R. (2017), *Microeconomics and Behavior* (9th ed.), New York: McGraw-Hill. [284]

Glewwe, P., Park, A., and Zhao, M. (2012), "Visualizing Development: Eyeglasses and Academic Performance in Rural Primary Schools in China," Working Paper WP12-2, Center for International Food and Agricultural Policy, University of Minnesota. [286]

Goodman, S. (2017), "Why is Eliminating P-Values So Hard? Reflections on Science and Statistics," *ASA Symposium on Statistical Inference*, Oct. 11–13, 2007.

——— (2002), "A Comment on Replication, P Values, and Evidence," *Statistics in Medicine*, 11, 875–879. [281]

Gosset, W. S. (1904), "The Application of the 'Law of Error' to the Work of the Brewery," *Laboratory Report*, 8, Arthur Guinness & Son, Ltd., Diageo, Guinness Archives, 3–16 and unnumbered appendix. [287]

——— (1905), Letter from W.S. Gosset to K. Pearson, Guinness Archives, GDB/BRO/1102 [283,284]

——— (1936), "Co-Operation in Large-Scale Experiments," *Supplement to the Journal of the Royal Statistical Society*, 3, 115–136.

——— (1962), *Letters of William Sealy Gosset to R.A. Fisher. Vols. 1–5*, Eckhart Library, University of Chicago. Private circulation.

Greenland, S., et al. (2016), "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician*, 70, 1–12. [281]

Hotelling, H. (1930), "British Statistics and Statisticians Today," *Journal of the American Statistical Association*, 25, 186–190. [282]

Jeffreys, H. (1939 [1961]), *Theory of Probability* (3rd ed.), London: Oxford University Press. [283,284,288]

Johnson, V. (2013), "Revised Standards for Statistical Evidence," *PNAS*, 110, 19313–19317. [288,289]

Kahneman, D. (2011), *Thinking Fast and Slow*, New York: Farrar, Straus and Giroux. [281]

Lavine, M., and Schervish, M. (1999), "Bayes Factors: What They Are and What They Are Not," *The American Statistician*, 53, 119–122 [289]

Lew, M. (2012), "Bad Statistical Practice in Pharmacology (And Other Basic Biomedical Disciplines): You Probably Don't Know P," *British Journal of Pharmacology*, 166, 1559–1567.

Lindley, D. (1991), *Making Decisions*, New York: Wiley. [284,285]

Manski, C. (2018), "Treatment Choice with Trial Data: Statistical Decision Theory should Supplant Hypothesis Testing," *The American Statistician*, this issue, DOI: 10.1080/00031305.2018.1513377. [284]

McCloskey, D. N., and Ziliak, S. T.. (1996), "The Standard Error of Regressions," *Journal of Economic Literature*, 34, pp. 97–114.

——— (2010), *Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents, Matrixx Initiatives Inc. et al. v*. Siracusano et al. (vol. No. 09–1156, pp. 22). Washington, DC: Supreme Court of the United States. (Ed.) Edward Labaton et al. Counsel of Record.

McCloskey, D. N., and Ziliak, S. T. (2009), "Signifying Nothing: Reply to Hoover and Siegler,"*Journal of Economic Methodology*, 15, 39–55

Meyers, J., Sacks, G., van Es, H., and Vanden Heuvel, J. (2011) "Improving Vineyard Sampling Efficiency via Dynamic Spatially Explicit Optimisation," *Australian Journal of Grape and Wine Research*, 17, 306–315. [285]

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, 558–625. [285]

Pearson, E. S. (1939), "'Student' as Statistician," *Biometrika*, 30, 210–250. [286]

——— (1990), *Student: A Statistical Biography of William Sealy Gosset*, Eds. R. L. Plackett and G. A. Barnard. Oxford: Clarendon Press. [282]

Pierson, S., et al. (2017), "Recommendations to Funding Agencies for Supporting Reproducible Research," *American Statistical Association*, available at *https://www.amstat.org/ASA/News/ASA-Develops-Reproducible-Research-Recommendations.aspx* [281,286]

Press, S. J. (2003), *Subjective and Objective Bayesian Statistics*, New York: Wiley. [288]

Press, S. J. (1972 [2005]), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd ed.), Mineola, NY: Dover. [284,285]

Ramsey, F. P. (1926), "Truth and Probability," reprinted in H. Kyburg and H. Smoker, eds., *Studies in Subjective Probability* (New York: R. E. Krieger, 1980), 25–52. [284,285]

Rothman, K. J., Johnson, E. S., and Sugano, D. S.. (1999), "Is Flutamide Effective in Patients with Bilateral Orchiectomy?" *Lancet*, 353, 1184. [287,288]

Savage, L. (1954), *The Foundations of Statistics*, New York: Dover. [285]

——— (1971 [1976]), "On Re-Reading R. A. Fisher," *Annals of Statistics*, 4, 441–500.

Soyer, E., and Hogarth, R. (2011), "The Illusion of Predictability: How Regression Statistics Mislead Experts," *International Journal of Forecasting*, 28, 695–711. [288]

Student (1907), "On the Error of Counting with a Haemacytometer," *Biometrika*, 5, 351–360.

——— (1908a), "The Probable Error of a Mean," *Biometrika*, VI, 1–24. [281,287]

——— (1908b), "The Probable Error of a Correlation Coefficient," *Biometrika*, 2/3, 300–310.

——— (1923), "On Testing Varieties of Cereals," *Biometrika*, 15, 271–293. [284,285]

——— (1925), "New Tables for Testing the Significance of Observations," *Metron*, V, 105–108. [281]

——— (1938), "Comparison between Balanced and Random Arrangements of Field Plots," *Biometrika*, 29, 363–378. [284,288]

——— (1942), Student's *Collected Papers*, eds. Pearson, E. S. and Wishart, J., London: Biometrika Office. [282]

Supreme Court of the United States (2011), "Matrixx Initiatives, Inc., et al., No. 09–1156, Petitioner v. James Siracusano et al.," *On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit*, March 22nd, 25 pp., syllabus.

Taleb, N. N. (2018), *Skin in the Game: Hidden Asymmetries in Daily Life*, New York: Random House. [282,284]

The Guinness Archives (Diageo), Guinness Storehouse Museum, Dublin; Special Collections Library, University College London; Cork County (Ireland) Archives; Museum of English Rural Life, National Library of Ireland; University of Oxford, Bodleian Library and New College Library; Winchester College (UK) Archives; and University of Chicago, Crerar Library, Eckhart Library, and Regenstein Library.

Wasserstein, R., and Lazar, N., eds. (2016), "ASA Statement on Statistical Significance and P-Values", *The American Statistician*, 70, 129–132. [281]

Ziliak, S. (2008), "Guinnessometrics: The Economic Foundation of 'Student's' *t*," *Journal of Economic Perspectives*, 22, 199–216. [281,282,287]

——— (2010a), "The *Validus Medicus* and a New Gold Standard," *The Lancet*, 376, 324–325. [282,283]

——— (2010b), "Significant Errors—Reply to Stephen Senn," *The Lancet*, 376, 1391. [282,283]

——— (2014), "Balanced versus Randomized Field Experiments in Economics: Why W.S. Gosset Matters," *Review of Behavioral Economics*, 1, 167–208. [282]

——— (2016), "The Significance of the ASA Statement on Statistical Significance and P-Values," *The American Statistician*, 70, 1–2.

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. [281,282,283,287]

Ziliak, S., and Teather-Posadas, E.. (2016), "The Unprincipled Randomization Principle in Economics and Medicine," in *Oxford Handbook of Professional Economic Ethics*, eds. G. DeMartino and D. McCloskey, Oxford: Oxford University Press, 423–452. [282,286]

Archival Sources

The American Statistician

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

Predictive Inference and Scientific Reproducibility

Dean Billheimer

To cite this article: Dean Billheimer (2019) Predictive Inference and Scientific Reproducibility, The American Statistician, 73:sup1, 291-295, DOI: 10.1080/00031305.2018.1518270

To link to this article: https://doi.org/10.1080/00031305.2018.1518270

© 2019 The Authors. Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 5852

View related articles

View Crossmark data

Citing articles: 10 View citing articles

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, 291–295: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518270

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Predictive Inference and Scientific Reproducibility

Dean Billheimer

Department of Biostatistics and Epidemiology, University of Arizona, Tucson, AZ

ABSTRACT

Most statistical analyses use hypothesis tests or estimation about parameters to form inferential conclusions. I think this is noble, but misguided. The point of view expressed here is that observables are fundamental, and that the goal of statistical modeling should be to predict future observations, given the current data and other relevant information. Further, the prediction of future observables provides multiple advantages to practicing scientists, and to science in general. These include an interpretable numerical summary of a quantity of direct interest to current and future researchers, a calibrated prediction of what's likely to happen in future experiments, a prediction that can be either "corroborated" or "refuted" through experimentation, and avoidance of inference about parameters; quantities that exists only as convenient indices of hypothetical distributions. Finally, the predictive probability of a future observable can be used as a standard for communicating the reliability of the current work, regardless of whether confirmatory experiments are conducted. Adoption of this paradigm would improve our rigor for scientific accuracy and reproducibility by shifting our focus from "finding differences" among hypothetical parameters to predicting observable events based on our current scientific understanding.

ARTICLE HISTORY
Received March 2018
Revised June 2018

KEYWORDS
Observables, Prediction, Predictive distribution, Scientific inference

"The only useful function for a statistician is to make predictions, and thus provide a basis for action."
W. Edwards Deming

1. Introduction

Our current disciplinary dilemma over p-values is attributable, at least in part, to a wide-spread concern in science about the reproducibility of research results. Efforts such as the article "Redefine Statistical Significance" (Benjamin et al. 2018) address long-recognized problems with conventional hypothesis tests, and make progress toward greater scientific reproducibility. However, I believe the continued focus on ''fixing" hypothesis testing ignores much larger and deeper inferential problems. Among these are the dichotomization of results as either "significant" or "nonsignificant the failure to incorporate the consequences of different types of testing errors, and an unclear definition of what it means to "reproduce" a result. Rather than infer the value of a parameter that can never be observed, our inferential focus should be the prediction of future observable quantities. *Surely, a scientist is interested in the probability that a future experiment will reproduce the current result*, (Aitchison and Dunsmore 1975). Indeed I suspect that many nonstatistical scientists believe, incorrectly, that this future experimental result is somehow indicated in our traditional inference methods of hypothesis testing and confidence intervals. As a consequence, they, and others attempting to replicate their work, are confused and frustrated when a follow-on experiment fails to provide a "significant" result. They see it as a failing of statistics. And they are right.

The title of this special issue of *The American Statistician* is "Statistical Inference in the 21st Century: A World Beyond $P < 0.05$". I propose that the statistical currency of this brave new world should be the prediction of observable events or quantities. Instead of inference about parameters, the real importance of any treatment or explanatory group effect is only realized through the distribution of observables. Further, the primary purpose of statistical inference should be to predict realizable values not yet observed, based on values that were observed. Adoption of this paradigm would improve our rigor for scientific accuracy and reproducibility by shifting our focus from "finding differences" among hypothetical parameters to predicting observable events based on our current scientific understanding. A focus on prediction allows the comparison of competing theories according to the quality of predictions they make, thus leading to better scientific inference. The perspective proposed here is also related to the concepts of reliability and validity in educational and psychometric testing theory. Reliability refers to the stability of measurement, and validity to the meaningfulness of the measurement. Probabilistic prediction of observables can be used to quantify the future confirmation of current results, and constitutes a type of "measurement reliability" for inferences. In addition, the focus on observable events, rather than unobservable parameters, ensures the "meaningfulness" of results for future investigators.

CONTACT Dean Billheimer dean.billheimer@arizona.edu Department of Biostatistics and Epidemiology, University of Arizona.
© 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 prediction of observables confers multiple additional advantages to scientists, including

- focus on quantities that are important for the specific problem under study, and thus avoiding standard, but less informative, inferential summaries that arise historically because of their convenient sampling properties,
- interpretation of "causes," for example, regression parameter estimates, based on the entire distribution of observables, thus encouraging the identification of relevant changes,
- the practical significance conferred to individuals, perhaps summarized across groups, avoiding paradoxes of practical vs. statistical significance,
- the interpretation of experimental results as one in a sequence, and acknowledgment that experiments late in the sequence are related to those that occurred earlier. This feature both aids scientific reproducibility and avoids the "cult of the single experiment" (Nelder 1999).
- full accounting of the variability contributing to, or inherent in, the observed data, avoiding uncertainty laundering (Gelman 2016).

Finally, a more fundamental advantage is that prediction of observable events is understandable to humans. We are faced with uncertainty everyday. Probability judgments about things that may or may not occur are common to all of us in our daily lives.

The next section provides a selective history of previous work advocating the use of predictive inference in science, and further justification for this position. Section 3 provides an abbreviated summary of Bayesian methods for constructing predictive distributions, while Section 4 illustrates these approaches for the Gaussian two-sample case, common in laboratory and translational research. Finally, we conclude with an exhortation for a greater scientific emphasis on the prediction of events that can actually be observed.

2. Why Predictive Inference?

Prediction of future observables has long been included as an aspect of statistics, but has largely been a weak sister to parametric statistical inference. A "predictivist" approach to inference was advocated by de Finetti (1937, 2017), Aitchison and Dunsmore (1975), and Geisser (1993), among others. All of these authors stress that observables are fundamental to the statistical reasoning process, and that the purpose of statistics is to infer about realizable values not observed, based on values that were observed. Thus, the proper use of statistical models is the prediction of future observations. More recently, Briggs (2016) argues that parametric statistical modeling is misguided and misleading for reasoning about processes that generate data, and that proper expressions of uncertainty involve only the predictive distribution of future observables, given previous data and specified values of covariates.

Our current parametric statistical modeling paradigm is a descendant of measurement error models,

$$X = \theta + \epsilon$$

in which the goal is to estimate a physical constant (θ) for example, Planck's constant or the mass of a specific object, in the presence of measurement uncertainty (ϵ). In most modern applications, however, the physical, biological, and/or psychological systems generating the data are inherently complex. The processes contributing to variation are often more complicated than the measurement error variety (Geisser 1988). Thus, model parameters are quantities associated with hypothetical distributions, and their use is mostly motivated by modeling convenience. There is little reason to think they exist outside of our imagination. In some instances, a parameter may be conceived as the limit of a function of an infinite sequence of observations (Geisser 1993), and can be a useful concept when our object of interest is approximated as a large sample limit.

In many applications, however, parametric inference is problematic for at least two reasons. First, future results based on a finite sample are more uncertain than those based on an hypothetical infinite sample. Thus parametric inference amounts to a type of "uncertainty laundering" (Gelman 2016) and serves to mask the actual uncertainty inherent in inference. A second concern is that most studies in laboratory and translational science are not "stand-alone" but instead are one in a sequence, each building on the results of those done previously. Thus, the natural quantity to consider in the *next experiment* is conditional on the finite sample inference from the previous experiment. Inherent in this reasoning is the idea of repeatability of the previous study. "If the results of my current study are reliable, what is the probability that I observe $\mathcal{X}$ in the follow on study? where $\mathcal{X}$ is an important feature of the study. It is important to note that the most relevant experimental quantity of interest may not be the usual reported summary statistics. For example, many rodent studies rely on and expect large treatment effects from experimental manipulations. A treatment that produces less than a 50% increase in response may be irrelevant to the investigator. Thus, a reasonable predictive quantity is, "the probability that a repeated future experiment produces at least a 50% increase in mean response, given the current data." I believe that most scientists think in terms of observables; indeed, they observe them every day. As such, it makes sense to focus inference on things that can be observed. Observation is the currency of science.

From a statistical perspective, the prediction of a future single observation may seem most natural. For many problems, especially those related to personal decision-making, it will be the appropriate event to consider. For a scientist considering reproducibility, however, other future sample sizes may be more relevant. Which sample sizes to consider? At least four situations come to mind (for future sample size M, and current sample size N)

- Prediction of the next observation ($M = 1$).
- Replication of results of the previous experiment ($M = N$).
- Prediction of results in a future targeted experiment—this depends on the anticipated next step (e.g., probability of success in a targeted phase III drug trial, based on phase II trial results.).
- Prediction of results in a large (infinite) sample ($M = \infty$).

For many laboratory, translational, and clinical scientists, a predictive statement about a replicating study is a useful summary of evidence. A high probability that a scientifically meaningful

event will occur (say > 0.95) can be a convincing argument. Scientifically meaningful, of course, depends on the context of the problem. The key idea is that the quantity to be predicted should be relevant to the experimental setting, and should be potentially observable. Note that a confirmatory experiment need not actually be conducted to provide a useful inferential statement. However, if a confirmatory experiment is conducted, the new result can be judged as "corroborating" or "refuting" the original result. Inferential statements about future events must have meaning to consumers of the analysis results; statements about hypothetical parameters often do not.

3. Predictive Methods for Introductory Problems

We begin with the standard statistical approach that $\mathbf{x}_n' = (x_1, x_2, \ldots, x_n)$ are modeled as originating independently (exchangeably) from some distribution (probability density) $f(\mathbf{x}_n|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of unknown parameters. The "usual approach" (whether frequentist, likelihoodist, or Bayesian) then proceeds to estimate the unknown value of $\boldsymbol{\theta}$ and the associated uncertainty of the estimate, often testing that some component is consistent with an hypothesized value. I find it somewhat puzzling that once the data are observed and the likelihood is specified, that $\mathbf{x}_n$ is mostly ignored, except perhaps for diagnostic checks for f.

Instead, I propose we focus on modeling the predictive distribution for the *next observation*, x_{n+1}, given the data observed thus far, $\mathbf{x}_n$, and use this as the basis of our inference. While predictive distributions are possible via frequentist, likelihood, and Bayesian methods, my approach is "convenient Bayesian" which relies on the de Finetti representation theorem (de Finetti 1937) and implies a lurking parametric model to assist in the construction. The rules of probability stipulate that the predictive distribution can be written as

$$g(x_{n+1}|\mathbf{x}_n) = \frac{\int_\Theta f(\mathbf{x}_{n+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta f(\mathbf{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{h(\mathbf{x}_{n+1})}{h(\mathbf{x}_n)} \tag{1}$$

where $h(\mathbf{x}_k)$ is the marginal density of $\mathbf{x}_k$, $\boldsymbol{\theta}$ has prior distribution, $\pi(\boldsymbol{\theta})$, and Θ is the parameter space for $\boldsymbol{\theta}$. Then, the predictive density $g(\cdot)$ may be written as follows:

$$g(x_{n+1}|\mathbf{x}_n) = \int_\Theta f(x_{n+1}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta} \mid \mathbf{x}_n)d\boldsymbol{\theta}, \tag{2}$$

where x_{n+1} is conditionally independent of $\mathbf{x}_n$ given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta} \mid \mathbf{x}_n)$ is the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}_n$. In passing, note that the expected utility decision structure can easily be adapted to utilize predictive distributions of observables, rather than posterior distributions for parameters. Thus, a coherent decision-making framework is available for predictions (Aitchison and Dunsmore 1975; Geisser 1993).

To make ideas more concrete, in the examples that follow, we suppose $X_i \sim N(\mu_x, \sigma^2)$, and independently, $Y_i \sim N(\mu_y, \sigma^2)$, with N_1 observations from population X and N_2 observations from population Y. We further specify reference prior distributions for μ_x, μ_y, and σ^2 as

$$p(\mu_x, \mu_y, \sigma^2) = p(\mu_x, \mu_y|\sigma^2) \times p(\sigma^2) \propto \text{const.} \times \frac{1}{\sigma^2} \tag{3}$$

Now, for a future sample of size M from each of the two populations, it is easy to show (see, e.g., Geisser 1993, p. 120) that

$$\frac{(\bar{X}_M - \bar{Y}_M) - (\bar{x}_{N_1} - \bar{y}_{N_2})}{\left[(1/N_1) + (1/N_2) + (2/M)\right]^{1/2} s} \tag{4}$$

has a Student's t distribution with $N_1 + N_2 - 2$ degrees of freedom, where

$$s^2 = \frac{\sum_{i=1}^{N_1}(x_i - \bar{x})^2 + \sum_{i=1}^{N_2}(y_i - \bar{y})^2}{N_1 + N_2 - 2}.$$

Thus, the predictive distribution for the difference of sample means is a familiar t distribution, but with a factor $2s^2/M$ added to the squared scale parameter. Note that this has similar form to the usual posterior distribution for the difference in mean parameters, except that $\mu_x - \mu_y$ has been replaced with $\bar{X}_M - \bar{Y}_M$. Clearly, for $M = 1$, we obtain addition of twice the pooled sample variance estimate. Conversely, as M gets large, the expression reverts to the usual two-sample t distribution for population mean differences.

Regression problems can use a similar predictive summary of the associative relationship between differing values of an explanatory variable (X) and the resulting predictive distributions of the dependent variable, Y. For example, for the simple linear model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i\,; \quad \epsilon_i \sim N(0, \sigma^2); \\ i &= 1, 2, \ldots, N, \text{ independently} \end{aligned} \tag{5}$$

with reference prior distribution

$$p(\beta_0, \beta_1, \sigma^2) \propto \frac{1}{\sigma^2} \tag{6}$$

the predictive distribution for a new observation Y^* at $X = x^*$ is

$$Y^* - \hat{y}^* \sim t_{N-2}\ s\sqrt{1 + \frac{1}{N} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}, \tag{7}$$

where $\hat{y}^*$ is the usual least squares point estimate for Y at $X = x^*$, s is computed as $\sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/(N-2)}$, and t_{N-2} is Student's t distribution with $N - 2$ degrees of freedom (see e.g., Clarke and Clarke 2012, p. 18).

If X is an important explanatory variable, our object of inference is the effect on the predictive distributions for $Y|X = x_1$ and $Y|X = x_2$, where x_1 and x_2 are meaningful to the experimenter. The importance of the explanatory variable can be summarized by graphing the predictive distributions for key values of X, while holding other covariates constant at meaningful values. As before, the "significance" of differences in predictions must be interpreted in the context of the scientific problem.

Predictive distribution results for standard (simple) models are summarized in Aitchison and Dunsmore (1975) and Geisser (1993). A modern and expanded compendium of results is included in Clarke and Clarke (2012). Finally, I do not (in general) advocate the use of reference prior distributions for scientific inference, and instead propose to capture relevant scientific information in prior distributions of either parameters or

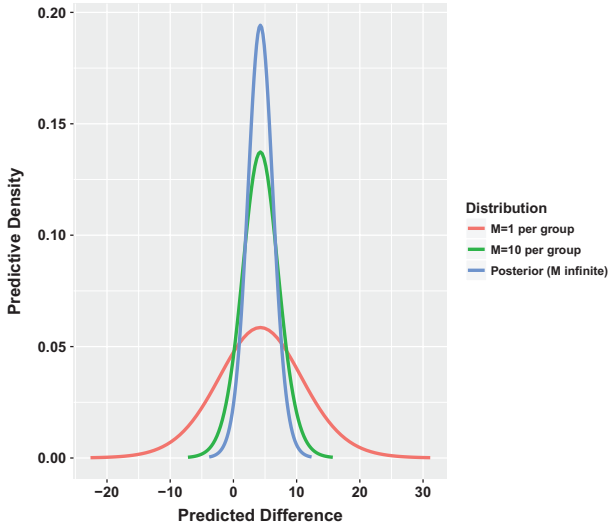


Figure 1. Posterior and conditional predictive distributions for Y - X.

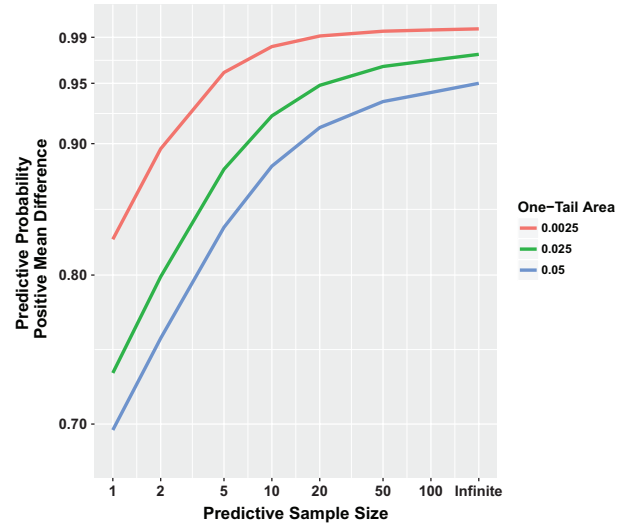


Figure 2. Predictive probability of a positive mean difference given initial one-sided p-value and sample size N=10 per group. Curves show the predictive probability of a positive mean difference for different values of predictive sample size and one-tail area (single sided p-value) in the original sample. Thus, an original experiment resulting in p=0.025 (one tail, green curve), has about 0.92 probability of yielding a positive mean difference in a repeated experiment of M=10 observations per group.

of potential observations (see, e.g., Geisser 1993; Bedrick, Christensen, and Johnson 1996). In either case, analytical solutions may not exist for many (most) problems of interest. However, for the Bayesian approach outlined above, Monte Carlo sampling methods or other approximations are simple to implement using well-established algorithms.

4. Illustrations

To illustrate the predictive inference procedure, we consider a "made-up" problem comparing two treatment groups from the section above, with $N = 10$ mice (observations) per group. As described, we model these observations as normal with unknown means, and common unknown variance. Suppose we observe means $\bar{x} = 14$ from the control group, $\bar{y} = 18.3$ from the treated group, and pooled standard deviation $s = 4.5$. This results in an effect size of 0.94 and a p-value of 0.05 from a two-sample t-test (using a two-sided alternative, so about 0.025 area in each tail of the distribution).

Now, using the predictive result from equation (4), we observe the posterior distribution and predictive distributions for a repeated experiment of $M = 10$ mice per group, and single observations, $M = 1$, shown in Figure 1. As expected, the posterior distribution for difference in means ($M = \infty$) produces a result similar to the hypothesis test (albeit with a very different interpretation) with about 0.025 probability of a negative mean difference. Conversely a repeated experiment with $M = 10$ per group is predicted to exhibit a negative *sample mean difference* with probability 0.08 (three times greater). If instead, we consider a single pair of observations, the probability of a negative difference of $Y - X$ is about 0.27; that is, the probability that a control-treated animal (X) is greater than a treated animal (Y). Switching our focus now to *positive* mean differences, Figure 2 shows the probability of a positive difference for selected tail areas in the original experiment and future sample sizes M. We observe that for small future sample sizes, there remains considerable uncertainty regarding even the sign of the difference in sample means.

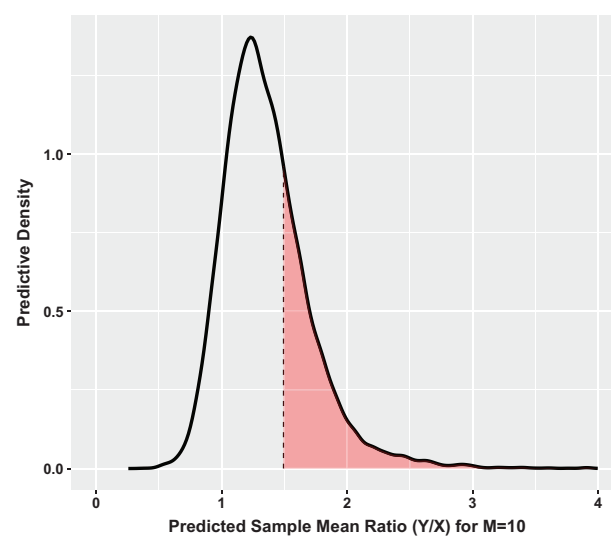


Figure 3. Predicted ratio of sample means for a repeated experiment with M=10 mice per group. The high probability region for the ratio of means ranges from about 0.8 to 2. The red shaded region shows the event that the predicted ratio is greater than 1.5, the minimum important fold change.

Finally, returning to the wishes of my laboratory collaborators, we would like to predict the probability of a big effect in a future experiment. For the current experiment, the observed treatment fold-change is 1.3 (ratio of treatment means). If we repeat the experiment, what is the probability of at least a 50% treatment-associated increase in sample mean?

Although an analytical solution to this question is not available, a Monte Carlo approximation is easy to program by sampling from scaled-shifted t distributions for each of the two treatment groups. Figure 3 shows the estimated predictive density for a ratio of sample means in a repeated experiment with $M = 10$ animals per group. We see that the estimated probability of a ratio greater than 1.5 is 0.28. While this value does provide some evidence of a robust treatment effect, it does not

suggest that a replication of the experiment will reliably produce the desired result. Note that the predictive approach encourages the use of summaries that are dictated by scientific considerations. We simply summarize the Monte Carlo realizations based on the selected measure.

5. Conclusion

Most statistical analyses use hypothesis tests about parameters (or occasionally confidence intervals) to form conclusions. They implicitly assume that if the "true parameters" were known, then the problem would be finished (Geisser 1993). Recent proposals, such as simply changing the accepted p-value threshold (Benjamin et al. 2018) to require stronger evidence, do not directly address reproducibility concerns. While a p-value (or Bayes factor) may be useful as a simple numerical summary, it should not be enshrined as the "mother of all statistics" for scientific decision-making.

The point of view expressed here is that parametric modeling can be a useful approximation, and that the values of parameters are important only through their effects on the distribution of future observables. Further, any conclusions or decisions based on these predictions must be judged in the context of the specific problem under study. The predictive inference paradigm encourages scientists (and statisticians) to characterize important results in terms of quantities that can be observed, and to predict the probability of these quantities in future studies.

Moreover, our discipline should move away from context-free decision-making—without a clear sense of losses associated with incorrect decisions. Predictive inference of future observables, given the current data and other related information, can be used as a basis for making coherent decision. This approach brings multiple advantages including

- observed values are easier to interpret than (hypothetical) parameters
- consideration of future observations, whether $M = 1$ or $M = N$ or $M = \infty$, encourages practitioners to consider how results will be used in future studies and/or in decision-making.
- for settings with finite populations, such as the treatment of rare diseases—inference based on finite sample prediction may be the only appropriate and ethical approach.

Finally, the predictive probability of a scientifically relevant, observable event can be used as a standard for communicating the reliability of the current work, regardless of whether the confirmatory experiment is conducted. If it is conducted, then the future observation can serve to support or detract from the original scientific model. Moreover, predictive inference allows the comparison of competing scientific theories according to the quality of predictions they make, thus leading to better science.

References

Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, Cambridge, UK: Cambridge University Press. [291,292,293]

Bedrick, E. J., Christensen, R., and Johnson, W. (1996), "A New Perspective on Priors for Generalized Linear Models," *Journal of the American Statistical Association*, 91, 1450–60. [294]

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B., ..., and Johnson, V. E. (2018), "Redefine Statistical Significance," *Nature Human Behavior*, 2, 6–10. [291,295]

Briggs, W. (2016), *Uncertainty. The Soul of Modeling, Probability and Statistics*, Berlin, Germany: Springer International. [292]

Clarke, B., and Clarke, J. (2012), "Prediction in Several Conventional Contexts," *Statistical Surveys*, 6, 1–73. [293]

de Finetti, B. (1937), "La prevision: ses lois logiques, ses sources subjectives," *Annals Institute Henri Poincare*, 7, 1–68. [292,293]

——— (2017), *Theory of Probability*, vol. I and II, New York: Wiley. [292]

Geisser, S. (1988), "The Future of Statistics in Retrospect," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 147–158. [292]

——— (1993), *Predictive Inference: An Introduction*, London: Chapman and Hall/ CRC. [292,293,294,295]

Gelman, A. (2016), "The Problems with p-Values are Not Just with p-Values," *The American Statistician.* Supplemental material to the ASA statement on p-values and statistical significance, by Ronald L. Wasserstein & Nicole A. Lazar (2016). The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70, 129–133, DOI: 10.1080/00031305.2016.1154108. [292]

Nelder, J. A. (1999), "Statistics for the Millenium," *The Statistician*, 48, 257–269. [292]

The American Statistician

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

Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing

Charles F. Manski

To cite this article: Charles F. Manski (2019) Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing, The American Statistician, 73:sup1, 296-304, DOI: 10.1080/00031305.2018.1513377

To link to this article: https://doi.org/10.1080/00031305.2018.1513377

© 2019 The Authors. Published with license by Taylor & Francis Group, LLC.

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 6532

View related articles

View Crossmark data

Citing articles: 11 View citing articles

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, 296–304: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1513377

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing

Charles F. Manski

Department of Economics and Institute for Policy Research, Northwestern University, Evanston, IL

Abstract

A central objective of empirical research on treatment response is to inform treatment choice. Unfortunately, researchers commonly use concepts of statistical inference whose foundations are distant from the problem of treatment choice. It has been particularly common to use hypothesis tests to compare treatments. Wald's development of statistical decision theory provides a coherent frequentist framework for use of sample data on treatment response to make treatment decisions. A body of recent research applies statistical decision theory to characterize uniformly satisfactory treatment choices, in the sense of maximum loss relative to optimal decisions (also known as maximum regret). This article describes the basic ideas and findings, which provide an appealing practical alternative to use of hypothesis tests. For simplicity, the article focuses on medical treatment with evidence from classical randomized clinical trials. The ideas apply generally, encompassing use of observational data and treatment choice in nonmedical contexts.

ARTICLE HISTORY
Received December 2017
Revised August 2018

KEYWORDS
Analysis of treatment response; Medical decisions; Minimax regret; Randomized clinical trials

1. Introduction

A central objective of empirical research on treatment response is to inform treatment choice. Identification problems combine with the necessity of inference from sample data to limit the informativeness of studies. Unfortunately, researchers commonly use concepts of statistical inference whose foundations are distant from the problem of treatment choice. It has been particularly common to use hypothesis tests to compare treatments.

The Wald (1950) development of statistical decision theory provides a coherent frequentist framework for use of sample data on treatment response to make treatment decisions. A body of recent research applies statistical decision theory to characterize uniformly satisfactory treatment choices, in the sense of maximum loss relative to optimal decisions (also known as maximum regret). This article describes the basic ideas and findings, which provide an appealing practical alternative to use of hypothesis tests.

To keep the exposition simple and framed in a familiar setting, I focus on medical treatment with evidence from classical randomized clinical trials. Trials have long enjoyed a favored status within evidence-based medicine, often being called the "gold standard" for collection of data on treatment response. The influential Cochrane system for grading the quality of evidence ordinarily reserves its highest rating for evidence from randomized trials (Higgins and Green 2011, Sec. 12.2.1). The drug approval process of the U.S. Food and Drug Administration (FDA) ordinarily considers only experimental evidence when making decisions on drug approval. While I focus on the use of trial data in medical decision making, the broad ideas discussed here are general, encompassing use of observational data and treatment choice in nonmedical contexts.

Section 2 reviews the use of hypothesis tests to compare medical treatments and the basic principles of statistical decision theory. Section 3 describes the recent research on uniformly satisfactory treatment choice using trial data. Section 4 concludes.

Readers may be aware that it has become increasingly common to express concern that evaluation of empirical research by the outcome of hypothesis tests generates publication bias and diminishes the reproducibility of findings. See, for example, Ioannidis (2005) and Wasserstein and Lazar (2016). This concern is important, but it is distinct from the theme of the present article.

The article relates directly to other important issues covered in the *ASA Statement on Statistical Significance and p-Values* (Wasserstein and Lazar 2016). Two of the six Principles of the Statement are

"3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

"5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result."

Treatment choice using statistical decision theory is not based at all on whether a p-value passes a threshold. Statistical decision theory clearly distinguishes between the statistical and clinical significance of empirical estimates of treatment effects.

CONTACT Charles F. Manski *cfmanski@northwestern.edu* Department of Economics, Northwestern University, Evanston, IL 60208.

© 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.

2. Background

2.1. Using Hypothesis Tests to Compare Treatments

A long-standing practice in medicine has been to use trial data to test a specified null hypothesis against an alternative and to use the outcome of the test to compare treatments. A common procedure when comparing two treatments in a trial is to view one as the status quo and the other as an innovation. The usual null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better. If the null hypothesis is not rejected, it is recommended that the status quo treatment continue to be used in clinical practice. If the null is rejected, it is recommended that the innovation become the treatment of choice. This type of test is institutionalized in the FDA drug approval process, which calls for comparison of a new drug with a placebo or a previously approved treatment. Approval of the new drug normally requires rejection of the null hypothesis of zero average treatment effect in two independent trials (Fisher and Moyé 1999).

The convention has been to perform a test that fixes the probability of rejecting the null hypothesis when it is correct, the probability of a Type I error. Then sample size determines the probability of rejecting the alternative hypothesis when it is correct, the probability of a Type II error. The power of a test is defined as one minus the probability of a Type II error. The convention has been to choose a sample size that yields specified power at some value of the effect size deemed clinically important. For example, International Conference on Harmonisation (1999) has provided guidance for the design and conduct of trials evaluating pharmaceuticals, stating (p. 1923)

> Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of type II error is conventionally set at 10% to 20%.

Manski and Tetenov (2016) observed that there are several reasons why hypothesis testing may yield unsatisfactory results for medical decisions and other forms of treatment choice. These include

1. *Use of Conventional Asymmetric Error Probabilities*: It has been standard to fix the probability of Type I error at 5% and the probability of Type II error at 10–20%. The theory of hypothesis testing gives no rationale for selection of these conventional error probabilities. It gives no reason why a clinician concerned with patient welfare should find it reasonable to make treatment choices that have a substantially greater probability of Type II than Type I error.
2. *Inattention to Magnitudes of Losses to Welfare When Errors Occur*: A clinician should care about more than the probabilities of Type I and II error. He should care as well about the magnitudes of the losses to patient welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between treatments is larger, but the theory of hypothesis testing does not take this into account.
3. *Limitation to Settings with Two Treatments*: A clinician often chooses among several treatments and many clinical trials compare more than two treatments. Yet the standard theory of hypothesis testing only contemplates choice between two treatments. Statisticians have struggled to extend it to deal sensibly with comparisons of multiple treatments.

The third issue is well-appreciated, but the first two are often overlooked. A simple example shows why they may matter for patient care.

2.1.1. Example

Suppose that a typically terminal form of cancer may be treated by a status quo treatment or an innovation. It is known from experience that mean patient life span with the status quo treatment is 1 year. Prior to use of the innovation, medical researchers see two possibilities for its effectiveness. It may be less effective than the status quo, yielding a mean life span of only 1/3 of a year, or it may be much more effective, yielding a mean life span of 5 years.

Suppose that a classical randomized trial is performed to learn the effectiveness of the innovation. Let the trial data be used to perform a conventional hypothesis test comparing the innovation and the status quo. The null hypothesis is that the innovation is no more effective than the status quo and the alternative is that the innovation is more effective. The probability of a Type I error is set at 0.05 and that of a Type II error is 0.20. The test result is used to choose between the treatments.

A Type I error occurs with frequentist probability 0.05 and reduces mean patient life span by 2/3 of a year (1 year minus 1/3 year). A Type II error occurs with frequentist probability 0.20 and reduces mean patient life span by 4 years (5 years minus 1 year). Thus, use of the test to choose between the status quo and the innovation implies that society is willing to tolerate a large (0.20) chance of a large welfare loss (4 years) when making a Type II error, but only a small (0.05) chance of a small welfare loss (2/3 of a year) when making a Type I error. The theory of hypothesis testing does not motivate this asymmetry.

2.2. Principles of Statistical Decision Theory

2.2.1. Basic Ideas

The standard formalization of decision making under uncertainty supposes that a decision maker must choose among a set of feasible actions. The welfare achieved by an action depends on an unknown feature of the environment, called the *state of nature*. The decision maker wants to choose an action that maximizes welfare, but he cannot do this with certainty because the state of nature is unknown. The decision maker lists all states of nature that he believes could possibly occur. This list, the *state space*, expresses partial knowledge. The larger the state space, the less the decision maker knows about the outcomes of actions.

For example, the decision maker may be a clinician and the actions may be treatments for a patient. A state of nature may characterize how a patient would respond to alternative treatments, which may be incompletely known. Welfare may be a health outcome of interest, perhaps patient life span or quality of life, that would occur when a specified treatment is administered to a patient. The clinician might ideally want to choose a treatment that optimizes the patient's health outcome

but, having incomplete knowledge of treatment response, he cannot do this with certainty.

Suppose that a sampling process generates observable sample data that are informative about the true state; for example, data on how each member of a sample of patients has responded to the assigned treatment. Wald (1950) considered the general problem of using such sample data to make decisions. He posed the task as choice of a*statistical decision function*, which maps potentially available data into a choice among the feasible actions. Wald's seminal book is abstract, making it a difficult read. Ferguson (1967), Berger (1985), and Parmigiani and Inoue (2009) provide comprehensive expositions.

Wald recommended ex ante evaluation of statistical decision functions as *procedures* applied as the sampling process is engaged repeatedly to draw independent data samples. The idea of a procedure transforms the original statistical problem of induction from a single sample into the deductive problem of assessing the probabilistic performance of a statistical decision function across realizations of the sampling process. Thus, the theory is frequentist.

Wald proposed that the decision maker evaluate a statistical decision function by the mean welfare it yields across realizations of the sampling process. His presentation differed semantically from the one that I use to describe treatment choice in that he defined loss to be the negative of welfare, took the objective to be minimization of loss rather than maximization of welfare, and used the term *risk* to denote mean loss across realizations of the sampling process. With these semantic distinctions, he prescribed a three-step decision process

1. Specify the set of feasible actions, the loss (or welfare) function, and the state space. These basic concepts of decision theory are context specific. The set of feasible actions is commonly considered to be predetermined. The loss function and the state space are subjective. The former formalizes what the decision maker wants to achieve and the latter expresses what states of nature he believes could possibly occur.
2. Eliminate inadmissible statistical decision functions. A decision function is inadmissible if there exists another that yields at least as good mean sampling performance in every state of nature and strictly better mean performance in some state.
3. Use some criterion to choose an admissible statistical decision function. Wald focused on the *minimax* criterion and on minimization of a subjective mean of the risk function (called *Bayes risk*). Savage (1951) proposed *minimax regret*.

2.2.2. Decision Criteria

What are reasonable ways to choose an admissible decision function? The Bayesian approach is particularly well known. It has been common to think of the Bayesian process of transforming a prior into a posterior distribution as antithetical to frequentist statistics, but Wald provided a clear frequentist perspective on Bayes decisions. He showed that minimization of Bayes risk, a frequentist decision criterion, yields the same decisions as would occur if one performs Bayesian inference, combining the prior distribution with the data to form a posterior subjective distribution, and then chooses an action to minimize the posterior mean of expected loss. Berger (1985, Sec. 4.4.1) gives an accessible proof.

Bayesian decision making is compelling when one feels able to place a credible subjective prior distribution on the state space. There exists a considerable body of work ranging across multiple disciplines that develops methods to help persons conceptualize uncertainty and express themselves in subjective probabilistic terms. See, for example, Savage (1971), Koriat, Lichtenstein, and Fischhoff (1980), Morgan and Henrion (1990), Manski (2004b), and Garthwaite, Kadane, and O'Hagan (2005).

Nevertheless, Bayesians have long struggled to provide guidance on specification of priors to be used in evidence-based medicine and the matter continues to be controversial. See, for example, the spectrum of views regarding Bayesian analysis of randomized trials expressed by the authors and discussants of Spiegelhalter, Freedman, and Parmar (1994). The controversy suggests that inability to express a credible prior is common in actual decision settings.

When one finds it difficult to assert a credible subjective distribution, Bayesian statisticians who believe it essential to use a probability distribution to express uncertainty may suggest use of some default distribution that is variously called a "reference" or "conventional" or "objective" prior; see, for example, Berger (2006). However, there is no consensus on what prior should play this role. The choice made matters for decision making.

An alternative is to abandon the notion that one must use a probability distribution to express uncertainty. In the absence of a prior, Wald argued that a reasonable way to act is to use a decision criterion that achieves uniformly satisfactory results, whatever the true state of nature may be. There are multiple ways to formalize the idea of uniformly satisfactory results. The two most commonly studied are embodied in the minimax and minimax-regret (MR) criteria.

The minimax criterion chooses an action that minimizes the maximum risk that might possibly occur, across all feasible states of nature. The minimax-regret criterion considers each state of nature and computes the incremental risk that occurs if one chooses a specified action rather than the one that minimizes risk in this state. This quantity, called *regret*, measures the nearness to optimality of the specified action in the state of nature. The decision maker must choose without knowing the true state. To achieve a uniformly satisfactory result, he computes the maximum regret of each action; that is, the maximum distance from optimality that the action would yield across all possible states of nature. The MR criterion chooses an action that minimizes maximum regret.

The minimax and MR criteria are sometimes confused with one another, but they yield the same choice only in certain special cases. Although the minimax criterion considers only the worst outcome that an action may yield, MR considers the worst outcome relative to what is achievable in a given state of nature. Savage (1951) distinguished the minimax criterion sharply from MR, writing that the former criterion is "ultra-pessimistic" while the latter is not. Maximum regret quantifies how uncertainty—lack of knowledge of the true state of nature—potentially diminishes the quality of decisions.

It is important to understand that use of the minimax or the MR criteria does not eliminate all subjectivity in decision making. As discussed earlier, decision theory begins with specification of a welfare function and a state space, both of which are subjective. Bayes decision theory goes a step further by placing a

subjective distribution on the state space. The minimax and MR criteria do not embrace this further element of subjectivity, but they still require the decision maker to specify a welfare function and state space.

The reader may have noticed that, to introduce this discussion, I asked what are "reasonable" decision criteria rather than what is an "optimal" criterion. Statistical decision theorists recognized from the outset that there is no singularly optimal way to choose an admissible decision function. There at most are reasonable ways. Wald (1950), who was particularly concerned with decision making in the absence of a prior distribution on the state space, motivated his focus on the minimax criterion in part by stating (p. 18)

> a minimax solution seems, in general, to be a reasonable solution of the decision problem when an a priori distribution …does not exist or is unknown to the experimenter.

Ferguson (1967) wrote (p. 28)

> It is a natural reaction to search for a 'best' decision rule, a rule that has the smallest risk no matter what the true state of nature. Unfortunately, *situations in which a best decision rule exists are rare and uninteresting*. For each fixed state of nature there may be a best action for the statistician to take. However, this best action will differ, in general, for different states of nature, so that no one action can be presumed best overall.

He went on to write (p. 29): "A *reasonable* rule is one that is better than just guessing."

2.2.3. Some History, Post Wald

The Wald framework has breathtaking generality. In principle, it enables comparison of all statistical decision functions whose risk functions exist. It enables comparison of alternative sampling processes as well as decision rules. It uses no asymptotic approximations. It applies whatever information the decision maker may have. The state space may be finite dimensional or larger; that is, nonparametric. The true state of nature may be point or partially identified. Settings with partial identification are ones where the sampling process generating the data incompletely reveals the true state asymptotically; see Manski (2003, 2007a).

Given the appeal of statistical decision theory, one might anticipate that it would play a central role in modern statistics. However, this has not occurred. After publication of Wald (1950), a surge of important extensions and applications followed in the 1950s. Much research focused on best point prediction under square loss with sample data, analysis of which began with Hodges and Lehmann (1950). In this important case, regret is mean square error and the MR criterion yields a predictor that minimizes maximum mean square error across the state space.

However, this period of rapid development closed by the 1960s, with the exception of Bayesian statistical decision theory. Bayesian analysis has continued to develop, but as a self-contained field of study disconnected from the Wald frequentist framework. Recent research in Bayesian statistics has focused more on the computational problem of transformation of priors into posteriors than on use of posteriors in decision making.

Why did statistical decision theory lose momentum long ago? One reason may have been the technical difficulty of the subject. Wald's ideas are easy to describe abstractly, but applying them can be analytically and computationally demanding. Determination of admissible decision functions and minimax/minimax-regret rules is often difficult. Another reason may have been diminishing interest in decision making as the motivation for analysis of sample data. Modern statisticians tend to view their objectives as estimation and hypothesis testing rather than decision making.

I cannot be sure what role these or other reasons played in the vanishing of statistical decision theory from statistics in the latter part of the twentieth century. However, the near absence of the subject in mainstream journals and textbooks of the period is indisputable. I think this is unfortunate. The recent research described in the next sections aims to reinvigorate statistical decision theory, focusing on the important applied problem of treatment choice. Other recent research, not described here, provides new analysis of best point prediction under square loss when the sampling process is afflicted with missing data or other problems of imperfect data quality; see Dominitz and Manski (2017).

3. Recent Work on Statistical Decision Theory for Treatment Choice

Bayesian statistical decision theory has long been available to design trials and to choose treatments with trial data. DeGroot (1970) provides a classical treatise on the subject. Canner (1970), Spiegelhalter, Freedman, and Parmar (1994), Cheng, Su, and Berry (2003), and Spiegelhalter (2004) study various aspects. However, as mentioned above, Bayesians have struggled to provide guidance on specification of priors and the matter continues to be controversial. Perhaps as a result, Bayesian analysis is well known but seldom used in evidence-based medicine. A limited exception is that the FDA has provided guidance permitting the use of Bayesian statistics in the design and analysis of clinical trials evaluating new medical devices; see U. S. Food and Drug Administration (2010).

I describe here a recent body of research that avoids specification of priors and instead studies uniformly satisfactory treatment choice, using maximum regret to measure performance across states of nature. Contributions to this emerging literature include Manski (2004a; 2005; 2007a; 2007b), Schlag (2006), Hirano and Porter (2009), Stoye (2009; 2012), Tetenov (2012), Manski and Tetenov (2016), and Kitagawa and Tetenov (2018).

The recent research supposes that the objective of treatment choice is to maximize a social welfare function that sums treatment outcomes across a population of patients that may have heterogeneous treatment response. For example, the objective may be to maximize the five-year survival rate in a population of cancer patients or mean life span in a population with a chronic disease. In this setting, a statistical decision function uses the data to choose an allocation of patients to treatments. Using terminology introduced in Manski (2004a), such a function has been called a *statistical treatment rule* (*STR*).

The mean sampling performance of an STR across repeated samples is its *expected welfare*. This term means the negative

of Wald's risk and is used because the objective is to maximize welfare rather than minimize loss. Given that the objective is to maximize rather than minimize a function, the minimax decision criterion becomes maximin instead. The MR criterion remains as earlier except that regret in a state of nature now is the maximum welfare achievable in that state minus the expected welfare of a specified STR.

In what follows, Section 3.1 explains why the recent work has measured the performance of STRs by maximum regret. Section 3.2. discusses treatment choice with existing trial data. Section 3.3 considers the design of trials.

3.1. Measuring Performance by Maximum Regret

In the absence of a prior distribution on the state space, practical and conceptual reasons motivate measurement of the performance of STRs by maximum regret across the state space, rather than by minimum expected welfare. I explain here.

3.1.1. Practical Appeal

From a practical perspective, it has been found that MR decisions behave more reasonably than do maximin ones in the context of treatment choice. In common settings of treatment choice with trial data on outcomes that take a bounded range of values, it has been found that the MR rule is well approximated by the *empirical success* (ES) rule, which chooses the treatment with the highest observed average outcome in the trial. The ES rule provides a simple and plausible way to use the results of a trial. The performance of the ES rule from the perspective of maximum regret was initiated by Manski (2004a). Subsequently, Schlag (2006) and Stoye (2009) showed that this rule either exactly or approximately minimizes maximum regret in common settings with two treatments when sample size is moderate. Hirano and Porter (2009) showed that the ES rule is asymptotically optimal.

In contrast, the maximin rule commonly ignores the trial data, whatever they may be. When Savage (1951) stated that the minimax criterion is "ultrapessimistic," he went on to write (p. 63): "it can lead to the absurd conclusion in some cases that no amount of relevant experimentation should deter the actor from behaving as though he were in complete ignorance." Savage did not flesh out this statement but it is easy to show that this occurs with trial data. Manski (2004a) provides a simple example that will be discussed in Section 3.2.

3.1.2. Conceptual Appeal

The conceptual appeal of using maximum regret to measure performance is that maximum regret quantifies how lack of knowledge of the true state of nature diminishes the quality of decisions. While the term "maximum regret" has become standard in the literature, it is important to keep in mind that this term is a shorthand for the maximum suboptimality of a decision criterion across the feasible states of nature. An STR with small maximum regret is uniformly near-optimal across all states.

Maximum regret is well-defined in general settings with multiple treatments and when patients have heterogeneous observable covariates that may be used to differentiate treatment. However, the concept is especially transparent when there are two treatments and the members of the patient population are observationally identical, all having the same observable covariates.

Suppose there are two feasible treatments, say A and B. In a state of nature where A is better, the regret of an STR is the product of the probability across repeated samples that the rule commits a Type I error (choosing B) and the magnitude of the loss in expected welfare that occurs when choosing B. Similarly, in a state where B is better, regret is the probability of a Type II error (choosing A) times the magnitude of the loss in expected welfare when choosing A.

Recall the critique in Section 2.1 of the conventional use of hypothesis testing to choose a treatment. I called attention to the asymmetric treatment of Type I and Type II error probabilities and the inattention to magnitudes of losses when errors occur. Evaluating treatment rules by regret overcomes both problems. Regret considers Type I and II error probabilities symmetrically and it measures the magnitudes of the losses that errors produce.

3.1.3. Example

To illustrate, consider again the example of Section 2.1.1, in which a conventional hypothesis test is used as an STR to choose between a status quo treatment for cancer and an innovation. There are two feasible states of nature in the example, with the innovation yielding mean life span of 1/3 year in one state and 5 years in the other. In the first state, the regret of this conventional "test rule" equals 1/30 of a year; that is, a 0.05 chance of a Type I error times a 2/3 of a year reduction in mean life span with improper choice of the innovation. In the second state, the regret of the test rule equals 4/5 of a year; that is, a 0.20 chance of a Type II error times a 4-year reduction in mean life span with improper choice of the status quo. Thus, the maximum regret of the test rule is 4/5 of a year.

Rather than use the conventional test rule to choose between the status quo and the innovation, one could seek an STR that has smaller maximum regret. Given the available trial data, a simple option would be to reverse the conventional probabilities of Type I and Type II; thus, one might use a test with a 0.20 chance of a Type I error and a 0.05 chance of a Type II error. In the first state, the regret of this unconventional test rule STR equals 2/15 of a year; that is, a 0.20 chance of a Type I error times a 2/3 of a year reduction in mean life span with improper choice of the innovation. In the second state, the regret of the unconventional test rule equals 1/5 of a year; that is, a 0.05 chance of a Type II error times a 4-year reduction in mean life span with improper choice of the status quo. Thus, the maximum regret of the unconventional test rule is 1/5 of a year.

In this example, the unconventional test rule delivers much smaller maximum regret than does the conventional test rule. There may exist other STRs that perform even better.

3.2. Treatment Choice With Existing Trial Data

To move beyond verbal discussion and examples, I now formalize treatment choice as a statistical decision problem in the relatively simple setting of a classical trial with two treatments

and a population of observationally identical patients. The presentation in this section draws substantially on Manski (2007, Chap. 12). The research literature cited earlier also studies more general and complex settings with multiple treatments, patients who have heterogeneous observable covariates, and imperfect trials that only partially identify treatment response.

3.2.1. General Analysis

Suppose that a health planner must assign treatment A or B to each member of patient population J. Each patient $j \in J$ has response function $y_j(\bullet)$: $T \to Y$ mapping treatments $t \in T$ into individual outcomes $y_j(t) \in R$. Let P denote the distribution of treatment response in the population.

The members of the population may respond heterogeneously to treatment, but they are observationally identical to the planner. For any $\delta \in [0, 1]$, the planner can allocate a fraction δ of patients to treatment B and $1 - \delta$ to A. The planner wants to choose δ to maximize an additive welfare function

$$\begin{aligned} U(\delta, P) &= E[y(A)](1-\delta) + E[y(B)]\delta \\ &= \alpha(1-\delta) + \beta\delta = \alpha + (\beta - \alpha)\delta, \end{aligned} \tag{1}$$

where $\alpha \equiv E[y(A)]$ and $\beta \equiv E[y(B)]$ are the mean outcomes if everyone were to receive treatment A or B, respectively. The quantity $\beta - \alpha$ is the average treatment effect (ATE) in the population. It is optimal to set $\delta = 1$ if the ATE is positive and $\delta = 0$ if the ATE is negative. The problem of interest is treatment choice when incomplete knowledge of P makes it impossible to determine the sign of the ATE.

Suppose that sample data are available. Let Q be the sampling distribution and Ψ be the sample space. For example, the data may be treatment response observed in a randomized trial. A STR is a function $\delta(\bullet) : \Psi \to [0, 1]$ that maps sample data into a treatment allocation. The welfare realized with δ and data ψ is the random variable

$$U(\delta, P, \psi) = \alpha + (\beta - \alpha)\delta(\psi). \tag{2}$$

The state space $[(P_s, Q_s), s \in S]$ is the set of (P, Q) pairs that the planner deems possible. Expected welfare in state s, the mean sampling performance of rule δ in this state, is

$$W(\delta, P_s, Q_s) = \alpha_s + (\beta_s - \alpha_s)E_s[\delta(\psi)]. \tag{3}$$

Here $E_s[\delta(\psi)] \equiv \int_\Psi \delta(\psi)\mathrm{d}Q_s(\psi)$ is the expected allocation of patients to treatment B, across repeated samples.

Rule δ is admissible if there exists no rule δ' such that $W(\delta', P_s, Q_s) \geq W(\delta, P_s, Q_s)$ for all $s \in S$ and $W(\delta', P_s, Q_s) > W(\delta, P_s, Q_s)$ for some s. The Bayes, maximin, and MR rules are as follows:

$$\text{Bayes rule: } \max_{\delta \in [0,1]} \int_S W(\delta, P_s, Q_s)\mathrm{d}\pi(s), \tag{4}$$

where π is a subjective distribution on the state space.

$$\text{Maximin rule : } \max_{\delta \in [0,1]} \min_{s \in S} W(\delta, P_s, Q_s). \tag{5}$$

$$\text{Minimax-regret rule : } \min_{\delta \in [0,1]} \max_{s \in S}[\max(\alpha_s, \beta_s) - W(\delta, P_s, Q_s)]. \tag{6}$$

3.2.2. Illustration: Choice Between a Status Quo Treatment and an Innovation When Outcomes Are Binary

To illustrate in perhaps the simplest nontrivial setting, let the outcomes y be binary, taking the value zero if treatment fails and one if it succeeds. Let A be a status quo treatment and B be an innovation. Suppose that the planner knows the success probability $\alpha \equiv P[y(A) = 1]$ of the status quo treatment but not the success probability $\beta \equiv P[y(B) = 1]$ of the innovation. The planner wants to choose treatments to maximize the success probability.

A randomized trial is performed to learn about outcomes under the innovation, with N subjects randomly drawn from the population and assigned to treatment B. The observed trial outcomes are that n subjects realize outcome $y = 1$ and $N - n$ realize $y = 0$. In this setting, N indexes the sampling process and the number n of experimental successes is a sufficient statistic for the data.

The feasible STRs are functions $\delta(\bullet)$: $[0, \ldots, N] \to [0, 1]$ that map the number of experimental successes into a treatment allocation. The expected welfare of rule δ is

$$W(\delta, P, N) = \alpha + (\beta - \alpha)E[\delta(n)]. \tag{7}$$

n is distributed binomial $\mathbf{B}[\beta, N]$, so

$$E[\delta(n)] = \sum_{i=0}^{N} \delta(i) f(n = i; \beta, N), \tag{8}$$

where $f(n = i; \beta, N) \equiv N![i!\bullet(N-i)!]^{-1}\beta^i(1-\beta)^{N-i}$ is the probability of i successes. The only unknown determinant of expected welfare is β, so the state space S indexes the feasible values of β. Specifically, $\beta_s \equiv P_s[y(b) = 1]$.

It is reasonable in this setting to conjecture that admissible treatment rules should be ones in which the fraction of the population allocated to treatment B increases with n. It turns out that the admissible treatment rules are a simple subclass of these rules. A theorem of Karlin and Rubin (1956) shows that the admissible rules are the *monotone treatment rules*. Monotone rules assign all persons to the status quo if the experimental success rate is below some threshold and all to the innovation if the success rate is above the threshold. Thus, δ is admissible if and only if

$$\delta(n) = 0 \text{ for } n < n_0, \tag{9a}$$

$$\delta(n) = \lambda \text{ for } n = n_0, \tag{9b}$$

$$\delta(n) = 1 \text{ for } n > n_0, \tag{9c}$$

for some $0 \leq n_0 \leq N$ and $0 \leq \lambda \leq 1$.

The collection of monotone treatment rules is a mathematically "small" subset of the space of all feasible treatment rules. Nevertheless, it still contains a broad range of rules. These include the following:

Data-Invariant Rules: These are the rules $\delta(\bullet) = 0$ and $\delta(\bullet) = 1$, which assign all persons to treatment A or B, respectively, whatever n may be.

Empirical Success Rule: An optimal treatment rule allocates all persons to treatment A if $\beta < \alpha$ and all to B if $\beta > \alpha$. The empirical success rule emulates the optimal rule by replacing β with its sample analog, the empirical success rate n/N.

Table 1. *Minimax-regret rules for small sample sizes*

	$N=0$	$N=1$	$N=2$	$N=3$	$N=4$	$N=5$	$N=6$	$N=7$	$N=8$	$N=9$	$N=10$
n_0: threshold sample size											
$\alpha=0.10$	0	0	0	0	0	0	0	0	0	1	1
$\alpha=0.25$	0	0	0	1	1	1	1	2	2	2	2
$\alpha=0.50$	0	1	1	2	2	3	3	4	4	5	5
$\alpha=0.75$	0	1	2	2	3	4	5	5	6	7	8
$\alpha=0.90$	0	1	2	3	4	5	6	7	8	8	9
λ: threshold allocation											
$\alpha=0.10$	0.9	0.67	0.52	0.41	0.32	0.26	0.18	0.09	0	0.89	0.78
$\alpha=0.25$	0.75	0.36	0.17	0.93	0.67	0.42	0.18	0.93	0.67	0.43	0.18
$\alpha=0.50$	0.5	1	0.5	1	0.5	1	0.5	1	0.5	1	0.5
$\alpha=0.75$	0.25	0.64	0.83	0.07	0.33	0.58	0.82	0.07	0.33	0.57	0.82
$\alpha=0.90$	0.1	0.33	0.48	0.59	0.68	0.74	0.82	0.91	1	0.11	0.22
Minimax regret value											
$\alpha=0.10$	0.09	0.067	0.052	0.041	0.033	0.027	0.022	0.019	0.017	0.017	0.017
$\alpha=0.25$	0.19	0.09	0.052	0.039	0.038	0.035	0.03	0.027	0.027	0.025	0.023
$\alpha=0.50$	0.25	0.063	0.063	0.044	0.044	0.035	0.035	0.03	0.03	0.027	0.027
$\alpha=0.75$	0.19	0.09	0.052	0.039	0.038	0.035	0.03	0.027	0.027	0.025	0.023
$\alpha=0.90$	0.09	0.067	0.052	0.041	0.033	0.027	0.022	0.019	0.017	0.017	0.016

Source: Manski (2007, Table 12.1).

Bayes Rules: The form of the Bayes rule depends on the prior subjective distribution placed on β. Consider the class of Beta priors, which form the conjugate family for a Binomial likelihood. Let $(\beta_s, s \in S) = (0,1)$ and let the prior be Beta with parameters (c,d). Then, the posterior mean for β is $(c+n)/(c+d+N)$. The resulting Bayes rule is

$$\delta(n) = 0 \text{ for } (c+n)/(c+d+N) < \alpha, \tag{10a}$$

$$\delta(n) = \lambda \text{ for } (c+n)/(c+d+N) = \alpha, \quad \text{where } 0 \le \lambda \le 1, \tag{10b}$$

$$\delta(n) = 1 \text{ for } (c+n)/(c+d+N) > \alpha. \tag{10c}$$

Maximin Rule: Minimum expected welfare for rule δ is

$$\min_{s\in S} W(\delta, P_s, N) = \alpha + \min_{s\in S}(\beta_s - \alpha)E_s[\delta(n)]. \tag{11}$$

$E_s[\delta(n)] > 0$ for all $\beta_s > 0$ and for all monotone rules except $\delta(\bullet) = 0$. When S contains states with $\beta_s < \alpha$, the maximin rule is $\delta(\bullet) = 0$. Thus, the maximin rule ignores the trial data on treatment response, whatever they may turn out to be. This illustrates the Savage (1951) statement that using the maximin rule can induce one to entirely ignore available sample data.

Minimax-Regret Rule: The regret of rule δ in state s is

$$\begin{aligned}\max(\alpha, \beta_s) &- \{\alpha + (\beta_S - \alpha)E_s[\delta(n)]\} \\ &= (\beta_S - \alpha)\{1 - E_s[\delta(n)]\}1[\beta_s \ge \alpha] \\ &\quad + (\alpha - \beta_s)E_s[\delta(n)] \times 1[\alpha \ge \beta_s].\end{aligned} \tag{12}$$

Thus, regret is the mean welfare loss when a member of the population is assigned the inferior treatment, multiplied by the expected fraction of the population assigned this treatment. The minimax-regret rule does not have an analytical solution, but it can be determined numerically. When all values of β are feasible, the minimax-regret rule is well approximated by the empirical success rule.

3.2.3. Numerical Computation of the Minimax-Regret Rule for Small Samples

Manski (2007, Table 12.1) reports numerical computations of the minimax-regret rule for specified values of α and N when all values of β are feasible; that is, when $(\beta_s, s \in S) = [0,1]$. I reproduce the findings here in Table 1 The top two panels display the value of (n_0, λ) for this rule. The third panel displays the value of minimax regret.

The top panel of the table shows that the threshold n_0 of experimental successes for allocation of persons to treatment B increases with the sample size and with the success probability of treatment A. The inequality $|n_0 - \alpha N| \le 1$ holds everywhere in the table. Thus, the minimax-regret rule is well approximated by an empirical success rule.

The third panel shows that the value of minimax regret decreases by roughly an order of magnitude as the sample size increases from to 10. For example, when $\alpha = 0.50$, it falls from 0.25 to 0.027. Thus, even a sample size as small as 10 suffices to make maximum regret quite small.

3.3. Designing Trials to Enable Near-Optimal Treatment Choice

From the perspective of treatment choice, an ideal objective for the design of trials would be to collect data that enable subsequent implementation of an optimal treatment rule in the patient population of interest; that is, a rule for use of trial data that always selects the best treatment, with no chance of error. Optimality is too strong a property to be achievable with finite sample size. However, near-optimal rules—ones with small maximum regret—exist when classical trials are large enough.

Manski and Tetenov (2016) investigated trial design that enables near-optimal treatment choices. It is shown that, given any $\varepsilon > 0$, *ε-optimal* rules exist when trials have large enough sample size. An ε-optimal rule has expected welfare, across repeated samples, within ε of the welfare of the best treatment

Table 2. *Minimum sample sizes per treatment enabling ε-optimal treatment choice: binary outcomes, two treatments, balanced designs*

ε	ES rule	One-sided 5% z-test	One-sided 1% z-test
0.0	145	3488	7963
0.03	17	382	879
0.05	6	138	310
0.10	2	33	79
0.15	1	16	35

Source: Manski and Tetenov (2016, Table 1).

in every state of nature. Equivalently, it has maximum regret no larger than ε.

The article considers trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments. It reports exact results for cases of two treatments and binary outcomes; see Section 3.3.1 below for a summary. It gives sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when there are multiple treatments and outcomes are bounded. These conditions are obtained by application of large deviations inequalities to evaluate the performance of empirical success rules.

Choosing sample size to enable existence of ε-optimal treatment rules requires specification of a value for ε. The selected ε determines how much deviation from optimality a decision maker is willing to tolerate when making treatment choices. The value of ε should be specified by clinical researchers concerned with patient care rather than by some universal convention. Clinical researchers may, perhaps, find it congenial to let ε equal the *minimum clinically important difference* (MCID) in the average treatment effect comparing alternative treatments.

Medical research has long distinguished between the statistical and clinical significance of treatment effects. While the idea of clinical significance has been interpreted in various ways, many writers call an average treatment effect clinically significant if its magnitude is greater than a specified value deemed minimally consequential in clinical practice. International Conference on Harmonisation (1999) put it this way (p. 1923): "The treatment difference to be detected may be based on a judgment concerning the minimal effect which has clinical relevance in the management of patients."

3.3.1. Findings with Binary Outcomes, Two Treatments, and Balanced Designs

Determination of sample sizes that enable near-optimal treatment is simple in settings with binary outcomes (coded 0 and 1 for simplicity), two treatments, and a balanced design which assigns the same number of subjects to each treatment group. Manski and Tetenov (2016, Table 1), reproduced here as Table 2, provides exact computations of the minimum sample size that enables ε-optimality when a clinician uses one of three different treatment rules, for various values of ε.

The first column shows the minimum sample size (per treatment arm) that yields ε-optimality when a clinician uses the empirical success (ES) rule to make a treatment decision. The ES rule chooses the treatment with the better average outcome in the trial. The rule assigns half the population to each treatment if there is a tie. It is known that the ES rule minimizes maximum regret in settings with binary outcomes, two treatments, and balanced designs (Stoye 2009).

The second and third columns display the minimum sample sizes that yield ε-optimality of rules based on one-sided 5 and 1% hypothesis tests. Decisions made with these tests take the two treatments to be a status quo and an innovation, choosing the innovation if the estimated treatment effect is positive and statistically significant. There is no consensus on what hypothesis test should be used to compare two proportions. Results are reported based on the widely used one-sided two-sample z-test, which is based on an asymptotic normal approximation (Fleiss 1973).

The findings are remarkable. A sample as small as two observations per treatment arm makes the ES rule ε-optimal when $\varepsilon = 0.1$ and a sample of size 145 suffices when $\varepsilon = 0.01$. The minimum sample sizes required for ε-optimality of the test rules are orders of magnitude larger. If the z-test of size 0.05 is used, a sample of size 33 is required when $\varepsilon = 0.1$ and 3488 when $\varepsilon = 0.01$. The sample sizes must be more than double these values if the z-test of size 0.01 is used. See Manski and Tetenov (2016) for discussion of the factors that underlie these findings.

3.3.2. Implications for Practice

Based on their exact calculations and analytical findings using large-deviations inequalities, Manski and Tetenov conclude that sample sizes determined by clinically relevant near-optimality criteria tend to be much smaller than ones set by conventional statistical power criteria. Reduction of sample size relative to prevailing norms can be beneficial in multiple ways. Reduction of total sample size can lower the cost of executing trials, the time necessary to recruit adequate numbers of subjects, and the complexity of managing trials across multiple centers. Reduction of sample size per treatment arm can make it feasible to perform trials that increase the number of treatment arms and, hence, yield information about a wider variety of treatment options.

4. Conclusion

Science does not always progress monotonically. There are times when important, even fundamental, ideas are discovered and receive attention but then are neglected. This has occurred with statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end. A revival in the context of treatment choice began in the early 2000s and has been gathering force. I hope that the growing dissatisfaction of statisticians with ritual applications of hypothesis testing, exemplified by the ASA Statement in Wasserstein and Lazar (2016), will encourage statisticians to relearn statistical decision theory and use it when studying not only treatment choice with trial data but decision making with sample data more generally.

Acknowledgment

The author is grateful to the Editors and reviewers for their comments.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer. [298]

Berger, J. (2006), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1, 385–402. [298]

Canner, P. (1970), "Selecting One of Two Treatments When the Responses Are Dichotomous," *Journal of the American Statistical Association*, 65, 293–306. [299]

Cheng, Y., Su, F., and Berry, D. (2003), "Choosing Sample Size for a Clinical Trial Using Decision Analysis," *Biometrika*, 90, 923–936. [299]

DeGroot, M. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill. [299]

Dominitz, J., and Manski, C. (2017), "More Data or Better Data? A Statistical Decision Problem," *Review of Economic Studies*, 84, 1583–1605. [299]

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, San Diego, CA: Academic Press. [298,299]

Fisher, L., and Moyé, L. (1999), "Carvedilol and the Food and Drug Administration Approval Process: An Introduction," *Controlled Clinical Trials*, 20, 1–15. [297]

Fleiss, J. (1973), *Statistical Methods for Rates and Proportions*, New York: Wiley. [303]

Garthwaite, P., Kadane, J., and O'Hagan, A. (2005), "Statistical Methods for Eliciting Probability Distributions," *Journal of the American Statistical Association*, 100, 680–701. [298]

Higgings, J., and Green, S. (eds.) (2011), *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0, The Cochrane Collaboration, available at *http://handbook-5-1.cochrane.org*. [296]

Hirano, K., and Porter, J. (2009), "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701. [299,300]

Hodges, E., and Lehmann, E. (1950), "Some Problems in Minimax Point Estimation," *Annals of Mathematical Statistics*, 21, 182–197. [299]

International Conference on Harmonisation, ICH E9 Expert Working Group (1999), "Statistical Principles for Clinical Trials: ICH Harmonized Tripartite Guideline," *Statistics in Medicine*, 18, 1905–1942. [297,303]

Ioannidis, J. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, 696–701. [296]

Karlin, S., and Rubin, H. (1956), "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio," *Annals of Mathematical Statistics*, 27, 272–299. [301]

Kitagawa, T. and Tetenov, A. (2018), "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616. [299]

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980), "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–118. [298]

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer. [299]

——— (2004a), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221–246. [299,300]

——— (2004b), "Measuring Expectations," *Econometrica*, 72, 1329–1376. [298]

——— (2005), *Social Choice With Partial Knowledge of Treatment Response*, Princeton, NJ: Princeton University Press. [299]

——— (2007a), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press. [299]

——— (2007b), "Minimax-Regret Treatment Choice With Missing Outcome Data," *Journal of Econometrics*, 139, 105–115. [299]

Manski, C., and Tetenov, A. (2016), "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, 113, 10518–10523. [297,299,302,303]

Morgan, G. and Henrion, M. (1990), *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis*, New York: Cambridge University Press. [298]

Parmigiani, G. and Inoue, L. (2009), *Decision Theory: Principles and Approaches*, New York: Wiley. [298]

Savage, L. (1951), "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46, 55–67. [298,300,302]

Savage, L. (1971), "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66, 783–801. [298]

Schlag, K. (2006), "Eleven Tests Needed for a Recommendation," European University Institute Working Paper ECO No. 2006/2. [299,300]

Spiegelhalter, D., Freedman, L., and Parmar, M. (1994), "Bayesian Approaches to Randomized Trials" (with discussion), *Journal of the Royal Statistics Society*, Series A, 157, 357–416. [298,299]

Spiegelhalter, D. (2004), "Incorporating Bayesian Ideas Into Health-Care Evaluation," *Statistical Science*, 19, 156–174. [299]

Stoye, J. (2009), "Minimax Regret Treatment Choice With Finite Samples," *Journal of Econometrics*, 151, 70–81. [299,300,303]

——— (2012), "Minimax Regret Treatment Choice With Covariates or With Limited Validity of Experiments," *Journal of Econometrics*, 166, 138–156. [299]

Tetenov, A. (2012), "Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria," *Journal of Econometrics*, 166, 157–165. [299]

U.S. Food and Drug Administration (2010), *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. Available at *https://www.fda.gov/MedicalDevices/ucm071072.htm*. [299]

Wald, A. (1950), *Statistical Decision Functions*, New York: Wiley. [296,298,299]

Wasserstein, R., and Lazar, N. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [296,303]

The American Statistician

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

Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II

Charles F. Manski & Aleksey Tetenov

To cite this article: Charles F. Manski & Aleksey Tetenov (2019) Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II, The American Statistician, 73:sup1, 305-311, DOI: 10.1080/00031305.2018.1543617

To link to this article: https://doi.org/10.1080/00031305.2018.1543617

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 3206

View related articles

View Crossmark data

Citing articles: 6 View citing articles

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, 305–311: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543617

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II

Charles F. Manski[a] and Aleksey Tetenov[b]

[a]Department of Economics and Institute for Policy Research, Northwestern University, Evanston, IL; [b]Department of Economics, University of Bristol, Bristol, UK

Abstract

A convention in designing randomized clinical trials has been to choose sample sizes that yield specified statistical power when testing hypotheses about treatment response. Manski and Tetenov recently critiqued this convention and proposed enrollment of sufficiently many subjects to enable *near-optimal* treatment choices. This article develops a refined version of that analysis applicable to trials comparing aggressive treatment of patients with surveillance. The need for a refined analysis arises because the earlier work assumed that there is only a primary health outcome of interest, without secondary outcomes. An important aspect of choice between surveillance and aggressive treatment is that the latter may have side effects. One should then consider how the primary outcome and side effects jointly determine patient welfare. This requires new analysis of sample design. As a case study, we reconsider a trial comparing nodal observation and lymph node dissection when treating patients with cutaneous melanoma. Using a statistical power calculation, the investigators assigned 971 patients to dissection and 968 to observation. We conclude that assigning 244 patients to each option would yield findings that enable suitably near-optimal treatment choice. Thus, a much smaller sample size would have sufficed to inform clinical practice.

ARTICLE HISTORY
Received September 2018
Revised October 2018

KEYWORDS
Analysis of treatment response; Clinical trials; Medical decisions; Minimax regret; Sample size

1. Introduction

A core objective of randomized clinical trials is to inform treatment choice. The convention has been to choose sample sizes that yield specified statistical power when testing designated null hypotheses about treatment response against designated alternatives. However, statistical power and hypothesis testing are at most loosely connected to effective patient care.

Manski and Tetenov (2016) critiqued the use of power calculations to set sample size and developed an alternative principle that aims to inform patient care directly. They proposed enrollment of sufficiently many subjects to enable determination of near-optimal treatment choices and provided criteria to accomplish this. An optimal treatment rule would always select the best treatment, with no chance of error. This is infeasible to achieve with trials having finite sample size. *Near-optimal* treatment choices are ones that are suitably close to this ideal. The article gave numerical calculations of sufficient sample sizes for trials with binary outcomes.

A broad conclusion was that sample sizes determined by clinically relevant near-optimality criteria are much smaller than ones set using conventional statistical power calculations. A variety of factors contribute to this conclusion. One is that the near-optimality perspective considers Type I and Type II errors symmetrically. In contrast, power calculations are performed with the probability of a Type I error a priori constrained to equal 0.05 or another conventional value. Another is that the near-optimality perspective evaluates products of effect sizes and error probabilities. It allows larger error probabilities if the two treatments are nearly equivalent from a patient welfare perspective than if one treatment is substantially better for patients.

Reduction of sample size relative to prevailing norms can be beneficial in multiple ways. Reduction of total sample size can lower the cost of executing trials, the time necessary to recruit adequate numbers of subjects, and the need to perform trials across multiple centers. Reduction of sample size per treatment arm can make it feasible to perform trials that increase the number of treatment arms and, hence, yield information about more treatment options.

This article develops and applies a refined version of the analysis in Manski and Tetenov (2016) to trials that compare aggressive treatment of patients with surveillance. Patient care abounds with instances of this choice problem. Internists choose between prescription of pharmaceuticals and surveillance when treating patients at risk of heart disease or diabetes. Oncologists choose between surveillance and aggressive treatments such as surgery or chemotherapy when treating cancer patients at risk of metastasis. Aggressive treatment may be appealing to the extent that it better prevents onset or reduces the severity of illness. Surveillance may be attractive to the extent that it avoids side effects that may occur with aggressive treatment.

The need for a refined version of the analysis in Manski and Tetenov (2016) arises because this earlier work studied settings in which there is only a primary health outcome of interest, without any secondary outcomes. An important aspect

CONTACT Charles F. Manski *cfmanski@northwestern.edu* Department of Economics and Institute for Policy Research, Northwestern University, Evanston, IL 60208.

© 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.

of choice between surveillance and aggressive treatment is that the latter may have side effects. The prevailing approach to choice of sample size in trials has been to focus entirely on the primary outcome of a treatment, without considering secondary outcomes. This practice is reasonable when the primary outcome is the dominant determinant of patient welfare or, put another way, when there is little variation in secondary outcomes across treatments. It is not reasonable otherwise. When the side effects of aggressive treatments are serious, it is more reasonable to consider how the primary outcome and side effects jointly determine patient welfare. This requires new analysis of sample design, regardless of whether one adopts the perspective of near optimality or statistical power.

To provide a realistic case study illuminating the general issues, we consider a trial comparing *nodal observation* and *lymph node dissection (lymphadenectomy)* when treating patients with early-stage cutaneous melanoma at risk of metastasis. Nodal observation is surveillance of lymph nodes by ultrasound scan, a procedure that has negligible side effects. Lymph node dissection is a surgical procedure in which the lymph nodes in the relevant regional basin are removed. Dissection is commonly viewed as an aggressive treatment. A particularly concerning side effect is chronic swelling in the region of lymph node removal, a condition called lymphedema, which may reduce patient quality of life substantially (Cheville et al. 2010). Choice between nodal observation and lymph node dissection is a common decision faced in early treatment of melanoma, breast cancer, and other forms of localized cancer. We focus on melanoma because there has long been controversy about the merits of dissection relative to observation in this context (see Faries 2018).

The Multicenter Selective Lymphadenectomy Trial II (MSLT-II) compared dissection and observation for melanoma patients who had recently undergone sentinel lymph-node biopsy and who had obtained a positive finding of malignancy. The primary outcome was defined to be melanoma-specific survival for three years following the date of randomization. Findings were reported in Faries et al. (2017).

Our concern is choice of sample size in the trial. Using a statistical power calculation, the investigators assigned 971 patients to dissection and 968 to observation. Considering sample size from the perspective of near-optimal treatment choice, we conclude that assigning 244 patients to each treatment would yield findings that enable suitably near-optimal treatment choice. Thus, a much smaller sample size would have sufficed to inform clinical practice.

We perform the computations under the assumption that treatment choice will be made with the *empirical success* (ES) rule, which selects the treatment with the higher sample average welfare. When used to choose between two treatments, the ES rule approximately yields equal Type I and Type II error probabilities. This contrasts sharply with conventional hypothesis tests, which yield asymmetric error probabilities (typically 0.05 for Type I errors and 0.10–0.20 for Type II errors). The ES rule provides a simple and plausible way to use the results of a trial. Analysis of its performance from the perspective of near optimality was initiated by Manski (2004). Stoye (2009) shows that use of this rule either exactly or approximately minimizes the sample size needed to achieve near optimality in common settings with two treatments.

The analyses in Manski and Tetenov (2016) and in the present article focus on trials such as MSLT-II that are performed to help clinicians choose between treatments that are already used in practice, rather than for a regulatory purpose such as drug approval. Trials performed for regulatory purposes may be analyzed from the perspective of improving patient welfare, but the analysis may need to recognize legal constraints and the incentives of the firms or other entities that perform the trials. See Tetenov (2016) for discussion of these matters.

2. Methods

2.1. Critique of the Use of Power Calculations to Set Trial Size

The use of statistical power calculations to set trial size derives from the presumption that outcome data will be used to test a null hypothesis against an alternative. A common practice is to use a hypothesis test to recommend which of two treatments, say A and B, should be given to a patient population. The null hypothesis often is that treatment B is no better than A, and the alternative is that B is better. If the null hypothesis is not rejected, choice of A is recommended. If the null is rejected, B is recommended.

The standard practice has been to perform a test that fixes at a predetermined level the probability of rejecting the null hypothesis when it is correct (a Type I error). Then sample size determines the probability of rejecting the alternative hypothesis when it is correct (a Type II error). The power of a test is one minus the probability of a Type II error. The convention has been to choose a sample size that yields specified power at some value of the average treatment effect deemed clinically relevant. For example, International Conference on Harmonisation (1999) provides guidance for the design of trials evaluating pharmaceuticals, stating (p. 1923):

> "Conventionally the probability of Type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of Type II error is conventionally set at 10% to 20%."

Trials with samples too small to achieve these error probabilities are called "underpowered" and are criticized as scientifically useless and medically unethical (e.g., Halpern, Karlawish, and Berlin, 2002).

MSLT-II followed the standard practice of setting sample size to obtain statistical power. The investigators enrolled 1939 subjects, citing this reasoning (Faries et al. 2017, p. 2214): "We estimated that with a total sample of 1925 patients, the trial would have a power of 83% to detect a between-group difference of 5 percentage points in melanoma-specific survival." Specifically, the investigators contemplated using the trial data to perform a two-tailed hypothesis test. The null hypothesis is that both treatments yield the same rate of three-year melanoma-specific survival. The alternative is that the survival rates differ by at least 5 percentage points. The probabilities of Type I and

Type II errors are 0.05 and 0.17. The study investigators provide detailed statements of the study protocol and statistical analysis plan in a supplementary online document.[1]

Manski and Tetenov (2016) critique the standard practice, observing that there are several reasons why hypothesis testing may yield unsatisfactory results for medical decisions. These include

Use of conventional asymmetric error probabilities: It has been standard to set the probability of Type I error at 5% and the probability of Type II error at 10–20%. The theory of hypothesis testing gives no rationale for selection of these error probabilities. It gives no reason why a clinician concerned with patient welfare should find it reasonable to make treatment choices that have a substantially greater probability of Type II than Type I error.

Disregard of magnitudes of losses when errors occur: A clinician should care about more than the probabilities of Type I and II errors. He should care as well about the magnitudes of the losses to patient welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between treatments is larger, but the theory of hypothesis testing does not take this into account.

Limitation to settings with two treatments: A clinician often chooses among several treatments and many clinical trials compare more than two treatments. Yet the standard theory of hypothesis testing only contemplates choice between two treatments. Statisticians have long struggled to extend it to deal sensibly with comparisons of multiple treatments, without consensus on how to do this.

See Manski (2019) for further critique of the use of hypothesis tests to make treatment choices.

2.2. Setting Sample Size to Enable Near-Optimal Treatment

An ideal objective for trial design would be to collect data that enable optimal treatment choice in the patient population of interest, with no chance of error. Optimality is too strong a property to be achievable with finite sample size, but near-optimal rules exist when trials with perfect internal and external validity are large enough.

Near optimality was suggested as a principle for decision making under uncertainty by Savage (1951) within an essay commenting on the seminal Wald (1950) development of statistical decision theory. Statistical decision theory studies the broad problem of decision making when one has incomplete knowledge of the "state of nature" (i.e., how patients respond to treatment), but can collect informative sample data. For example, a clinician who has incomplete knowledge of treatment response may use trial data to become better informed.

Considering each possible state of nature, Savage proposed computation of the mean loss in welfare that would occur across repeated samples if one were to choose a specified treatment rule rather than the one that is best in this state. This quantity measures the nearness to optimality of the specified treatment rule in each state of nature. The actual decision problem requires choice of a treatment rule without knowing the true state of nature. The decision maker can evaluate a rule by the maximum distance from optimality that it yields across all possible states of nature. He can then choose a rule that minimizes the maximum distance from optimality. Doing so yields a rule that is uniformly near optimal. Statistical decision theory has used the term *regret* as a shorthand for nearness to optimality. Maximum nearness to optimality is called *maximum regret* and the rule that minimizes the maximum distance from optimality is called the *minimax-regret* rule. See Manski (2019) for further discussion of the Wald theory and minimax regret.

The concept of near optimality is applicable in general settings with multiple treatments, but it is easiest to explain when there are two treatments, say A and B. In states of nature where A is better, the nearness to optimality of a specified treatment rule is the product of the probability (across repeated samples) that the rule commits a Type I error (choosing B) and the magnitude of the loss in patient welfare that occurs when choosing B. Similarly, in states where B is better, nearness to optimality is the probability of a Type II error (choosing A) times the magnitude of the loss in welfare when choosing A. In contrast to the use of hypothesis testing to choose a treatment, evaluation by near optimality views Type I and II error probabilities symmetrically and it assesses the magnitudes of the welfare losses that errors produce.

Manski and Tetenov (2016) investigated trial design that enables near-optimal treatment. They supposed that the objective is to maximize average welfare across the relevant patient population. For example, the objective may be to maximize the five-year survival rate of a population of cancer patients or the average number of quality-adjusted life years of patients with a chronic disease.

They considered trials that draw predetermined numbers of subjects at random within groups stratified by treatments and observed patient risk factors. They showed that, given any specified positive value of a constant ε measuring nearness to optimality, ε-optimal rules exist when trials have large enough sample size. A ε-optimal rule is one whose mean value of average patient welfare, across repeated samples, is within ε of the optimum in every state of nature.

They reported exact numerical results for the case of two treatments with binary outcomes such as survival versus death. They gave simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when there are multiple treatments and outcomes are bounded.

2.3. Choosing the MSLT-II Sample Size to Enable Near-Optimal Treatment

The sample-size calculations in Manski and Tetenov (2016) concerned settings where treatments do not have side effects. Hence, they should not be applied to MSLT-II or other trials comparing surveillance and aggressive treatment. The present article develops a refined version of the earlier analysis that recognizes the possibility of side effects.

We assume a simple patient welfare function that transparently expresses patient concern with both survival, the primary outcome of treatment, and lymphedema, a possible secondary outcome. Patients may perhaps have more complex welfare

[1] *www.nejm.org/doi/suppl/10.1056/NEJMoa1613210/suppl_file/nejmoa1613210_protocol.pdf*

functions than the one posed here. Our methodology for setting sample size may be applied with any welfare function. The specific findings depend on the welfare function.

Let welfare with nodal observation equal 1 if a patient survives for three years and equal 0 otherwise. Welfare with dissection depends on whether a patient experiences lymphedema. When a patient does not experience lymphedema, welfare with dissection equals 1 if the patient survives for three years and equals 0 otherwise. When a patient experiences lymphedema, welfare is lowered by a specified fraction h, whose value expresses the harm associated with lymphedema. Thus, a patient who experiences lymphedema has welfare $1 - h$ if he survives and $-h$ if he does not survive.

When making the treatment decision, a clinician does not know whether a given patient will survive and/or experience lymphedema. To cope with uncertainty about patient-specific treatment response, medical economists have recommended that clinicians maximize average welfare in a relevant patient population. Implementation of this recommendation does not require that a clinician knows patient-specific treatment response. It does, however, require that the clinician know average treatment response in the patient population (see, e.g., Phelps and Mushlin 1988; Meltzer 2001).

A few symbols help to explain in the setting of choice between nodal observation and dissection. Let nodal observation be treatment A. Let $y(A)$ denote the primary outcome with treatment A. Thus, $y(A) = 1$ if a patient survives with observation and let $y(A) = 0$ if the patient does not survive. Let p denote the probability of a specified event. Then average patient welfare with observation is the survival probability $p[y(A) = 1]$.

Let lymph node dissection be treatment B. Let $y(B) = 1$ if a patient survives with dissection and $y(B) = 0$ otherwise. Let $s(B)$ denote the secondary outcome with treatment B. Thus, $s(B) = 1$ if a patient experiences lymphedema and $s(B) = 0$ otherwise. Then average patient welfare with dissection is the difference between the survival probability and h times the probability of lymphedema; that is, $p[y(B) = 1] - h \cdot p[s(B) = 1]$. The optimal treatment is the one yielding the higher average patient welfare. Thus, observation is optimal if $p[y(A) = 1]$ exceeds $p[y(B) = 1] - h \cdot p[s(B) = 1]$. Dissection is optimal if otherwise.

In this setting, a state of nature is a set of values for the four probabilities $p[y(A) = 1], p[y(B) = 1], p[s(B) = 1|y(B) = 0]$, and $p[s(B) = 1|y(B) = 1]$. Determination of the optimal choice is feasible with knowledge of these probabilities. A trial such as MSLT-II yields information about the probabilities of survival and lymphedema that clinicians need to know to choose treatments maximizing average patient welfare. The sample size determines the extent of the information. For any positive constant ε, a sample of size N per treatment arm enables ε-optimal treatment if N is sufficiently large.

The Appendix explains the computation of sample sizes that enable ε-optimal treatment for any values of h and ε, assuming use of the ES rule. We show computational results using two methods to determine maximum regret, one applying simulated annealing and the other using a normal approximation to the finite-sample distribution of ES. Our computations of maximum regret are conservative in the sense that we impose no a priori restrictions on the values of the probabilities of survival and lymphedema. One may believe that some values of these probabilities are implausible. If so, one may restrict attention to the values deemed plausible and maximize regret across these values. Maximum regret across a restricted set of probability values logically must be less than or equal to maximum regret across all values. Hence, performing our analysis with restrictions imposed on the plausible probabilities cannot weaken our findings on minimum sample sizes that suffice to enable ε-optimal treatment choice. It may yield smaller minimum sample sizes than those we report in Section 3.

3. Findings

We report the minimal sample size enabling near-optimal treatment when $h = 0.2$ and $\varepsilon = 0.0085$. Recall that we use a scale for patient welfare in which, absent lymphedema, a patient has welfare 1 if he survives three years and welfare 0 if he dies within three years. Setting $h = 0.2$ supposes that suffering from lymphedema reduces a patient's welfare by one-fifth relative to lymphedema-free survival. This quantification of the welfare loss produced by lymphedema is suggested by Cheville et al. (2010), who elicited from a group of patients their perspectives on the matter. See Basu and Meltzer (2007) for discussion of elicitation of patient treatment and health preferences more generally.

Setting $\varepsilon = 0.085$ follows naturally from how the MSLT-II investigators performed their power calculation. They judged a difference of 5 percentage points in melanoma-specific survival to be a clinically meaningful loss in patient welfare and they judged 0.17 to be an acceptable probability of Type II error. As discussed in Section 2, regret equals the magnitude of welfare loss times the probability that the loss will occur. Thus, the MSLT-II investigators judged $0.17 \times 0.05 = 0.0085$ to be an acceptable level of regret.

When $h = 0.2$ and $\varepsilon = 0.0085$, we find that near-optimal treatment is achievable if one assigns 244 patients to observation and 244 to dissection. This total sample size of 488 is much smaller than the 1939 subjects enrolled in MSLT-II.

4. Discussion

Our reconsideration of MSLT-II opens many possibilities that the investigators could have contemplated if they had approached trial design from the perspective of near-optimal treatment rather than statistical power. They could have achieved the declared study objective—comparison of observation and dissection for patients with a malignant sentinel lymph node—with a much smaller total sample size. Reducing total sample size would have lowered the cost of executing the trial, the time required to recruit subjects, and the need to perform trials across multiple centers. Or, maintaining enrollment of 1939 subjects, they could have expanded the study objective by performing a trial with more than two treatment arms, thus yielding information about more treatment options.

Our suggestion of potential alternatives to the MSLT-II design should not be interpreted as criticism of its investigators. They adhered to what has been the standard practice in setting sample size, proceeding as have thousands of other clinical

trials. The lessons of Manski and Tetenov (2016) and this article are intended to be forward looking, as trials are designed henceforth. The ideas developed in these papers have general potential application.

A coherent alternative to setting trial size to enable near-optimal treatment is application of Bayesian statistical decision theory. The Bayesian perspective would be attractive if trial designers and clinicians were able to place a credible consensus subjective prior distribution on treatment response. However, Bayesian statisticians have long struggled to provide guidance on specification of priors and the matter continues to be controversial. In the context of clinical trials, the authors and discussants of Spiegelhalter, Freedman, and Parmar (1994) express a spectrum of views. Bayesian decision theorists have occasionally recognized that inability to specify a credible subjective distribution may yield poor decisions. Berger (1985) cautions that (p. 121): "A Bayesian analysis may be 'rational' in the weak axiomatic sense, yet be terrible in a practical sense if an inappropriate prior distribution is used."

Appendix: Computation of Sample Sizes That Enable ε-Optimal Treatment

A.1 Setup and Notation

In the setting described in the article, the vector of unknown parameters is $\theta = (a, b_{00}, b_{01}, b_{10}, b_{11})$. The probability that treatment A yields a positive outcome (survival) is $a \equiv P_\theta(Y(A) = 1)$ and the probability of a negative outcome (death) is $1 - a$. Treatment B has four possible outcomes, with two values for the main outcome $Y(B) \in \{0, 1\}$ and two for the side effect $S(B) \in \{0, 1\}$. The probability of each outcome is $b_{ys} \equiv P_\theta(Y(B) = y, S(B) = s), y \in \{0, 1\}, s \in \{0, 1\}$. The parameter space for θ is $\Theta = [0, 1]^5$, with the restriction that $b_{00} + b_{01} + b_{10} + b_{11} = 1$. This specification of the parameter space imposes no assumptions on the joint distribution of primary outcomes and side effects.

The patient's expected gain/loss from treatment B compared to treatment A in state of nature θ equals

$$\tau_\theta = \mathbb{E}_\theta(Y(B) - hS(B)) - \mathbb{E}_\theta(Y(A)).$$

With N subjects per treatment arm, a sample analog of this quantity computed from the trial data is

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(B) - hS_i(B)) - \frac{1}{N}\sum_{i=N+1}^{2N} Y_i(A),$$

where observations of individuals randomly assigned to treatment B have indices $i = 1, \ldots, N$ and those assigned to treatment A have indices $i = N + 1, \ldots, 2N$.

We assume that the patient will choose between the treatment options using the *empirical success* (*ES*) rule, choosing treatment B if $\hat{\tau} > 0$, choosing treatment A if $\hat{\tau} < 0$, and choosing either treatment with probability 1/2 if $\hat{\tau} = 0$.

The regret of the ES rule in state θ is the product of the magnitude of error $|\tau_\theta|$ from choosing the suboptimal treatment and the probability of making that error:

$$R_N(\theta) = \begin{cases} -\tau_\theta[P_\theta(\hat{\tau} > 0) + 1/2P_\theta(\hat{\tau} = 0)], \tau_\theta \le 0, \\ \tau_\theta[P_\theta(\hat{\tau} < 0) + 1/2P_\theta(\hat{\tau} = 0)], \tau_\theta > 0, \end{cases}$$

To determine whether a particular sample size N is sufficient to enable ε-optimal treatment choice when using the ES treatment rule, we need to compute the maximum regret of the ES rule $\max_{\theta\in\Theta} R_N(\theta)$ and compare it to ε. Exact computation of $\max_{\theta\in\Theta} R_N(\theta)$ is challenging because the function $R_N(\theta)$ is not convex in θ. It generally has multiple local maxima, both over $\{\theta : \tau_\theta < 0\}$ and over $\{\theta : \tau_\theta > 0\}$.

We consider two approximations for $\max_{\theta\in\Theta} R_N(\theta)$. The first method maximizes regret with an asymptotic normal approximation to $P_\theta(\hat{\tau} < 0)$. The derivations in the following section of the Appendix show that maximizing regret with this normal approximation over θ reduces to a very simple one-dimensional numerical optimization problem. The second method computes exact values of $R_N(\theta)$ for given values of θ and searches for the maximum using the simulated annealing algorithm, described in detail in Press et al. (2007). This method is often used to search numerically for the global maximum of a function that may have multiple local maxima. It is computationally intensive and is not guaranteed to find the exact maximum, but its precision can be improved by increasing the computation time. With sufficient computation time this method is more accurate, since the normal approximation clearly underestimates maximum regret, especially in small samples.

The maximum regret of the ES rule approximated using both methods is reported in Table A1 for sample sizes ranging from $N = 10$ to $N = 250$ and for values of the side effect disutility ranging from $h = 0$ to $h = 0.5$. A table with calculations of maximum regret for other values of N and h is available from the authors.

Comparison of the results obtained using the two methods shows that it is important to compute regret exactly in very small sample sizes, where the normal approximation tends to substantially underestimate maximum regret. For sample sizes $N > 100$, both methods produce similar results. It is then more practical to use the normal approximation as it is computationally much simpler.

For $h = 0.2$, the maximum regret of the ES rule is greater than 0.0085 for all sample sizes $N \le 243$ and is smaller than 0.0085 for all sample sizes $N \ge 244$. Hence, assigning 244 subjects to each treatment is sufficient for near-optimal treatment choice with $\varepsilon = 0.0085$.

A.2. Details for the Normal Approximation

A simple analytical approximation to maximum regret can be obtained with a normal approximation to the finite-sample distribution of $\hat{\tau}$. In a trial with N randomly drawn subjects per treatment arm,

$$\hat{\tau} \sim^a \mathcal{N}(\tau_\theta, V_\theta/N),$$
$$V_\theta = \mathrm{var}_\theta(Y(B) - hS(B)) + \mathrm{var}_\theta(Y(A)).$$

The regret of the ES rule in state θ equals

$$R_N(\theta) = \begin{cases} -\tau_\theta P_\theta(\hat{\tau} > 0), \tau_\theta \le 0, \\ \tau_\theta P_\theta(\hat{\tau} \le 0), \tau_\theta > 0, \end{cases}$$

which is approximated by

$$R_N^a(\theta) = \begin{cases} -\tau_\theta \Phi(\tau_\theta\sqrt{N/V_\theta}), \tau_\theta \le 0, \\ \tau_\theta \Phi(-\tau_\theta\sqrt{N/V_\theta}), \tau_\theta > 0. \end{cases} \tag{A.1}$$

Note that for a given value of $\tau_\theta \le 0, \Phi(\tau_\theta\sqrt{N/V_\theta})$ is increasing in V_θ, hence $R_N^a(\theta)$ is increasing in V_θ. For a given value of $\tau_\theta > 0, \Phi(-\tau_\theta\sqrt{N/V_\theta})$ and $R_N^a(\theta)$ are also increasing in V_θ. To find the maximum of $R_N^a(\theta)$ we will focus on a simple subset of parameter values that maximize V_θ for any τ_θ.

For any parameter vector $\theta = (a, b_{00}, b_{01}, b_{10}, b_{11})$, there exists an alternative parameter vector θ^* with the same expected gain/loss $\tau_\theta^* = \tau_\theta$, but with higher variance of the estimate $V_\theta^* \ge V_\theta$. We construct θ^* by spreading the probability mass b_{00} (on outcome $Y(B) - hS(B) = 0$) between b_{01}(outcome $Y(B) - hS(B) = -h$) and b_{10}(outcome $Y(B) - hS(B) = 1$). Moving $\frac{1}{1+h}b_{00}$ to b_{01} and $\frac{1}{1+h}b_{00}$

Table A1. Near-optimality (maximum regret) of ES rules (N is the number of subjects per treatment arm).

	$h=0$		$h=0.1$		$h=0.2$		$h=0.3$		$h=0.4$		$h=0.5$	
N	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.
10	0.038209	0.037490	0.044905	0.039672	0.045017	0.041857	0.045794	0.044046	0.046704	0.046237	0.049236	0.048431
20	0.026947	0.026689	0.030401	0.028180	0.030479	0.029672	0.031212	0.031166	0.032803	0.032661	0.034487	0.034157
30	0.021983	0.021841	0.024046	0.023039	0.024516	0.024237	0.025874	0.025435	0.026805	0.026634	0.028039	0.027834
40	0.019029	0.018937	0.020710	0.019963	0.021105	0.020989	0.021930	0.022016	0.023172	0.023044	0.024218	0.024071
50	0.017016	0.016949	0.018217	0.017860	0.018829	0.018772	0.019865	0.019683	0.020688	0.020595	0.021621	0.021507
60	0.015530	0.015480	0.016640	0.016307	0.017170	0.017134	0.018019	0.017962	0.018859	0.018789	0.019709	0.019617
70	0.014376	0.014336	0.015231	0.015099	0.015890	0.015861	0.016708	0.016624	0.017444	0.017387	0.018227	0.018150
80	0.013447	0.013414	0.014291	0.014124	0.014861	0.014835	0.015612	0.015546	0.016306	0.016257	0.017034	0.016968
90	0.012677	0.012649	0.013371	0.013317	0.014009	0.013985	0.014690	0.014653	0.015365	0.015321	0.016048	0.015990
100	0.012025	0.012002	0.012724	0.012634	0.013287	0.013266	0.013952	0.013898	0.014570	0.014530	0.015215	0.015163
150	0.009817	0.009804	0.010330	0.010316	0.010841	0.010827	0.011359	0.011339	0.011876	0.011850	0.012395	0.012362
200	0.008501	0.008493	0.008941	0.008933	0.009384	0.009374	0.009826	0.009814	0.010274	0.010255	0.010720	0.010696
250	0.007603	0.007597	0.007995	0.007990	0.008390	0.008382	0.008786	0.008775	0.009183	0.009168	0.009580	0.009560

to b_{10} preserves the mean of $Y(B) - hS(B)$, and hence τ_θ. Similarly, spreading the probability mass b_{11} (on outcome $Y(B)-hS(B) = 1-h$) by moving $\frac{h}{1+h}b_{11}$ from b_{11} to b_{01} and by moving $\frac{1}{1+h}b_{11}$ from b_{11} to b_{10} also preserves the mean of $Y(B) - hS(B)$. In summary, $\theta^* = (a, 0, b^*_{01}, b^*_{10}, 0)$ where

$$b^*_{01} = b_{01} + \frac{1}{1+h}b_{00} + \frac{h}{1+h}b_{11},$$
$$b^*_{10} = b_{10} + \frac{h}{1+h}b_{00} + \frac{1}{1+h}b_{11}.$$

Since $\mathbb{E}_{\theta^*}(Y(B) - hS(B)) = \mathbb{E}_\theta(Y(B) - hS(B))$, to show that $\mathrm{var}_{\theta^*}(Y(B) - hS(B)) \geq \mathrm{var}_\theta(Y(B) - hS(B))$ it is sufficient to show that $\mathbb{E}_{\theta^*}((Y(B) - hS(B))^2) \geq \mathbb{E}_\theta((Y(B) - hS(B))^2)$.

$$\mathbb{E}_\theta((Y(B) - hS(B))^2) = h^2 b_{01} + b_{10} + (1-h)^2 b_{11},$$
$$\begin{aligned}\mathbb{E}_{\theta^*}((Y(B) - hS(B))^2) &= h^2\left(b_{01} + \frac{1}{1+h}b_{00} + \frac{h}{1+h}b_{11}\right)\\ &\quad + \left(b_{10} + \frac{h}{1+h}b_{00} + \frac{1}{1+h}b_{11}\right)\\ &= h^2 b_{01} + b_{10} + \frac{h^2+h}{1+h}b_{00} + \frac{h^3+1}{1+h}b_{11}.\end{aligned}$$

Since $\frac{h^3+1}{1+h} = 1 - h + h^2 > 1 - 2h + h^2 = (1-h)^2$ for $h > 0$, it follows that $\frac{h^3+1}{1+h}b_{11} \geq (1-h)^2 b_{11}$. Hence, $\mathbb{E}_{\theta^*}((Y(B) - hS(B))^2) \geq \mathbb{E}_\theta((Y(B) - hS(B))^2)$.

Under θ^*, the outcome values $Y(B) - hS(B)$ of treatment B have a binary distribution on $\{-h, 1\}$, with probabilities $b^*_{01} = 1 - b^*_{10}$ and b^*_{10}. Hence,

$$\mathrm{var}_{\theta^*}(Y(B) - hS(B)) = (1+h)^2 b^*_{10}(1 - b^*_{10}).$$

and

$$V_{\theta^*} = (1+h)^2 b^*_{10}(1 - b^*_{10}) + a(1-a), \tag{A.2}$$
$$\tau^*_\theta = b^*_{10} - h(1 - b^*_{10}) - a. \tag{A.3}$$

Thus,

$$\max_{\theta\in\Theta} R^a_N(\theta) = \max_{\theta\in\Theta^*} R^a_N(\theta),$$

where $\Theta^* = \{\theta \in \Theta \text{ s.t. } \theta = (a, 0, 1 - b^*_{10}, b^*_{10}, 0), a \in [0,1], b^*_{10} \in [0,1]\}$.

We can further constrain the set of values of (a, b^*_{10}) on which V_{θ^*} may attain its maximum. The value of the treatment effect $\tau^*_\theta(a, b^*_{10})$ as a function of $\begin{bmatrix} a \\ b^*_{10} \end{bmatrix}$ remains constant along the vector $\mathbf{u} = \begin{bmatrix} 1+h \\ 1 \end{bmatrix}$. The directional derivative of $V_{\theta^*}(a, b^*_{10})$ in direction $\mathbf{u}$ equals

$$\begin{aligned}\nabla_{\mathbf{u}} V_{\theta^*} = \nabla V_{\theta^*} \cdot \mathbf{u} &= [(1-2a) \quad (1+h)^2(1 - 2b^*_{10})] \cdot \begin{bmatrix} 1+h \\ 1 \end{bmatrix}\\ &= (1+h)((1-2a) + (1+h)(1 - 2b^*_{10})).\end{aligned}$$

Hence, over the set of (a, b^*_{10}) values on which τ^*_θ is constant, V_{θ^*} is increasing in direction $\mathbf{u}$ when

$$(1-2a) + (1+h)(1 - 2b^*_{10}) > 0$$

(for low values of a and b^*_{10}) and reaches its maximum at the point where

$$(1-2a) + (1+h)(1 - 2b^*_{10}) = 0,$$

that is,

$$b^*_{10} = \frac{1 + h/2 - a}{1+h},$$

if this point is feasible or at the closest feasible point. It follows that the set of parameter values (a, b^*_{10}) which maximize variance V_{θ^*} for different values of τ^*_θ consists of three line segments:

$$a = 0,\ b^*_{10} \in \left[\frac{1+h/2}{1+h}, 1\right], \tag{A.4}$$
$$a \in [0,1], b^*_{10} = \frac{1+h/2-a}{1+h}, \tag{A.5}$$
$$a = 1,\ b^*_{10} \in \left[0, \frac{h/2}{1+h}\right]. \tag{A.6}$$

Acknowledgments

We are grateful to the editor and reviewers for constructive comments.

References

Basu, A., and Meltzer, D. (2007), "Value of Information on Preference Heterogeneity and Individualized Care," *Medical Decision Making*, 27, 112–127. [308]

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer. [309]

Cheville, A., Almoza, M., Courmier, J., and Basford, J. (2010), "A Prospective Cohort Study Defining Utilities Using Time Trade-Offs and the Euroqol-5D to Assess the Impact of Cancer-Related Lymphedema," *Cancer*, 116, 3722–3731. [306,308]

Faries, M. (2018), "Completing the Dissection in Melanoma: Increasing Decision Precision," *Annals of Surgical Oncology*, 25, 585–587, DOI: 10.1245/s10434-017-6330-4. [306]

Faries, M. B., Thompson, J. F., Cochran, A. J., Andtbacka, R. H., Mozzillo, N., Zager, J. S., Jahkola, T., Bowles, T. L., Testori, A., Beitsch, P. D., and Hoekstra, H. J. (2017), "Completion Dissection or Observation for Sentinel-Node Metastasis in Melanoma," *New England Journal of Medicine*, 376, 2211–2222. [306]

Halpern, S., Karlawish, J., and Berlin, J. (2002), "The Continued Unethical Conduct of Underpowered Clinical Trials," *Journal of the American Medical Association*, 288, 358–362. [306]

International Conference on Harmonisation (1999), "ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonized Tripartite Guideline," *Statistics in Medicine*, 18, 1905–1942. [306]

Manski, C. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221–246. [306]

——— (2019), "Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," *The American Statistician*. [307]

Manski, C., and Tetenov, A. (2016), "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, 113, 10518–10523. [305,306,307,309]

Meltzer, D. (2001), "Addressing Uncertainty in Medical Cost-Effectiveness: Implications of Expected Utility Maximization for Methods to Perform Sensitivity Analysis and the Use of Cost-Effectiveness Analysis to Set Priorities for Medical Research," *Journal of Health Economics*, 20, 109–129. [308]

Phelps, C., and Mushlin, A. (1988), "Focusing Technology Assessment Using Medical Decision Theory," *Medical Decision Making*, 8, 279–289. [308]

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007), *Numerical Recipes: The Art of Scientific Computing* (3rd ed.), Cambridge: Cambridge University Press. [309]

Savage, L. (1951), "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46, 55–67. [307]

Spiegelhalter, D., Freedman, L., and Parmar, M. (1994), "Bayesian Approaches to Randomized Trials (with discussion)," *Journal of the Royal Statistical Society*, Series A, 157, 357–416. [309]

Stoye, J. (2009), "Minimax Regret Treatment Choice With Finite Samples," *Journal of Econometrics*, 151, 70–81. [306]

Tetenov, A. (2016), "An Economic Theory of Statistical Testing," Cemmap Working Paper CWP50/16. [306]

Wald, A. (1950), *Statistical Decision Functions*, New York: Wiley. [307]

The American Statistician

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

Frequentist, Bayes, or Other?

Michael Lavine

To cite this article: Michael Lavine (2019) Frequentist, Bayes, or Other?, The American Statistician, 73:sup1, 312-318, DOI: 10.1080/00031305.2018.1459317

To link to this article: https://doi.org/10.1080/00031305.2018.1459317

© 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: 6091

View related articles

View Crossmark data

Citing articles: 4 View citing articles

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, 312–318: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1459317

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Frequentist, Bayes, or Other?

Michael Lavine

Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA

ABSTRACT

Both philosophically and in practice, statistics is dominated by frequentist and Bayesian thinking. Under those paradigms, our courses and textbooks talk about the accuracy with which true model parameters are estimated or the posterior probability that they lie in a given set. In nonparametric problems, they talk about convergence to the true function (density, regression, etc.) or the probability that the true function lies in a given set. But the usual paradigms' focus on learning the true model and parameters can distract the analyst from another important task: discovering whether there are many sets of models and parameters that describe the data reasonably well. When we discover many good models we can see in what ways they agree. Points of agreement give us more confidence in our inferences, but points of disagreement give us less. Further, the usual paradigms' focus seduces us into judging and adopting procedures according to how well they learn the true values. An alternative is to judge models and parameter values, not procedures, and judge them by how well they describe data, not how close they come to the truth. The latter is especially appealing in problems without a true model.

ARTICLE HISTORY
Received July 2017
Revised March 2018

KEYWORDS
Likelihood; Statistical philosophy; Theory of inference

Many statistical analyses are framed as problems of model selection or parameter estimation. Frequentists have procedures for choosing models and estimating parameters; Bayesians have procedures for calculating posterior probabilities of models and parameters. Either type of procedure might make sense when there are true models and parameters. But often there are not, and pretending otherwise can lead to inferential errors because it diverts attention from the often crucial task of finding many good models. We illustrate with examples from the literature.

Example 1 (prostate cancer). Hastie, Tibshirani, and Friedman (2009) analyzed a dataset on prostate cancer that was originally reported by Stamey et al. (1989). According to Hastie, Tibshirani, and Friedman,

> "[Stamey et al.] examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45)."

After showing a pairwise scatterplot of all the variables, a correlation matrix of the predictor variables, and a simple linear regression of the level of prostate-specific antigen (lpsa) on the eight predictors, the next analysis by Hastie, Tibshirani, and Friedman is presented in their Table 3.3, reproduced here as Table 1 which shows results from linear models of the form $Y \sim \mathrm{N}(X\beta, \sigma^2)$ where Y is the level of antigen in a patient, X is a matrix of covariates, β is a vector of coefficients, and σ^2 is the error variance. They explain that the table:

> "shows the coefficients from a number of different selection and shrinkage methods. They are *best-subset selection* using an all-subsets search, *ridge regression*, the *lasso*, *principal components regression*, and *partial least squares*. Each method has a complexity parameter, and this was chosen to minimize an estimate of prediction error based on tenfold cross-validation."

There are 2^8 models $M_1, \ldots, M_{256}$ (we follow Hastie, Tibshirani, and Friedman and consider only linear models with an intercept) that indicate which coefficients are set to 0, and the focus is on choosing a model and estimating its parameters. But choosing a model and estimating parameters tell only part of the story about the data. We hope to tell another part by finding models that are nearly as good as the best. For a parametric model with parameter θ and density $p_\theta(y)$ the log-likelihood function is $\ell(\theta) = \log(p_\theta(y))$ treated as a function of θ for the fixed, observed y. We're interested in log-likelihood functions because we interpret them as quantifying how well the parameter value θ describes the data y. When a multidimensional θ is divided into two parts, say θ_1 and θ_2, the profile log-likelihood function is $\ell^P(\theta_1) = \sup_{\theta_2} \log(p_{(\theta_1,\theta_2)}(y))$ treated as a function of θ_1 for the fixed, observed y, which we interpret as quantifying how well the parameter value θ_1 describes the data y when paired with its best partner θ_2.

In the prostate cancer example we can consider the model M to be a parameter and examine the profile log-likelihood function $l^P(M) = \sup_{\beta \in M,\sigma} \log(p_M(y))$, where the notation $\sup_{\beta \in M,\sigma}$ means that the supremum is taken over σ and those β's in model M while the other β's are set to 0. We try to find features of M that allow $l^P(M)$ to be large relative to its maximum over the 256 models. Our story will depend not just on the maximum of $l^P(M)$, but on the whole function. We work with

CONTACT Michael Lavine lavine@math.umass.edu Department of Mathematics and Statistics, University of Massachusetts, LGRT. Box 34515, Amherst, MA 01003.

© 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.

Table 1. Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data.

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	− 0.141		− 0.046		− 0.152	− 0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	− 0.288		0.000		− 0.051	0.079
gleason	− 0.021		0.040		0.232	0.011
pgg45	0.267		0.133		− 0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

The blank entries correspond to variables omitted. Table and caption from Hastie, Tibshirani, and Friedman (2009).

the 67 cases Hastie, Tibshirani, and Friedman used for training and omit the 30 cases they used for testing.

Figure 1 is a density and rug plot of the 256 values $l^P(M_i)$. The maximum is $l^P(M) \approx -67.5$ which is for the full model and corresponds to the LS column of Table 1.

LASSO says that the model ought to include lcavol, lweight, lbph, and svi and exclude age, lcp, gleason, and pgg45. $l^P(M)$ for the LASSO model is about -71.2. Figure 1 shows there are many models with $l^P(M)$ as large or nearly as large as -71.2. Our view is that if any of the models with large $l^P(M)$, say larger than about -75 or so, exclude lcavol, lweight, lbph, or svi, then the inference for including those variables is weak because we can describe the data nearly as well without those variables as with them.

Figure 2 shows $l^P(M)$ labelled according to which variables are in the model, for the 80 models containing at least three of lcavol, lweight, lbph, and svi. The x-axis shows which other variables are included. The five curves show which of the four key variables is omitted. Inference for including lcavol is quite strong: removing lcavol and refitting any model decreases l^P by more than 10. But inference for including the other variables is weak: removing them and refitting decreases l^P by much less. In LASSO's preferred model, labelled "1" on the x-axis, removing any of lweight, lbph, or svi decreases l^P by less than 3.1, so the model without the variable describes the data nearly as well as a similar model with the variable. Our conclusion contrasts with convention: fitting LASSO's preferred model yields p-values of 9×10^{-7}, 0.017, 0.051, and 0.017 for lcavol, lweight, lbph, and

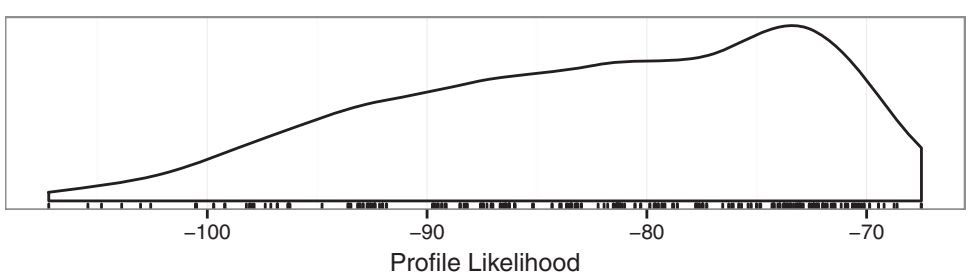


Figure 1. Density and rug plots of the profile likelihood for the 256 models of the prostate cancer data in Example 1.

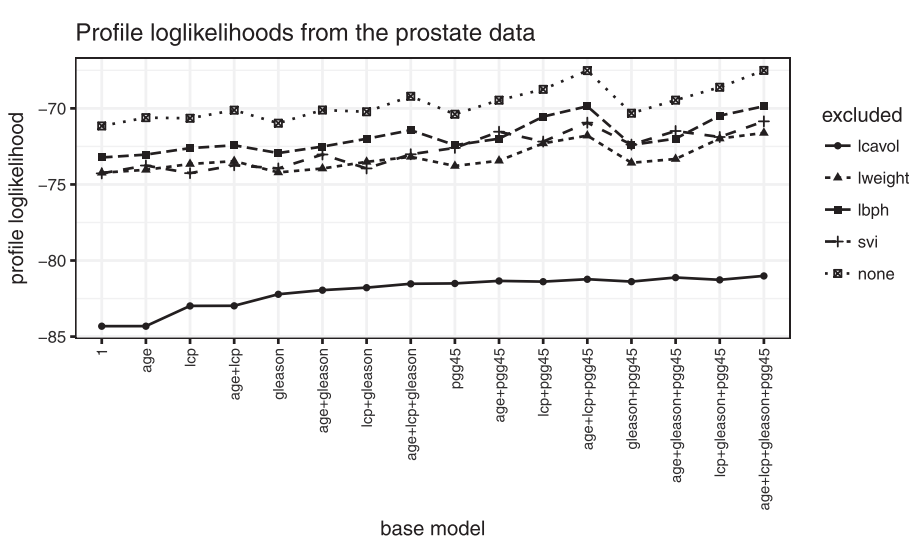


Figure 2. Profile log likelihoods of the 80 models in Example 1 that include at least three of lcavol, lweight, lbph, and svi. The base model indicates which other predictors are included. The five lines are for models that exclude either one or none of lcavol, lweight, lbph, and svi.

svi, respectively, and an apparently strong inference for including lweight and svi, at least if we don't correct for postselection inference. We think the inference for including lcavol is strong because it is consistent across many models and we did not find any good models that do not include lcavol. In contrast, we think the inference for including lweight, lbph, and svi is weak because we can find many good models that exclude them. Our criteria are about how well we can model the data with and without each of the covariates and differ from the usual criteria of p-values and effect size.

Though more could be done—for example, model diagnostics, removing more than one key variable, fitting nonlinear regressions—we will stop exploring the prostate data for now. Example 1 illustrates some themes of this article:

1. When exploring the question of whether a variable ought to be in a model, it is useful to see how much descriptive power is gained by including that variable. Example 1 used profile log-likelihood to make the comparison, but later examples will show the dangers of relying too heavily on likelihood.
2. It was useful to see how strongly the evidence for including a variable depends on which other variables were included. In Example 1 the dependence was weak, but that isn't always the case.
3. There are no true models and parameters. LPSA is not generated as an exact linear function of the predictors plus Normally distributed errors with constant variance. Just to be clear, even if we adopted a nonparametric model such as a spline, there would still be no true regression function and no true error distribution because the population to which we want to generalize is fuzzy and is changing through time and space. Because there are no true models and parameters, it is misleading to speak of estimating, finding confidence intervals for, or finding the posterior distribution of them and misleading to judge procedures according to how accurately they learn true values. But we can still judge models and parameter values according to how well they describe the data.
4. Analysis of the prostate data should *not* be framed as optimization. The data are not a perfectly drawn random sample from a well-defined population; the population of potential prostate patients is changing over time and space, as are the relationships among the variables; and the region of spacetime over which we want to generalize is fuzzy. For all these reasons, the parametric form is only an approximation to an imprecisely defined thing we want to study. Whatever target function we adopt and whatever its optimum may be, the answer may change if we adopt a slightly different formulation of the problem or collect data at a slightly different time in a slightly different way from a slightly different population. Therefore, we should be interested not only in the precise optimum of a precisely defined target function but should look for all (model, parameter) combinations at which our target function is large because those combinations are plausible descriptions of the data. They may even be optima under slightly different formulations.
5. Even though there is no true model and no true coefficients, it still makes sense to talk about how well (model, parameter) combinations describe the data.

Our point of view contrasts with that of many other statisticians. Here, for example, is how van der Laan (2015) explained an alternate view.

> "…it is not true that one must start with simplifying assumptions, as the field of statistics has the theoretical foundation that allows the precise formulation of the statistical estimation problem. …However, that is precisely what most statisticians blithely do, proudly referring to the quote, 'All models are wrong, but some are useful.' …
> "The consequences of giving up on formulating the actual statistical model are dramatic for our field, making statistics an art instead of a science. …if one gives up on the scientific standard that a statistical model has to be true, then any statistician can do whatever they want. …
> "It is complete nonsense to state that all models are wrong, so let's stop using that quote. For example, a statistical model that makes no assumptions is always true. …
> "…targeting the estimator toward the estimand so the resulting estimator is minimally biased and statistical inference is possible."

We disagree. It is not that we "give up on the scientific standard that a statistical model has to be true;" it's that in many problems, including Example 1, there is no true model, not even a nonparametric one. Because there is no true model, it is not possible to target the estimator toward the estimand so the resulting estimator is minimally biased: bias is not a relevant concept.

Instead of choosing a model, we prefer to keep many models in mind, along with the parameters of those models that provide good descriptions of the data. We agree with Lindley (1993) who said,

> "The ideal statistician holds all possible models in his head but the real statistician works with a subset of models, what Savage called a 'small world', the choice of which is governed by the data and other considerations."

and add that we try to keep our small world as large as possible for as long as possible. We turn next to another example where again there is no true model.

Example 2 (soil respiration). Soil respiration refers to the exchange of carbon between the soil and the atmosphere. This example draws on data analyzed by Giasson et al. (2013) of over 100,000 measurements of soil respiration in the Harvard Forest, in north-central Massachusetts. To quote from Giasson et al.:

> "[S]oil respiration (R_s) [is] the sum of belowground autotrophic (roots and associated mycorrhizae) and heterotrophic (mainly microbes, microfauna, and mesofauna) respiration. Estimates of global R_s range from 68 to 98 Gt C yr^{-1} [gigatons of carbon per year] (Raich and Schlesinger 1992; Schlesinger and Andrews 2000; Bond-Lamberty and Thomson 2010), or about two-thirds of all of the C emitted to the atmosphere by terrestrial ecosystems. The amount of C emitted through R_s is ~10 times more than that released through fossil fuel combustion and cement manufacturing (IPCC 2007; Peters et al. 2012), although, for the most part, R_s is closely coupled to a large photosynthetic uptake, leading to a much smaller net C exchange with the atmosphere (Schlesinger and Andrews 2000)."

Because R_s is generated by metabolic processes it is affected by soil temperature T_s, and much interest centers on the relationship between T_s and R_s. Figure 3 shows that modeling $\log(R_s)$ as

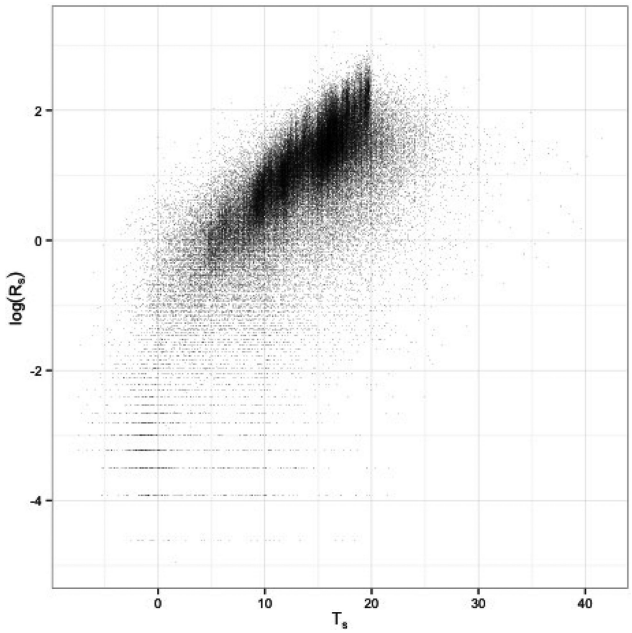


Figure 3. $\log(R_s)$ versus T_s in the Harvard Forest. Over 100,000 data points.

a roughly linear function of T_s is a good description of most of the Harvard Forest data, with the exception of a small fraction of data points with very low R_s or high T_s.

A good first description of the data is

Model A $\mathbb{E}[\log(R_s)]$ is approximately a linear function of T_s with slope about 0.127 ± 0.0004 or so and intercept -0.7809 ± 0.006 or so with Normal residuals whose SD is about 0.67 or so.

There are multiple enhancements of Model A that describe the data more accurately including the following.

Model B We could propose a nonlinear regression function such as, for example, a simple quadratic. Figure 3 doesn't suggest much curvature except for the points with low R_s or high T_s but the t-statistic for the quadratic term is over 100 and the p-value is minuscule. A nonparametric regression function would presumably fit even better than a quadratic.

Model C The relationship between R_s and T_s varies slightly by the type of site, that is, by whether the site is deciduous, hemlock, coniferous, swamp, bog, mixed, and so on, so we could propose a model in which different forest types have different slopes and intercepts, either random or fixed.

Model D Residuals from Model A are asymmetric, non-Normal, and vary by type of site, so we could propose a more accurate model for residuals.

Model E Giasson et al. (2013) further explain measurement of R_s:

> "R_s was measured in fixed locations on a given sampling day, generally where PVC or aluminum collars had been inserted and left in the soil …

There is much collar-to-collar variability in the response of R_s to T_s, and that was the main focus of Giasson et al. (2013). It is not our focus here, but we could propose a model in which parameters differ from collar to collar.

If we were writing a report to ecologists on Example 2 we would want our report to emphasize two ideas: (i) the bulk of the data, with some exceptions whose cause we don't yet understand, can be well-described by a linear relationship with a slope about 0.127 or so and a residual SD around 0.67 or so; and (ii) Models B through E describe the data modestly better than Model A. These ideas are *not* quantified by estimates, hypothesis tests, p-values, confidence intervals, or posterior probabilities. With over 100,000 data points, confidence and posterior intervals would presumably be small. That's inappropriate when there's no precise thing we're trying to estimate. There are no true models and no true coefficients in Example 2 for reasons similar to those of Example 1: the data are not a perfectly drawn random sample from a well-defined population; the population, such as it is, is changing over time and space; and the region of spacetime over which we want to generalize is fuzzy. We should not take a position on whether the relationship between $\log(R_s)$ and T_s is linear, quadratic, or nonparametric; we should present to the investigators multiple ways of describing the data. We should not select models; we should keep in mind many models and their relative strengths.

Even effect sizes are fundamentally imprecise. That is, there is no precise meaning to the effect of a deciduous rather than a hemlock forest. Deciduous forests differ from each other, as do hemlock forests. Within the Harvard Forest, different deciduous stands differ from each other. Even so, our data tell us something about forests in general and about the comparison between deciduous and hemlock forests. But the thing our data tells us is fuzzy and should not be accorded the apparent precision that 100,000 data points convey in our usual paradigms.

van der Laan (2015) said, "As a consequence [of giving up on the scientific standard that a statistical model has to be true], we will not be an intrinsic part of the scientific team, and the scientists will naturally have their doubts about the methods used, thus only accepting the answers we generate for them if it makes sense to them: …." We disagree. Rather, scientists *should* have doubts about methods that rely on there being a true statistical model, even if nonparametric, for this dataset. If we insist there is a true model, and if the answers it generates do not make sense to ecologists, then those answers should not be immediately accepted. They should be checked to see whether they lead to a new ecological understanding or whether they are an artifact of our insisting that there is a true model and that our algorithms are accurately estimating it.

The next example is another in which there is no true model. It illustrates one danger of taking the model and likelihood function too seriously.

Example 3 (ozone). Wang et al. (2012) reported,

> "Lower atmospheric ozone, a gaseous air pollutant, is toxic to many plants and threatens the health of forests across the United States. In the 1990s, the USDA Forest Service began a monitoring program to assess status and trends with respect to the presence of phytotoxic concentrations of ozone in our nation's forests. Ozone detection is based on the foliar (leaf) injury response of ozone-sensitive bioindicator species; a response that is conditioned primarily by the levels of ozone in the air and moisture in the soil. …
> "Ozone enters plants through stomata (pores) in the leaves. Once in the leaf, it damages the chloroplasts, leaving marks that are visible to the naked eye. …

> "The USDA Forest Service administers a nation-wide ozone biomonitoring program to address national concerns about ozone impacts on forest health (Smith, Coulston, and O'Connell 2008). Information on plant response to ozone in forested landscapes is collected on a network of 1,200 field sites in 46 states. …
> "There are five environmental covariates. Three — N100, W126, and SUM06 — are measures of ozone. Two — MI and PDSI — are measures of plant moisture stress. MI provides an indication of plant moisture deficit (precipitation input vs. transpiration output) while PDSI is a measure of soil moisture stress (i.e., drought severity). Both measures are important because ozone damages leaves by entering open stomata, and stomates are closed when plant moisture is limiting or the soil is too dry. …. Because W126 and SUM06 are nearly collinear, we omit W126 in what follows.
> "In addition to the environmental covariates, there are four spatial covariates — state, site, climate division, and ecological section — and year. …
> "To answer our questions, we fit multiple models, all of the form
>
> $$\text{logit}(\Pr[Y_i = 1]) = \text{various terms."}$$

Those models and two goodness of fit measures, AIC and BIC, are shown in Table 2. AIC, BIC, and their variants are information criteria of the form $-(\text{log-likelihood}) + k$ where k is a penalty. They are used to select the model with the smallest value.

Wang et al. (2012) presented a thorough comparison of the models covering considerations such as the ecological relevance of the predictors, the a priori plausibility of the models' fitted values, and the improvement in either AIC or BIC of one model over another. AIC and BIC are used as informal guides, not formal model selection criteria. Wang et al. summarized:

> " …familiar parametric regression curves for binary data (logistic, probit, complementary log-log) are not sufficiently flexible to model the data well, at least not without high-degree polynomials of the predictors. Some sort of semiparametric models are needed. We used splines in generalized additive models, but other choices might work equally well. …
> " …good models should include at least one ozone variable. Whether that variable is SUM06 or N100 makes just a small difference. …Similarly, good models should include at least one moisture variable.
> …
> "[Our analysis] also tells us something about our ability to predict ozone damage. Many fitted values are below 0.2 or 0.1 — and the data agree — which shows that there are some environmental conditions — for example low levels of ozone combined with low plant available moisture — under which we can be confident there will be little ozone damage. In contrast, there are no fitted values greater than about 0.7, which shows that we have not been able to identify environmental conditions under which we can reliably expect plants to be damaged."

The story Wang et al. thought it important to tell consists of three main points: use flexible models; include both ozone and moisture as predictors; we can predict when damage won't occur but not when it will. That may not be the whole story about this dataset, but telling this part of the story does not require selecting a model, estimating parameters, testing hypotheses, or calculating posterior probabilities. Instead, it relies on informal measures of goodness of fit to convey how well various models describe the data.

Wang et al. used AIC and BIC informally because they realized that information criteria can be sensitive to how details of model specifications interact with data. As Wang et al. note, the log-likelihood, being a sum, can be sensitive to outliers, so it pays to look at each y_i's contribution to see whether a small number of data points have large influence in model comparison. It is a familiar point that sums and averages can be sensitive to outliers but one that is often ignored in model selection and comparison, especially by people who treat statistical analysis as an optimization problem. Wang et al. explore that point in the context of their models. Rather than repeat their analysis, we present a simpler example that illustrates the same point.

Example 4 (logistic regression). Our example will be a logistic regression. We (1) generated 1000 points with $X \in [-20, 20]$ from a model in which $\Pr[y_i = 1] = \exp(x_i)/(\exp(x_i) + 1)$, (2) added one additional point $(x, y) = (-20, 1)$, and (3) compared a logistic regression in which X enters linearly to one in which

Table 2. Models and measures of fit as calculated by Wang et al. (2012) in Example 2.

		Model structures			
Type	Model no.	Fixed effects	Random effects	AIC	BIC
Logistic	1	MI × SUM06		7086	7112
	2	MI × SUM06 + N100 + PDSI		7069	7109
GAM	3	s(MI) + s(SUM06)		6946	7073
	4	s(MI) + s(SUM06, k = 5)		6950	7043
	5	s(MI) + s(SUM06, k = 5) + s(N100) + s(PDSI)		6884	7098
	6	s(MI, by = REG) + s(SUM06, k = 5, by = REG)		6936	7116
Mixed	7	s(MI) + s(SUM06, k = 5)	(1|CLM_DIV)	6864	6904
	8	s(MI) + s(SUM06, k = 5)	(1|SITE)	6626	6666
	9	s(MI) + s(SUM06, k = 5)	(1|STATE)	6311	6351
	10	s(MI) + s(SUM06, k = 5)		6310	6350
	11	s(MI) + s(N100)		6306	6346
	12	s(PDSI) + s(SUM06, k = 5))	(1|ECO_SEC)	6359	6399
	13	s(PDSI) + s(N100)		6355	6395
	14	s(MI) + s(SUM06, k = 5) + s(PDSI)		6295	6348
	15	s(MI) + s(SUM06, k = 5) + s(N100) + s(PDSI)		6290	6357

SUM06 is the "sum of all hourly average ozone concentrations ≥ 60 ppb;" N100 is the "number of hours with O_3 concentrations at or above 100 ppb;" REG refers to four regional units, Northern, Southern, Interior West, and Pacific Northwest; CLM_DIV is a categorical variable indicating areas of similar temperature and precipitation; SITE identifies an ozone biosite; STATE identifies a US state; and ECO_SEC is a subdivision of a larger domain of similar climate based on vegetation, terrain features, and elevation. $s(\cdot)$ indicates a nonparametric smoother, in this case a spline, of the enclosed term.

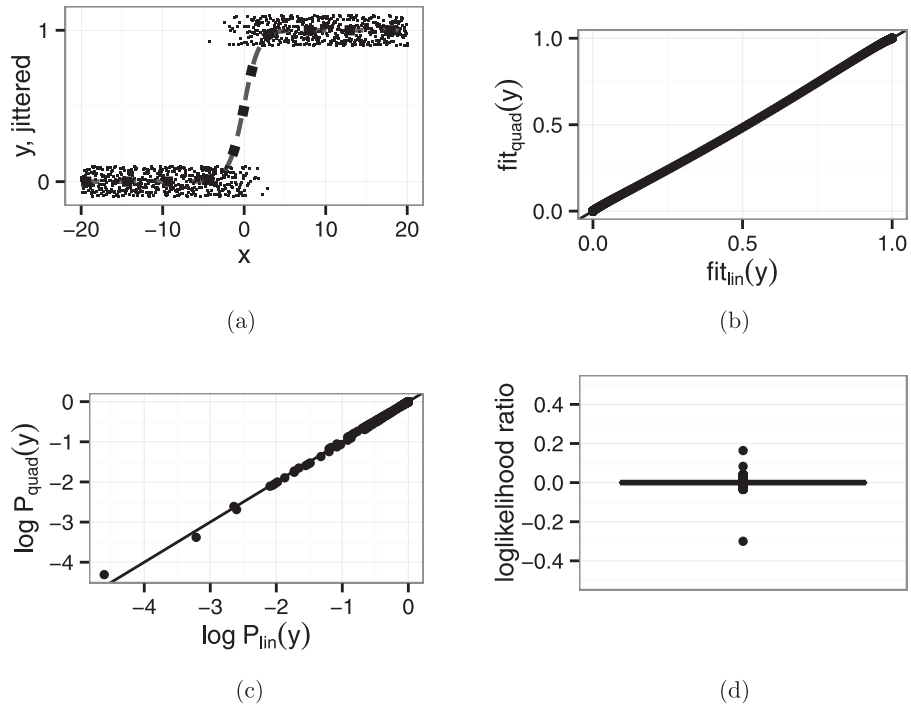


Figure 4. Example 4. Data generated from linear logistic model. Comparison of linear and quadratic logistic regression. (a) Jittered y_i vs. x_i. Small points are data. Dashed curve is the linear logistic fit. Dotted curve is the quadratic logistic fit. (b) Points are fitted values from quadratic and linear logistic regression. Line is $y = x$. (c) $\log P_{\hat{\theta}}(y_i)$ from quadratic and linear logistic regression. Line is $y = x$. (d) Boxplot of log-likelihood ratio $\log \frac{P_{\hat{\theta}_{\text{lin}}}(y_i)}{P_{\hat{\theta}_{\text{quad}}}(y_i)}$

it enters quadratically to see how the additional point affects the comparison.

Figure 4 depicts the model comparison when the data do not include the additional point. Figure 4(a) plots the 1000 original points as small dots, jittered vertically for legibility. The dashed curve is the fit when X enters linearly and the large dotted curve is the fit when X enters quadratically. Figure 4(b) plots the fitted values from the quadratic model against those from the linear model. The fits are nearly identical, as they should be because the smaller model is true.

4c plots $\log \Pr_{\hat{\theta}_{\text{quad}}}(y_i)$ vs. $\log \Pr_{\hat{\theta}_{\text{lin}}}(y_i)$, the individual points' log likelihoods, and 4d is a boxplot of the points' contribution to the log likelihood ratio. The contributions are all less than ± 0.4 and no point dominates the sum. R tells us that AIC for the linear and quadratic fits are 156.44 and 158.08, respectively; the linear fit wins by a little bit.

In contrast, Figure 5 compares the two models when the additional point $(-20, 1)$ is included. Figures 5(a) and 5(b) show that the fitted values of the two models are still nearly identical. Even so, Figure 5(c) shows that the new point's probability is much less, on the log scale, under the linear model than under the quadratic model. The boxplot in 5d shows that the new point's contribution to the log-likelihood is close to -15, much larger than any other point's. The AIC values are 194.64 for the linear fit and 167.99 for the quadratic. The quadratic fit wins by a lot.

The explanation for the large influence of the added point in Example 4 can be seen in the lower left corner of Figure 5(b) but perhaps even more clearly in Figure 6, which shows the fitted values from the quadratic model plotted against the fitted values from the linear model separately for $y_i = 0$ and $y_i = 1$. (Wang et al. 2012, have a similar plot.) The new point has a fitted value very close to zero under both models, but its fitted value is slightly larger under the quadratic model so its point in Figure 5(b) is slightly above the line. When $y_{\text{new}} = 1$ both models are surprised. But though the quadratic model's fitted value is only slightly larger, its surprise, as measured by log-likelihood, is much less. Any point near the lower left corner of the $y_i = 1$ panel or the upper-right corner of the $y_i = 0$ panel has potential for high influence in model comparison because it can yield a large likelihood ratio, which is an artifact of treating the model too seriously.

This article was prompted by the author's observation that data description receives too little attention in statistical research, application, and teaching. We tried to demonstrate through examples that quality of description is a useful way to report at least some aspects of statistical analyses. If the description is explanatory, in the sense that it fits with plausible mechanisms in physics, medicine, sociology, or whatever, so much the better, but good models should describe the data well. We agree with Lindley (1993), who said, "In the earlier stages … consideration of models can be done informally by inspection of the data. Only when the issues become more delicate is there a need for a formal system …." When we explore data our usual goal is to find many models and parameters that describe the data and report those to the investigator, whose response may influence which formal system is applied.

Our primary criterion for good description is likelihood, but likelihood can sometimes mislead. One problem occurs when there are outliers, so we may make reports such as, "a linear logistic model describes most of the data well, but there is an outlying point" (Example 4) or "A linear model describes most of the data well but there are exceptions at the extremes"

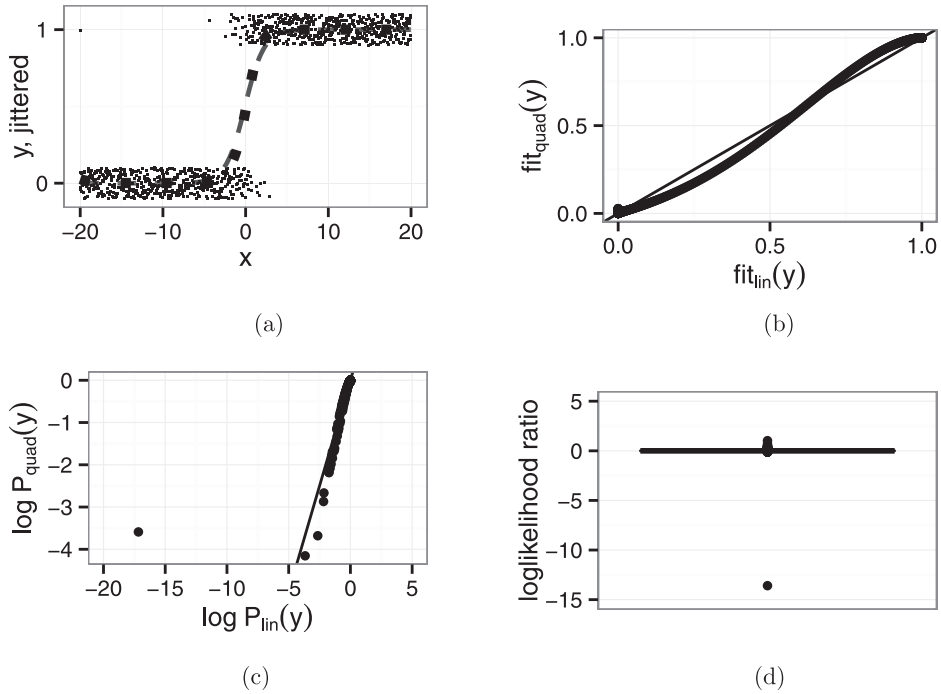


Figure 5. Example 4. Data generated from linear logistic model with one additional point $(-20, 1)$. Comparison of linear and quadratic logistic regression. (a) Jittered y_i vs. x_i. Small points are data. Dashed curve is the linear logistic fit. Dotted curve is the quadratic logistic fit. (b) Points are fitted values from quadratic and linear logistic regression. Line is $y = x$. (c) $\log P_{\hat{\theta}}(y_i)$ from quadratic and linear logistic regression. The additional point is the one near $(-17, -3.5)$. Line is $y = x$. (d) Boxplot of log-likelihood ratio $\log \frac{P_{\hat{\theta}_{lin}}(y_i)}{P_{\hat{\theta}_{quad}}(y_i)}$.

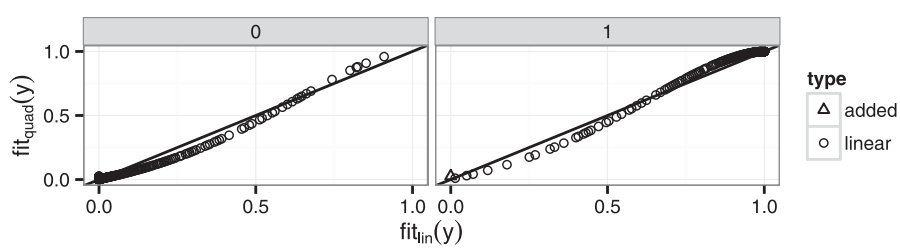


Figure 6. Fitted values from the quadratic model vs. fitted values from the linear model in Example 4. Left panel: $y = 0$; right panel: $y = 1$. Circles: points generated from the linear model. Triangle: the added point.

(Example 2). Another problem occurs because, when the sample size is large, one model or set of parameters may have a much larger likelihood than another, even though the two models are nearly identical, similarly to the idea that statistical significance does not equal practical significance.

To summarize, we think statistics education and practice should pay more attention to data description, even if that comes at the expense of formality and algorithms.

References

Bond-Lamberty, B., and Thomson, A. (2010), "Temperature-Associated Increases in the Global Soil Respiration Record," *Nature*, 464, 579–582. [314]

Giasson, M.-A., et al. (2013), "Soil Respiration in a Northeastern US Temperate Forest: A 22-Year Synthesis," *Ecosphere*, 4.11. [314]

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer Series in Statistics, New York: Springer. [312]

IPCC (2007), "Changes in Atmospheric Constituents and in Radiative Forcing," *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC*, eds. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. TIgnor, and H. K. Miller. Cambridge University Press, chap. 2, section 2.3.1. [314]

Lindley, D. V. (1993), "Discussion of Exchangeability and Data Analysis," (with discussion), *Journal of the Royal Statistical Society*, Series A, 156, 29–30. [314,317]

Peters, G., et al. (2012), "CO2 Emissions Rebound After the Global Financial Crisis," in *Nature Climate Change*, 2, 2–4. [314]

Raich, J. W., and Schlesinger, W. H. (1992), "The Global Carbon Dioxide Flux in Soil Respiration and Its Relationship to Vegetation and Climate," in *Tellus B*, 44.2, 81–99. [314]

Schlesinger, W. H., and Andrews, J. A. (2000), "Soil Respiration and the Global Carbon Cycle," *English Biogeochemistry*, 48, 7–20. [314]

Smith, G. C., Coulston, J. W., and O'Connell, B. M. (2008), "Ozone Bioindicators and Forest Health: A Guide to the Evaluation, Analysis, and Interpretation of the Ozone Injury Data in the Forest Inventory and Analysis Program," Tech. Rep. General Technical Report NRS-34, United States Department of Agriculture. [316]

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989), "Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate II Radical Prostatectomy Treated Patients," *Journal of Urology*, 16, 1076–1083. [312]

van der Laan, M. J. (2015), "Statistics as a Science, Not an Art: The Way to Survive in Data Science," *Amstat News*, 452, 29–30. [314,315]

Wang, P., Baines, A., Lavine, M., and Smith, G. (2012), "Modelling Ozone Injury to U.S. Forests," *Environmental and Ecological Statistics*, 19, 461–472. [316,317]

The American Statistician

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

Inference and Decision Making for 21st-Century Drug Development and Approval

Stephen J. Ruberg, Frank E. Harrell Jr., Margaret Gamalo-Siebers, Lisa LaVange, J. Jack Lee, Karen Price & Carl Peck

To cite this article: Stephen J. Ruberg, Frank E. Harrell Jr., Margaret Gamalo-Siebers, Lisa LaVange, J. Jack Lee, Karen Price & Carl Peck (2019) Inference and Decision Making for 21st-Century Drug Development and Approval, The American Statistician, 73:sup1, 319-327, DOI: 10.1080/00031305.2019.1566091

To link to this article: https://doi.org/10.1080/00031305.2019.1566091

© 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: 7576

View related articles

View Crossmark data

Citing articles: 16 View citing articles

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, 319–327: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1566091

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Inference and Decision Making for 21st-Century Drug Development and Approval

Stephen J. Ruberg[a], Frank E. Harrell, Jr.[b], Margaret Gamalo-Siebers[c], Lisa LaVange[d], J. Jack Lee[e], Karen Price[c], and Carl Peck[f]

[a]Analytix Thinking, LLC, Indianapolis, IN; [b]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN; [c]Lilly Corporate Center, Eli Lilly & Company, Indianapolis, IN; [d]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC; [e]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX; [f]NDA Partners, LLC, San Luis Obispo, CA

ABSTRACT

The cost and time of pharmaceutical drug development continue to grow at rates that many say are unsustainable. These trends have enormous impact on what treatments get to patients, when they get them and how they are used. The statistical framework for supporting decisions in regulated clinical development of new medicines has followed a traditional path of frequentist methodology. Trials using hypothesis tests of "no treatment effect" are done routinely, and the p-value < 0.05 is often the determinant of what constitutes a "successful" trial. Many drugs fail in clinical development, adding to the cost of new medicines, and some evidence points blame at the deficiencies of the frequentist paradigm. An unknown number effective medicines may have been abandoned because trials were declared "unsuccessful" due to a p-value exceeding 0.05. Recently, the Bayesian paradigm has shown utility in the clinical drug development process for its probability-based inference. We argue for a Bayesian approach that employs data from other trials as a "prior" for Phase 3 trials so that synthesized evidence across trials can be utilized to compute probability statements that are valuable for understanding the magnitude of treatment effect. Such a Bayesian paradigm provides a promising framework for improving statistical inference and regulatory decision making.

ARTICLE HISTORY
Received March 2018
Accepted December 2018

KEYWORDS
Bayesian statistics; Clinical trial; Prior and posterior probability; p-value; Statistical inference; Substantial evidence

1. Introduction

There has been nearly uniform acceptance of $p < 0.05$ in most fields of research since R.A. Fisher first made his statements about what he thought might be considered "statistically significant" evidence (Fisher 1926). Following the passage of the Kefauver-Harris Amendment to the Food, Drug and Cosmetic Act of 1962, drug manufacturers have been required to demonstrate "evidence [of effectiveness] consisting of adequate and well-controlled investigations, including clinical investigations … to evaluate the effectiveness of the drug involved" (Drug Amendment Act 1962). With the passage of the 1962 legislation, there were Congressional hearings and debates about the need for "adequate and well-controlled" trials, including topics such as randomization, whether placebo-controlled trials were ethical, and why investigators should be blinded to the study medication. During the years following the passage of the Act, FDA promulgated regulations that defined and set standards for these trial characteristics, while hiring statisticians to assess the adequacy of the evidence. The use of the null hypothesis significance testing (NHST) approach to compare treatments was adopted according to the prevailing customs in the canon of statistical methods.

In regulated drug development, the worldwide approach is to conduct a sequence of clinical studies, first establishing the safety of a new drug in small trials (Phase 1 trials). This is followed by treatment of larger groups of patients with various doses of the treatment under various durations and conditions to learn what might be most useful for a disease/condition of interest, referred to as Phase 2. Once there is an understanding of the right patient population(s) to treat and the doses that are safe and efficacious, clinical development moves into large-scale confirmatory trials (Phase 3) with the intent being to demonstrate clearly that the treatment is safe and effective when used as intended. Because of the Kefauver-Harris Amendment and its reference to "clinical investigations" (note the plural term), it generally is expected that drug sponsors will conduct at least two independent Phase 3 trials. Each trial is evaluated independently and declared a success or failure based on whether the trial achieves a statistically significant result (i.e., $p < 0.05$ in most cases). Both trials must be "successful" for a drug to be considered by a regulatory agency for review and approval. While the above is a general framework for drug development that helps put this work into context, there are a wide variety of drug development programs that may deviate from the Phase 1-2-3 paradigm, although the standard of evidence ($p < 0.05$ in two adequate and well-controlled trials) remains in place most often.

The adoption of a p-value < 0.05 for concluding "success" was also done as a convention and a matter of convenience in the paper and pencil era of the late 1960s and 1970s. This convenience was first noted by Fisher himself in 1925 when he

CONTACT Stephen J. Ruberg AnalytixThinking@gmail.com Analytix Thinking, LLC, Indianapolis, IN.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.
© 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.

wrote, "The value for which p=0.05 is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not" (Fisher 1925, p. 45). However, there is nothing that any of the authors of this report can find, after diligent review of Congressional hearings and documents or the FDA archives, that discusses or debates the level of evidence needed for drug approval or the standard by which that evidence is judged. Thus, the acceptance of fundamental concepts for what constitutes adequate, well-controlled investigations—use of placebo, blinding and randomization—that we take for granted today, were heavily debated 50 years ago. While these trial design characteristics are accepted today as the gold standard for generating credible evidence, the acceptable level of statistical evidence ($p < 0.05$) was, AND STILL IS, just a matter of tradition and convenience.

Using a decision criterion of $p < 0.05$, NHST ignores available evidence from sources external to each individual trial as well as the clinical context in which a decision is made. However, is it always reasonable to conclude that a treatment does not work after obtaining, for example, a p-value of 0.01 in one Phase 3 trial and a p-value of 0.08 in a second Phase 3 trial of the same treatment, that is, one "successful" trial and one "unsuccessful" trial? Furthermore, should the interpretation of two such trials—one success and one failure—be the same regardless of whether the trials are for a chronic, debilitating disease (e.g., rheumatoid arthritis) or a transient treatment of symptoms (e.g., dental pain)? In contrast to inferring whether a treatment works or not using NHST (and requiring $p < 0.05$ in two independent trials), imagine a world in which the approval of a new treatment—drug or biologic—is truly based on the probability that the treatment works according to the totality of the evidence and the clinical context. "That the treatment works" is subject to judgment, as is the probability level that is needed for approval. That takes some very hard thinking and lively discussion. But we should not shy away from this diligence and hide behind convenience. We posit that patients would accept a treatment that has a 50% probability of curing pancreatic cancer, while most might prefer a 95% probability that a treatment produces a 10 kg weight loss in one year. Those statements are much better suited for describing a treatment effect in a Bayesian paradigm. Moreover, the Bayesian paradigm seems ideally suited for synthesizing evidence from not only two Phase 3 trials but also the other trials conducted over many years in the drug development process, as well as other scientifically relevant evidence.

This paper sets the stage for Bayesian thinking for regulated drug development in Section 2. Section 3 provides a real and recent industry example with some specific details of how the Bayesian paradigm could be implemented when analyzing and interpreting Phase 3 clinical trials. Current uses of Bayesian approaches within FDA, as well as possible extensions of the Bayesian paradigm for regulatory decision making, are covered in Section 4. We end with further discussion in Section 5.

2. Bayesian Thinking for Drug Development

The debates about the suitability of a frequentist paradigm versus the Bayesian paradigm have been long-standing and, at times, acrimonious. There is abundant literature comparing and contrasting both approaches (Bayarri and Berger 2004; Christensen 2005; Inoue et al. 2005). Goodman (1999) provides an excellent description of "the p-value fallacy," and posits that the over-reliance on p-values "has almost eliminated our ability to distinguish between statistical results and scientific conclusions." This could certainly be a criticism of current regulatory practice in which a p-value < 0.05 in pivotal trials is almost always a requirement for regulatory consideration, let alone approval, despite other supportive clinical trial evidence from earlier phases of drug development.

The Bayesian way of thinking and formal analytical approach seems ideally suited for the drug development process. The learn and confirm paradigm has been espoused in drug development (Sheiner 1997). The idea is to learn as much as one can about a new treatment, followed by well informed, confirmatory trials. This approach acknowledges that clinical drug development proceeds over many years. It is a continual sequence of learning studies and confirming, using ever larger samples sizes and different populations of subjects, gradually accumulating knowledge about the efficacy and safety of a novel treatment. Since this development proceeds over many years, additional scientific knowledge is also accumulating during that time—knowledge of the disease state, knowledge of molecular biology surrounding the mechanism of action of the treatment, genomics and additional data from animal experiments or other companies developing similar treatments. The advantage of the Bayesian approach is that it can formally incorporate prior data or external information into the analysis of the confirmatory trials.

It is important to note that this is not an attempt to "lower the bar" for substantial evidence of a treatment effect and may in some instances increase the evidentiary barrier to approval. For regulatory decision making, Bayesian analytical approaches can be rigorously prespecified to provide a formal mechanism utilizing information that comes from outside a single Phase 3 trial or can be used to synthesize information across Phase 3 trials. Information external to the Phase 3 study can be incorporated into a vague/diffuse prior or an informative prior, depending on the strength of evidence from such information related to the treatment effect. Such priors would need agreement and formal documentation with regulators in advance of Phase 3 trials. This is no different than the current practice for clinical trial analysis using frequentist approaches for which models and assumptions are agreed in advance. It is worth noting, however, that in the case of some difficult to treat diseases, such as Alzheimer's Disease or sepsis, the history of failed clinical trials could result in a more skeptical prior that may lead to a more rigorous standard being applied to determine trial success.

The focus of this article is on the evaluation of efficacy at the completion of two Phase 3 trials. However, Bayesian approaches also seem ideally suited for use in interim analysis of clinical trials where data are accumulating within a clinical trial and decisions are made whether to continue or terminate that trial based on the data/evidence accumulated at that time. Also, safety considerations and analysis of adverse events plays a major role in drug development and regulatory decisions, but often is not the subject of rigorous hypothesis testing. The evaluation of adverse effects of a treatment seem amenable to

Bayesian approaches, especially for less frequent adverse events. In any single clinical trial, the sample size may be insufficient to elucidate a treatment effect on an adverse outcome. Nevertheless, the accumulation of safety data across all phases of drug development can be used in formal Bayesian approaches to assess the probability that any particular adverse outcome is the result of an experimental treatment as well as the probability associated with any magnitude of that effect.

In summary, the Bayesian approach makes direct probability statements about the treatment effect itself, given the observed data. The inferential framework is more relevant to information needs of patients and prescribers, and in fact, all stakeholders of our healthcare system.

3. Industry Example

The frequentist paradigm is engrained in our scientific culture and implementation of regulated clinical research for new treatments. A Bayesian perspective in this realm is novel, and the following example illustrates how the drug development process and decision making could be modified to take this approach. The example is taken from real and recent experience in a pharmaceutical company. The sponsor asked that additional details about the drug and disease not be shared since there are proprietary elements of this example. We do not believe that withholding such information detracts substantially from the key points that we are seeking to illustrate.

Because the Bayesian approach requires the definition of a prior, its application should be considered upon initiation of the first efficacy clinical trials—generally before Phase 2. However, for simplicity and to illustrate how the Bayesian approach could replace NHST for confirmatory Phase 3 trials, the following example focuses on the analysis and interpretation of two Phase 3 trials, the priors for which are described below. While much of the methodology presented here is described in Spiegelhalter et al. (2004), this example illustrates how the Bayesian approach can lead to better understanding and interpretation of clinical trial results that are considered conflicting or confusing under the frequentist paradigm. In that sense, the approach described herein serves as an informative contribution to the comparison of frequentist and Bayesian paradigms in a practical application with important consequences in drug development and approval.

The prior distribution of effect size for Phase 3 can be quantified from a variety of sources, which is one of the benefits of the Bayesian approach. Phase 2 results of the new treatment provide some direct evidence of its effect. In addition to randomized Phase 2 trials that include a comparator, either a placebo or an active treatment, qualified data from other trials outside the clinical development program of the new treatment can also be used. For simplicity, we will focus on placebo-controlled trials, although the concepts can be easily extended to trials with active comparators. Bayesian network meta-analysis (Greco et al. 2016) can be employed to synthesize data into a single posterior distribution of effect for both the new treatment as well as the placebo. Care must be taken to account for differences in study populations, duration of treatment, study design and other relevant factors when doing such meta-analysis.

The scientific context for this example is that a treatment from a competitor sponsor was approved recently by the FDA and on the market. That treatment had a very similar mechanism of action and was approved for the treatment of a disease and population of patients similar to what was intended by the sponsor for the experimental treatment in this example. The sponsor wanted to leverage data from the competitor treatment and scientific knowledge that was emerging in this disease area to skip Phase 2 and proceed with two Phase 3 trials.

When there is scant or no clinical data on the experimental treatment of interest, one can consider using appropriately discounted data from other treatments in the same class (e.g., treatments that have a similar mechanism of action, such as inhibition of the same/similar enzyme or binding to the same/similar cell surface receptor). There are also well-developed methods for "prior elicitation" (O'Hagan et al. 2006; Morris et al. 2014) in which experts can use animal data, disease state knowledge and other biological or medical knowledge to express their beliefs about the treatment effect in the form of a prior distribution of effect size. Finally, there are historical evaluations of a large number of new treatments and the success rates of such treatments in different disease states (Hay et al. 2014) that can provide insight into the development of a prior distribution.

In this example, the primary response variable for this disease is the percentage of patients responding with a prespecified and clinically meaningful degree of improvement across a variety of disease symptoms. As noted previously, there were no Phase 2 studies of the experimental treatment, but two Phase 3 studies were available of an FDA approved and marketed competitor treatment with the same mechanism of action in the same disease state and patient population. There were also many studies in this disease state of other treatments compared to placebo. Using a Bayesian network meta-analysis, the posterior distribution of responses from those studies served as a foundation for development of a prior distribution of treatment responses for the Phase 3 studies of the experimental treatment versus placebo, as well as a very good estimate for the prior distribution of the placebo response (Figure 1). Because placebo has been used in many studies, one can see that its posterior distribution of response has a much sharper peak (smaller variance) than the competitor treatment, which included only two studies.

Credible prior distribution of response for the Phase 3 studies cannot, in general, be taken directly from the posterior distributions derived in Figure 1. For example, the only experimental treatments that progress to Phase 3 studies are those that show a substantial effect in Phase 2 (i.e., statistically significant or medically meaningful). There is well-documented regression to the mean in which estimates of the treatment effect in Phase 3 studies tend to be smaller than those observed in Phase 2 studies of the same compound (Chuang-Stein and Kirby 2014). Also, there may be other causes for differences between the studies used for the posterior distribution described previously and the planned Phase 3 study of the experimental treatment that are not easily measured or defined. They include, but are not limited to, changes in medical care/practice that occur over time, differences in investigative sites chosen for the study (including a potentially different mix of countries in global Phase 3 trials), changes in patient inclusion/exclusion criteria or the precise definition of the primary endpoint.

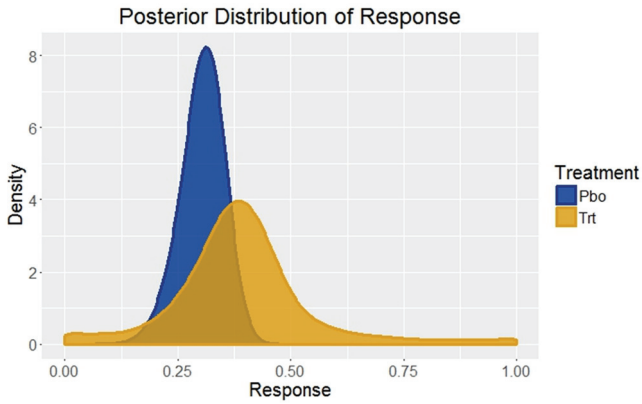


Figure 1. Distributions of treatment responses for placebo and an FDA approved and marketed competitor treatment based on systematic literature review and Bayesian network meta-analysis. For placebo (Pbo), the response rate is 30% with 95% credible interval (20%, 39%); for the competitor treatment (Trt) the response rate is 39% with 95% credible interval (7%, 82%). These are the starting points for developing the priors to be used for the Phase 3 trials of placebo versus the experimental treatment.

Cumulatively, the reasons given above make it advisable to discount the derived posterior distribution of response (Figure 1) when creating the prior distribution of response for the planned Phase 3 trail, denoted $\pi'(m_E)$. This can be accomplished easily with two different approaches. The first is to use a robust prior (Berger 1990) in which a mixture is used combining the posterior distributions derived from the data, $\pi(\mathrm{m}_{E,D})$, and a more skeptical or diffuse prior, $\pi(\mathrm{m}_{E,S})$. The latter prior of the mixture is chosen so that it is centered closer to the null hypothesis of no effect (skeptical) and has greater uncertainty or variability (diffuse). This is done to counter-balance the possibility of using a prior that is too optimistic, that is, too closely dependent on historical data. This can be written as

$$\pi'(m_{\mathrm{E}}) = \varepsilon\pi\left(m_{\mathrm{E,D}}\right) + (1-\varepsilon)\pi\left(m_{\mathrm{E,S}}\right) \tag{1}$$

where ε is the borrowing factor or mixing proportion. Another discounting approach is to use a power prior (Ibrahim and Chen 2000)

$$\pi'(m_{\mathrm{E}}) = \pi\left(m_{\mathrm{E,D}}\right)^{\varepsilon} \tag{2}$$

where $\varepsilon \leq 1$.

The choice of ε is based on the degree to which the Phase 3 study is similar to the other studies that were used to create $\pi(m_{\mathrm{E,D}})$, which is subjective. While the subjectivity may be considered a drawback to the Bayesian approach, the construction of the Phase 3 prior, including the choice of ε , can be debated and adjudicated by agreement with regulatory agencies prior to the commencement of the Phase 3 trials. For this analysis, ε was chosen to be 0.4 since there were many similarities between the studies of the competitor treatment and the experimental treatment, but no direct data on the experimental treatment. While analyses were done using both borrowing approaches, we present the results using the robust prior approach (Equation 1) since the results did not differ much between the two methods.

Lastly, the prior effective sample size for placebo and the competitor treatment are approximately 83 and 6, respectively. This is obtained by approximating the posterior distribution of the response rate of placebo and a competitor treatment by a beta distribution. The sum of the shape parameters of the beta distribution is used as the approximate effective sample size for the prior. The effective sample size can be used as an indicator of the "informativeness" of the prior, and the larger value for the placebo treatment shows that there is considerably more information to be gained using the prior for the placebo than using the prior for the competitor treatment (Morita et al. 2008).

Table 1. Results from two randomized, controlled, double-blind trials of two dose levels (every 4-week dosing—Q4W—and every 2-week dosing—Q2W) of an experimental treatment versus placebo. *N* = sample size; Responders were defined prospectively based on improvement in a variety of clinical symptoms. *P*-values were derived from logistic regression, taking prespecified covariates into account.

	Study 1			Study 2		
Measure	Placebo	Q4W	Q2W	Placebo	Q4W	Q2W
N	379	378	371	376	376	372
Responders (%)	29.3	35.2	31.8	27.7	34.8	38.4
p-value		0.052	0.490		0.051	0.002

In this example, there were two Phase 3 studies of the experimental treatment that included two different dose levels based on frequency of dosing—every 2-week dosing (Q2W) and every 4-week dosing (Q4W). The Q2W dose will be referred to as the high dose since twice as much of the experimental treatment is given in each 4-week period of treatment. The summary results are presented in Table 1. The placebo (PBO) responses for the two Studies were consistent with historical data that were used for creating the Phase 3 prior. The Q4W dose (lower dose) of the experimental treatment behaved consistently across trials and was very close to traditional statistical significance at the 0.05 level. The Q2W dose (higher dose) performed erratically; in Study 1, Q2W was slightly better than placebo and slightly worse than Q4W (i.e., not consistent with a dose-response relationship), while in Study 2, Q2W outperformed placebo and Q4W, exhibiting a dose–response relationship. Also, Q2W of this experimental treatment performed similarly in Study 2 to the synthesized effect size derived from the Bayesian network meta-analysis of the competitor treatment (38.4% versus 39%, respectively), as was anticipated prior to the initiation of the Phase 3 program.

In the frequentist paradigm, such results present a conundrum. Strictly speaking, Study 1 did not achieve statistical significance and, therefore, is a failed study. Study 2 did achieve statistical significance for Q2W only, and therefore, was a successful study since the objective of the study was to demonstrate that at least one dose was significantly different from placebo. With the regulatory requirement to show replication in two adequate and well-controlled trials, the Phase 3 program, and thus, the experimental treatment, is technically a failure. Yet, there are very similar results for the Q4W effect across the two trials, and the *p*-values are so close to 0.05 that it is difficult to argue that the treatment has no effect.

Using the Bayesian paradigm with the prior distribution of treatment effect as described previously (Figure 1) with borrowing parameter $\varepsilon = 0.4$ (Equation 1), the posterior distributions of treatment effect are shown in Figure 2a (Study 1) and Figure 2b (Study 2). From these distributions based on the Phase 3 data, the probability that there is a positive treatment effect can be computed easily, and the results are given in Table 2. At present,

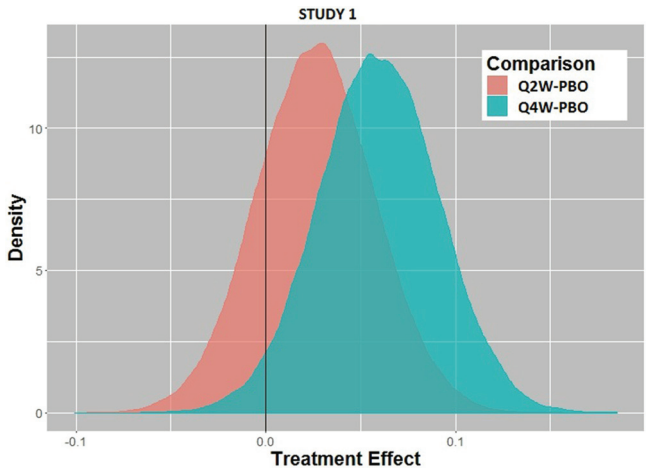


Figure 2a. Posterior distributions of treatment effect for every 2-week dosing (Q2W) versus placebo (PBO) and every 4-week dosing (Q4W) versus placebo for Study 1.

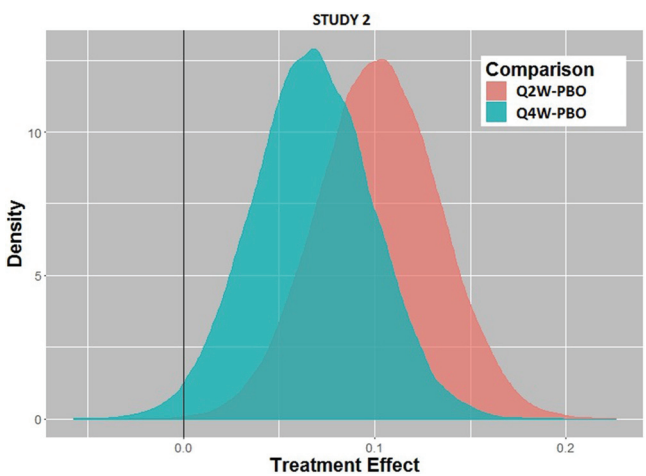


Figure 2b. Posterior distributions of treatment effect for every 2-week dosing (Q2W) versus placebo (PBO) and every 4-week dosing (Q4W) versus placebo for Study 2.

Table 2. Posterior probabilities of a positive treatment effect calculated from the posterior distributions of effect size for Study 1 and Study 2 using the prespecified borrowing factor ($\varepsilon = 0.4$) and sensitivity analysis using smaller borrowing factors ($\varepsilon = 0.1, 0.2, 0.3$).

	Study 1		Study 2	
Borrowing Factor	Q4W versus Placebo	Q2W versus Placebo	Q4W versus Placebo	Q2W versus Placebo
0.4	0.967	0.805	0.986	0.999
0.3	0.969	0.803	0.987	0.999
0.2	0.972	0.807	0.983	0.999
0.1	0.970	0.804	0.983	1.000

there are no conventions related to acceptable probability levels for a treatment effect to be considered a success, which could and should vary depending on the disease state severity or unmet medical need. Nonetheless, with probability levels exceeding 0.98, one could safely conclude that Study 2 shows a clear treatment effect at both dose levels, while Study 1 gives an inconsistent result between the two dose levels, especially since the higher dose (Q2W) has a lower probability of a treatment effect than the lower dose (Q4W) (0.805 versus 0.967,

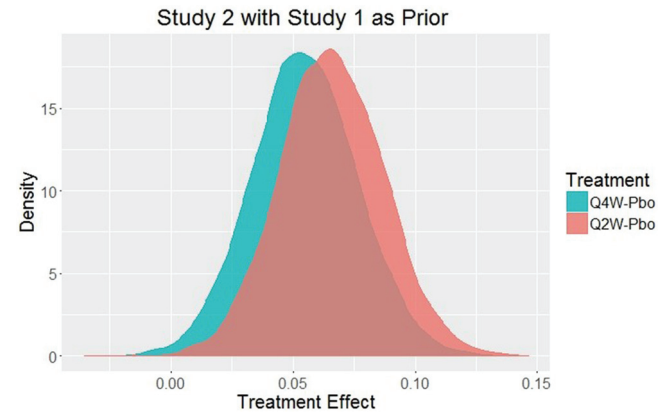


Figure 3. Posterior distributions of treatment effect for every 2-week dosing (Q2W) and every 4-week dosing (Q4W) for Study 2 using the posterior distribution of treatment effect from Study 1 as the prior distribution for Study 2 with borrowing factor ε =0.9.

respectively). Also, in Table 2 are results using smaller borrowing factors (i.e., relying less on historical data as a prior) for a sensitivity analysis. The results are unchanged, indicating that the prespecified borrowing factor is not a substantial determinant of the outcome.

In our quest to answer the fundamental question "What is the probability that the experimental treatment has a response different from the placebo response?" it is natural to consider the synthesis of data from both Phase 3 trials. This overall synthesis of the prior information and all Phase 3 data is quite natural in the Bayesian approach and allows for a single evaluation of the Phase 3 program and interpretation of the treatment effect that can be used for a regulatory decision. In this case, the posterior distribution from Study 1 could be used as the prior distribution for Study 2 to obtain an overall assessment of the treatment effect. If a borrowing factor of 0.9 is used for that prior distribution (consistent with the fact that Study 2 was identical to Study 1, but had a slightly different set of geographic sites), then the posterior distribution of treatment effect for each dose is shown in Figure 3.

For Q4W versus placebo, the probability that there is a positive treatment effect is 0.995 while for Q2W that same probability is 0.999. Thus, it is unequivocal that both doses produce a positive treatment effect with this approach. Using the traditional NHST approach, the regulatory criteria are not officially met, and the frequentist approach struggles to have a comprehensive and coherent decision paradigm. In this real example, the sponsor decided not to spend the considerable time and effort of submitting this treatment for approval, in part, due to its low chance of clearing the regulatory hurdle. In contrast, what appeared to be confusing results from a frequentist analysis of two independent clinical studies is rendered more interpretable with the holistic Bayesian analysis [median effect (95% credible interval): Q4W − PBO = 0.0544 (0.0129, 0.0972); Q2W − PBO = 0.0656 (0.0249, 0.1080)].

It should be noted that if Study 2 was used as the prior for Study 1, with the same borrowing factor of 0.9, the posterior probabilities for the synthesized treatment effect are also greater than 0.99. There may be other approaches to the synthesis of all data/information to arrive at a single probability for the

Table 3. The probability of exceeding a treatment effect (E) for each dose group based on the posterior probability distribution for the combined Phase 3 trials.

Treatment effect (E)	Pr(Q4W effect > E)	Pr(Q2W effect > E)
0.06	0.4011	0.6152
0.07	0.2353	0.4170
0.08	0.1192	0.2531
0.09	0.0537	0.1275
0.10	0.0192	0.0545

treatment effect, but we leave that to other creative thinkers involved in drug development and regulatory decision making.

There are other aspects of the treatment effect that are quantified easily in the Bayesian paradigm, once the posterior distribution of that treatment effect has been derived. For example, it is important to know the likelihood that the more frequent dose (Q2W) is more effective than the less frequent dose (Q4W). While the frequentist approach and the current regulatory paradigm assess trials separately and independently thereby giving a mixed result, using the Bayesian paradigm and the synthesized Phase 3 data, Pr(Q2W effect > Q4W effect) = 0.6532. This probability for greater effectiveness of Q2W can be weighed against increased risk of adverse events possibly seen with the high dose. Also, the probability of any chosen effect size can be calculated for clinicians to determine the likelihood of a clinically meaningful effect. Table 3 shows the probabilities for a range of treatment effect sizes. From historical data on the FDA approved treatment for this disease (used in the meta-analysis previously described), the treatment effect was approximately 0.09. The results in Table 3 show that this experimental treatment has a low probability of being more efficacious than the currently approved and marketed product.

Finally, in our example, some may be concerned with the disparity of the results for the Q2W dose group across the two studies, and such an observation alone may cast doubt on the credibility of the studies. The differences between Study 1 and Study 2 results for the Q2W dose were examined in great detail (e.g., all treatment groups were very well-balanced with respect to baseline demographics; drop-out patterns and missing data assumptions could not explain such a discrepancy). It is completely unclear as to why the placebo and Q4W responses were consistent across studies and the Q2W responses were not. With enough experience over many clinical trials, one will recognize that there are times when two identical studies run contemporaneously produce different results. One may be willing to accept the outcome of Study 2 as more consistent with expectations (e.g., similarity with historical data used in the network meta-analysis, consistent dose response, biologic plausibility) and, therefore, more representative of the truth. Based on that perspective, we have calculated the probability of observing an average response as small as that observed in Study 1 for the Q2W dose using the posterior distribution of effect for Q2W from Study 2. That probability is 0.0025, indicating that the Q2W response in Study 1 is very unusual and perhaps beyond the expectations of sampling variation from one study to the next.

In the authors' views of this clinical development program, the totality of the evidence indicates with a high degree of certainty (>99%) that the experimental treatment produces a positive treatment effect. Of course, there are other considerations that are necessary for the approval of a new treatment: Is the treatment safe? Does the high dose (i.e., more frequent dosing regimen in this example) produce sufficiently more efficacy to warrant any potential increase in adverse effects that might accompany using the higher dose? However, the first question that needs to be answered is, "Does this treatment work?" or stated differently, "Does this treatment produce a response different from the placebo response?" An affirmative answer to this question opens the door for regulatory submission and review; if answered in the negative, then no such submission and review are possible. As such, obtaining the best answer to this question is very important to the sponsor and ultimately the patients who are waiting for new treatments. Lastly, sponsors must make the decision about seeking approval for a new treatment based on the treatment effect and its clinical utility. Even treatments that are shown to be highly likely to work (i.e., be different from placebo) based on rigorous clinical trials may not be worth the cost or adverse effects of the treatment in the context of clinical practice. Here again, the Bayesian paradigm can give quantitative probabilities related to a range of effect sizes as a natural interpretation of the posterior distribution of effect size (Table 3).

4. Regulatory Perspective

The acceptability of Bayesian approaches within the FDA differs across and within the various Centers. The Center for Devices and Radiologic Health (CDRH) has issued a formal guidance document (US Food and Drug Administration 2010) on the topic and has encouraged and engaged in numerous submissions of new products using a Bayesian approach as the definitive analysis to support decision making. The Center for Biologics Evaluation and Research (CBER) cosigned the CDRH guidance, and submissions using the Bayesian approach as the primary analysis have occurred and been accepted. CDRH regulates devices, and a strong argument can be made for accepting Bayesian approaches in that setting. If a device is approved for use, for example, then subsequent versions of that device involving engineering changes or software changes can rely heavily on the fact that the previous device worked as it was purported to do. Borrowing data/information from previous versions of a device for approval of subsequent versions of the same device is both scientifically sound and legal under the governing Federal statute and regulations (FDA 2018b), which differ somewhat from the statutes governing approval of new drugs. The use of Bayesian approaches is also used within CBER, which regulates vaccines. Approval of new vaccines may rely on well-established clinical data and repeatable manufacturing processes created from previous vaccine approvals. In fact, one might consider ignoring prior evidence when evaluating minor changes in a product—but still requiring approval—as unethical. This is not to say that Bayesian methods are used exclusively within these Centers; when novel devices or biologics are developed and submitted to FDA for approval, more traditional frequentist methods are often utilized.

The Center for Drug Evaluation and Research (CDER) has, to date, issued no guidance on Bayesian approaches, and rarely

employs Bayesian inference as the definitive basis for concluding whether a drug has a positive treatment effect or not. In CDER, new therapeutic treatments for a wide variety of diseases are reviewed. Some of these treatments have biological mechanisms that are similar to others as multiple sponsors develop treatments based on their common knowledge of disease pathology. Other treatments have novel biological mechanisms of action as scientific discovery unveils new insights into disease processes. How or when to use prior information in this setting is less obvious, and decision-making criteria are more complex. Consequently, the application of Bayesian approaches has been controversial.

Space does not permit a full elaboration of specific possible uses of Bayesian approaches in clinical drug development, but we acknowledge some areas where such approaches have emerged or are emerging.

- FDA has recognized the utility of Bayesian approaches in earlier phases of clinical drug development, the results of which are not directly involved in regulatory approval as with Phase 3 trials.
- FDA has researched possible application in areas such as noninferiority trials (Gamalo et al. 2014). Noninferiority trials seek to confirm that an experimental treatment is no worse than an approved control treatment. The use of Bayesian approaches in noninferiority trials seems quite natural because the comparison of an experimental treatment versus an active control requires synthesizing data from historical trials of the active control agent. The final FDA Guidance on noninferiority trials acknowledges the potential utility of a Bayesian approach in this setting (US Food and Drug Administration 2016).
- Bayesian methods for adaptive trials are of interest in some regulatory settings. The PREVAIL trial in Ebola provides one such example (Dodd et al. 2016; Proschan, Dodd and Price 2016). Furthermore, Bayesian trial designs are referenced in the recently released draft guidance from CDER on adaptive clinical trial designs, possibly opening the door to such methods in confirmatory trials (US Food and Drug Administration 2018a).
- The use of Bayesian approaches in pediatric trials using adult data as a source of prior information is a logical opportunity (Goodman and Sladky 2005; Greenhouse and Seltman 2005). However, children are not merely smaller adults. When creating prior distributions of effect, one must consider not only dose adjustments due to smaller body mass, but also how children may differ in their absorption, metabolism and excretion of a treatment.
- Using historical data on the natural progression of rare diseases as a source of prior information can be quite helpful since it is often difficult to recruit patients or even find sufficient numbers of patients to participate in trials of rare diseases.

Interest in Bayesian methods to inform regulatory decision making has been increasing in recent years, particularly when relevant prior data sources are available or the acquisition of new data carries difficulties, as in the situations mentioned above.

5. Discussion

The use of hypothesis testing and subsequent reliance on p-values for determining the success of a clinical trial has been engrained in the drug development and approval process for 50 years. Arguments in favor of this approach include having an objective, albeit arbitrary, criteria that create a level playing field for all sponsors seeking regulatory approval for marketing of their products. This well-worn path has created a cross-industry consistency that does streamline discussions between sponsors and regulators since it obviates the need for more complex and detailed conversations about definition of priors and a suitable posterior probability that the treatment works required for approval. Furthermore, time has taught us that the vast majority of treatments approved by FDA and other regulatory agencies around the world are effective. Indeed, the reason for withdrawal of treatments from the market are most often due to adverse events that emerge only after many thousands or even millions of patients take a new treatment. Thus, some regulators have argued that the system and inferential approach based on the frequentist paradigm has served our society well and supplanting it with a new paradigm should only be done with caution and careful study.

Such perspectives give less weight to the cost and time it takes to conduct Phase 3 trials that are treated as separate and independent entities. In an era where drug development is enormously expensive and takes many years, why would our research community throw out the body of knowledge about an experimental treatment and act as if Phase 3 trials are isolated entities unto themselves? If our society truly understood this process, would they not consider it puzzling and perhaps even wasteful? Furthermore, statisticians and others have acknowledged for decades that a p-value is not a measure of treatment effect since a large enough sample size can produce statistically significant results for arbitrarily small differences between an experimental treatment and a control. The advantage of the Bayesian approach is that it makes probability statements about the quantity of greatest interest to all members of society—the probability associated with any magnitude of the treatment effect, whether that magnitude be a difference from placebo or any agreed medically meaningful effect.

However, one needs to consider the operating characteristics of any paradigm in terms of producing not only Type 1, but also Type 2 errors—nonapproval of treatments that are indeed useful for patients—the latter being rarely a consideration in present day regulatory reviews. Balancing Type 1 and Type 2 errors presents a difficult challenge for the regulatory decision-making process. Perhaps our society would benefit from a more rigorous and formal use of a practical metric of public health value or a utility function. The utility function is the bridge between "what do I believe about the treatment effect" (expressed as a probability statement about the size of the treatment effect) and "what is the optimal decision" balancing the probability and cost of Type 1 and Type 2 errors (Müller et al. 2017; Ventz and Trippa 2015). From a regulatory approval perspective, such a utility function basis for approval decisions may be based on the societal cost of false-positive and false-negative decisions. Furthermore, it is logically dependent on the severity and burden of the disease on individual outcomes, the safety risks that patients are willing to

take, and the overall social impact of the disease state. From a payer perspective (individual, insurance company or any other form of coverage), cost implications play a role as well. The probabilistic framework inherent in the Bayesian paradigm is well suited for such quantitative evaluations.

In our example, there were no Phase 2 studies of the experimental treatment that could have provided an informative prior for Phase 3 because the Sponsor chose not to do such studies influenced by cost and time considerations. This scenario is not uncommon in drug development. A sponsor may have approval for their treatment in one disease state, and for a second related disease state, the sponsor wants to proceed directly to Phase 3. Another situation is when a sponsor may be trailing a competitor in the race to the market and may assume the risk of selecting doses for Phase 3 without the benefit of Phase 2 trials to inform that selection. In such cases, sponsors routinely look to other treatments with the same mechanism of action for insights into the magnitude of the possible efficacy of the new treatment. The value of the Bayesian approach is to quantify explicitly how such information should be incorporated into a prior distribution. In doing so, we are not considering that the new treatment has the same efficacy as another treatment, but rather use the results on the other treatment as a starting point for borrowing the appropriate level of information from which we evaluate the current experimental treatment.

Some argue that there are subjective elements to the Bayesian approach, but that is one of the distinguishing positive features—it requires stating subjective assumptions clearly and quantifying them. Also, we should not be blind to the fact that there are assumptions and subjective choices related with frequentist methods as well (e.g., normal distribution, homoscedasticity, model selection, inclusion of covariates in a model as well as the arbitrary p-value itself). For those involved in clinical drug development, there is awareness of the many subjective decisions and agreements that are made with FDA or other global regulatory agencies through the numerous meetings and correspondences that happen over many years of development of an experimental treatment.

We present a practical approach that could be undertaken within the existing sponsor-regulatory interaction framework. Discussions and decisions about use of Phase 2 data, borrowing factors and external information from the evolving field of science or medicine could be thoughtfully reviewed and agreed before the initiation of a Phase 3 trial. Specifically, sponsors could propose which clinical trials of the experimental treatment or of treatments from other sponsors or even government research that may be included in a meta-analysis in order to develop a prior distributions of treatment response. The details of such meta-analysis, including statistical models and assumptions, could be agreed jointly by the sponsor and regulators. Based on the design of the Phase 3 trial (patient population, doses, duration, etc.), the similarity to previous trials could be evaluated and appropriate borrowing factors could be agreed upon for each treatment in the Phase 3 trial. Nonclinical data—from the sponsor, other pharmaceutical companies or the academic community—generated about the mechanism of action of the new drug, genetic factors or other molecular biology knowledge could be used to strengthen or weaken such borrowing factors. A different borrowing factor could be applied for placebo, active control treatment and the experimental treatment based on the strength of evidence for each. In this way, just as in current practice, statistical design and analysis methodologies for Phase 3 trials would be prespecified and levels of evidence agreed upon by all parties involved. Furthermore, prespecified sensitivity analyses using various priors, defined by the borrowing factor ε, could help evaluate the robustness of any conclusions. Thus, the current paradigm using Phases of drug development needs little overhaul, and the existing framework for regulatory interactions could remain the same, but the statistical content of such meetings would shift from frequentist to Bayesian approaches. There would be no major disruptions to existing practice, laws or regulations—merely a change in thinking, which, to date, has proven to be the most difficult of all.

Frequentist-based inference has dominated clinical trial design, conduct, analysis, and reporting since the advent of modern clinical trials in the 1940s. This paradigm has served a very useful role in bringing rigor to the evaluation of experimental treatments. Bayesian approaches offer the next generation of inferential thinking for decision making based on the probability of an effect size that synthesizes the most relevant available evidence and is calibrated by the severity of the disease state and the burden of disease on individual patients and on society as a whole. Thanks to breakthroughs in efficient computing algorithms, faster computers and more accessible software tools, Bayesian computations are now tractable (Lee and Chu 2012). The use of Bayesian approaches by the statistical and medical communities in the biopharmaceutical industry, medical product regulatory agencies and academia are only constrained by our history and our comfort in operating the way we have always done so. The conversion to a Bayesian paradigm is possible through meaningful dialog amongst industry and regulatory scientists within the current legal and regulatory framework of the FDA and other global regulatory agencies. We hope this article has helped to highlight the value of a Bayesian approach as well as the practical steps on a path forward.

Acknowledgments

We are thankful to the Reviewers, Associate Editor and Editor who have made substantive comments that have improved the content and the flow of the manuscript considerably.

References

Bayarri, M. J., and Berger, J.O. (2004), "The Interplay of Bayesian and Frequentist Analysis," *Statistical Science*, 19(1), 58–80. [320]

Berger, J. O. (1990), "Robust Bayesian Analysis: Sensitivity to the Prior," *Journal of Statistical Planning and Inference*, 25(3), 303–328. [322]

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59 (2), 121–126. [320]

Chuang-Stein, C., and Kirby, S. (2014), "The Shrinking or Disappearing Observed Treatment Effect," *Pharmaceutical Statistics*, 13 (5), 277-280. [321]

Dodd, L. E., Proschan, M. A., Neuhaus, J., Koopmeiners, J. S., Neaton, J., Beigel, J. D., Barrett, K., Lane, H. C., Davey, R. T., Jr. (2016), "Design of a Randomized Controlled Trial for Ebola Virus Disease Medical Countermeasures: PREVAIL II, the Ebola MCM Study," *The Journal of Infectious Diseases*, 213, 1906–1913. [325]

Drug Amendments Act of 1962, Public Law 87–781, 76 STAT 780, available at *http://prescriptiondrugs.procon.org/sourcefiles/1962Amendments.pdf* [319]

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh, Scotland: Oliver & Boyd, p. 45. [320]

Fisher, R. A. (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. [319]

Gamalo, M. A., Tiwari, R. C., and LaVange, L. M. (2014), "Bayesian Approach to the Design and Analysis of Non-inferiority Trials for Anti-infective Products," *Pharmaceutical Statistics*, 13(1), 25–40. [325]

Goodman, S. N. (1999), "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy," *Annals of Internal Medicine,* 130, 995–1004. [320]

Goodman, S. N., and Sladky, J. T., (2005), "A Bayesian Approach to Randomized Controlled Trials in Children Utilizing Information From Adults: The Case in Guillian-Barré Syndrome," *Clinical Trials,* 2, 305–310. [325]

Greco, T., Landoni, G., Biondi-Zoccai, G., D'Ascenzo, F., and Zangrillo, A. (2016), "A Bayesian Network Meta-analysis for Binary Outcome: How To Do It," *Statistical Methods in Medical Research,* 25(5), 1757–1773. [321]

Greenhouse, J. B., and Seltman, H. (2005), "Using Prior Distributions to Synthesize Historical Evidence: Comments on the Goodman-Sladky Case Study of IVIg in Guillian-Barré Syndrome," *Clinical Trials* 2, 311–318. [325]

Hay, M., Thomas, D. W., Craighead, J. L., Economides. C., and Rosenthal. J. (2014). "Clinical Development Success Rates for Investigational Drugs," *Nature Biotechnology*, 32(1), 40–51. [321]

Ibrahim, J. G., and Chen, M. H. (2000), "Power Prior Distributions for Regression Models," *Statistical Science*, 15(1), 46–60. [322]

Inoue, L.Y.T., Berry, D.A., and Parmigiani, G. (2005), "Relationship Between Bayesian and Frequentist Sample Size Determination. *The American Statistician*, 59 (1), 79-87. [320]

Lee, J. J., and Chu, C. T. (2012), "Bayesian Clinical Trials in Action," *Statistics in Medicine*, 31(25), 2955–2972. [326]

Morita, S., Thall, P.F., Mueller, P. (2008), "Determining the Effective Sample Size of a Parametric Prior." *Biometrics*, 64, 595–602. [322]

Morris, D. E., Oakley, J. E. and Crowe, J. A. (2014), "A Web-based Tool for Eliciting Probability Distributions From Experts," *Environmental Modelling & Software*, 52, 1–4. [321]

Müller, P., Xu, Y., and Thall, P. F. (2017), "Clinical Trial Design as a Decision Problem," *Applied Stochastic Models in Business and Industry*, 33 (3), 296–301. [325]

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006), *Uncertain Judgements: Eliciting Experts' Probabilities*. West Sussex, England: Wiley. [321]

Proschan, M. A., Dodd, L. E., and Price, D. (2016), "Statistical Considerations for a Trial of Ebola Virus Disease Therapeutics," *Clinical Trials*, 13(1), 39–48. [325]

Sheiner, L. B. (1997), "Learning Versus Confirming in Clinical Drug Development," *Clinical Pharmacology and Therapeutics*, 61, 275–291. [320]

Spiegelhalter, D. J., Abrams, K. R., Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluations,* West Sussex, England:Wiley. [321]

US Food and Drug Administration (2010), *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*, available at *http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm* [324]

US Food and Drug Administration (2016), *Non-Inferiority Clinical Trials to Establish Effectiveness*, available at *http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm* [325]

US Food and Drug Administration (2018a), *Adaptive Designs for Clinical Trials of Drugs and Biologics*,available at *https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf* [325]

US Food and Drug Administration (2018b), Labeling for Inter Fix Threaded Fusion Device, available at *https://www.accessdata.fda.gov/cdrh_docs/pdf/P970015C.pdf* [324]

Ventz, S., and Trippa, L. (2015), "Bayesian Designs and the Control of Frequentist Characteristics: A Practical Solution," *Biometrics*, 71 (1), 218–226. [325]

The American Statistician

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

Multiple Perspectives on Inference for Two Simple Statistical Scenarios

Noah N. N. van Dongen, Johnny B. van Doorn, Quentin F. Gronau, Don van Ravenzwaaij, Rink Hoekstra, Matthias N. Haucke, Daniel Lakens, Christian Hennig, Richard D. Morey, Saskia Homer, Andrew Gelman, Jan Sprenger & Eric-Jan Wagenmakers

To cite this article: Noah N. N. van Dongen, Johnny B. van Doorn, Quentin F. Gronau, Don van Ravenzwaaij, Rink Hoekstra, Matthias N. Haucke, Daniel Lakens, Christian Hennig, Richard D. Morey, Saskia Homer, Andrew Gelman, Jan Sprenger & Eric-Jan Wagenmakers (2019) Multiple Perspectives on Inference for Two Simple Statistical Scenarios, The American Statistician, 73:sup1, 328-339, DOI: 10.1080/00031305.2019.1565553

To link to this article: https://doi.org/10.1080/00031305.2019.1565553

© 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: 8395

View related articles

View Crossmark data

Citing articles: 14 View citing articles

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, 328–339: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1565553

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Multiple Perspectives on Inference for Two Simple Statistical Scenarios

Noah N. N. van Dongen[a], Johnny B. van Doorn[b], Quentin F. Gronau[b], Don van Ravenzwaaij[c], Rink Hoekstra[c], Matthias N. Haucke[c], Daniel Lakens[d], Christian Hennig[e], Richard D. Morey[f], Saskia Homer[f], Andrew Gelman[g], Jan Sprenger[a], and Eric-Jan Wagenmakers[b]

[a]Department of Philosophy and Education Sciences, University of Turin, Turin, Italy; [b]Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands; [c]Department of Psychology, University of Groningen, Groningen, The Netherlands; [d]Department of Industrial Engineering & Innovation Sciences, Technical University Eindhoven, Eindhoven, The Netherlands; [e]Department of Statistical Science, University College London, London, UK; [f]School of Psychology, Cardiff University, Cardiff, Wales, UK; [g]Department of Statistics and Department of Political Science, Columbia University, New York, NY

ABSTRACT
When data analysts operate within different statistical frameworks (e.g., frequentist versus Bayesian, emphasis on estimation versus emphasis on testing), how does this impact the qualitative conclusions that are drawn for real data? To study this question empirically we selected from the literature two simple scenarios—involving a comparison of two proportions and a Pearson correlation—and asked four teams of statisticians to provide a concise analysis and a qualitative interpretation of the outcome. The results showed considerable overall agreement; nevertheless, this agreement did not appear to diminish the intensity of the subsequent debate over which statistical framework is more appropriate to address the questions at hand.

ARTICLE HISTORY
Received August 2018
Accepted October 2018

KEYWORDS
Frequentist or Bayesian; Multilab analysis; Statistical paradigms; Testing or estimation

1. Introduction

When analyzing a specific dataset, statisticians usually operate within the confines of their preferred inferential paradigm. For instance, frequentist statisticians interested in hypothesis testing may report *p*-values, whereas those interested in estimation may seek to draw conclusions from confidence intervals. In the Bayesian realm, those who wish to test hypotheses may use Bayes factors and those who wish to estimate parameters may report credible intervals. And then there are likelihoodists, information-theorists, and machine-learners—there exists a diverse collection of statistical approaches, many of which are philosophically incompatible.

Moreover, proponents of the various camps regularly explain why their position is the most exalted, either in practical or theoretical terms. For instance, in a well-known article 'Why Isn't Everyone a Bayesian?', Bradley Efron claimed that "The high ground of scientific objectivity has been seized by the frequentists" (Efron 1986, p. 4), upon which Dennis Lindley replied that "Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit that he might have been wrong" (Lindley 1986, p. 7). Similarly spirited debates occurred earlier, notably between Fisher and Jeffreys (e.g., Howie 2002) and between Fisher and Neyman. Even today, the paradigmatic debates show no sign of stalling, neither in the published literature (e.g., Benjamin et al. 2018; McShane et al. in press; Wasserstein and Lazar 2016) nor on social media.

The question that concerns us here is purely pragmatic: "does it matter?" In other words, will reasonable statistical analyses on the same dataset, each conducted within their own paradigm, result in qualitatively similar conclusions (Berger 2003)? One of the first to pose this question was Ronald Fisher. In a letter to Harold Jeffreys, dated on March 29, 1934, Fisher proposed that "From the point of view of interesting the general scientific public, which really ought to be much more interested than it is in the problem of inductive inference, probably the most useful thing we could do would be to take one or more specific puzzles and show what our respective methods made of them" (Bennett 1990, p. 156; see also Howie 2002, p. 167). The two men then proceeded to construct somewhat idiosyncratic statistical "puzzles" that the other found difficult to solve. Nevertheless, three years and several letters later, on May 18, 1937, Jeffreys stated that "Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (Bennett 1990, p. 162). Similarly, Edwards, Lindman, and Savage (1963) suggested that well-conducted experiments often satisfy Berkson's *interocular traumatic test*—"you know what the data mean when the conclusion hits you between the eyes" (p. 217). Nevertheless, surprisingly little is known about the extent to which, in concrete scenarios, a data analyst's statistical plumage affects the inference.

CONTACT Eric-Jan Wagenmakers *EJ.Wagenmakers@gmail.com* University of Amsterdam, Department of Psychological Methods, University of Amsterdam, Nieuwe Achtergracht 129B, Amsterdam, 1018 VK, The Netherlands; Noah N. N. van Dongen *nnnvandongen@gmail.com* Department of Philosophy and Education Sciences, University of Turin, Turin, Italy.

© 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.

Table 1. Variable names and their description.

Variable	Description
Cetirizine exposure	Whether exposed to cetirizine
Birth defect	Whether birth defects were detected
Counts	Count data for each cell

Here we revisit Fisher's challenge. We invited four groups of statisticians to analyze two real datasets, report and interpret their results in about 300 words, and discuss these results and interpretations in a round-table discussion. The datasets are provided online at *https://osf.io/hykmz/* and described below. In addition to providing an empirical answer to the question "does it matter?", we hope to highlight how the same dataset can give rise to rather different statistical treatments. In our opinion, this method variability ought to be acknowledged rather than ignored (for a complementary approach see Silberzahn et al. in press).[1]

The selected datasets are straightforward: the first dataset concerns a 2×2 contingency table, and the second concerns a correlation between two variables. The simplicity of the statistical scenarios is on purpose, as we hoped to facilitate a detailed discussion about assumptions and conclusions that could otherwise have remained hidden underneath an unavoidable layer of statistical sophistication. The full instructions for participation can be found online at *https://osf.io/dg9t7/*.

2. DataSet I: Birth Defects and Cetirizine Exposure

2.1. Study Summary

Cetirizine is a nonsedating long-acting antihistamine with some mast-cell stabilizing activity. It is used for the symptomatic relief of allergic conditions, such as rhinitis and urticaria, which are common in pregnant women. In the study of interest, Weber-Schoendorfer and Schaefer (2008) aimed to assess the safety of cetirizine during the first trimester of pregnancy when used. The pregnancy outcomes of a cetirizine group ($n = 181$) were compared to those of the control group ($n = 1685$; pregnant women who had been counseled during pregnancy about exposures known to be nonteratogenic). Due to the observational nature of the data, the allocation of participants to the groups was nonrandomized. The main point of interest was the rate of birth defects.[2] Variables of the dataset[3] are described in Table 1 and the data are presented in Table 2.

2.2. Cetirizine Research Question

Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects? In the next sections each of four data analysis teams will attempt to address this question.

[1] In contrast to the current approach, Silberzahn et al. (in press) used a relatively complex dataset and did not emphasize the differences in interpretation caused by the adoption of dissimilar statistical paradigms.

[2] The original study focused on cetirizine-induced differences in major birth defects, spontaneous abortions, and preterm deliveries. We decided to look at all birth defects, because the sample sizes were larger for this comparison and we deemed the data more interesting.

[3] The dataset is made available on the OSF repository: *https://osf.io/hykmz/*.

Table 2. Cetirizine exposure and birth defects.

	Birth defects		
Cetirizine exposure	No	Yes	Total
No	1588	97	1685
Yes	167	14	181
Total	1755	111	1866

2.3. Analysis and Interpretation by Lakens and Hennig

2.3.1. Preamble

Frequentist statistics is based on idealised models of data-generating processes. We cannot expect these models to be literally true in practice, but it is instructive to see whether data are consistent with such models, which is what hypothesis tests and confidence intervals allow us to examine. We do appreciate that automatic procedures involving for example fixed significance levels allow us to control error probabilities assuming the model, but given that the models do never hold precisely, and that there are often issues with measurement or selection effects, in most cases we think it is prudent to interpret outcomes in a coarse way rather than to read too much meaning into, say, differences between p-values of 0.047 and 0.062. We stick to quite elementary methodology in our analyses.

2.3.2. Analysis and Software

We performed a Pearson's chi-squared test with Yates' continuity correction to test for dependence between exposure of pregnant women exposed to cetirizine and birth defects using the `chisq.test` function in R software version 3.4.3 (R Development Core Team 2004). However, because Weber-Schoendorfer and Schaefer (2008) wanted "to assess the safety of cetirizine during the first trimester of pregnancy," their actual research question is whether we can reject the presence of a meaningful effect. We therefore performed an equivalence test on proportions (Chen, Tsong, and Kang, 2000) as implemented in the TOSTtwo.prop function in the TOSTER package (Lakens 2017).

2.3.3. Results and Interpretation

The chi-squared test yielded $\chi^2(1, N = 1866) = 0.817$, $p = 0.366$, which suggests that the data are consistent with an independence model at any significance level in general use. The answer to the question whether the drug is safe depends on a smallest effect size of interest (when is the difference in birth defects considered too small to matter?). This choice, and the selection of equivalence bounds more generally, should always be justified by clinical and statistical considerations pertinent to the case at hand. In the absence of a discussion of this essential aspect of the study by the authors, and in order to show an example computation, we will test against a difference in proportions of 10%, which, although debatable, has been suggested as a sensible bound for some drugs (see Röhmel 2001, for a discussion).

An equivalence test against a difference in proportions ($M_{\text{dif}} = 0.02$, 95% CI$[-0.02; 0.06]$) of 10% based on Fisher's exact z-test was significant, $z = -3.88$, $p < 0.001$. This means that we can reject differences in proportions as large, or larger, than 10%, again at any significance level in general use.

Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects? Based on the current study,

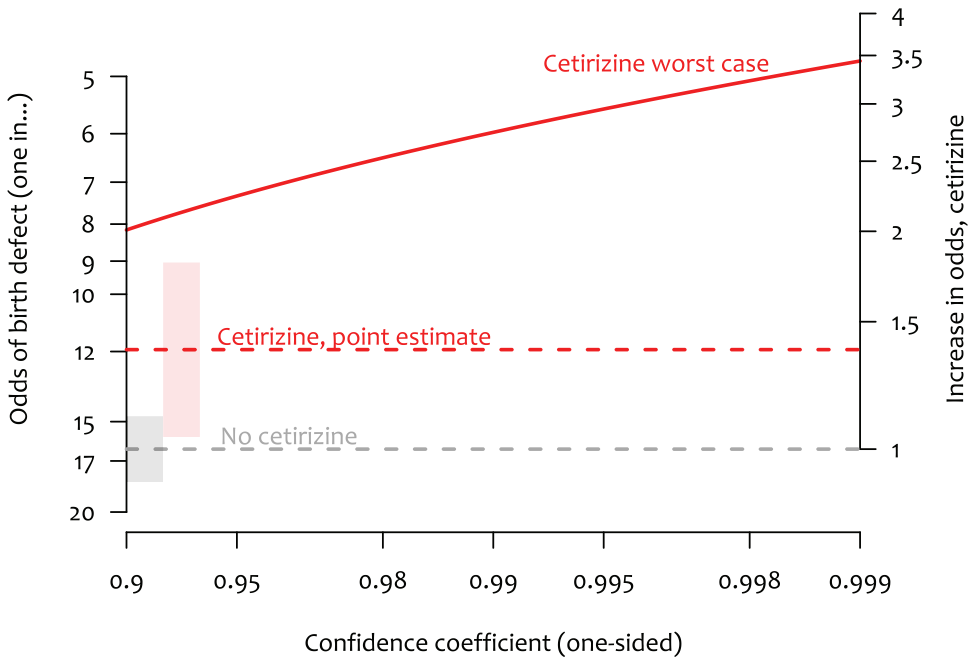


Figure 1. Estimates of the odds of a birth defect when no cetirizine (control) was taken during pregnancy and when cetirizine was taken. Horizontal dashed lines and shaded regions show point estimates and standard errors. The solid line labeled "Cetirizine worst case" shows the upper bound of the one-sided CI as a function of the confidence coefficient (*x*-axis). The right axis shows the estimated increase in odds of a birth defect for the cetirizine group compared to the control group.

there is no evidence that cetirizine exposure during pregnancy is associated with a higher incidence of birth defects. Obviously, this does not mean that cetirizine is safe; in fact the observed birth defect rate in the cetirizine group is about 2% higher than without exposure, which may or may not be explained by random variation. Is cetirizine during the first trimester of pregnancy "safe"? If we accept a difference in the proportion of birth defects of 10%, and desire a 5% long run error rate, there is clear evidence that the drug is safe. However, we expect that a cost–benefit analysis would suggest proportions of 5% to be unacceptably high, which is in the 95% confidence interval and therefore well compatible with the data. Therefore, we would personally consider the current data inconclusive.

2.4. Analysis and Interpretation by Morey and Homer

Fitting a classical logistic model with the binary birth defect outcome predicted from the cetirizine indicator confirmed the nonsignificant relationship ($p = 0.287$). The point estimate of the effect of taking cetirizine is to increase the odds of the birth defect by only 37%. At the baseline levels of birth defects in the non-cetirizine-exposed sample (approximately 6%), this would amount to about an extra two birth defects in every hundred pregnancies in the cetirizine-exposed group.

There are several problems with taking these data as evidence that cetirizine is safe. The first is the observational nature of the data. We have no way of knowing whether an apparent effect—or lack of effect—reflects confounds. Suppose, though, that we set this question aside and assess the evidence that birth defects are not more common in the cetirizine group. We can use a classical one-sided CI to determine the size of the differences we can rule out. We call the upper bound of the $100(1-\alpha)\%$ CI the "worst case" for that confidence coefficient. Figure 1 shows that at 95%, the worst-case odds increase is for the cetirizine group is 124%. At 99.5%, the worst case increase is 195%. We can translate this into more a more intuitive metric of numbers of birth defects: at baseline rates of birth defects, these would amount to additional 6 and 10 birth defects per 100, respectively (Figure 2).

The large p-value of the initial significance test suggests that we cannot rule out that cetirizine group has lower rates of birth defects; the one-sided test assuming a decrease in birth defects as the null would not yield a rejection except at high α levels. But also, the "worst-case" analysis using the upper bound of the one-sided CI suggests we also cannot rule out a substantial *increase* in birth defects in the cetirizine group.

We are unsure whether cetirizine is safe, but it seems clear to us that these data do not provide much evidence of its relative safety, contrary to what Weber-Schoendorfer and Schaefer suggest.

2.5. Analysis and Interpretation by Gronau, van Doorn, and Wagenmakers

We used the model proposed by Kass and Vaidyanathan (1992):

$$\begin{aligned} \log\left(\frac{p_1}{1-p_1}\right) &= \beta - \frac{\psi}{2}, \\ \log\left(\frac{p_2}{1-p_2}\right) &= \beta + \frac{\psi}{2}, \\ y_1 &\sim \text{Binomial}(n_1, p_1), \\ y_2 &\sim \text{Binomial}(n_2, p_2). \end{aligned} \tag{1}$$

Here, $y_1 = 97$, $n_1 = 1,685$, $y_2 = 14$, and $n_2 = 181$, p_1 is the probability of a birth defect in the control group, and p_2 is that probability in the cetirizine group. Probabilities p_1 and p_2 are functions of model parameters β and ψ. Nuisance parameter

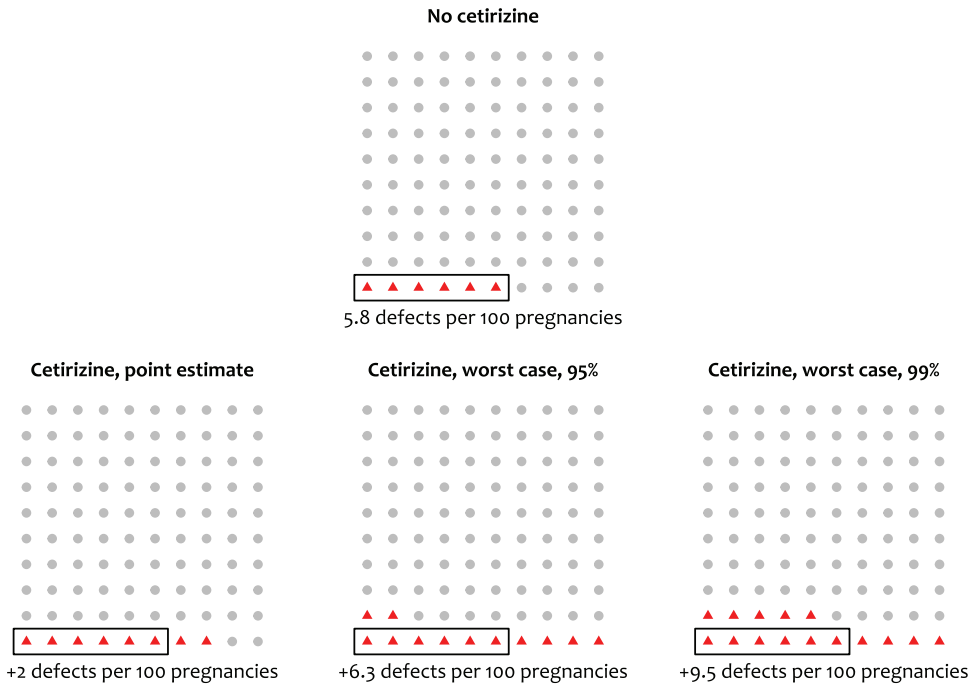


Figure 2. Frequency representations of the number of birth defects expected under various scenarios. Top: Expected frequency of birth defects when cetirizine was not taken (control). Bottom-left: Point estimate of the expected frequency of birth defects when cetirizine is taken. Bottom-middle (bottom-right): Upper bound of a one-sided 95% (99%) CI for the expected frequency of birth defects when cetirizine was taken. Because the analysis is intended to be comparative, in the bottom panels the no-cetirizine estimate was assumed to be the truth when calculating the increase in frequency.

β corresponds to the grand mean of the log odds, whereas the test-relevant parameter ψ corresponds to the log odds ratio. We assigned β a standard normal prior and used a zero-centred normal prior with standard deviation σ for the log odds ratio ψ. Inference was conducted with `Stan` (Carpenter et al. 2017; Stan Development Team 2016) and the `bridgesampling` package (Gronau, Singmann, and Wagenmakers in press). For ease of interpretation, the results will be shown on the odds ratio scale.

Our first analysis focuses on estimation and uses $\sigma = 1$. The result, shown in the left panel of Figure 3, indicates that the posterior median equals 1.429, with a 95% credible interval ranging from 0.793 to 2.412. This credible interval is relatively wide, indicating substantial uncertainty about the true value of the odds ratio.

Our second analysis focuses on testing and quantifies the extent to which the data support the skeptic's $\mathcal{H}_0 : \psi = 0$ versus the proponent's $\mathcal{H}_1$. To specify $\mathcal{H}_1$ we initially use $\sigma = 0.4$ (i.e., a mildly informative prior; Diamond and Kaul 2004), truncated at zero to respect the fact that cetirizine exposure is hypothesized to cause a *higher* incidence of birth defects: $\mathcal{H}_+ : \psi \sim N^+(0, 0.4^2)$.

As can be seen from the right panel of Figure 3, the observed data are predicted about 1.8 times better by $\mathcal{H}_+$ than by $\mathcal{H}_0$. According to Jeffreys (1961, Appendix B), this level of support is "not worth more than a bare mention." To investigate the robustness of this result, we explored a range of alternative prior choices for σ under $\mathcal{H}_+$, varying it from 0.01 to 2. The results of this sensitivity analysis are shown in Figure 4 and reveal that across a wide range of priors, the data never provide more than anecdotal support for one model over the other. When σ is selected post hoc to maximize the support for $\mathcal{H}_+$ this yields $\mathrm{BF}_{+0} = 1.84$, which, starting from a position of equipoise, raises the probability of $\mathcal{H}_+$ from 0.50 to about 0.65, leaving a posterior probability of 0.35 for $\mathcal{H}_0$.

In sum, based on this dataset we cannot draw strong conclusions about whether or not cetirizine exposure during pregnancy is associated with a higher incidence of birth defects. Our analysis shows an "absence of evidence," not "evidence for absence."

2.6. Analysis and interpretation by Gelman

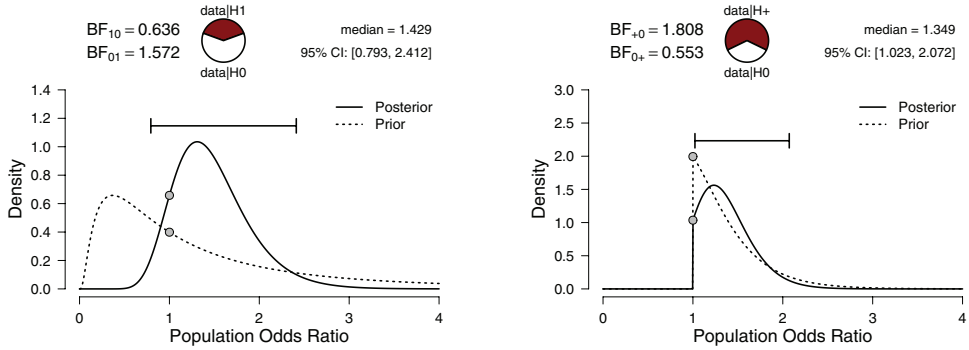
I summarized the data with a simple comparison: the proportion of birth defects is 0.06 in the control group and 0.08 in the cetirizine group. The difference is 0.02 with a standard error of 0.02. I got essentially the same result with a logistic regression predicting birth defect: the coefficient of cetirizine is 0.3 with a standard error of 0.3. I performed the analyses in R using rstanarm (code available at *https://osf.io/nh4gc/*).

I then looked up the article, "The safety of cetirizine during pregnancy: A prospective observational cohort study," by Weber-Schoendorfer and Schaefer (2008) and noticed some interesting things:

(a) The published article gives $N = 196$ and 1686 for the two groups, not quite the same as the 181 and 1685 for the "all birth defects" category. I could not follow the exact reasoning.[4]

(b) The two groups differ in various background variables: most notably, the cetirizine group has a higher smoking rate (17% compared to 10%).

[4] Clarification: the original article does not provide a rationale for why several participants were excluded from the analysis.



(a) Estimation results. (b) Testing results.

Figure 3. Gronau, van Doorn, and Wagenmakers' Bayesian analysis of the cetirizine dataset. The left panel shows the results of estimating the log odds ratio under $\mathcal{H}_1$ with a two-sided standard normal prior. For ease of interpretation, results are displayed in the odds ratio scale. The right panel shows the results of testing the one-sided alternative hypothesis $\mathcal{H}_+ : \psi \sim \mathcal{N}^+(0, 0.4^2)$ versus the null hypothesis $\mathcal{H}_0 : \psi = 0$. Figures inspired by JASP (*jasp-stats.org*).

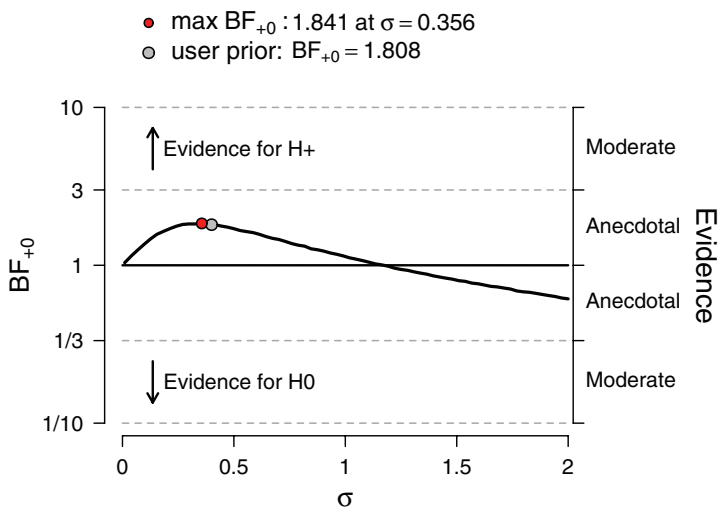


Figure 4. Sensitivity analysis for the Bayesian one-sided test. The Bayes factor BF_{+0} is a function of the prior standard deviation σ. Figure inspired by JASP.

(c) In the published article, the outcome of focus was "major birth defects," not the "all birth defects" given for us to study.

(d) The published article has a causal aim (as can be seen, for example, from the word "safety" in its title); our assignment is purely observational.

Now the question, "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?" I have not read the literature on the topic. To understand how the data at hand address this question, I would like to think of the mapping from prior to posterior distribution. In this case, the prior would be the distribution of association with birth defects of all drugs of this sort. That is, imagine a population of drugs, $j = 1, 2, \ldots,$ taken by pregnant women, and for each drug, define θ_j as the proportion of birth defects among women who took drug j, minus the proportion of birth defects in the general population. Just based on my general understanding (which could be wrong), I would expect this distribution to be more positive than negative and concentrated near zero: some drugs could be mildly protective against birth defects or associated with low-risk pregnancies, most would have small effects and not be strongly associated with low- or high-risk pregnancies, and some could cause birth defects or be taken disproportionately by women with high-risk pregnancies. Supposing that the prior is concentrated within the range $(-0.01, +0.01)$, the data would not add much information to this prior.

To answer, "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?," the key question would seem to be whether the drug is more or less likely to be taken by women at a higher risk of birth defects. I am guessing that maternal age is a big predictor here. In the reported study, average age of the exposed and control groups was the same, but I do not know if that is generally the case or if the designers of the study were purposely seeking a balanced comparison.

3. DataSet II: Amygdalar Activity and Perceived Stress

3.1. Study Summary

In a recent study published in the *Lancet*, Tawakol et al. (2017) tested the hypothesis that perceived stress is positively associated with resting activity in the amygdala. In the second study reported in Tawakol et al. (2017), $n = 13$ individuals with an increased burden of chronic stress (i.e., a history of post-

Table 3. Variable names and their description.

Variable	Description
Perceived stress scale	Participant score on the PSS
Amygdalar activity	Intensity of amygdalar resting state activity

Table 4. Raw data as extracted from Figure 5 in Tawakol et al. (2017), with help of Jurgen Rusch, Philips Research Eindhoven.

Perceived Stress Scale	Amygdalar Activity
12.0103	5.2418
32.0350	6.8601
22.0296	6.4402
20.0079	5.4620
24.0155	5.4439
24.0155	5.3349
24.0155	5.4216
26.0082	5.5176
28.0120	5.1615
21.9872	4.7114
21.9872	4.1844
20.0138	4.3079
16.0088	3.3015

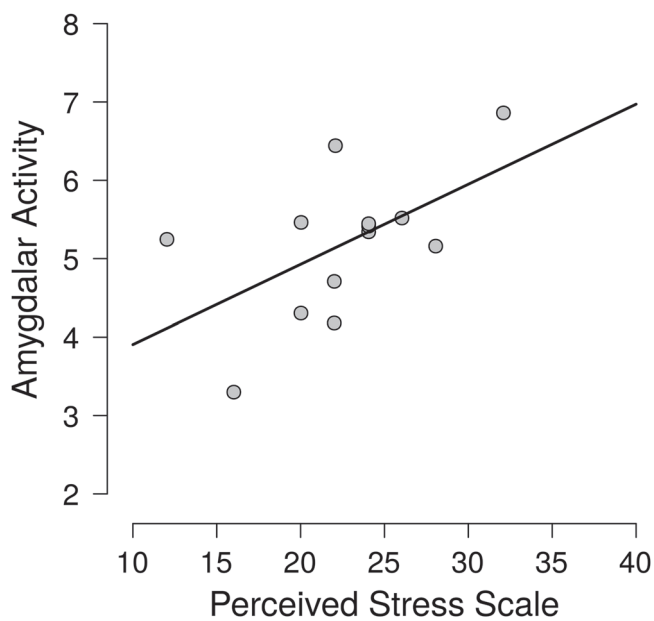


Figure 5. Scatterplot of amygdalar activity and perceived stress in 13 patients with PTSD. Data extracted from Figure 5 in Tawakol et al. (2017), with help of Jurgen Rusch, Philips Research Eindhoven.

traumatic stress disorder or PTSD) were recruited from the community, completed a Perceived Stress Scale (i.e., the PSS-10; Cohen, Kamarck, and Mermelstein 1983) and had their amygdalar activity measured. Variables of the dataset[5] are described in Table 3, the raw data are presented in Table 4, and the data are visualized in Figure 5.

3.2. Amygdala Research Question

Do PTSD patients with high resting state amygdalar activity experience more stress? In the next sections each of four data analysis teams will attempt to address this question.

[5] The dataset is made available on the OSF repository: *https://osf.io/hykmz/*.

3.3. Analysis and Interpretation by Lakens and Hennig

3.3.1. Analysis and Software

We calculated tests for uncorrelatedness based on both Pearson's product-moment correlation and Spearman's rank correlation using the `cor.test` function in the stats package in R 3.4.3.

3.3.2. Results and interpretation

The Pearson correlation between perceived stress and resting activity in the amygdala is $r = 0.555$, and the corresponding test yields $p = 0.047$. Although this is just smaller than the conventional 5% level, we do not consider it as clear evidence for nonzero correlation. From the appendix it becomes clear that the reported correlations are exploratory: "Patients completed a battery of self-report measures that assessed variables that may correlate with PTSD symptom severity, including comorbid depressive and anxiety symptoms (MADRS, HAMA) and a well-validated questionnaire Perceived Stress Scale (PSS-10)." Therefore, corrections for multiple comparisons would be required to maintain a given significance level. The article does not provide us with sufficient information to determine the number of tests that were performed, but corrections for multiple comparisons would thus be in order. Consequently, the fairly large observed correlation and the borderline significant p-value can be interpreted as an indication that it may be worthwhile to investigate the issue with a larger sample size, but do not give conclusive evidence. Visual inspection of the data does not give any indication against the validity of using Pearson's correlation, but with $N = 13$ we do not have very strong information regarding the distributional shape. The analogous test based on Spearman's correlation yields $p = 0.062$, which given its weaker power is compatible with the qualitative interpretation we gave based on the Pearson correlation.

Do PTSD patients with high resting state amygdalar activity experience more stress? Based on the current study, we cannot conclude that PTSD patients with high resting state amygdalar activity experience more stress. The single $p = 0.047$ is not low enough to indicate clear evidence against the null hypothesis after correcting for multiple comparisons when using an alpha of 0.05. Therefore, our conclusion is: Based on the desired error rate specified by the authors, we cannot reject a correlation of zero between amygdalar activity and participants' score on the perceived stress scale. With a 95% CI that ranges from $r = 0$ to $r = 0.85$, it seems clear that effects that would be considered interesting cannot be rejected in an equivalence test. Thus, the results are inconclusive.

3.4. Analysis and Interpretation by Morey and Homer

The first thing that should be noted about this dataset is that it contains a meager 13 data points. The linear correlation by Tawakol et al. (2017) depends on assumptions that are for all intents and purposes unverifiable with this few participants. Add to this the difficulty of interpreting the independent variable—a sum of ordinal responses—and we are justified being skeptical of any hard conclusions from these data.

Suppose, however, that the relationship between these two variables was best characterized by a linear correlation, and set aside any worries about assumptions. The large observed

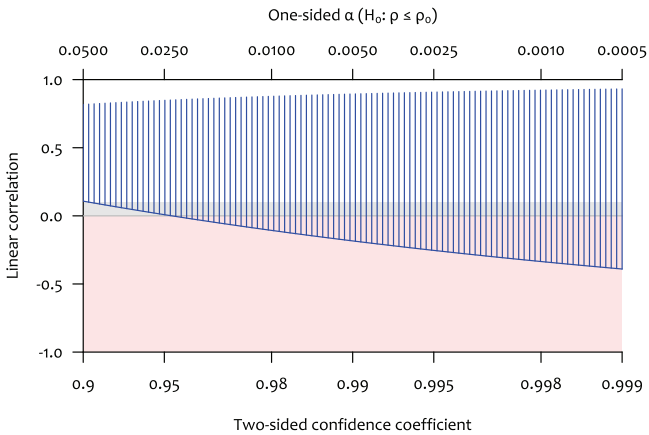


Figure 6. Confidence intervals and one-sided tests for the Pearson correlation as a function of the confidence coefficient. The vertical lines represent the confidence intervals (confidence coefficient on lower axis), and the curve represents the value that is just rejected by the one-sided test (α on the upper axis).

correlation coupled with the marginal p-value should signal to us that a wide range of correlations are not ruled out by the data. Consider that the 95% CI on the Pearson correlation is [0.009,0.848]; the 99.5% CI is [−0.254,0.908]. Negligible correlations are not ruled out; due to the small sample size, any correlation from "essentially zero" to "the correlation between height and leg length" (i.e., very high) is consistent with these data. The solid curve in Figure 6 shows the lower bound of the confidence interval on the linear correlation for a wide range of confidence levels; they are all negligible or even negative.

Finally, the authors did not show that this correlation is selective to the amygdala; it seems to us that interpreting the correlation as evidence for their model requires selectivity. It is important to interpret the correlation in the context of the relationship between amygdala resting state activity, stress, and cardiovascular disease. If one could not show that amygdala resting-state activation showed a substantially higher correlation with stress than other brain regions not implicated in the model, this would suggest that the correlation cannot be used to bolster their case. Given the uncertainty in the estimate of the correlation, there is little chance of being able to show this.

All in all, we are not sure that the information in these thirteen participants is enough to say anything beyond "the correlation doesn't appear to be (very) negative."

3.5. Analysis and Interpretation by Van Doorn, Gronau, and Wagenmakers

We applied Harold Jeffreys's test for a Pearson correlation coefficient ρ (Jeffreys 1961; Ly, Marsman, and Wagenmakers 2018) as implemented in JASP (*jasp-stats.org*; JASP Team 2018).[6] Our first analysis focuses on estimation and assigns ρ a uniform prior from −1 to 1. The result, shown in the left panel of Figure 7, indicates that the posterior median equals 0.47, with a 95% credible interval ranging from −0.01 to 0.81. As can be expected with only 13 observations, there is great uncertainty about the size of ρ.

[6] JASP is an open-source statistical software program with a graphical user interface that supports both frequentist and Bayesian analyses.

Our second analysis focuses on testing and quantifies the extent to which the data support the skeptic's $\mathcal{H}_0 : \rho = 0$ versus the proponent's $\mathcal{H}_1$. To specify $\mathcal{H}_1$, we initially use Jeffreys's default uniform distribution, truncated at zero to respect the directionality of the hypothesized effect: $\mathcal{H}_+ : \rho \sim U[0, 1]$.

As can be seen from the right panel of Figure 7, the observed data are predicted about 3.9 times better by $\mathcal{H}_+$ than by $\mathcal{H}_0$. This degree of support is relatively modest: when $\mathcal{H}_+$ and $\mathcal{H}_0$ are equally likely a priori, the Bayes factor of 3.9 raises the posterior plausibility of $\mathcal{H}_+$ from 0.50 to 0.80, leaving a nonnegligible 0.20 for $\mathcal{H}_0$.

To investigate the robustness of this result, we explored a continuous range of alternative prior distributions for ρ under $\mathcal{H}_+$; specifically, we assigned ρ a stretched Beta(a,a) distribution truncated at zero, and studied how the Bayes factor changes with $1/a$, the parameter that quantifies the prior width and governs the extent to which $\mathcal{H}_+$ predicts large values of r. The results of this sensitivity analysis are shown in Figure 8 and confirm that the data provide modest support for $\mathcal{H}_+$ across a wide range of priors. When the precision is selected post hoc to maximize the support for $\mathcal{H}_+$ this yields $\mathrm{BF}_{+0} = 4.35$, which—under a position of equipose– raises the plausibility of $\mathcal{H}_+$ from 0.50 to about 0.81, leaving a posterior probability of 0.19 for $\mathcal{H}_0$.

A similar sensitivity analysis could be conducted for $\mathcal{H}_0$ by assuming a "perinull" (Tukey 1995, p. 8)—a distribution tightly centered around $\rho = 0$ rather than a point mass on $\rho = 0$. The results will be qualitatively similar.

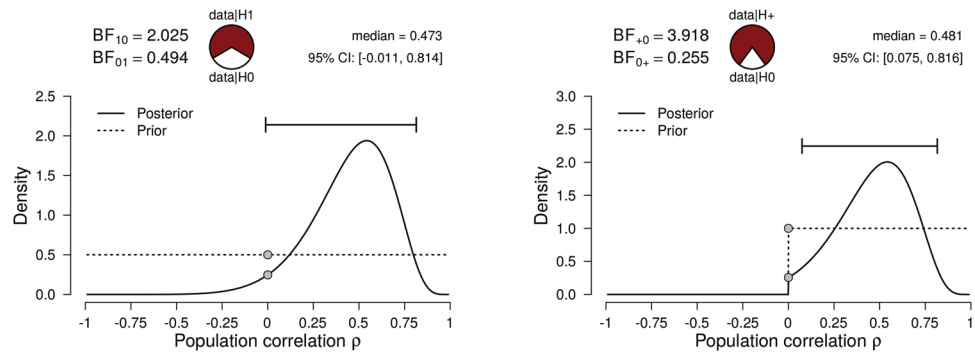
In sum, the claim that "PTSD patients with high resting state amygdalar activity experience more stress" receives modest but not compelling support from the data." The 13 observations do not warrant categorical statements, neither about the presence nor about the strength of the hypothesized effect.

3.6. Analysis and interpretation by Gelman

I summarized the data with a simple scatterplot and a linear regression of logarithm of perceived stress on logarithm of amygdalar activity, using log scales because the data were all-positive and it seemed reasonable to model a multiplicative relation. The scatterplot revealed a positive correlation and no other striking patterns, and the regression coefficient was estimated at 0.6 with a standard error of 0.4. I performed the analyses in R using rstanarm (code available at *https://osf.io/nh4gc/*). and the standard error is based on the median absolute deviation of posterior simulations (see help("mad") in R for more on this).

I then looked up the article, "Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study," by Tawakol et al. (2017). The goal of the research is "to determine whether [the amygdala's] resting metabolic activity predicts risk of subsequent cardiovascular events." Here are some items relevant to our current task:

(a) Perceived stress is an intermediate outcome, not the ultimate goal of the study.
(b) Any correlation or predictive relation will depend on the reference population. The people in this particular study are "individuals with a history of posttraumatic stress disorder" living in the Boston area.



(a) Estimation results (b) Testing results

Figure 7. van Doorn, Gronau, and Wagenmakers' Bayesian analysis of the amygdala dataset. The left panel shows the result of estimating the Pearson correlation coefficient ρ under $\mathcal{H}_1$ with a two-sided uniform prior. The right panel shows the result of testing $\mathcal{H}_0 : \rho = 0$ versus the one-sided alternative hypothesis $\mathcal{H}_+ : \rho \sim U[0, 1]$. Figures from JASP.

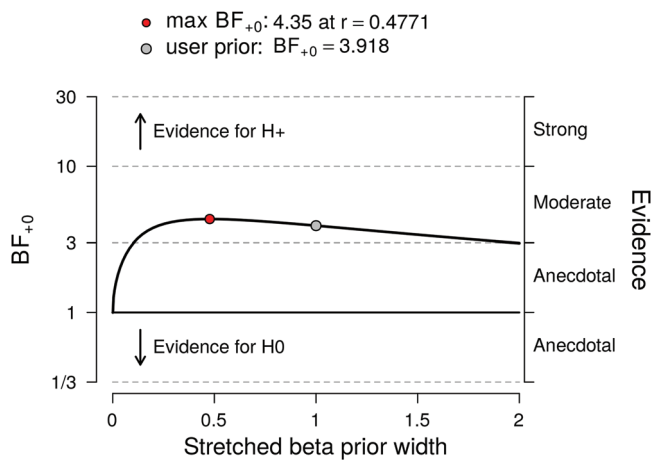


Figure 8. Sensitivity analysis for the Bayesian one-sided correlation test. The Bayes factor BF_{+0} is a function of the prior width parameter $1/a$ from the stretched Beta(a,a) distribution. Figure from JASP.

(c) The published article reports that "Perceived stress was associated with amygdalar activity ($r = 0.56$; $p = 0.0485$)." Performing the correlation (or, equivalently, the regression) on the log scale, the result is not statistically significant at the 5% level. This is no big deal given that I do not think that it makes sense to make decisions based on a statistical significance threshold, but it is relevant when considering scientific communication.

(d) Comparing my log-scale scatterplot to the raw-scale scatterplot (Figure 5A in the published article), I would say that the unlogged scatterplot looks cleaner, with the points more symmetrically distributed. Indeed, based on these data alone, I would move to an unlogged analysis–that is, the estimated correlation of 0.56 reported in the article.

To address the question, "Do PTSD patients with high resting state amygdalar activity experience more stress?," we need two additional decisions or pieces of information. First, we must decide the population of interest; here there is a challenge in extrapolating from people with PTSD to the general population. Second, we need a prior distribution for the correlation being estimated. It is difficult for me to address either of these issues: as a statistician my contribution would be to map from assumptions to conclusions. In this case, the assumptions about the prior distribution and the assumptions about extrapolation go together, as in both cases we need some sense of how likely it is to see large correlations between the responses to a subjective stress survey and a biomeasurement such as amygdalar activity. It could well be that there is a prior expectation of positive correlation between these two variables in the general population, but that the current data do not provide much information beyond our prior for this general question.

4. Round-Table Discussion

As described above, the two datasets have each been analyzed by four teams. The different approaches and conclusions are summarized in Table 5. The discussion was carried out via E-mail and a transcript can be found online at *https://osf.io/f4z7x/*. Below we highlight and summarize the central elements of a discussion that quickly proceeded from the data analysis techniques in the concrete case to more fundamental philosophical issues. Given the relative agreement among the conclusions reached by different methodological angles, our discussion started out with the following deliberately provocative statement:

> In statistics, it doesn't matter what approach is used. As long as you do conduct your analysis with care, you will invariably arrive at the same qualitative conclusion.[7]

In agreement with this claim, Hennig stated that "we all seem to have a healthy skepticism about the models that we are using. This probably contributes strongly to the fact that all our final interpretations by and large agree. Ultimately our conclusions all state that "the data are inconclusive." I think the important point here is that we all treat our models as tools for thinking that can only do so much, and of which the limitations need to

[7] The statement is based on Jeffreys' claims "[a]s a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached" (Jeffreys 1939, p. 365) and "it would only be once in a blue moon that we [Fisher and Jeffreys] would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (Bennett 1990, p. 162).

Table 5. *Overview of the approaches and results of the research teams.*

Research team	DataSet I: Cetirizine and birth defects	DataSet II:Amygdalar Activity
Lakens and Hennig	• Frequentist test of equivalence • 10% equivalence region • Data deemed inconclusive	• Frequentist correlation, $p = .047$ • Concerns about multiple comparisons • Data deemed inconclusive
Morey and Homer	• Frequentist logistic model, $p = .287$ • Observational, so possible confounds • Data deemed inconclusive	• Frequentist correlation, $p = .047$ • Small *n* means assumptions unverifiable • Is the effect specific for amygdala? • Data deemed inconclusive
Gronau, Van Doorn, and Wagenmakers	• Default Bayes factor $BF_{01} = 1.6$ • Evidence "not worth more than a bare mention" • Data deemed inconclusive	• Default Bayes factor $BF_{10} = 2$ • Small *n* • Data deemed inconclusive
Gelman	• Bayesian analysis needs good prior • Data likely to be inconclusive • A key question is who takes the drug in the population	• Bayesian analysis needs good prior • Problems with generalizing to population • Data likely to be inconclusive

be explored, rather than 'believing in' our relying on any specific model (Email 29). On the other hand, Hennig wonders "whether differences between us would've been more pronounced with data that wouldn't have allowed us to sit on the fence that easily" (Email 29) and Lakens wonders "about what would have happened if the data were clearer" (Email 32). In addition, Morey points out that "none of us had anything invested in these questions, whereas almost always analyses are published by people who are most invested" (Email 33). Wagenmakers responds that we [referring to the group that organized this study] wanted to use simple problems that would not pose immediate problems for either one of the paradigms...[and] we tried to avoid data sets that met Berkson's 'inter-ocular traumatic test' (when the data are so clear that the conclusion hits you right between the eyes) where we would immediately agree. (Email 31).

Moreover, the focus was on the analyses and discussion as free as possible from other consideration (e.g., personal investment in the questions).

However, differences between the analyses were emphasized as well. First, Morey argued that the differences (e.g., research planning, execution, analysis, etc.) between the Bayesian and frequentist approach are critically important and not easily inter-translatable (Email 2). This gave rise to an extended discussion about the frequentist procedures' dependence of the sampling protocol, which Bayesian procedures lack. While Bayesians such as Wagenmakers see this as a critical objection against the coherent use of frequentist procedures (e.g., in cases where the sampling protocol is uncertain), Hennig contends that one can still interpret the *p*-value as indicating the compatibility of the data with the null model, assuming a hypothetical sampling plan (see Email 5–11, 16–18, 23, and 24). Second, Lakens speculated that, in the cases where the approaches differ, there "might be more variation within specific approaches, than between" (Email 1). Third, Hennig pointed out that differences in prior specifications could lead to discrepancies between Bayesians (Email 4) and Homer pointed out that differences in alpha decision boundaries could lead to discrepancies between frequentists' conclusions (Email 11). Finally, Gelman disagreed with most of what had been said in the discussion thus far. Specifically, he said: "I don't think 'alpha' makes any sense, I don't think 95% intervals are generally a good idea, I don't think it's necessarily true that points in 95% interval are compatible with the data, etc etc." (Email 15).

A concrete issue concerned the equivalence test used by Lakens and Hennig for the first dataset. Wagenmakers objects that it does not add relevant information to the presentation of a confidence interval (Email 12). Lakens responds that it allows to reject the hypothesis of a >10% difference in proportions at almost any alpha level, thereby avoiding reliance on default alpha levels, which are often used in a mindless way and without attention to the particular case (Email 13).

A more foundational point of contention with the two datasets and their analysis was about the question of how to formulate Bayesian priors. For these concrete cases, Hennig contends that the subject-matter information cannot be smoothly translated into prior specifications (Email 27), which is the reason why Morey and Homer choose a frequentist approach, while Gelman considers it "hard for me to imagine how to answer the questions without thinking about subject-matter information" (Email 26).

Lakens raised the question of how much the approaches in this paper are representative of what researchers do in general (Email 43 and 44). Wagenmakers' discussion of *p*-values echoes this point. While Lakens describes *p*-values as a guide to "deciding to act in a certain way with an acceptable risk of error" and contends many scientists conform to this rationale (Email 32), Wagenmakers has a more pessimistic view. In his experience, the role of *p*-values is less epistemic than social: they are used to convince referees and editors and to suggest that the hypothesis in question is true (Email 37). Also Hennig disagrees with Lakens, but from a frequentist point of view: *p*-values should not guide binary accept/reject-decisions, they just indicate the degree to which the observed data is compatible with the model specified by the null hypothesis (Email 34).

The question of how data analysis relates to learning, inference and decision making was also discussed regarding the merits (and problems) of Bayesian statistics. Hennig contends that there can be "some substantial gain from them [priors] only

if the prior encodes some relevant information that can help the analysis. However, here we don't have such information" and the Bayesian "approach added unnecessary complexity" (Email 23). Wagenmakers reply is that "the prior is not there to help or hurt an analysis: it is there because without it, the models do not make predictions; and without making predictions, the hypotheses or parameters cannot be evaluated" and that "the approach is more complex, but this is because it includes some essential ingredients that the classical analysis omits" (Email 24). In fact, he insinuates that frequentists learn from data through "Bayes by stealth": the observed p-values, confidence intervals and other quantities are used to update the plausibility of the models in an "inexact and intuitive" way. "Without invoking Bayes' rule (by stealth) you can't learn much from a classical analysis, at least not in any formal sense." (Email 24) According to Hennig, there is more to learning than "updating epistemic probabilities of certain parameter values being true. For example I find it informative to learn that 'Model X is compatible with the data' " (Email 25). However, Wagenmakers considers Hennig's example of learning as a synonymous to observing. Though he agrees that "it is informative to know when a specific model is or is not compatible with data; to learn anything, however, by its very definition requires some form of knowledge updating" (Email 30). This discussion evolved, finally, into a general discussion about the philosophical principles and ideas underlying schools of statistical inference. Ironically, both Lakens (decision-theoretically oriented frequentism) and Gelman (falsificationist Bayesianism) claim the philosophers of science Karl Popper and Imre Lakatos, known for their ideas of accumulating scientific progress through successive falsification attempts, as one of their primary inspirations, although they spell out their ideas in a different way (Email 42 and 45).

Hennig and Lakens also devoted some attention to improving statistical practice and either directly or indirectly questioned the relevance of foundational debates. Concerning the above issue with using p-values for binary decision making, Hennig suspects that "if Bayesian methodology would be in more widespread use, we'd see the same issue there ... and then 'reject' or 'accept' based on whether a posterior probability is larger than some mechanical cutoff" (Email 34) and "that much of the controversy about these approaches concerns naive 'mechanical' applications in which formal assumptions are taken for granted to be fulfilled in reality" (Email 29). In addition, Lakens points out that

> whether you use one approach to statistics or another doesn't matter anything in practice. If my entire department would use a different approach to statistical inferences (now everyone uses frequentist hypothesis testing) it would have basically zero impact on their work. However, if they would learn how to better use statistics, and apply it in a more thoughtful manner, a lot would be gained. (Email 32)

Homer provides an apt conclusion to this topic by stating "I think a lot of problems with research happen long before statistics get involved. E.g. Issues with measurement; samples and/or methods that can't answer the research question; untrained or poor observers" (Email 35).

Finally, an interesting distinction was made between a prescriptive use of statistics and a more pragmatic use of statistics. As an illustration of the latter, Hennig has a more pragmatic perspective on statistics, because a strong prescriptive view (i.e., fulfillment of modeling assumptions as a strict requirement for statistical inference) would often mean that we can't do anything in practice (Email 2). To clarify this point, he adds: "Model assumptions are never literally fulfilled so the question cannot be whether they are..., but rather whether there is information or features of the data that will mislead the methods that we want to apply" (Email 23). The former is illustrated by Homer:

> I think that assumptions are critically important in statistical analysis. Statistical assumptions are the axioms which allow a flow of mathematical logic to lead to a statistical inference. There is some wiggle room when it comes to things like 'how robust is this test, which assumes normality, to skew?' but you are on far safer ground when all the assumptions are/appear to be met. I personally think that not even checking the plausibility of assumptions is inexcusable sloppiness (not that I feel anyone here suggested otherwise). (Email 11)

From what has been said in the discussion, there is general consensus that not all assumptions need to be met and not all rules need to be strictly followed. However, there is great disagreement about which assumptions are important; which rules should be followed and how strictly; and what can be interpreted from the results when (it is uncertain if) these rules and assumptions are violated. The interesting subtleties of this topic and the discussants' views on use of statistics can be read in the online supplement (model assumptions: Email 4, 11, 23, and 24; alpha-levels, p-values, Bayes factors, and decision procedures: Email 2, 9, 11, 14, 15, 24, and 31–40; sampling plan, optional stopping, and conditioning on the data: Email 2, 5–11, 16–18, 23, and 24).

In summary, dissimilar methods were used that resulted in similar conclusions and varying views were discussed on how statistical methods are used and should be used. At times it was a heated debate with interesting arguments from both (or more) sides. As one might expect, there was disagreement about particularities of procedures and consensus on the expectation that scientific practice would be improved by better general education on the use of statistics.

5. Closing Remarks

Four teams each analyzed two published datasets. Despite substantial variation in the statistical approaches employed, all teams agreed that it would be premature to draw strong conclusions from either of the datasets. Adding to this cautious attitude are concerns about the nature of the data. For instance, the first dataset was observational, and the second dataset may require a correction for multiplicity. In addition, for each scenario, the research teams indicated that more background information was desired; for instance, "when is the difference in birth defects considered too small to matter?"; "what are the effects for similar drugs?"; "is the correlation selective to the amygdala?"; and "what is the prior distribution for the correlation?". Unfortunately, in the routine use of statistical procedures such information is provided only rarely.

It also became evident that the analysis teams not only needed to make assumptions about the nature of the data and any relevant background knowledge, but they also needed to interpret the research question. For the first dataset, for instance, the question was formulated as "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?". What the original researchers wanted to know, however, is whether or not cetirizine is safe—this more general question opens up the possibility of applying alternative approaches, such as the equivalence test, or even a statistical decision analysis: should pregnant women be advised not to take cetirizine? We purposefully tried to steer clear from decision analyses because the context-dependent specification of utilities adds another layer of complexity and requires even more background knowledge than was already demanded for the present approaches. More generally, the formulation of our research questions was susceptible to multiple interpretation: as tests against a point null, as tests of direction, or as tests of effect size. The goals of a statistical analysis can be many, and it is important to define them unambiguously—again, the routine use of statistical procedures almost never conveys this information.

Despite the (unfortunately near-universal) ambiguity about the nature of the data, the background knowledge, and the research question, each analysis team added valuable insights and ideas. This reinforces the idea that a careful statistical analysis, even for the simplest of scenarios, requires more than a mechanical application of a set of rules; a careful analysis is a process that involves both skepticism and creativity. Perhaps popular opinion is correct, and statistics is difficult. On the other hand, despite employing widely different approaches, all teams nevertheless arrived at a similar conclusion. This tentatively supports the Fisher–Jeffreys conjecture that, regardless of the statistical framework in which they operate, careful analysts will often come to similar conclusions.

Funding

This work was supported in part by a Vici grant from the Netherlands Organisation of Scientific Research awarded to EJW (016.Vici.170.083) and a Vidi grant awarded to DL (452.17.013). NvD's and JS's work was supported by ERC Starting Investigator Grant No. 640638.

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., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T.-H., Hoijtink, H., Jones, J. H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018), "Redefine Statistical Significance." *Nature Human Behaviour*, 2:6–10. [328]

Bennett, J. H., editor (1990), *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford. [328,335]

Berger, J. O. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science*, 18:1–32. [328]

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A probabilistic Programming Language." *Journal of Statistical Software*, 76:1–32. [331]

Chen, J. J., Tsong, Y., and Kang, S.-H. (2000), "Tests for Equivalence or Noninferiority Between Two Proportions." *Drug Information Journal*, 26:569–578. [329]

Cohen, S., Kamarck, T., and Mermelstein, R. (1983), "A Global Measure of Perceived Stress." *Journal of Health and Social Behavior*, 24:385–396. [333]

Diamond, G. A., and Kaul, S. (2004), "Prior Convictions: Bayesian Approaches to the Analysis and Interpretation of Clinical Megatrials," *Journal of the American College of Cardiology*, 43:1929–1939. [331]

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70:193–242. [328]

Efron, B. (1986), "Why Isn't Everyone a Bayesian?" *The American Statistician*, 40:1–5. [328]

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (in press), "Bridge-sampling: An R Package for Estimating Normalizing Constants," *Journal of Statistical Software*. [331]

Howie, D. (2002). *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*. Cambridge: Cambridge University Press. [328]

JASP Team (2018). JASP (Version 0.9)[Computer software]. [334]

Jeffreys, H. (1939). *Theory of Probability* (1st ed.). Oxford: Oxford University Press . [335]

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press. [331,334]

Kass, R. E., and Vaidyanathan, S. K. (1992), "Approximate Bayes Factors and Orthogonal Parameters, With Application to Testing Equality of Two Binomial Proportions," *Journal of the Royal Statistical Society*, Series B, 54:129–144. [330]

Lakens, D. (2017), "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-analyses," *Social Psychological and Personality Science*, 8:355–362. [329]

Lindley, D. V. (1986). "Comment on 'Why Isn't Everyone a Bayesian?' By Bradley Efron." *The American Statistician*, 40:6–7. [328]

Ly, A., Marsman, M., and Wagenmakers, E.-J. (2018), "Analytic Posteriors for Pearson's Correlation Coefficient," *Statistica Neerlandica*, 72:4–13. [334]

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (in press), "Abandon Statistical Significance," *The American Statistician*. [328]

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3–900051–00–3. [329]

Röhmel, J. (2001), "Statistical Considerations of FDA and CPMP Rules for the Investigation of New Anti-Bacterial Products," *Statistics in Medicine*, 20:2561–2571. [329]

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, u., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., Fong, N., Gamez–Djokic, M., Glenz, A., Gordon–McKeon, S., Heaton, T. J., Hederos, K., Heene, M., Hofelich Mohr, A. J., F., H., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D. M., Lei, R., Lindsay, T. A., Liverani, S., Madan, C. R., Molden, D., Molleman, E., Morey, R. D., Mulder, L. B., Nijstad, B. R., Pope, N. G., Pope, B., Prenoveau, J. M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F. D., Sherman, M. F., Sommer, S. A., Sotak, K., Spain, S., C., S., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S., and Nosek, B. A. (in press), "Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results," *Advances in Methods and Practices in Psychological Science*. [329]

Stan Development Team (2016). *rstan: The R interface to Stan*. R package version 2.14.1. [331]

Tawakol, A., Ishai, A., Takx, R. A. P., Figueroa, A. L., Ali, A., Kaiser, Y., Truong, Q. A., Solomon, C. J. E., Calcagno, C., Mani, V., Tang, C. Y., Mulder, W. J. M., Murrough, J. W., Hoffmann, U., Nahrendorf, M., Shin, L. M., Fayad, Z. A., and Pitman, R. K. (2017),

"Relation Between Resting Amygdalar Activity and Cardiovascular Events: A Longitudinal and Cohort Study," *The Lancet*, 389: 834–845. [332,333,334]

Tukey, J. W. (1995), "Controlling the Proportion of False Discoveries for Multiple Comparisons: Future Directions," in *Perspectives on Statistics for Educational Research: Proceedings of a Workshop*, eds. V. S. L. Williams, L. V. Jones, and I. Olkin, Research Triangle Park, NC: National Institute of Statistical Sciences, pp. 6–9. [334]

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. [328]

Weber-Schoendorfer, C. and Schaefer, C. (2008), "The Safety of Cetirizine During Pregnancy: A Prospective Observational Cohort Study," *Reproductive Toxicology*, 26:19–23. [329,331]

The American Statistician

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

Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < 0.05$ Universe

David Trafimow

To cite this article: David Trafimow (2019) Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p<0.05$ Universe, The American Statistician, 73:sup1, 340-345, DOI: 10.1080/00031305.2018.1537888

To link to this article: https://doi.org/10.1080/00031305.2018.1537888

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 4699

View related articles

View Crossmark data

Citing articles: 11 View citing articles

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, 340–345: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1537888

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < 0.05$ Universe

David Trafimow

Department of Psychology, New Mexico State University, Las Cruces, NM

ABSTRACT

The American Statistical Association's Symposium on Statistical Inference (SSI) included a session on how editorial practices should change in a universe no longer dominated by null hypothesis significance testing (NHST). The underlying assumptions were first, that NHST is problematic; and second, that editorial practices really should change. The present article is based on my talk in this session, and on these assumptions. Consistent with the spirit of the SSI, my focus is not on what reviewers and editors should not do (e.g., NHST) but rather on what they should do, with an emphasis on changes that are not obvious. The recommended changes include a wider consideration of the nature of the contribution than submitted manuscripts usually receive; a greater tolerance of ambiguity; more of an emphasis on the thinking and execution of the study, with a decreased emphasis on the findings; replacing NHST with the a priori procedure; and a call for reviewers and editors to recognize that there are many cases where the basic assumptions of inferential statistical procedures simply are not met, and that inferential statistics (even the a priori procedure) may consequently be inappropriate.

ARTICLE HISTORY
Received December 2017
Revised September 2018

KEYWORDS
A priori procedure; Editorial practices; Editorial changes; Journal practices

It is no secret that editorial practices can and should be improved. Some potential improvements are obvious. These include rendering data accessible for others to perform their own analyses, increased transparency, disclosure of conflicts of interests, and others. It is difficult to imagine that many researchers would disagree with these. However, there are other changes that, while not necessarily surprising, would less obviously elicit universal agreement, and might even elicit some disagreement. My aim is to introduce these less obvious changes and make the case that they nevertheless are worth discussing and possibly accepting.

1. More Consideration of the Nature of the Contribution

At first glance, this seems (a) obvious and (b) what editors and reviewers are already doing. Of course, editors and reviewers consider the nature of the contribution! But perhaps there can be more consideration with respect to the following issues.

The first issue is that there are many ways of contributing, including new theory, new application, integration, unification, new data, and others. Most journals tend to feature only two or three kinds of contributions, but editors and reviewers might be more open to contributions that are less traditional for the journal. Consider, for example, a contribution by Albert Einstein (related by Einstein 1961). Einstein presented a formula that describes the shrinkage of objects in their direction of motion, depending on their relative velocity.[1] It is entertaining to imagine a hypothetical scenario where Einstein attempts to publish the formula in a journal whose editorial practices are those prevalent in contemporary academic journals. As Einstein admitted, there is a major potential problem with the formula, which is that Lorenz published the formula long before Einstein thought of it. As journals are looking for *novel* contributions, Einstein's formula should be rejected, and would be rejected. Such are contemporary editorial practices.

But wait! There is a subtlety to consider. In addition to suggesting the formula and admitting that Lorenz got there first, Einstein also pointed out a major difference in the context of the formula, for himself and Lorenz. For Lorenz, the formula was a *special hypothesis* designed to save the data. In contrast, for Einstein, the formula was a derivation from a larger theory, relativity theory. Therefore, the formula was unifying in the Einstein context but was not unifying in the Lorenz context. From a unification perspective, the formula was an important contribution after all because it provided a demonstration of the ability of relativity theory to unify physics. This hardly can be considered trivial, and yet the formula would be difficult to publish under contemporary editorial practices (Trafimow and

[1] $L = L_O\sqrt{\frac{1-v^2}{c^2}}$, where L is the observed length, L_O is the proper length, v is the relative velocity between observer and moving object, and c is the speed of light.

CONTACT David Trafimow dtrafimo@nmsu.edu Department of Psychology, New Mexico State University, MSC 3452, P. O. Box 30001, Las Cruces, NM 88003-8001.
© 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.

Rice 2009). Of course, not all reviewers would fail to recognize such contributions, but it would be a positive development if more reviewers and editors were open to subtle, but nevertheless important, contributions, such as those exemplified by the Einstein and Lorenz episode.

Another contribution issue pertains to whether the submission potentially benefits the journal, the area, the field, or all of science. To take an example from my own area of social psychology, and the journal I edit, *Basic and Applied Social Psychology* (BASP), consider the following items:

- The submission contributes to BASP.
- The submission contributes to the attitude area.
- The submission contributes to the general field of social psychology.
- The submission contributes to all of science.

Hopefully, the bullet-listed items go well together, and a submission would contribute to the journal, the area, the field, and perhaps even to all of science.

And yet, the bullet-listed items might not go together well. For example, as an editor, I have seen reviewers recommend the rejection of manuscripts based on perceptions of too much overlap with previous literature. From a journal-centric point of view, it seems sensible to recommend rejection to avoid using up valuable journal space on a manuscript that fails to contribute much to the journal because its components are not novel. But what if, despite the lack of novelty in the components, the manuscript nevertheless potentially contributes to the area, possibly by integrating the components in a way that suggests new implications? Or better yet, as in the Einstein example, there might be a unifying principle. In such cases, what seems best for the journal (from a narrow perspective) and what is best for science are not necessarily in complete accord. I urge reviewers and editors to place more weight on what is best for science, even if it comes at a seeming cost to what is best for the journal.

As a personal example of practicing what I am preaching, consider recent BASP editorials (Trafimow 2014; Trafimow and Marks 2015, 2016). These editorials have little to do with social psychology—the substantive topic of BASP—but have a lot to do with methodological and statistical practices. From a journal-centric perspective, I arguably should not have published the editorials,[2] but a broader perspective suggests the opposite conclusion. In summary, there are many ways of contributing, and different levels and scopes of contributions; and reviewers and editors could make recommendations and decisions based on wider, more sophisticated, and more philosophically informed evaluation processes.

2. Tolerate Some Ambiguity

Having read countless reviews, little kills a submitted manuscript more assuredly than ambiguous findings. Reviewers see the ambiguities quickly, emphasize them in their reviews, and perforce recommend rejection. In turn, based on the negative reviewer recommendations, the editor inevitably decides to reject the manuscript. And the rejection makes sense. Why commit valuable journal pages to an article with ambiguous findings when the pages could be committed to an article that contains unambiguous findings? Nevertheless, I urge greater toleration of ambiguity.

Suppose that an author submits a manuscript, using a typical sample size, and the findings are either statistically significant or not. There are several issues with this scenario, to be addressed later, but one matter is particularly relevant to ambiguity tolerance. To see it, imagine that we had access to Laplace's Demon, who knows everything, including whether the sample effect size is close to the population effect size. Let us further imagine that the Demon guarantees that the nonsignificant sample effect size is close to the population effect size; that is, the population effect size is small but greater than zero in the direction hypothesized by the researcher. Assuming all is well with respect to theory and experimental design, should the finding be published?

The question is a difficult one. On the positive side, the small population effect size is in the predicted direction, and consequently may support the theory (but see Trafimow 2017 for a potential discrepancy). On the negative side, the fact that the effect size is small renders the support unimpressive, as it is easier to predict or account for small than large effect sizes, ceteris paribus (all else being equal). With only a small effect size to account for, the reviewers and editor likely can come up with alternative explanations. A way out might be to simply ask the Demon if the theory is true, but the Demon refuses to impart more information. The upshot is that it really is not clear what to believe about the theory, and all we can be sure of is a small population effect size in the "right" direction. Despite the Demon's valuable help in guaranteeing the population effect size, and even acknowledging the positive reason for publishing the manuscript, the case for acceptance is thus far weak.

But there is another consideration. The blunt fact of the matter is that most population effect sizes in research really are small, though it is not difficult to think of experiments that would result in large effect sizes. For a large effect size thought experiment, suppose participants in the experimental condition were instructed to write an X in the upper right-hand corner of their questionnaires and participants in the control condition were not. Doubtless, the population effect size would be gigantic (and the sample effect size too). The problem, of course, is that finding that participants can comply with a simple instruction would be of trivial importance at both the theoretical and applied levels. Because researchers wish to test *interesting* ideas, and interesting ideas, by their nature, engender some subtlety, the population effect sizes likely will be small. In addition, even well-tested dependent measures are not perfectly reliable and valid, and even well-tested manipulations likely do not take for every participant, thereby lowering the effect size still further from what it otherwise would be. Thus, although large effect sizes might be superior to small effect sizes, ceteris paribus; this ceteris paribus condition rarely applies. To really hammer home the general lack of applicability of the ceteris paribus condition, consider what some believe is the most important experiment in the history of science—the work by Michelson and Morley (1887)—that failed to detect the existence of the hypothesized luminiferous ether that was thought to provide the medium by which light waves could propagate, and reach Earth from the

[2] In fact, the editorials worked out well for BASP, though I had no way to know this beforehand.

stars. Today, scientists agree that there is no luminiferous ether, and so the population effect size is 0. But the population effect size of 0 does not detract from the extremely high value of the research. Small—or even zero—effect sizes can matter.[3]

The bottom line, then, is that some small effect sizes are very important. One should assess the theory, auxiliary assumptions connecting non-observational terms in the theory to observational terms in empirical hypotheses, quality of the experimental design, and implications for applications, as important components of one's evaluation of a manuscript. Because most experimental predictions (e.g., in psychology, medicine, marketing, and so on) are directional rather than being point predictions and given the foregoing assertion that many interesting ideas involve small rather than large population effect sizes, it is an inevitability that there will be ambiguity. This is a fact of life in science, and reviewers, and editors would do well to admit it and tolerate the ambiguity that comes with it. When the sample effect size is small, reviewers and editors should nevertheless accept it, regardless of *p*-values, if it is of basic or applied importance and the sample size is large enough to engender confidence in its accuracy. Moreover, the present scenario dramatically underestimates the degree of ambiguity in normal science, where researchers do not have access to Laplace's Demon. When addressing the normal science scenario in the next section, we will see yet more reason to tolerate ambiguity in the results, support for an increased focus on theoretical and design issues, and a decreased focus on how the results come out.

3. Emphasize Thinking and Execution, Not Results

Reviewers and editors often attend, insufficiently, to the fact that a *p*-value is a sample statistic.[4] It is not difficult to imagine a scenario where the same experiment is performed repeatedly to obtain a distribution of *p*-values for that experiment. Of course, in normal science, the experiment is performed only once. But the value of imagining a *p*-value distribution is that it immediately becomes clear that the obtained *p*-value is only one of the many *p*-values that could have been obtained.

In the previous section, we imagined a small population effect size, and a small sample effect size that accurately estimates the small population effect size. But in real research, the sample effect size might be large, small, or even in the wrong direction. Unfortunately, science has been dominated by null hypothesis significance testing (NHST), that uses a threshold (usually 0.05) for statistical significance. If the sample effect size is sufficiently large to render a *p*-value coming in under 0.05, the manuscript is deemed publishable (if other aspects pass muster); otherwise, it is not. Because there is a distribution of sample effect sizes, and a distribution of associated *p*-values, it should be obvious that getting lucky (picking a large sample effect size and low *p*-value from the distributions of sample effect sizes and *p*-values) greatly aids in publishing one's work. But as was made salient by a recent discussion in BASP (Hyman 2017; Kline 2017; Locascio 2017a, 2017b; Marks 2017), it should be clear that publishing predominantly the "lucky" findings results in a literature of published effect sizes that exceeds the effect sizes that would be obtained if all findings were published. Empirical support for such effect size overestimation was obtained by the Open Science Collaboration (2015), who found that the average effect size in the original cohort of published articles was 0.403 whereas it was only 0.197 in the replication cohort. To avoid scientific literatures littered with overly optimistic sample effect sizes, Locascio suggested that publication decisions should be based on the worth of the theory and the execution of the experiment, rather than on the findings. Whether or not one is willing to go as far as Locascio wishes to go is a matter for serious discussion, but it is at least possible to move in that direction, which would be good for science.[5]

An additional way to address this problem is to emphasize replications. If a finding is lucky, it likely would not hold up when submitted to replication attempts, especially not multiple replication attempts. In contrast, greater trust can be placed in findings that do replicate. Some caveats are as follows. First, findings can be unlucky as well as lucky, and so even a correct finding might not always replicate when subjected to replication attempts. Second, it is possible for lucky events to happen again, which is a reason for multiple replication attempts rather than single replication attempts. Third, there may be outside reasons (e.g., lack of money, inaccessible participant populations, and so on) why replications are not feasible with respect to particular findings. Despite the caveats, replications are desirable when practical, though what is meant by replicability can be a complicated topic (see Trafimow 2018 for a discussion).

4. Replace NHST with A Priori Thinking

Invoking Laplace's Demon again, suppose that the Demon warned us that our sample statistics would be completely unrepresentative of their corresponding population parameters. In that case, scientists likely would no longer be interested in sample statistics. Put another way, absent the Demon, the reason scientists are interested in sample statistics is because they believe that the sample statistics they intend to obtain will provide reasonable estimates of corresponding population parameters. In fact, researchers often are taught to collect the largest feasible sample size because it is well-known that, under the usual assumptions of random and independent sampling, the larger the sample, the more the sample resembles the population. The larger the sample, the more *confident* the researcher can be that the sample statistics to be obtained will be *close* to their corresponding population parameters. But if

[3] It is interesting that Carver (1933) reanalyzed the Michelson and Morley (1887) data using the null hypothesis significance testing procedure. Carver obtained a statistically significant effect! Had Michelson and Morley performed a significance test, as journal editors would insist on today, they would have concluded that they had supported the existence of the luminiferous ether. The negative consequences for science, had this happened, are incalculable. For example, the equation in Footnote 1 is an outgrowth of Michelson and Morley's disconfirmation of the luminiferous ether.

[4] Although there are many criticisms of how researchers use *p*-values to perform null hypothesis significance tests, I call the reader's attention to two recent books (Hubbard 2016; Ziliak and McCloskey 2016). Both are noteworthy because they are extremely readable, place the discussion in larger conceptual backgrounds than is typical, and provide rich contexts.

[5] One way to deal with optimistic effect sizes statistically is to use regression equations to estimate the extent to which the effect size could be expected to shrink upon a replication attempt. A potential drawback, however, is that the researcher might not know the values to instantiate into the equations.

the goal is to be *confident* that the sample statistics will be *close* to their corresponding population parameters, it makes sense to ask: "How *close* do we want to be?" and "How *confident* do we want to be that we are that close?" Given specifications of closeness and confidence, it is possible to calculate the necessary sample size. It is worthwhile to pause for a moment to emphasize that the goal is not to obtain sufficient participants to obtain a p-value of a particular size, or a confidence interval of a particular width; but rather to obtain sample statistics that the researcher can be confident are close to their corresponding population parameters. A consequence of this subtle, but nevertheless important change from typical statistical thinking, is that the expected effect size plays no part in the calculations (Trafimow 2017; Trafimow and MacDonald 2017), as will become clear with the subsequent example.

Consider the extremely simple example where there is only one group, each participant is randomly and independently sampled from a normally distributed population, and the researcher is interested in the sample mean as an estimate of the population mean. Trafimow (2017) provided an accessible proof of Equation (1), where f is the fraction of a standard deviation defined as "close," Z_C is the z-score that corresponds to the probability the researcher wants to have of being close (i.e., confidence), and n is the necessary sample size to meet the specifications for closeness and confidence:

$$n = \left(\frac{Z_C}{f}\right)^2. \tag{1}$$

To see the implications, suppose that the researcher wishes to have a 95% probability that the sample mean to be obtained will be within 0.2 standard deviations of the population mean. The z-score that corresponds to a 95% probability is 1.96, and so the result is as follows: $n = \left(\frac{1.96}{0.2}\right)^2 = 96.04$ Rounding to the nearest whole number, the researcher needs to collect 96 participants to have a 95% probability of obtaining a sample mean within 0.2 standard deviations of the corresponding population parameter. Note that the researcher does not need to know or guess the population mean, standard deviation, nor effect size to make the calculation. Also, if the sample size is not under the experimenter's control, and hence is given, algebraic manipulation of Equation (1) can be made to render the probability of obtaining a sample mean, given the sample size, within the desired distance of the population mean.

Thus, using the a priori procedure, the researcher makes specifications of closeness and confidence before collecting data, finds the necessary sample size, collects that sample size or a larger one, calculates the sample statistics of interest, and then comes the easy part. The researcher simply believes that the sample statistics accurately estimate the corresponding population parameters. The reason such belief is justified is that the researcher sets up the conditions, a priori, for such belief. Using a priori equations, of which Equation (1) is a simple example, costs practically nothing but can provide a useful way to decide on sample sizes. Or from the viewpoint of a reviewer or editor who is faced with already collected data, these equations can be used to estimate, in a posteriori fashion, how likely the sample statistics presented are to be good estimates of the corresponding population parameters at the desired level of precision.

Lest the a priori procedure be confused with traditional power analysis, two obvious differences are worth relating. First, the goal of traditional power analysis is to find the sample size needed so the researcher can have a good chance of obtaining a statistically significant finding. Regarding NHST, it is interesting to ponder the overwhelming tendency of speakers at the SSI (2017) to eschew it. But if we are to reject NHST, then power analysis designed to facilitate NHST does not make much sense either. In contrast, the goal of the a priori procedure is to find the sample size needed to obtain sample estimates of population parameters at specified levels of precision and confidence. Second, the effect size—or expected effect size—plays a crucial role for power analysis computations whereas it plays no role whatsoever for a priori procedure computations. Philosophically, this is because the goal is to obtain sample statistics that accurately estimate population parameters and not to obtain sufficiently low p-values to reject hypotheses. Mathematically, this is because the standard deviation cancels out in the derivation of a priori equations (see Trafimow 2017 for an accessible proof). Given the effect size lesson exemplified by the Michelson and Morley (1887) experiment, it is a plus that the a priori procedure is uninfluenced by the effect size or expected effect size.

A final point with respect to the a priori procedure pertains to the replication crisis that has become such a concern as of late. Interestingly, this procedure suggests the possibility of computing the probability of replication, also in an a priori way, though under an assumption of ideal replication conditions. Imagine an ideal replication that is the same as the original experiment, with the sole exception of randomness. In that case, the probability that both replications will have sample statistics within the specified distances of the corresponding population parameters, at the sample size used, is simply the square of the probability that this will be so for one of the experiments (Trafimow 2018). A caveat, however, is that the probability of replication, when computed in this way, should be considered an ideal probability. In real science, the replication will not be exact (Hubbard 2016), and so the a priori probability of replication should be considered an upper bound (Trafimow 2018). But this is nevertheless useful. When confronted with a manuscript, reviewers and editors can run their own calculations using a priori equations. If even the ideal probability of replication is low (and it often is low at typical sample sizes), the reviewers and editors can be assured that the real probability of replication is even lower than that. And if probability of replication is important for the journal, a low ideal probability of replication might become a strong criterion for rejection. Going the other way, if the ideal probability of replication is impressive, there is at least some reason for optimism, and reviewers and editors can use it as a starting point for their subjective judgments of how likely the findings would be to replicate in the real scientific universe.

There is one last point to be made pertaining to the a priori procedure. Equation (1) provides the simplest possible case, but the simplicity of the example does not imply that the procedure will not work for more complex cases. On the contrary, Trafimow and MacDonald (2017) expanded the procedure to work for however many groups the experimenter wishes to use. In addition, Trafimow, Wang, and Wang (2018) have expanded

the a priori procedure to apply to skewed distributions and to locations as well as means. Equations not yet published also have been invented concerned with estimating differences between means or locations as opposed to the means or locations themselves, and for estimating standard deviations, scales, and shape parameters.[6] Work is in progress that pertains to the estimation of correlation coefficients, proportions, and others.[7] Consequently, we expect that soon, researchers will be able to make a priori calculations for practically anything they find to be of interest, under a variety of possible assumptions. And, of course, reviewers and editors also will be able to make these calculations to aid in their manuscript evaluations.

5. The Assumptions of Random and Independent Sampling Might Be Wrong

Berk and Freedman (2003) argued that it is the rare study that uses sampling that is completely random and where each participant sampled is independent of the other participants sampled. Without making assertions about exactly how often these assumptions are importantly violated, there can be little doubt that the violation prevalence is considerable. In that case, no inferential statistical procedures, including even the a priori procedure discussed in the foregoing section, are completely valid. When no inferential statistical procedures are valid, reviewers and editors may be doing a disservice to science by either allowing, or even insisting on, researchers performing them. There are times when science is better served by reviewers and editors simply admitting that the assumptions of random and independent sampling are inapplicable. In those cases, an option is for authors to report only descriptive statistics, and reviewers and editors should be open to that.[8] The notion that inferential statistical procedures may sometimes, and even often, be inappropriate, may be tough medicine for reviewers and editors to swallow. But the medicine nevertheless is therapeutic. Another option is to use methods for addressing violations of independence or random sampling (e.g., Liu and Singh 1995), while keeping in mind that these have their own assumptions and drawbacks.

That inferential statistical assumptions are practically never perfectly accurate is well-known, but perhaps the imperfections are addressed by the famous quotation by Box and Draper (1987): "Essentially, all models are wrong, but some are useful" (p. 424). But at least with respect to p-values, this quotation does not work as well as some might hope. To see why, consider that if one admits that the model, including all assumptions, is never exactly right, then that means that the model is always wrong, though it may be close to right and may even be close enough to right to be useful. Well, then, if the model is known wrong as Box and Draper admit, why compute a p-value to obtain evidence against a known wrong model? The model is wrong no matter what p-value is obtained! Nor does the p-value give a valid indication of how close the model is to being right nor to how useful it is. Then, too, the p-value does not validly indicate the amount of evidence against the model being close to right, only against the model being exactly right.

The issue of not being exactly right is not necessarily as problematic with respect to alternative inferential procedures. For example, consider again the a priori procedure. Although the model again likely is not exactly right, remember that the goal is not to test a (known wrong) model but rather to obtain a sample size that engenders confidence that the sample statistics to be obtained are precise estimates of corresponding population parameters. Well, then, suppose that the model is close to being right, though it remains wrong. The slight wrongness of the model implies that a priori calculations will result in the researcher collecting a sample size that either is slightly insufficient to meet objectives or slightly in excess. In the latter case, little harm is done except that the researcher undergoes a bit more effort than is needed and enjoys the compensation of extra precision. In the former case, the researcher will have slightly too much confidence in the precision of the sample statistics. These problems need not be fatal, whereas for p-values, they are fatal as the known wrongness of the model really does render p-values pointless at best, and harmful at worst.

6. Conclusion

Like many people who work, academic researchers are interested in their careers. Because promotions in academia depend largely on publications, academic researchers are strongly motivated to publish. Thus, journal editors have much power. If journal editors insist on practices that are good for science, authors will comply, and science will benefit accordingly. In contrast, a failure to insist on practices that are good for science, and even an insistence on practices that are bad for science (such as NHST), not surprisingly work to the detriment of science. In some cases, it is obvious what constitute good or bad scientific practices, and in those cases, editors likely will insist on good ones. I know of no journal editors who would encourage scientific practices that they believe to be detrimental to science. But what is good or bad for science is not always obvious. The five recommendations that provided the present focus are some of them, though there are many more. In addition, however, consistent with the focus of the SSI session on journal editing, there is a general need to greatly expand the discussion of editorial issues. As academia becomes increasingly ruled by a publish-or-perish ethos; and journal editors consequently continue to gain in collective influence, though not necessarily in wisdom; it is increasingly vital to improve the evaluation procedures of reviewers and editors. Hopefully, the five present recommendations provide a useful continuation of the focus of the SSI session on editorial practices in a post $p < 0.05$ scientific universe.

References

Berk, R. A., and Freedman, D. A. (2003), "Statistical Assumptions as Empirical Commitments." in *Law, Punishment, and Social Control: Essays in*

[6] I take this opportunity to thank Tonghui Wang, Cong Wang, and Hunter Myüz for their high-quality help, without which these advances would not have been made.

[7] I again thank Tonghui Wang, Cong Wang, and Hunter Myüz for their invaluable aid.

[8] In general, whether inferential statistical procedures are valid or not, a strong case could be made for expanded descriptive statistics, possibly accompanied by improved visual displays. These points have been elaborated by Valentine, Aloe, and Lau (2015).

Honor of Sheldon Messinger, eds. T. G. Blomberg and S. Cohen (2nd ed), New York: Aldine de Gruyter, pp. 235–254. [344]

Box, G. E. P., and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, New York: Wiley. [344]

Carver, R. P. (1993), "The Case Against Statistical Signi?cance Testing, Revisited," *Journal of Experimental Education*, 61, 287–292. [342]

Einstein, A. (1961), *Relativity: The Special and the General Theory*, Robert W. Lawson, Trans., New York: Crown Publishers. [340]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Los Angeles, CA: Sage Publications. [342,343]

Hyman, M. (2017), "Can 'Results Blind Manuscript Evaluation' Assuage 'Publication Bias'?" *Basic and Applied Social Psychology*, 39(5), 247–251. [342]

Kline, R. (2017). "Comment on Locascio, Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39(5), 256–257. [342]

Locascio, J. (2017a). "Results Blind Publishing," *Basic and Applied Social Psychology*, 39(5), 239–246. [342]

——— (2017b), "Rejoinder to Responses to 'Results Blind Publishing,'" *Basic and Applied Social Psychology*, 9(5), 258–261. [342]

Liu, R. Y., and Singh, K. (1995), "Using i.i.d, Bootstrap Inference for General non-i.i.d, Models," *Journal of Statistical Planning and Inference*, 43, 67–75. [344]

Marks, M. J. (2017), "Commentary on Locascio 2017," *Basic and Applied Social Psychology*, 39(5), 252–253. [342]

Michelson, A. A., and Morley, E. W. (1887), "On the Relative Motion of Earth and Luminiferous Ether," *American Journal of Science, Third Series*, 34, 233–245, available at *http://history.aip.org/exhibits/gap/PDF/michelson.pdf* [341,342,343]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [342]

Trafimow, D. (2014), "Editorial," *Basic and Applied Social Psychology*, 36, 1–2. [341]

——— (2017), "Using the Coefficient of Confidence to Make the Philosophical Switch from a Posteriori to a priori Inferential Statistics," *Educational and Psychological Measurement*, 77(5), 831–854. [341,343]

——— (2018), "An *a priori* Solution to the Replication Crisis," *Philosophical Psychology*, 31, 1188–1214. [342,343]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [341]

——— (2016), "Editorial," *Basic and Applied Social Psychology*, 38, 1–2. [341]

Trafimow, D., and MacDonald, J. A. (2017), "Performing Inferential Statistics Prior to Data Collection," *Educational and Psychological Measurement*, 77(2), 204–219. [343]

Trafimow, D., and Rice, S. (2009), "What If Social Scientists Had Reviewed Great Scientific Works of the Past?" *Perspectives in Psychological Science*, 4, 65–78. [341]

Trafimow, D., Wang, T., and Wang, C. (2018), "From a Sampling Precision Perspective, Skewness is a Friend and Not an Enemy!" *Educational and Psychological Measurement*, 79, 129–150. [343]

Valentine, J. C., Aloe, A. M., and Lau, T. S. (2015), "Life After NHST: How to Describe Your Data Without 'p-ing' Everywhere," *Basic and Applied Social Psychology*, 37, 260–273. [344]

Ziliak, S. T., and McCloskey, D. N. (2016), *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*, Ann Arbor, MI: The University of Michigan Press. [342]

The American Statistician

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

The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration

Joseph J. Locascio

To cite this article: Joseph J. Locascio (2019) The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration, The American Statistician, 73:sup1, 346-351, DOI: 10.1080/00031305.2018.1505658

To link to this article: https://doi.org/10.1080/00031305.2018.1505658

© 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: 4747

View related articles

View Crossmark data

Citing articles: 2 View citing articles

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, 346–351: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1505658

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration

Joseph J. Locascio

Neurology Department, Harvard Medical School, Boston, MA

ABSTRACT

The author has previously proposed results blind manuscript evaluation (RBME) as a method of ameliorating often cited problems of statistical inference and scientific publication, notably publication bias, overuse/misuse of null hypothesis significance testing (NHST), and irreproducibility of reported scientific results. In RBME, manuscripts submitted to scientific journals are assessed for suitability for publication without regard to their reported results. Criteria for publication are based exclusively on the substantive importance of the research question addressed in the study, conveyed in the Introduction section of the manuscript, and the quality of the methodology, as reported in the Methods section. Practically, this policy is implemented by a two stage process whereby the editor initially distributes only the Introduction and Methods sections of a submitted manuscript to reviewers and a provisional decision regarding acceptance is made, followed by a second stage in which the complete manuscript is distributed for review but only if the decision of the first stage is for acceptance. The present paper expands upon this recommendation by addressing implications of this proposed policy with respect to statistical consultation and collaboration in research. It is suggested that under RBME, statisticians will become more integrated into research endeavors and called upon sooner for their input.

ARTICLE HISTORY
Received February 2018
Revised April 2018

KEYWORDS
Publishing; Publication bias; Results-Blind; Significance testing

1. Introduction

In earlier articles (Locascio 1999, 2011, 2015, 2017a, 2017b), I suggested and defended an approach to evaluating manuscripts submitted to scientific journals (especially in behavioral and medical sciences), I referred to as "Results Blind Manuscript Evaluation" (RBME). In the present article, I would like to expand on implications of RBME with respect to statistical inference and methodology, focusing mostly on the role of statistical consultants and collaborators in research and in preparing reports thereof. I will first briefly summarize what RBME is and preliminary to that, the problems it is intended to ameliorate.

2. Problems in Scientific Publication and Statistical Inference

Part of the impetus for this special issue of *The American Statistician* is the intensified controversies that have recently arisen regarding problems in statistical inference and science publishing, especially the interrelated concerns of publication bias, the overuse and misuse of null hypothesis significance testing (NHST), and claimed irreproducibility of published scientific findings.

Publication bias, as meant here, is the widely recognized and documented practice in which, all else being equal, null findings of even well conducted studies are not as likely as positive findings of studies to be reported and published in scientific journals, either because manuscripts reporting such are not as likely to be accepted by journals (Fanelli 2010; Van Assen et al. 2014) or because of an inhibition among investigators to even write up or submit such reports to journals because of this assumed bias at the journals (the "file drawer" problem discussed by Rosenthal 1979). As the result of publication bias, to the degree that true effects are negligible and this bias is operative at the author submission and/or journal acceptance stage, the literature becomes contaminated with a misleadingly higher proportion of false positive findings than would occur by chance alone if there were no such bias. Further, awareness of publication bias can create a temptation among investigators to deliberately doctor study results to show positive findings in order to get the publications seen as vital to career advancement. And what may be far more commonly practiced are the many ways that unintentional and unconscious processes and choices may be at play in how researchers collect and analyze their data in order to obtain the "statistically significant" findings they feel they need for publication (e.g., "p hacking" and "opportunistic biases"; DeCoster and Sparks 2015). The implications of publication bias in meta-analysis especially has often been cited, and a number of methods that attempt to assess and correct for it have been proposed (e.g., "funnel plots"; Begg and Mazumdar 1994; Egger et al. 1997; Light and Pillemer 1984), though a better solution would seem to be to avoid the bias at its source.

"Statistically significant" results as obtained by NHST have to a great degree become collectively accepted in behavioral and medical sciences as the gold standard indicator of what

CONTACT Joseph J. Locascio, Ph.D. *JLocascio@partners.org* Harvard Medical School, and Memory and Movement Disorders Units, Massachusetts Alzheimer's Disease Research Center, Neurology Dept., Massachusetts General Hospital (MGH), 50 Staniford St., Office 560-H, Boston, MA 02114.

© 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.

constitutes the positive findings that journals are biased toward accepting. However, serious and valid concerns about the overuse and misuse of NHST have been cited by statisticians for decades. (There is no space to go into these criticisms again here, but you can see, e.g., Cohen 1994, Hunter 1997, and Seife 2015, for a sampling of them.) More highly regarded alternatives to NHST have been widely recommended (e.g., confidence intervals, effect size estimates, Bayesian statistics, graphical analysis, etc.) though none of these has come close to replacing NHST as the workhorse of statistical inference.

There has also been recent concern about a recognized and documented frequent failure to replicate findings in scientific reports ("the crisis of irreproducibility"; see, e.g., Open Science Collaboration 2015). The reasons for these failures have been heavily debated with publication bias being commonly alleged to be one of the factors responsible, which includes biases at the investigator level whereby conscious and unconscious stratagems are employed to get "significant" results. It has also been claimed that, specifically, the bias of publishing only results with NHST p-values below the widely accepted 0.05 or 0.01 cutoffs has made much failure to replicate inevitable given statistical "regression to the mean" (Trafimow and Earp 2017).

3. Results blind Manuscript Evaluation (RBME)

As a method of mitigating the problems of statistical inference and scientific publication noted above, I have suggested that behavioral science and medical journals consider instituting a policy of results blind manuscript evaluation (RBME) of the suitability of submitted manuscripts for publication. For a detailed description of RBME see Locascio (2017a, 2017b) and also Locascio (1999, 2011, 2015). By RBME, I mean that in deciding whether to publish a manuscript, journal reviewers and editors would give no weight to reported results in making this judgment. Reports would be judged *exclusively* on the basis of the perceived: (1) *importance* of the substantive issues addressed, which is generally communicated in the "Introduction" section of the manuscript, and (2) soundness of the *methodology* employed, which would include among other things, appropriate: materials, measurement methods with e.g., up-front reliability/validity estimates, subject selection, study layout/experimental design, proposed data analysis techniques, and sample size, all conveyed in the "Methods" section. Regarding appropriate data analysis methodology, many proposed methods recommended as alternatives to NHST could potentially be employed, instead of or in addition to NHST. Use of NHST, if judged applicable, and correctly used and interpreted, could possibly be permissible as providing one supplementary piece of evidence that might be pertinent to research questions, at the discretion of reviewers. However, importantly, obtaining an NHST p-value less than an arbitrary cutoff would no longer be a necessary precondition for publication – it can't be because neither p-values *nor any other result* would be considered in the decision to publish. This RBME editorial policy would be made explicit and prominently stated as a part of journals' guidelines for authors so that investigators considering manuscript submission would be fully aware of it and not mistakenly censor or inhibit their own submissions because null findings were obtained, and they think lack of positive or statistically significant findings precludes publication or lessens the likelihood of it.

Note that I am recommending results blind *evaluation* of manuscripts for suitability for publication, not a complete results blind *review*. Reported results sometimes indicate some aspects of methodology and data analysis techniques which could not be fully elaborated in the Methods/Data Analysis sections and reflect good or poor methodology. Further, Results and Discussion sections certainly have to be edited and reviewed prior to the manuscript going to print, for presentation as well as for substantive reasons. Data analysis in the Results section has to be checked for execution, and correct interpretation of results in the Discussion section has to be assessed as well as a statement of limitations of the study. By "results blind evaluation", I merely mean that the nature of the observed findings of a study, or whether or not the study found any effects at all, should be given little or no weight and have no direct bearing on whether the manuscript is to be *accepted for publication*.

As a method of practically implementing RBME, I recommend a *two stage* procedure whereby authors would submit a manuscript in its entirety to a journal, just as they normally do now, seamlessly without change. But, upon receipt of it, the editor would distribute it to reviewers after having stripped it of everything except the Introduction and Methods sections (and any appendices or supplementary information relevant to methods), in order to obtain a "first stage *evaluation*" of suitability for publication. If the decision of reviewers at this stage is to reject, the authors would be informed that their manuscript was rejected and why. However, manuscripts passing this first stage would then have their Abstract, Results and Discussion sections re-attached by the editor and submitted to reviewers for a "second stage *review*". A manuscript surviving the first stage as acceptable or judged as requiring no more than minor modifications/revisions would be considered "conditionally accepted" for publication, that is, conditional on the absence of any new evidence of flawed methodology discovered in the second stage sufficiently grievous to override the first stage decision. Note the second stage review merely serves a disconfirmatory or veto function, but where still no weight is given to what the reported results are per se, in the decision to publish. After completion of the second stage, a formal decision is finalized for acceptance, acceptance with revision, or rejection (in the unlikely event rejection is decided at the second stage), and if not rejected, recommendations for minor revision, editing, wording and cosmetic alterations, etc. would be suggested or made as pertinent. Such a process would arguably involve about the same amount of labor for reviewers as is currently the case, and possibly less, given that reviewers will sometimes only have to review half a manuscript if the first stage evaluation results in rejection. If it is accepted during the first stage, and the same reviewers who conducted the first stage evaluation, perform the second step review, there is no more additional workload than had they reviewed the entire manuscript at once from the onset.

RBME would I believe ameliorate (though not entirely solve) all three of the problems with scientific publication cited above, that is, publication bias, misuse of NHST, and irreproducibility. That it would reduce publication bias is essentially self-evident. There would be no results based bias at the journal because reviewers cannot decide acceptance based on results if

they do not know what the results were. There would be little or no self-censoring bias among investigators because, being fully informed of the RBME journal policy, they would presumably have no inhibition in reporting the null findings of a well conducted study, knowing that the absence of positive findings per se will in no way reduce their chances for publication and getting just credit and recognition for their hard work.

Although RBME is somewhat orthogonal to many of the claimed problems with NHST, indirectly, many of these endlessly criticized and lamented issues will be mitigated or made nonissues. Although NHST need not necessarily be explicitly banned by anyone, authors would feel no compulsion to employ that particular methodology, or any other for that matter, beyond what they deem as most relevant, fitting, appropriate, methodologically sound, justifiable, and defensible for their study. Controversy will no doubt continue over whether NHST has some limited utility in science given proper circumstances of relevancy, and correct application, implementation and interpretation; however the widely criticized overuse, misuse and over-emphasis of NHST p-value cut-offs as the predominant gate-keeper of communication of scientific results will end. Although cautions will still be cited (and should be) regarding problems with misuse of NHST, the *practical imperative* of NHST, that is, the widely perceived necessity of employing NHST and getting $p < 0.05$ for publication, which implicitly fuels and exacerbates many of these problems, will be over. Researchers will no longer feel any particular need to get $p < 0.05$ or $p < 0.01$ or any other p-value. However, if they are using NHST, it will now become important for them to defend that they are using it correctly and, unlike before, justify why they are using it at all.

The problem of irreproducibility of results will also presumably be mitigated as well because there will likely be a reduction in published false positives given that positivity of findings per se will no longer consciously or unconsciously influence any decision to report study results or publish them. Null findings judged to be of equal validity to any positive results because they were based on equally sound methodology applied to equally important research questions will have the same chance of being published as the positive results. Thus, the scientific literature will convey a representative, balanced sample of findings containing a proportion of negative and positive findings duly reflective of what is actually likely to be true. Any future, well conducted replication work would likely reproduce what has already been generally found by previously published well conducted studies, and contribute to accumulating supportive evidence in further reinforcing truthful scientific theories and refuting false ones.

RBME is not claimed to be panacea for all problems of statistical inference and publication. Some forms of bias in publishing will still remain, e.g., bias by reviewers and editors as to what is an "important" research question and good methodology, and notably biases that do not occur at the point of journal *reviewer evaluation*. For example, bias at point of author *submission* can still occur to the extent that the decision to submit is based on study results being supportive of personally favored and espoused theories, or their being in line with monetary interests, or because of other nonscientific incentives. A pharmaceutical company may only submit reports of studies whose results are seen as favorable to their product. Or investigators may only submit reports of studies that ostensibly confirm their pet views. Pre-registering studies help avoid some of these problems, but of course they involve more elaborate procedures than RBME which is more practical as a widespread journal practice. Further, publication bias toward positive results can still be exercised by the journal *editor* with the pre-decision to decline to forward a manuscript for further review. However, in my experience, a "preemptive rejection" such as this is usually due to a manuscript not meeting author guideline requirements or because the substantive content of the paper is judged to clearly not be a good fit for the given journal. An editor who has already made the decision to adopt an RBME policy for the journal, would not seem likely to then consciously turn down manuscripts based purely on results in direct contradiction of the policy. And finally, bias contingent on observed results is still possible to some extent by reviewers at *stage 2* of RBME. However, under RBME this can happen only *after* a reviewer has already committed to a conditional acceptance for publication, a decision that could not have been based on results because they were not seen. Results can influence the reviewer's decision regarding publication only in a *negative* manner, that is, to subsequently want to reject that for which one has already formally stated an acceptance. The results wouldn't be reviewed without the initial acceptance. Bias based on results cannot work in an *affirmative* manner, that is, results cannot influence the reviewer to accept a report based largely on results of a study that was viewed as otherwise not worthy of publication. Furthermore, by the rules of RBME, a second-stage rejection will require the reviewer to justify to the editor that his/her change of mind is based on specifically *methodological* problems discovered in the Results or Discussion sections, problems serious enough to override the reviewer's original decision and not being amenable to revision.

There are other limitations of RBME, many of which are stated and addressed in detail in Locascio (2017a, b). For example, some serious methodological flaws may only be evident in the Results and Discussion sections of manuscripts. However, rejection at the second stage of RBME, noted above, would hopefully catch those. Further, not every contingency can be anticipated. For example, it's conceivable that a compelling result relevant to a medical emergency might supersede minor methodological considerations. I do not advocate that RBME be enforced rigidly without occasionally considering, on a case-by-case basis, possible exceptional circumstances, and there is nothing to preclude authors from arguing in a cover letter to the editor such extenuating circumstances. Lastly, one of the positive consequences of RBME would be to accentuate the need for a sample size adequate to address a study's goals, but this may be seen as unfairly favoring large universities and medical centers with access to large subject pools. In a sense, this is unfortunate, but it would provide an incentive for researchers to develop resources for greater participant access, and besides, if we have to be unfair in some way, is being unfair in favoring better studies so bad? I can think of worse ways of being "unfair".

Naturally, because of these and possibly other unforeseen practical and theoretical problems with RBME, it is recommended that RBME be introduced in a gradual, experimental manner with on-the-fly modifications as might be warranted.

4. How RBME Would Impact Statistical Consultation and Collaboration

Researchers want to do good research and they want to get published. Unfortunately, our existing publication systems sometimes put these two goals at odds. With RBME, I believe they would become better aligned. Claimed findings being reported would no longer per se have any bearing on publication and dissemination across the scientific community. This criterion for publication would instead be replaced by whether what is being reported is likely to be true, and the latter depends mostly on sound methodology, described largely in the Methods section of manuscripts. (The judged substantive importance of research questions addressed in a study, conveyed for the most part in the Introduction section, which is presumably a factor in publication now, would also continue to be so under RBME.) The switch of emphasis from *what* is claimed to have been found to *how* it was found, and from what was claimed as the answer to a question, to what the question was and how accurately it appears to have been answered in determining potential for publication, would have a profound impact on the role of statistical analysis, methodology, and statistical and methodological consultation/collaboration in research. At least the first provisional decision regarding acceptance for publication can no longer be based on the results of a study because they were not looked at in making the decision. The importance of the described methodology would, however, become magnified. It would now be critical that the study's methods are sound, thoroughly described and well justified, including the proposed data analysis techniques to be employed, and justification for the sample size. This means statisticians and methodologists will become increasingly needed as a very practical necessity, even if, regrettably, not for more elevated reasons.

4.1. Statisticians Would Become More Involved in Research Studies

I am a member of a university based biostatistics consulting group, which provides statistical consultation and analysis assistance for studies conducted across research facilities throughout a large metropolitan area. In this role, it is more often the case than not, that I'm not merely asked to answer specific questions on an otherwise fully developed research study. Rather, oftentimes, the investigators have a somewhat vague idea for a study they wish to conduct, or are articulating it poorly, and I find myself trying to solve an equation in two unknowns. First, I need to clarify for myself, and for the investigators, what exactly the research question and hypotheses are, and then having done that, translate this into the specifics of their study situation, and, finally, recommend statistical tools that seem to be a good fit for addressing those questions and hypotheses in light of anticipated or collected data. Before long, I feel myself, often with mutual consent, being pulled into the guts of the study, not merely serving on the periphery to answer a few specific technical statistical questions. Without my having asked, I'm sometimes invited to be a co-author on an anticipated manuscript or even to get on board a grant.

If I assume the researchers with whom I consult are somewhat representative of many others I've never met in the field, it's clear to me that many investigators in medical and behavioral sciences could benefit from statistical collaboration, beyond just spot consultation and analysis. Under known RBME journal policies, these investigators may feel a more imperative need to seek such in-depth statistical and methodological assistance. Statisticians might be more frequently brought on board in a study as an integral part of the research team, not just an outside "technician". This change would be to everyone's benefit.

Under RBME, researchers will at least appreciate the practical benefit of integrating a statistician into their research project. As such, I predict that the statistician will often need to struggle with the investigators to get a clear idea of what the basic research questions are and help articulate them well. In the past, for this purpose, I've often found it helpful to sketch rough causal models as path analysis diagrams (or if you prefer, graphs of structural equation models, SEM, directed acyclic graphs, DAG, etc.). These models force everyone to be very explicit, specific and thorough on just what are the underlying causal mechanisms that are at the root of their hypotheses or research questions, what are the relevant exogenous and endogenous variables, and what covariates might be important and why. Once there is consensus on underlying models to be investigated, it would be primarily the statistician's task to select from his/her hopefully diverse toolbox of statistical techniques, those that seem most congruent with the study's analytic requirements.

As a statistician myself, predicting a greater need for statisticians under RBME may seem self-serving, but I truly believe that such increased involvement will ultimately be of benefit to the research community in general and the advancement of science.

4.2. Statisticians Would Become Involved in Studies Earlier

In my role as statistical consultant for research, I'm often asked for assistance very late in the game, so to speak. Someone has designed a study and collected a lot of data, and now wants to talk about statistical analysis. Or else, revisions to a manuscript submitted to a journal have been requested by reviewers including or especially regarding data analysis, and the investigators need the advice of a statistician and/or help in analysis. In these cases, oftentimes I feel what the study really needs is a new design and new data, more fitting to the purpose of the study, once I discover what that is, but it's too late for that. I try to be sensitive and do the best I can by recommending and conducting ad hoc statistical patch-ups, making clear the limitations of these methods, but these situations can be frustrating. Under RBME, investigators will know in a very practical sense that they must seek needed statistical help *early* in the research process, i.e., as they begin to design a study, if possible before data are collected, or at least before they submit a manuscript for publication, because under RBME they know the proposed research design and data analysis methods described in their report will heavily influence the decision to publish or not. The obtained results will not be looked at in the initial decision and have little or no influence on it. Being informed by an RBME rejection that there are fundamental problems in a study's design, may unfortunately sometimes be too late for any corrective action to be taken for that particular study, but the situation will at

least be a learning experience for future reference and promote greater and more timely care in subsequent work.

The statistician will also more likely become involved in a formal justification of sample size at the beginning of a study. This justification would not necessarily entail conventional power analysis, which assumes NHST, but could also involve estimating the sample size needed for a given pre-specified precision in estimating effects. There are also other formal methods of deriving sample sizes that avoid standard power analysis and NHST (see, e.g., Trafimow and MacDonald 2017). To whatever degree possible, the statistician should also monitor, oversee, and help with preparing, proofing, checking (with graphs, programming checks, etc.), and cleaning the data set, right from the beginning, instead of finding out accidentally, much later after analyses are run, that there are obvious data level problems.

In fairness to principal investigators who are not statisticians, the potential of RBME to require statistical advice earlier in the study development, may not be seen as a boon to those investigators, especially if funds for statistical consulting are not available until later in the process. However, this may create a greater impetus for investigators to sharpen their knowledge of research design and data analysis methods, including power analysis. In this internet age, there are a proliferation of good online tutorials and videos for self-education in these areas, that would hopefully provide a groundwork until formal statistical assistance can be procured.

4.3. The Statistical Analysis Section of Manuscripts Would Expand

Under RBME, authors of research manuscripts will be fully aware that the decision to publish their paper will be heavily influenced by their reported data analysis strategy. As a result, I believe it will be quite likely that the portion of the Methods section that describes statistical analysis methods will become more expanded and/or, if space limitations are a problem, more detail will be provided in an associated appendix or supplementary materials to that end (to also be reviewed in stage one of RBME). More attention will be paid by investigators to improving the quality of the statistical methodology and reporting it thoroughly. This will mean that investigators will first need to be clear, explicit, and specific about what their research questions and/or hypotheses are, and then in consultation with statisticians and methodologists, make sure they choose a research design appropriate to addressing those questions and hypotheses, which in turn should partly dictate the kinds of data analyses they need. A statistical analysis section might be further subdivided into subsections describing (a) research questions and hypotheses, (b) research design, (c) sample size justification, (d) data analysis techniques, and (e) assumptions of analysis methods and checks to verify they were met. On a related point, the section describing measurement tools employed in the study, will have to give special attention to establishing the reliability and validity of measures, preferably by using data from the current study, instead of or in addition to assessments reported in the literature. Note that results of biometric/psychometric analyses, as well as other preliminary "results", e.g., tests to show covariates were well balanced across study groups, manipulation checks, treatment implementation checks, etc. might more appropriately go in the Methods section to be reviewed in the first stage of RBME. Such information bears on the quality of the methodology, not the answer to the research questions. An expanded data analysis section such as described of course presupposes more care in data analysis and active consultation/collaboration by statisticians and research methodologists.

Authors submitting a manuscript to a journal should be fully aware of a journal's RBME policy, given that it should be explicitly stated and emphasized in the journal's guidelines for authors. Thus, authors should be motivated to explain more of the details of their data analysis methods, and other design issues, e.g., how attrition or nonresponse is to be handled, in the Methods-Data Analysis section, not just in the Results. This should lessen the likelihood of reviewers discovering methodological problems for the first time at Stage 2 of RBME.

4.4. Statistical Knowledge and the Role of Statisticians as Informal Educators Would Become more Important

Under RBME, I believe technical issues such as controversies regarding the use of Bayesian versus frequentist approaches to statistical inference will no longer be debated only among statisticians and/or in exclusively statistical journals whose contents seem esoteric to many other researchers. Investigators will of necessity have to at least familiarize themselves with core issues of current statistical controversies and form some educated opinions. A statistician can help them do that in a role of informal educator and mentor. As a statistical consultant, I try to informally teach substantive investigators the basics of what I'm recommending or doing as an analyst. I never "dump" esoteric printouts on people with whom I consult, without offering and recommending to them that I walk them through the statistical output, at least the first time through for new output they haven't seen before. I also try to provide information-rich graphical illustrations when possible that show in a more salient and intuitive way an otherwise very technical result (e.g., a graph illustrating a model predicted interaction effect or a curved quadratic function of predicted values, rather than just the corresponding regression coefficients). Many of these graphs I produce are for internal use only and are never published or seen by outside audiences. Further, I write descriptions for investigators of statistical methods I'm using and results I've obtained that are understandable to nonstatisticians, and the nonstatistical audience who may ultimately also read them, without omitting important statistical details. Aside from informal education, with RBME, statisticians may also become in more demand as formal teachers in academic settings.

4.5. Statisticians would become more Needed by Journals to Review Manuscripts

Since under RBME, the initial decision regarding acceptance for publication will be heavily based on the statistical analysis methods proposed in the Methods section (among other things included under Methods and the Introduction), "statistical reviews" of manuscripts at journals will become more often a necessity. Statistical considerations must be addressed such as justification of sample size, whether the statistical methods are a good fit for the study objective, the design and data, whether

statistical assumptions are going to be checked, and how, justification of reliability and validity of measures, justification of use of NHST and/or a Bayesian approach, etc. I am a member of the statistical review boards for two academic journals and I often find myself requesting large scale makeovers of the design and data analysis of a study, suggesting to me that there was no involvement by a statistician in the study. Under RBME, authors will know that reviewers with advanced statistical training will likely be examining their submitted manuscript and will have a big say in deciding whether to publish it or not.

5. Conclusion

In addition to other cited benefits of RBME in mitigating publication bias, problems with NHST, and irreproducibility of research findings (Locascio 2017a, 2017b), I believe RBME will provide an inducement to better quality of scientific research methodology and data analysis, greater self-education of researchers in statistics and methodology, and/or increased collaboration between principal investigators and statisticians. The emphasis of research reports will shift to a more equitable balance between reporting what was found and establishing that what is reported is likely to be true. And this shift will necessitate greater input of statisticians in research projects. As technology advances, specialization tends to do so also, and it soon may become commonplace to include a statistical and methodological specialist on research teams for virtually all studies. Most researchers cannot be expected to keep abreast of all the rapid and complex advances in their substantive field while simultaneously staying up to date on the latest equally rapid and complex advances in data analysis methods. RBME should fit right into these general trends.

Acknowledgements

Thanks to Dr. Ron Wasserstein for suggesting the idea for this paper, i.e., implications of RBME on statistical consultation.

ORCID

Joseph J. Locascio http://orcid.org/0000-0003-3439-1209

References

Begg, C. B., and Mazumdar, M. (1994), "Operating Characteristics of a Rank Correlation Test for Publication Bias," *Biometrics*, 50, 1088–1101. [346]

Cohen, J. (1994), "The World is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [347]

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., and Sparks, C.W. (2015), "Opportunistic Biases: Their Origins, Effects, and an Integrated Solution," *American Psychologist*, 70, 499–514. [346]

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997), "Bias in Meta-Analysis Detected by a Simple Graphical Test," *British Medical Journal*, 315, 629–634. [346]

Fanelli D. (2010), "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data," *PLoS ONE*, 5, e10271. doi:10.1371/journal.pone.0010271 [346]

Hunter, J. E. (1997), "Needed: A Ban on The Significance Test," *Psychological Science*, 8, 3–7. [347]

Light, R. J., and Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research*, Cambridge, MA: Harvard University Press. [346]

Locascio, J. J. (1999), "Significance Tests and 'Results-Blindness,' " *American Psychological Association (APA) Monitor*, May, 1999, pg. 11. [346,347]

——— (2011), "A Statistical Suggestion," *Scientific American Magazine*, Letters, 305, 8–10. [346,347]

——— (2015), "Psychology and Statistics: Readers Respond to the BASP Ban on *p*-Values," *Significance Magazine*, 12, 35–36. [346,347]

——— (2017a), "Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39, 239–246. [346,347,348,351]

——— (2017b), "Rejoinder to Responses to 'Results Blind Publishing'," *Basic and Applied Social Psychology*, 39, 258–261. [346,347,348,351]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, 943. [347]

Rosenthal, R. (1979), "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin*, 86, 638–641. [346]

Seife, C. (2015), "Statistical Significance," in *This Idea Must Die: Scientific Theories That Are Blocking Progress*, ed. J. Brockman, New York: HarperCollins Perennial, pp. 519–522. [347]

Trafimow, D., and Earp, B.D. (2017), "Null Hypothesis Significance Testing and Type I Error: The Domain Problem," *New Ideas in Psychology*, 45, 19–27. [347]

Trafimow, D., and MacDonald, J. A. (2017), "Performing Inferential Statistics Prior to Data Collection," *Educational and Psychological Measurement*, 77, 204–219. [350]

Van Assen M.A.L.M., Van Aert R.C.M, Nuijten M.B., and Wicherts J.M. (2014), "Why publishing everything is more effective than selective publishing of statistically significant results," *PLoS ONE*, 9, e84896. [346]

The American Statistician

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

Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

Stuart H. Hurlbert, Richard A. Levine & Jessica Utts

To cite this article: Stuart H. Hurlbert, Richard A. Levine & Jessica Utts (2019) Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires, The American Statistician, 73:sup1, 352-357, DOI: 10.1080/00031305.2018.1543616

To link to this article: https://doi.org/10.1080/00031305.2018.1543616

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 11016

View related articles

View Crossmark data

Citing articles: 60 View citing articles

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, 352–357: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543616

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

Stuart H. Hurlbert[a], Richard A. Levine[b], and Jessica Utts[c]

[a]Department of Biology, San Diego State University, San Diego, CA; [b]Department of Statistics, San Diego State University, San Diego, CA; [c]Department of Statistics, University of California, Irvine, CA

ABSTRACT

Many controversies in statistics are due primarily or solely to poor quality control in journals, bad statistical textbooks, bad teaching, unclear writing, and lack of knowledge of the historical literature. One way to improve the practice of statistics and resolve these issues is to do what initiators of the 2016 ASA statement did: take one issue at a time, have extensive discussions about the issue among statisticians of diverse backgrounds and perspectives and eventually develop and publish a broadly supported consensus on that issue. Upon completion of this task, we then move on to deal with another core issue in the same way. We propose as the next project a process that might lead quickly to a strong consensus that the term "statistically significant" and all its cognates and symbolic adjuncts be disallowed in the scientific literature except where focus is on the history of statistics and its philosophies and methodologies. Calculation and presentation of accurate *p*-values will often remain highly desirable though not obligatory. Supplementary materials for this article are available online in the form of an appendix listing the names and institutions of 48 other statisticians and scientists who endorse the principal propositions put forward here..

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
NeoFisherian significance assessment; Statistical significance; Type I error; *p*-values; Dichotomized language; Teaching of statistics

An error does not become truth by reason of multiplied propagation, nor does truth become error because nobody will see it.

– Mahatma Gandhi

1. Introduction

Its cogency, clarity, and sharp focus on a limited set of important issues make the *ASA Statement on Statistical Significance and p-Values* (Wasserstein and Lazar 2016) of great potential value. It should 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.

The "call for papers" for this issue advises potential contributors to avoid "lengthy discussions of 'Don'ts,' which are already addressed effectively in the ASA statement and supplementary materials." We demur slightly. Principle 3 in the ASA statement reads: "Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold." That and accompanying explanatory text warrant being supplemented to make its operational imperatives clearer to statisticians, researchers, and editors. We believe those most knowledgeable about this issue may be close to a consensus that scientific conclusions and business or policy decisions should almost *never* be based on whether a *p*-value passes a specific threshold because in weighing the strength of evidence there is no need for arbitrarily defined thresholds.

We acknowledge with thanks the many useful suggestions on this manuscript provided by editors, referees, and several of the endorsers listed in Appendix A, especially Sander Greenland.

2. The Bull

An apt metaphor for the phrase "statistically significant" and its relatives is that of a *toro bravo*, a champion bull raised for bullfighting who is now on his last legs and awaiting only the coup de grâce. This particular bull was "bred" a century ago by British and Polish gentlemen to "fight," ostensibly, for more objectivity in data interpretation and decision-making. Since then the uncontrolled rampages of this toro bravo have damaged many on the arena floor, victims of bad statistical advice, reasoning, and decision-making. He remains a tough character having been periodically steroid-doped by his caretakers—influential statisticians, editors, textbook authors, and teachers. However, he also has been weakened by the logical critiques of generations of unsuccessful matadors and their assistants. For a swift and clean coup de grâce all that is required now is for that small portion of the scientific community who write the "instructions to authors" for journals to educate itself and translate those logical critiques into some gentle new prescriptions.

3. Some History, Briefly

The birth and career of "statistically significant" and his cousins ("significantly different," "nonsignificant," "critical *p*-values," "fixed alpha," "$p < \alpha$," etc.) have been reviewed, with greater or

CONTACT Stuart H. Hurlbert hurlbert@sdsu.edu Department of Biology Emeritus, San Diego State University, San Diego, CA 92812.
Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/TAS.
© 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.

lesser degrees of accuracy and clarity, in dozens if not hundreds of articles in the published literature and on social media. So we know in particular (1) how Fisher codified earlier ideas about tests of significance based on standardized critical p-values, (2) how Neyman and Pearson developed a superficially similar decision theoretic framework based on critical alphas or Type I error rates, (3) how others, especially publishers and early textbook writers, created awkward hybrids of those paradigms, and (4) how almost the whole statistical world "went dichotomous" and helped make "statistically significant" the *toro bravo* that he remains to this day. The 38-page review of Hurlbert and Lombardi (2009) describes that history in detail, also discussing *in extenso* several closely related issues and controversies beyond the scope of this note.

All the sound and fury, all the resulting confusion, errors, bad teaching, and poor advice from reviewers and editors have taken a large toll. Their effects are impossible to estimate. But many writers, starting more than half a century ago, have provided strong and widely accepted arguments for concluding that it is time to abandon "statistically significant" and dichotomization of the p-scale.

Perhaps we needed that dichotomy for a time. The proposition of concrete evidentiary standards such as critical p-values or fixed alphas (of, say, 0.05), combined with a powerful label ("statistically significant"), may have been a critical and "brilliant stroke of simplification that opened the arcane domain of statistical calculation to a world of experimenters and research workers" who were confronting the new statistical methodologies with some trepidation (Stigler 2008; Hurlbert and Lombardi 2009). Neophytes at the beginning of the 20th century likely were neither amenable to nor capable of nuanced interpretations of new statistical constructs. Nor did they have the computational power to churn out exact p-values or tables comprehensive enough to allow their estimation by interpolation.

But now statistical neophytes are a smaller portion of the scientific community, this and related issues have been thrashed out in the literature for half a century, and simple reason and logic should have a better chance of carrying the day.

4. Solution: A Proscription and a Prescription

For the narrowly defined problem at hand, we propose a two-part solution. It *could* be implemented quickly. It will not *require* any immediate large increase in the statistical sophistication of scientists collectively; but it will *promote* just such an increase over time.

Brian Caffo (pers. comm.) has suggested to us that our two propositions might be regarded as calls for community policing and editorial policing, respectively.

First, we propose that in research articles all use of the phrase "statistically significant" and closely related terms ("nonsignificant," "significant at $p = 0.xxx$," "marginally significant," etc.) be disallowed on the solid grounds long existing in the literature. Just present the p-values without labeling or categorizing them. Every professional statistician, every scientist who uses statistics, and every statistics instructor is in a position to help build momentum for this improvement and proscribe "statistically significant" in those situations where they have individual decision-making authority.

Second, we propose that direct formal requests be made to the editors and editorial boards of journals to modify their instructions to authors to include a disallowance of manuscripts that do not adhere to the above proscription. This task *might* be accomplished under the aegis of ASA, the Royal Statistical Society and perhaps other statistical societies, with a statement that is much simpler and briefer than the 2016 ASA statement, but that has a much larger number of endorsers. This process would build on the sendout of the 2016 statement recommended above.

For a journal an additional "instruction to authors" could read something like the following:

> *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.*

These recommendations should be well received. Occasionally in the past, strong statements have been put out unilaterally by a few individual editors or editorial boards of journals as to what statistical procedures must or must not be used. In contrast, an instruction like that suggested above neither proscribes nor requires use of particular methodologies. It only stipulates a simple matter of language.

A community grassroots effort might advance implementation of both propositions. A large contingent of reputable statisticians and other scientists endorsing them could approach the editors and editorial boards of journals directly urging them to develop new guidelines internally via their own established procedures for doing so. That process by itself might cause some individual editors and editorial board members to join the "community policing" movement, even if their journals or societies decline to formally modify their "instructions to authors."

Curious about how much support our propositions would have, we decided to test reaction to them while this article was under review. We forwarded the manuscript to about 100 scientists who have authored statistics textbooks or reference books, who have served as lead editors for journals (statistical or otherwise), who have served as society presidents, or who have published significant critiques of statistical practice. We asked each whether they were willing to publicly endorse the first sentences of the "proscription" and "prescription" above. We were pleasantly surprised to receive endorsements from 47 scientists (in addition to ourselves), from 10 countries and a wide variety of disciplines (e.g., medicine, psychology, sociology, economics, environmental sciences, etc.). Several endorsers also prompted useful changes in the manuscript itself. The list of endorsers is given in supplementary file Appendix A.

Let us consider briefly some objections to our two propositions that might be raised.

The text under principle #3 in the 2016 ASA statement includes, "Pragmatic considerations often require binary, 'yes–no' decisions but this does not mean that p-values alone can ensure that a decision is correct or incorrect." We would say that the "often" is unwarranted and that situations requiring

binary decisions solely on the basis of individual p-values are vanishingly rare in both basic and applied research. And even where some sort of concrete or physical action is to be taken, the terms "statistically significant" and "nonsignificant" will remain unneeded and actions taken by any decision-makers will not be determined solely by the results of a single test or p-value any more than they are now.

Others may be concerned about how we can justify and determine or fix set-wise or family-wise Type I error rates when multiple tests or comparisons are being conducted if we abandon critical p-values and fixed α's for individual tests. The short and happy answer is: "You can't. And shouldn't try!" Backed up by a 20-page review of the topic, Hurlbert and Lombardi (2012) recommended: "Whatever statistical tests are dictated by the objectives and design of a study are best carried out one-by-one without any adjustments for multiplicities, whether the latter derive from there being multiple treatments, multiple monitoring dates or multiple response variables. Clarity and interpretability of results will be favoured." Despite the obsession over set-wise Type I error rates in certain quarters, for example, some statisticians advising clinical trials, other statisticians have been pointing out their ponderous and arbitrary nature for half a century (e.g., Wilson 1962; Cox 1965; J. A. Nelder [in O'Neill and Wetherill 1971]; Carmer and Walker 1982; Perry 1986; Finney 1988; Mead 1988; Rothman 1990; Keppel 1991; Pearce 1993; Stewart-Oaten 1995; Nakagawa 2004; Schulz and Grimes 2005).

It would seem that any real problems created by our propositions would be so rare or unique as to not constitute any counter argument of weight. Benefits should surely outweigh any potential cost. And authors are always free to make "special-case" pleas to editors.

The statistical use of the word "significant" will never be understood by those without statistical training to mean anything less than its synonyms, "important" and "influential." According to the Merriam Webster dictionary (*https://www.merriam-webster.com/dictionary/significant*), the English word dates back to at least 1579, and it is among the 10% most frequently used words in the English language. So it is time for the discipline of statistics to give it up and return it to its natural roots. Even if statisticians and scientists understand that the statistical meaning of significance differs from its synonyms, scientific findings often are presented to the public by the media. Consumers of them almost surely are misled by equating "significant" with its synonyms, "important," and "influential," even when a journalist is careful to say "a small but significant effect…." In data analysis contexts use of the term should be minimized lest scientist readers automatically infer that it refers to p-values lower than some alpha.

5. Earlier Matadors

Many matadors over many decades have pointed out the desirability of doing away with critical p-values and fixed alphas and the attendant dichotomized terminology. Key excerpts from a few of them will show that our proposals are not revolutionary but are in fact quite moderate and easily defended ones. Here are nine sets of authors, selected in part for the explicitness with which they advocate abandoning the terms "significant" and "nonsignificant" and the dichotomized thinking they reflect. Historical readings must allow for the fact that continuous p-values were traditionally called "significance levels" by Fisher and those he instructed (like D. R. Cox), yet Neyman and his successors adopted the same term for the fixed alpha level against which a calculated p-value would be compared.

> "It is customary to take arbitrary p values, such as .05 and .01 and use them to dichotomize this continuum into a *significant* and an *insignificant* portion. This habit has no obvious advantage, if what is intended is merely a restatement of the probability values these are already given in any case and are far more precise than a simple dichotomous statement. … If the verbal dichotomous scale is not satisfactory—as it clearly is not—the answer surely is to keep to the continuous p scale, rather than subdivide the verbal scale." (Eysenck 1960)

> "Tradition notwithstanding, there seems to be little justifiable reason [for dichotomizing our interpretation of the P scale, so scientists should]… do away with arbitrary levels of significance, and the calling of one test result 'significant' and another 'not significant'." (Skipper, Guenther, and Nass 1967)

> "In this note we confine ourselves to one of the simplest models, which already calls for the exercise of judgment in ways which will not appeal to those anxious to reduce the practice of statistics to the mechanical application of mathematical rules. … The statistician analyzing results of this kind should evaluate the P value. It is not for him, nor for the individual experimenter, alone to impose an α value on other persons. … The use of fixed significance levels, $\alpha = 0.05$ or 0.01, was introduced by Fisher for largely accidental reasons connected with Pearson's copyright in the tables of X^2. Although it has many advantages, especially the dangerously seductive one of saving us the effort of thinking, for the reasons indicated above and below we now ought to abandon it." (Barnard 1982).

> "It is ridiculous to interpret the results of a study differently according to whether the P value obtained was, say, 0.055 or 0.045. These P values should lead to very similar conclusions, not diametrically opposed ones… In recent years there has been a welcome move away from regarding the P value as significant or not significant, according to which side of the arbitrary 0.05 value it is, towards quoting the actual P value… Forcing a choice between significant and non-significant obscures the uncertainty present whenever we draw inferences from a sample." (Altman 1991)

> "The decision-theoretic approach to hypothesis testing suggested by Neyman and Pearson is disappearing from use in major medical journals, and the practice of dividing results of hypothesis tests into 'significant' and 'non-significant' is outdated and unhelpful. … It used to be the convention to say that there was 'significant' evidence against the null hypothesis if $P < 0.05$, and to categorize results as 'significant' or 'non-significant'. This is outdated: it is much better to report the precise P-value." (Sterne 2002)

> "In Fisherian testing, the p value is actually a more fundamental concept than the α level. … A reasonable view would be that an α level should never be chosen; that a scientist should simply evaluate the evidence embodied in the p value." (Christensen 2005)

"We will advocate discarding this [Neyman-Pearsonian] framework for most significance testing situations and replacing it with an explicitly neoFisherian one that 1) does not fix α, 2) does not describe P values as 'significant' or 'non-significant,' 3) does not accept null hypotheses on the basis of high P values but only suspends judgment, 4) recognizes the obvious, near universal need to present effect size information in conjunction with significance tests, and 5) acknowledges the frequent utility of confidence intervals (and other adjunct statistics helpful to interpretation) as well as the fact that they are often unneeded. …. From discussion of this issue with other scientists, it seems the biggest psychological impediment to the acceptance of the neoFisherian paradigm is a reluctance to throw out that deceptive crutch, the phrase 'statistically significant.' As Stoehr (1999) points out, we all would like 'a quick, objective and automatic way' to evaluate our results, but there is none that also meets the additional requirements of 'logical' and 'useful.' We must simply apply the same sorts of nuanced thinking and nuanced language we use in other contexts involving gradations in strength of evidence." (Hurlbert and Lombardi 2009).

"Thus, it is a mark of good practice to present the P-value itself rather than to report whether or not the result was statistically significant. Because significance and nonsignificance are simply degraded descriptions of a P-value, they do not have most of the meanings implied or ascribed to them by some experts." (Greenland and Poole 2011)

"Never, ever, use the word 'significant' in a paper. It is arbitrary, and, as we have seen, deeply misleading. Still less should you use 'almost significant', 'tendency to significant' or any of the hundreds of similar circumlocutions listed by Matthew Hankins [2013] on his *Still not Significant* blog." (Colquhoun 2014)

Many more recent papers have been published that strongly support these authors and the neoFisherian paradigm generally and warn of other problems in interpreting statistical analyses. Fortunately, a recent authoritative, comprehensive and detailed review by Greenland et al. (2016) makes it easy for readers to catch up with the best thinking on these topics. It is important reading for any person drawing statistical inferences or trying to understand the literature.

Given the illustrious string of authorities who have taken to task "statistically significant" and the distortions it entails, why does use of the term persist? Indeed, why did the ASA statement (Wasserstein and Lazar 2016) not recommend the phrase be abandoned? Some authors ascribe the problem to human cognitive biases that have been created or aggravated by elementary statistical training (McShane and Gal 2017; Greenland 2017). More prosaically, few statisticians have a deep knowledge of the historical literature of statistics or realize how much of the most cogent criticism of statistical practice is found in the journals of other fields they rarely consult. And few have time to seek it. Hurlbert and Lombardi (2009) thoroughly reviewed the history of "statistically significant," explicitly recommended its abandonment, and anticipated virtually all the conclusions of the ASA statement. The drafters of the ASA statement did not cite that paper perhaps feeling its blunt recommendations were too radical. But readers who consult it will find a wealth of information that supports and expands the conclusions in the ASA statement, including additional justification for the recommendations in this article.

6. Nuanced Reporting

How then does a mere researcher account for all these lamentations in practice? How should we write up analyses if we can no longer say what effects or correlations are "statistically significant"? What do Hurlbert and Lombardi (2009) mean by "nuanced thinking and nuanced language"? A new set of terminological baggage? Heavens no.

We continue to recommend the use of p-values, confidence intervals, Bayes factors and all of the other tools available. We simply wish to remove the dichotomous use of the term "significant" as an accompaniment to them. Two examples will help illustrate how well this can work.

Kristen Reifel and her colleagues provided one good example with their paper on the spatial distribution of plankton in the Salton Sea, California's largest lake (Reifel et al. 2007). The results (p-values, coefficients) of many regression analyses are fully incorporated into Table 2 and Figure 6 of the paper. There is no verbal characterization of any p-value, nor mention of a p-value in the text. The text includes only discussions of apparent trends or effects. Readers (including editors and reviewers) are expected to look at the graphs, p-values, and coefficients and make their own judgments about whether the authors over-interpreted their results, under-interpreted them, or got it just right.

In a quite different type of study, Pocock et al. (2016) reported how results of a clinical trial on the effects of renal denervation on patients with high blood pressure were clarified by use of ANCOVA. In their tables they report mean effect sizes, and p-values, 95% confidence intervals and standard errors for them with no recourse, with one minor exception, in the paper to labels such as "significant" or "nonsignificant."

We encourage the reader to download copies of these papers and confirm that it really is that simple. A few quotes from these papers will not suffice: it is examination of the entirety of their results and discussion sections that will confirm our claim most clearly. Studies conducting statistical analyses are too diverse in type, size, scope, objective, diversity of statistical procedures used, and other summary statistics accompanying p-values for generally applicable *specific* guidelines to be possible. One exception concerns the content of abstracts: in reporting the main subject matter findings, emphasis should always be on effect sizes, broadly defined. Especially in the past, many journals have been willing to accept abstracts, and even entire manuscripts, where no explicit information on effect sizes was given so long as sufficiently low p-values were cited.

Abstracts for the two articles cited above show the wide range of possibilities for focusing on effect sizes. In their abstract, Reifel et al. (2007) state, very partially, "Several diatom species increased up to 800-fold in abundance by ca. 20 km downcurrent from inflow points in September. …. Zooplankton abundances did not show regular trends downcurrent of river inflows

except for the larvae of [the barnacle], which increased in density ca. 100-fold." That was sufficient for an abstract; readers can consult the body of the paper for the statistical methods, graphs, precise effect size estimates and p-values forming the basis for those conclusions.

In their abstract, Pocock et al. (2016) summarize the key finding in this more precise way: "Analysis of covariance was performed on the 6-month change in systolic blood pressure, estimating a mean treatment difference of −4.11 mm Hg (95% confidence interval: −8.44 to 0.22 mm Hg; $p = 0.064$), which was similar to the unadjusted difference but with a smaller confidence interval." More precise information is given than Reifel et al. (2007) provided, but that is feasible in the abstract only because Pocock et al. were dealing with a single dependent variable. Reifel et al. were dealing with about thirty on each of two sampling dates.

Few actions would more encourage authors to give primary emphasis to effect sizes in their abstracts than would editors disallowing the use of "statistically significant." Especially for large studies involving many different statistical analyses, decisions as to which effect sizes are sufficiently important and conclusively demonstrated as to merit mention in an abstract will necessarily be subjective ones, as will be decisions about how to describe and discuss results in the body of a paper.

7. Conclusion

Just as the initiators of the 2016 ASA statement achieved success by keeping the objective narrow and the focus sharp, so we believe that the next most feasible and concrete step is to implement the two steps outlined.

What could be simpler or more productive than persuading statisticians and other scholars to analyze and write up their data while forgoing "statistically significant" and related terms? Many benefits will flow from this practice, as so many other statistical controversies have been driven by the ancient dichotomy (e.g., misuse of one-tailed tests: Lombardi and Hurlbert 2009; irrational advocacy of set-wise Type I error rates: Hurlbert and Lombardi 2012). And how much work would it be to start getting editorial boards "with the program?" All of us can only try and see.

We are heartened that the three editors of this major special issue of The American Statistician, after thrashing all sorts of matters out with 45 sets of authors, have come out in full support of our thesis. Their introductory editorial states: "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" (Wasserstein et al. 2019).

Supplementary Materials

Appendix A: Statisticians and other scientists endorsing the propositions (1) that in research articles all use of the phrase "statistically significant" and closely related terms ("nonsignificant," "significant at $p = 0.xxx$," "marginally significant," etc.) be disallowed on the solid grounds long existing in the literature; and (2) that direct formal requests be made to the editors and editorial boards of journals to modify their instructions to authors to include a disallowance of manuscripts that do not adhere to the above proscription.

References

Altman, D. G. (1991), *Practical Statistics for Medical Research*, London: Chapman and Hall. [354]

Barnard, G. A. (1982), "Conditionality Versus Similarity in the Analysis of 2 × 2 Tables," in *Statistics and Probability: Essays in Honor of C.R. Rao*, eds. G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, New York: North Holland Publishing, pp. 59–65. [354]

Carmer, S. G., and Walker, W. M. (1982), "Baby Bear's Dilemma: A Statistical Tale," *Agronomy Journal*, 74, 122–124. [354]

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 121–126. [354]

Colquhoun, D. (2014), "An Investigation of the False Discovery Rate and the Misinterpretation of p-Values," *Royal Society Open Science*, 1, 140216. [355]

Cox, D. R. (1965), "A Remark on Multiple Comparison Methods," *Technometrics*, 7, 223–224. [354]

Eysenck, H. J. (1960), "The Concept of Statistical Significance and the Controversy About One-Tailed Tests," *Psychological Review*, 67, 269–271. [354]

Finney, D. J. (1988), "Was This in Your Statistics Textbook? III. Design and Analysis," *Experimental Agriculture*, 24, 421–432. [354]

Greenland, S. (2017), "The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. [355]

Greenland, S., and Poole, C. (2011), "Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony," *Jurimetrics Journal*, 51, 113–129. [355]

Greenland, S., Senn, S. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [355]

Hankins, M. C. (2013), "Still not significant," available at *http://mchankins.wordpress.com/2013/04/21/still-notsignificant-2/*. [355]

Hurlbert, S. H., and Lombardi, C. M. (2009), "Final Collapse of the Neyman-Pearson Decision-Theoretic Framework and Rise of the NeoFisherian," *Annales Zoologici Fennici*, 46, 311–349. [353,355]

——— (2012), "Lopsided Reasoning on Lopsided Tests and Multiple Comparisons," *Australian & New Zealand Journal of Statistics*, 54, 23–42. [354,356]

Keppel, G. (1991), *Design and Analysis: A Researcher's Handbook* (3rd ed.), Englewood Cliffs, NJ: Prentice Hall. [354]

Lombardi, C. M., and Hurlbert, S. H. (2009), "Misprescription and Misuse of One-Tailed Tests," *Austral Ecology*, 34, 447–468. [356]

McShane, B. B., and Gal, D. (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112, 885–908. [355]

Mead, R. (1988), *The Design of Experiments*, Cambridge: Cambridge University Press. [354]

Nakagawa, S. (2004), "A Farewell to Bonferroni: The Problems of Low Statistical Power and Publication Bias," *Behavioral Ecology*, 15, 1044–1045. [354]

O'Neill, R., and Wetherill, G. B. (1971), "The Present State of Multiple Comparison Methods," *Journal of the Royal Statistical Society*, Series B, 33, 218–250. [354]

Pearce, S. C. (1993), "Data Analysis in Agricultural Experimentation. III. Multiple Comparisons," *Experimental Agriculture*, 29, 1–8. [354]

Perry, J. N. (1986), "Multiple-Comparison Procedures: A Dissenting View," *Journal of Economic Entomology*, 79, 1149–1155. [354]

Pocock, S. J., Bakris, G., Rhatt, D. L., Brar, S., Fahy, M., and Gersh, B. J. (2016), "Regression to the Mean in SYMPLICITY HTN-3: Implications for Design and Reporting of Future Trials," *Journal of the American College of Cardiology*, 68, 2016–2025. [355,356]

Reifel, K. M., Trees, C. C., Olivo, E., Swan, B. K., Watts, J. M., and Hurlbert, S. H. (2007), "Influence of River Inflows on Spatial Variation of Phytoplankton Around the Southern End of the Salton Sea, California," *Hydrobiologia*, 576, 167–183. [355,356]

Rothman, K. J. (1990), "No Adjustments Are Needed for Multiple Comparisons," *Epidemiology*, 1, 43–46. [354]

Schulz, K. F., and Grimes, D. A. (2005), "Multiplicity in Randomized Trials I: Endpoints and Treatments," *Lancet*, 365, 1591–1595. [354]

Skipper, K. S. Jr., Guenther, A. L., and Nass, G. (1967), "The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science," *American Sociologist*, 2, 16–18. [354]

Sterne, J. A. C. (2002), "Teaching Hypothesis Tests—Time for Significant Change?," *Statistics in Medicine*, 21, 985–994. [354]

Stewart-Oaten, A. (1995), "Rules and Judgments in Statistics: Three Examples," *Ecology*, 76, 2001–2009. [354]

Stigler, S. (2008), "Fisher and the 5% Level," *Chance*, 21, 12. [353]

Stoehr, A. M. (1999), "Are Significance Thresholds Appropriate for the Study of Animal Behaviour?," *Animal Behaviour*, 57, F22–F25. [355]

Wasserstein, R., and Lazar, N. (2016), "ASA Statement on Statistical Significance and p-Values," *The American Statistician*, 70, 131–133. [352,355]

Wasserstein, R., Lazar, N., and Schirm, A. (2019), "Editorial: Moving to a World Beyond $p < 0.05$," *The American Statistician*, 73(1). [356]

Wilson, W. (1962), "A Note on the Inconsistency Inherent in the Necessity to Perform Multiple Comparisons," *Psychological Bulletin*, 59, 296–300. [354]

The American Statistician

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

The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication

Harlan Campbell & Paul Gustafson

To cite this article: Harlan Campbell & Paul Gustafson (2019) The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication, The American Statistician, 73:sup1, 358-373, DOI: 10.1080/00031305.2018.1555101

To link to this article: https://doi.org/10.1080/00031305.2018.1555101

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

View supplementary material

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 4767

View related articles

View Crossmark data

Citing articles: 4 View citing articles

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, 358–373: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1555101

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication

Harlan Campbell and Paul Gustafson

Department of Statistics, University of British Columbia, Vancouver, Canada

Abstract

In response to growing concern about the reliability and reproducibility of published science, researchers have proposed adopting measures of "greater statistical stringency," including suggestions to require larger sample sizes and to lower the highly criticized "$p < 0.05$" significance threshold. While pros and cons are vigorously debated, there has been little to no modeling of how adopting these measures might affect what type of science is published. In this article, we develop a novel optimality model that, given current incentives to publish, predicts a researcher's most rational use of resources in terms of the number of studies to undertake, the statistical power to devote to each study, and the desirable prestudy odds to pursue. We then develop a methodology that allows one to estimate the reliability of published research by considering a distribution of preferred research strategies. Using this approach, we investigate the merits of adopting measures of "greater statistical stringency" with the goal of informing the ongoing debate.

ARTICLE HISTORY
Received March 2018
Revised November 2018

KEYWORDS
Meta-research; Null hypothesis significance Testing; Publication; Reliability; Reproducibility; Statistical power

1. Introduction

> It is to be remarked that the theory here given rests on the supposition that the object of the investigation is the ascertainment of truth. When an investigation is made for the purpose of attaining personal distinction, the economics of the problem are entirely different. But that seems to be well enough understood by those engaged in that sort of investigation.
>
> *Note on the Theory of the Economy of Research,*
> Charles Sanders Peirce, 1879

In a highly cited essay, Ioannidis (2005) uses the Bayes theorem to claim that more than half of published research findings are false. While not all agree with the extent of this conclusion (e.g., Goodman and Greenland 2007; Leek and Jager 2017), recent large-scale efforts to reproduce published results in a number of different fields (economics, Camerer et al. 2016; psychology, OpenScienceCollaboration 2015; oncology, Begley and Ellis 2012), have also raised concerns about the reliability and reproducibility of published science. Unreliable research not only reduces the credibility of science, but is also very costly (Freedman, Cockburn, and Simcoe 2015) and as such, addressing the underlying issues is of "vital importance" (Spiegelhalter 2017). Many researchers have recently proposed adopting measures of "greater statistical stringency," including suggestions to require larger sample sizes and to lower the highly criticized "$p < 0.05$" significance threshold. In statistical terms, this represents selecting lower levels for acceptable Type I and Type II error rates. The main argument against "greater statistical stringency" is that these changes would increase the costs of a study and ultimately reduce the number of studies conducted.

Consider the debate about lowering the significance threshold in response to the work of Johnson (2013), who, based on the correspondence between uniformly most powerful Bayesian tests and classical significance tests, recommends lowering significance thresholds by a factor of 10 (e.g., from $p < 0.05$ to $p < 0.005$). Gaudart et al. (2014), voicing a common objection, contend that such a reduction in the allowable Type I error rate will result in inevitable increases to the Type II error rate. While larger sample sizes could compensate, this can be costly: "increasing the size of clinical trials will reduce their feasibility and increase their duration" (Gaudart et al. 2014). In the view of Johnson (2014), this may not necessarily be such a bad thing, pointing to the excess of false positives (FPs) and the idea that (in the context of clinical trials) "too many ineffective drugs are subjected to phase III testing [...] wast[ing] enormous human and financial resources."

More recently, a highly publicized call by over seven dozen authors to "redefine statistical significance" has made a similar suggestion: lower the threshold of what is considered "significant" for claims of new discoveries from $p < 0.05$ to $p < 0.005$ (Benjamin et al. 2018). This has prompted a familiar response (e.g., Wei and Chen 2018). Amrhein, Korner-Nievergelt, and Roth (2017) review the arguments for and against more stringent thresholds for significance and conclude that: "[v]ery possibly, more stringent thresholds would lead to even more results being left unpublished, enhancing publication bias. [...] [W]hile aiming at making our published claims more reliable, requesting more stringent fixed thresholds would achieve quite the opposite."

CONTACT Harlan Campbell harlan.campbell@stat.ubc.ca PO Box 1882, Rossland, BC, V0G 1Y0.
© 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.

There is also substantial disagreement about suggestions to require larger sample sizes. In some fields, showing that a study has a sufficient sample size (i.e., high power) is common practice and an expected requirement for funding and/or publication, while in others it rarely occurs. For example, Charles et al. (2009) found that 95% of randomized controlled trials (RCTs) report sample size calculations. In contrast, only a tiny fraction of articles in some fields –about 3% for psychology, and about 2% for toxicology—justify their sample size (Bosker, Mudge, and Munkittrick 2013; Fritz , Scherndl, and Kühberger 2013). The number is only marginally higher in conservation biology at approximately 8% (Fidler et al. 2006).

One argument is that, once a significant finding is achieved, the size of a study is no longer relevant. Aycaguer and Galbán (2013) explain as follows: "If a study finds important information by blind luck instead of good planning, I still want to know the results." Another viewpoint is that, while far from ideal, underpowered studies should be published since cumulatively, they can contribute to useful findings (Walker 1995). Others disagree and contend that small sample sizes undermine the reliability of published science (Button et al. 2013a; Dumas-Mallet et al. 2017; Nord et al. 2017). In the context of clinical trials, IntHout, Ioannidis, and Borm (2016) review the many conflicting opinions about whether trials with suboptimal power are justified and conclude that, in circumstances when evidence for efficacy can be effectively combined across a series of trials (e.g., via meta-analysis), small sample sizes might be justified.

Despite the long-running and ongoing debates on significance thresholds and sample size requirements, there has been little to no modeling of how changes to a publication policy might affect what type of studies are pursued, the incentive structures driving research, and ultimately, the reliability of published science. One example is Borm, den Heijer, and Zielhuis (2009) who conclude, based on simulation studies, that the consequences of publication bias do not warrant the exclusion of trials with low power. Another recent example is Higginson and Munafò (2016), who, based on results from an optimality model of the "scientific ecosystem," conclude that in order to "improve the scientific value of research" peer-reviewed publications should indeed require larger sample sizes, lower the α significance threshold, and give "less weight to strikingly novel findings." Our work here aims to build upon these modeling efforts to better inform the ongoing discussion on the reproducibility of published science.

The model and methodology we present seeks to add three features absent from the model of Higginson and Munafò (2016). While these authors consider how researchers balance available resources between exploratory and confirmatory studies, this simple dichotomy does not allow for a detailed assessment of the willingness of researchers to pursue high-risk studies (those studies that are, a priori, unlikely to result in a statistically significant finding). Our approach addresses this issue by considering a continuous spectrum of a priori risk, that is, the "pre-study probability." Second, Higginson and Munafò (2016) define the "total fitness of a researcher" (i.e., the payoff for a given research strategy) with diminishing returns for confirmatory studies, but not for exploratory studies. This choice, however well intended, has problematic repercussions for their optimality model. (Under their framework, the optimal research strategy will depend on T, an arbitrary total budget parameter.) Finally, by failing to incorporate the number of correct studies that go unpublished within their metric for the value of scientific research, many potential downsides of adopting measures to increase statistical stringency are ignored. Other differences between our approach and previous ones will be made evident and include: considering outcomes in terms of distributional differences, and specific modeling of how sample size requirements are implemented.

The remainder of this article is structured as follows. In Section 2, we describe the model and methodology proposed to evaluate different publication policies. We also list a number of metrics of interest and look to recent meta-research analyses to calibrate our model. In Section 3, we use the proposed methodology to evaluate the potential impact of lowering the significance threshold; and in Section 4, the impact of requiring larger sample sizes. Finally, in Section 5, we conclude with suggestions as to how publication policies can be defined to best improve the reliability of published research.

2. Methods

Recently, economic models have been rather useful for evaluating proposed research reforms (Gall, Ioannidis, and Maniadis 2017). However, modeling of how resources ought to be allocated among research projects is not new. See, for example, the work of Greenwald (1975), Dasgupta and Maskin (1987), McLaughlin (2011), and Miller and Ulrich (2016). As noted in the introduction, our framework for modeling the scientific ecosystem is closest in spirit to that of Higginson and Munafò (2016) who formulate a relationship between a researcher's strategy and their payoff, with the strategy involving a choice of mix between exploratory and confirmatory studies, and a choice of pursuing fewer studies with larger samples or more studies with smaller samples.

The publication process is complex and includes both objective and subjective considerations of efficacy and relevance. The title of this article was chosen specifically to emphasize this point (Gornitzki, Larsson, and Fadeel 2015). A large, complicated human process like that of scientific publication cannot be entirely reduced to metrics and numbers: there are often financial, political and even cultural reasons for an article being accepted or rejected for publication. With this in mind, the model presented here should not be seen as an attempt to precisely map out the peer-review process, but rather, as a useful tool for determining the consequences of implementing different publication policies.

Within our optimality model, many assumptions and simplifications are made. Most importantly, we assume that each researcher must make decisions consisting of only two choices: what statistical power (i.e., sample size) to adopt and what "prestudy probability" to pursue. Before elaborating further, let us briefly discuss these two concepts.

2.1. Statistical Power

Increasing statistical power by conducting studies with larger sample sizes would undeniably result in more published

research being true. However, these improvements may only prove modest, given current publication guidelines. When we consider the perspectives of both researchers and journal editors, it is not surprising that statistical power has not improved substantially (Smaldino and McElreath 2016; Lamberink et al. 2018) despite being highlighted as an issue over five decades ago (Cohen 1962).

In practice, multiple factors can influence sample size determination (Lenth 2001; Hazra and Gogtay 2016; Roos 2017). Strictly in terms of publication prospects however, there is little incentive to conduct high-powered studies: basic logic suggests that the likelihood of publication is only minimally affected by power. To illustrate, consider a large number of hypotheses tested, out of which 10% are truly nonnull. Under the assumption that only (and all) positive results are published with $\alpha = 0.05$ (which may in fact be realistic in certain fields, Fanelli 2011), simple analytical calculation shows that increasing average power from an "unacceptably low" 55% to a "respectable" 85% (at the cost of more than doubling sample size), results in only a minimal increase in the likelihood of publication: from 10% to 13%. Moreover, the proportion of true findings amongst those published is only increased modestly: from 55% to 64%. Indeed, a main finding of Higginson and Munafò (2016) is that the rational strategy of a researcher is to "carry out lots of underpowered small studies to maximize their number of publications, even though this means around half will be false positives." This result is in line with the views of many; see, for example, Bakker, van Dijk, and Wicherts (2012), Button et al. (2013b), and Gervais et al. (2015).

From a journal's perspective, there is also little incentive to require larger sample sizes as a requirement for publication. There are minimal consequences from publishing false claims. Fraley and Vazire (2014) review the publication history of six major journals in social-personality psychology and find that "journals that have the highest impact [factor] also tend to publish studies that have smaller samples." This finding is in agreement with Szucs and Ioannidis (2016) who conclude that, in the fields of cognitive neuroscience and psychology, journal impact factors are negatively correlated with statistical power; see also Brembs, Button, and Munafò (2013).

2.2. Prestudy Probability

We use the term "prestudy probability" (*psp*) as shorthand for the a priori probability that a study's null hypothesis is false. In this sense, highly exploratory research will typically have very low *psp*, whereas confirmatory studies will have a relatively high *psp*. Studies with low *psp* are not problematic per se. To the contrary, there are undeniable benefits to pursuing "long-shot" novel ideas that are very unlikely to work out, see Cohen (2017). While replication studies (i.e., studies with higher *psp*) may be useful to a certain extent, there is little benefit in confirming a result that is already widely accepted. As Button et al. (2013a) note: "As R [the pre-study odds] increases [...] the incremental value of further research decreases." Most scientific journals no doubt take this into account in deciding what to publish, with more surprising results more likely to be published. This state of affairs persists, despite recent calls for more replication studies (e.g., Moonesinghe, Khoury, and Janssens 2007). Indeed, replication studies are still often rejected on the grounds of "lack of novelty;" see Makel, Plucker, and Hegarty (2012), Yeung (2017), and Martin and Clarke (2017). As such, researchers deciding which hypotheses to pursue toward publication will likely emphasize those with lower *psp*.

Recognize that the lower the *psp*, the less likely a "statistically significant" finding is to be true. As such, we are bound to a "seemingly inescapable trade-off" (Fiedler 2017) between the novel and the reliable. Journal editors face a difficult choice. Either publish studies that are surprising and exciting yet most likely false, or publish reliable studies which do little to advance our knowledge. Based on a belief that both ends of this spectrum are equally valuable, Higginson and Munafò (2016) conclude that, in order to increase reliability, current incentive structures should be redesigned, "giving less weight to strikingly novel findings." This is in agreement with the view of Hagen (2016) who writes: "If we are truly concerned about scientific reproducibility, then we need to reexamine the current emphasis on novelty and its role in the scientific process."

2.3. Model Framework

We describe our model framework in five simple steps.

(1) We assume, for simplicity, that all studies test a null hypothesis of equal population means against a two-sided alternative, with a standard two-sample Student t-test. Each study has an equal number of observations per sample ($n_1 = n_2$; $n = n_1 + n_2$). Furthermore, let us assume that a study can have one of only two results: (1) *positive* (p-value $< \alpha$), or (2) *negative* (p-value $\geq \alpha$). Given the true effect size, $\mu_d = \mu_2 - \mu_1$ (the difference in population means), σ^2, the common population variance, and the sample size, n, we can easily calculate the probability of a significant result, using the standard formula for power.[1]

Then, for a given true effect size of δ or zero, we have the probability of a TP, FN, FP, and TN, equal to: $\Pr(\text{TP}) = \Pr(\text{Positive}|\mu_d = \delta)$, $\Pr(\text{FN}) = \Pr(\text{Negative}\ |\mu_d = \delta)$, $\Pr(\text{FP}) = \Pr(\text{Positive}|\mu_d = 0)$, and $\Pr(\text{TN}) = \Pr(\text{Negative}|\mu_d = 0)$, respectively.

(2) Next, we consider a large number of studies, n_S, each with a total sample size of n. Of these n_S studies, only a fraction, *psp* (where *psp* is the prestudy probability), have a true effect size of $\mu_d = \delta$. For the remaining $(1 - psp) \cdot n_S$ studies, we have $\mu_d = 0$. Note that for a given sample size, n, these n_S studies are each "powered" at level $pwr = Pr(\text{TP})$. Throughout this article, we focus on scenarios, where $\sigma^2 = 1$ and $\delta = 0.20$ or

[1] The probability of a positive result is equal to

$$\Pr(\text{Positive}|\mu_d) = (1 - F_{n-2,\frac{\mu_d}{\sigma^*}}(t^*_{\alpha/2})) + F_{n-2,\frac{\mu_d}{\sigma^*}}(-t^*_{\alpha/2}), \quad (1)$$

where $t^*_{\alpha/2}$ is the upper $100\cdot\frac{\alpha}{2}$-th percentile of the t-distribution with $n - 2$ degrees of freedom, $\sigma^* = \sigma\sqrt{(1/n_1 + 1/n_2)}$, and $F_{df,ncp}(x)$ is the cdf of the noncentral t distribution with df degrees of freedom and noncentrality parameter ncp. We calculate the minimum required sample size, n, to obtain a desired power, pwr, as follows:

$$n = \text{argmin}_n(|pwr - (1 - F_{n-2,\frac{\mu_d}{\sigma^*}}(t^*_{\alpha/2})) + F_{n-2,\frac{\mu_d}{\sigma^*}}(-t^*_{\alpha/2})|) \quad (2)$$

$\delta = 0.43$. These choices reference the analyses of Lamberink et al. (2018) and Richard, Bond Jr, and Stokes-Zoota (2003). Lamberink et al. (2018), based on an empirical analysis of over one hundred-thousand clinical trials conducted between 1975 and 2014, estimate that the median effect size of clinical trials is approximately a Cohen's $d = 0.20$. In a similar analysis based on data from over 25,000 social/personality studies, Richard, Bond Jr, and Stokes-Zoota (2003) estimate that the mean effect size in social psychology research is of a Cohen's $d = 0.43$.

(3) We also label each study as either published (PUB) or unpublished (UN) for a total of 8 distinct categories (= 2 (positive, negative) × 2 (true and false) × 2 (published and unpublished)). One can determine the expected number of studies (out of a total of n_S studies) in each category by simple arithmetic. Table 1 lists the equations for each of the eight categories with A equal to the probability of publication for a positive result, and B equal to the probability of publication for a negative result. Throughout this article, we will always assume that only positive studies are published, hence, we set $B = 0$. Initially, we will let the probability of publication of a positive study, A, depend on the *psp* according to the function $A = (1 - psp)^m$, where $m \in (0, \infty)$ is a tuning parameter. This simple, decreasing function of *psp* represents a system in which positive results with lower *psp* are more likely to be published on the basis of novelty. A higher value of m indicates that higher *psp* studies (i.e., low-risk hypotheses) are less likely to be published. Consequently, a higher value of m implies a lower overall publication rate. In Section 4, we will define A differently, such that the probability of publication depends on the values of both *psp* and *pwr*.

(4) We determine the total number of studies, n_S, based on three parameters: T, the total resources available (in units of observations sampled); k, the fixed cost per study (also expressed in equivalent units of observations sampled); and n, the total sample size per study. Consequently, as in Higginson and Munafò (2016), $n_S = (k + n)^{-1}T$. Then for any given level of power, *pwr*, we can easily obtain the necessary sample size per study, n, (using established sample size formula for two-sample t-tests, see equations 1 and 2), and a resulting total number of studies, n_S. Keep in mind that n_S, the number of studies, is a function of *pwr*. Throughout this article, when necessary, we take T=100,000. However, note that when comparing the outcomes of different publication policies, this choice is entirely irrelevant. Consider that when k is small, the cost of data, relative to the total cost of a study, is large. Conversely, when k is large, the relative cost of increasing the study sample size is small. For example, suppose k=20 and $\delta = 0.20$, then increasing the power from 0.55 to 0.85, implies doubling the total cost of the study. If instead $k = 500$ (all else being equal), this increase in power corresponds to only a 50% increase in the total study cost.

(5) Finally, let us define a "research strategy" to be a given pair of values for (*psp*, *pwr*) within ([0,1] × [α,1]). Then, for a given research strategy we can easily calculate the total expected number of publications (*ENP*):

$$\mathrm{ENP}(psp, pwr) = \mathrm{TP}_{\mathrm{PUB}} + \mathrm{FP}_{\mathrm{PUB}} + \mathrm{TN}_{\mathrm{PUB}} + \mathrm{FN}_{\mathrm{PUB}}. \quad (3)$$

Figure 1 illustrates how ENP is calculated for different values of *psp* and *pwr*, with fixed $\alpha = 0.05$, $k = 100$, and $m = 3$.

With this set-up in hand, suppose now a researcher pursues -consciously or unconsciously- strategies that maximize the expected number of publications (Charlton and Andras 2006). This may not be an entirely unreasonable assumption considering the influence of Goodhart's Law ("Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes") on academia, see Fire and Guestrin (2018) and van Dijk, Manor, and Carey (2014). Figure 2 shows the analytical calculations of how ENP changes over a range of values of *psp* and *pwr* and under different fixed values for k and m. (Note how the central panel of Figure 2 ($k = 100$, $m = 3$) corresponds to Figure 1.) Depending on k and m, the value of (*psp*, *pwr*) that maximizes ENP can change substantially. With larger k, higher powered strategies will yield a greater ENP; with smaller m, optimal strategies are those with higher *psp*. It is interesting to observe how the optimal strategy changes under different scenarios. However, it may be more informative to consider a *distribution* of preferred strategies. This may also be a more realistic approach. While rational researchers may be drawn toward optimal strategies, surely scientists are not willing and/or able to precisely identify these.

Taking the incentive to publish as a starting point, we assume scientists are more likely to spend resources on studies whose *psp* and *pwr* produce a higher expected number of publications. So we consider a *distribution* of resource allocation across study characteristics whose density is proportional to the expected number of publications, that is, $f_{\mathrm{RES}}(psp, pwr) \propto \mathrm{ENP}(psp, pwr)$. We emphasize that from a funder's viewpoint, $f_{\mathrm{RES}}()$ describes where dollars are being spent on the (*psp*, *pwr*) plane. For example, consider the nine strategies illustrated in Figure 1. Given the relative ENP values, we expect the expenditure on $(psp, pwr) = (0.2, 0.2)$ studies to be 18 times greater than that on $(psp, pwr) = (0.8, 0.5)$ studies.

In the appendix, we give expressions for two distributions that follow on from $f_{\mathrm{RES}}()$. The first is $f_{\mathrm{ATM}}(psp, pwr)$ which describes where *attempted* studies (not resources) fall on the (*psp*, *pwr*) plane, presuming that resources are allocated according to $f_{\mathrm{RES}}()$. Since higher powered studies are most costly, $f_{\mathrm{RES}}()$ and $f_{\mathrm{ATM}}()$ are not the same. Hence, $f_{\mathrm{ATM}}()$ describes what kinds of studies are attempted more frequently.

Table 1. Equations for the expected number of studies (out of a total of n_S studies) for each of the eight categories; with A = prob. of publication for a positive result and B = prob. of publication for a negative result. The number of studies n_S, changes for different values of *pwr*. We have that: $n_S = k + n(pwr)^{-1}T$, where $n(pwr)$ is the required sample size to obtain a power of *pwr*.

For a fixed resources, T, and a given (*psp*, *pwr*), the expected number of...	**Equation (note: throughout this article, we take $B = 0$)**
True positives (TPs)	
published	$\mathrm{TP}_{\mathrm{PUB}}(psp, pwr) = psp \cdot n_S \cdot A \cdot \Pr(\mathrm{TP})$
TPs unpublished	$TP_{UN}(psp, pwr) = psp \cdot n_S \cdot (1 - A) \cdot \Pr(\mathrm{TP})$
False negatives (FNs)	
published	$FN_{\mathrm{PUB}}(psp, pwr) = psp \cdot n_S \cdot B \cdot \Pr(\mathrm{FN})$
FNs unpublished	$\mathrm{FN}_{UN}(psp, pwr) = psp \cdot n_S \cdot (1 - B) \cdot \Pr(\mathrm{FN})$
FPs published	$\mathrm{FP}_{\mathrm{PUB}}(psp, pwr) = (1 - psp) \cdot n_S \cdot A \cdot Pr(FP)$
FPs unpublished	$\mathrm{FP}_{\mathrm{UN}}(psp, pwr) = (1 - psp) \cdot n_S \cdot (1 - A) \cdot \Pr(\mathrm{FP})$
True negatives (TNs)	
published	$\mathrm{TN}_{\mathrm{PUB}}(psp, pwr) = (1 - psp) \cdot n_S \cdot B \cdot \Pr(\mathrm{TN})$
TNs unpublished	$\mathrm{TN}_{\mathrm{UN}}(psp, pwr) = (1 - psp) \cdot n_S \cdot (1 - B) \cdot \Pr(\mathrm{TN})$
Publications	$\mathrm{ENP}(psp, pwr) = TP_{\mathrm{PUB}} + FN_{\mathrm{PUB}} + FP_{\mathrm{PUB}} + TN_{\mathrm{PUB}}$

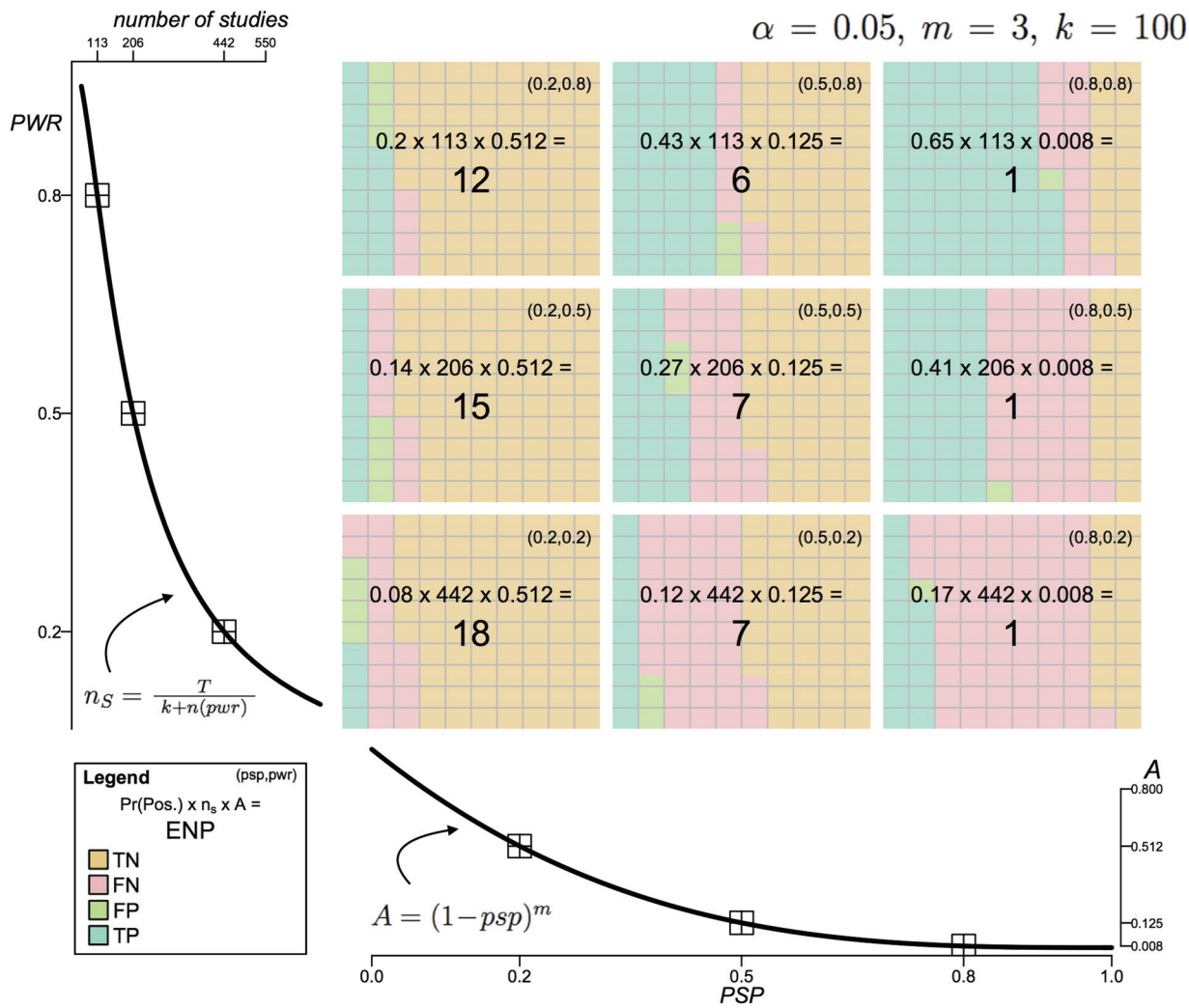


Figure 1. The nine-gridded squares represent nine different research strategies; from left to right, we have $psp = 0.2$, 0.5, and 0.8; from bottom to top, we have $pwr = 0.2$, 0.5, and 0.8. For each pair of values for (psp, pwr), the different colored areas represent the probabilities of a TP, a FP, a TN, and a FN. The Expected Number of Publications (ENP) is calculated by multiplying (1) the probability of a positive finding (Pr(Positive) $= psp \cdot \Pr(\text{TP}) + (1-psp) \cdot \Pr(\text{FP})$), (2) the number of studies ($n_S = \{k+n(pwr)\}^{-1}T$), and (3) the probability of publication ($A = (1-psp)^m$). The figure corresponds to an ecosystem defined by $\alpha = 0.05$, $m = 3$, $k = 100$, $B = 0$, $\mu = 0.20$, $\sigma^2 = 1$, and T=100,000.

Of course not all attempted studies are published. Thus, in the appendix we also give an expression for $f_{\text{PUB}}(pwp, pwr)$, the density of (psp, pwr) across *published* studies that is implied when $f_{\text{ATM}}()$ describes the characteristics of attempted studies. Note then that each of $f_{\text{RES}}()$, $f_{\text{ATM}}()$, and $f_{\text{PUB}}()$ describes relatively favored and disfavored (psp, pwr) combinations. However, the distinction between the three is important. While $f_{\text{RES}}()$ describes resource deployment, $f_{\text{ATM}}()$ describes the resulting constellation of attempted studies, and $f_{\text{PUB}}()$ describes the resulting constellation of published studies.

Armed with $f_{\text{RES}}()$, $f_{\text{ATM}}()$, and $f_{\text{PUB}}()$, we can investigate the properties of a given scientific ecosystem, and how these properties vary across ecosystems. Specifically, an ecosystem is specified by choices of α, k, m, A, and B.

For any specification, properties of the three distributions are readily computed via two-dimensional numerical integration using a fine (200 by 200) grid of (psp, pwr) values. We do not require any simulation to obtain our results (except for those results relating to the accuracy of published research, see details in the appendix). To illustrate, consider the much coarser 3 by 3 grid plotted in Figure 1 and see how *ENP* could be easily calculated by summation across the two dimensions.

2.4. Ecosystem Metrics

We will evaluate the merits of different publication policies on the basis of the following six ecosystem metrics of interest. In the appendix, we provide further details on these metrics and introduce some compact notation that will be useful for their calculation from $f_{\text{RES}}()$ and its by-products.

1. *The Publication Rate (PR).* The ratio of the number of published studies over the number of attempted studies is of evident interest.
2. *The Reliability (REL).* The proportion of published findings that are correct is a highly relevant metric for the scientific ecosystem. Ideally, we wish to see a literature with as few FPs as possible.
3. *The Breakthrough Discoveries (DSCV).* The ability of the scientific ecosystem to produce breakthrough findings is

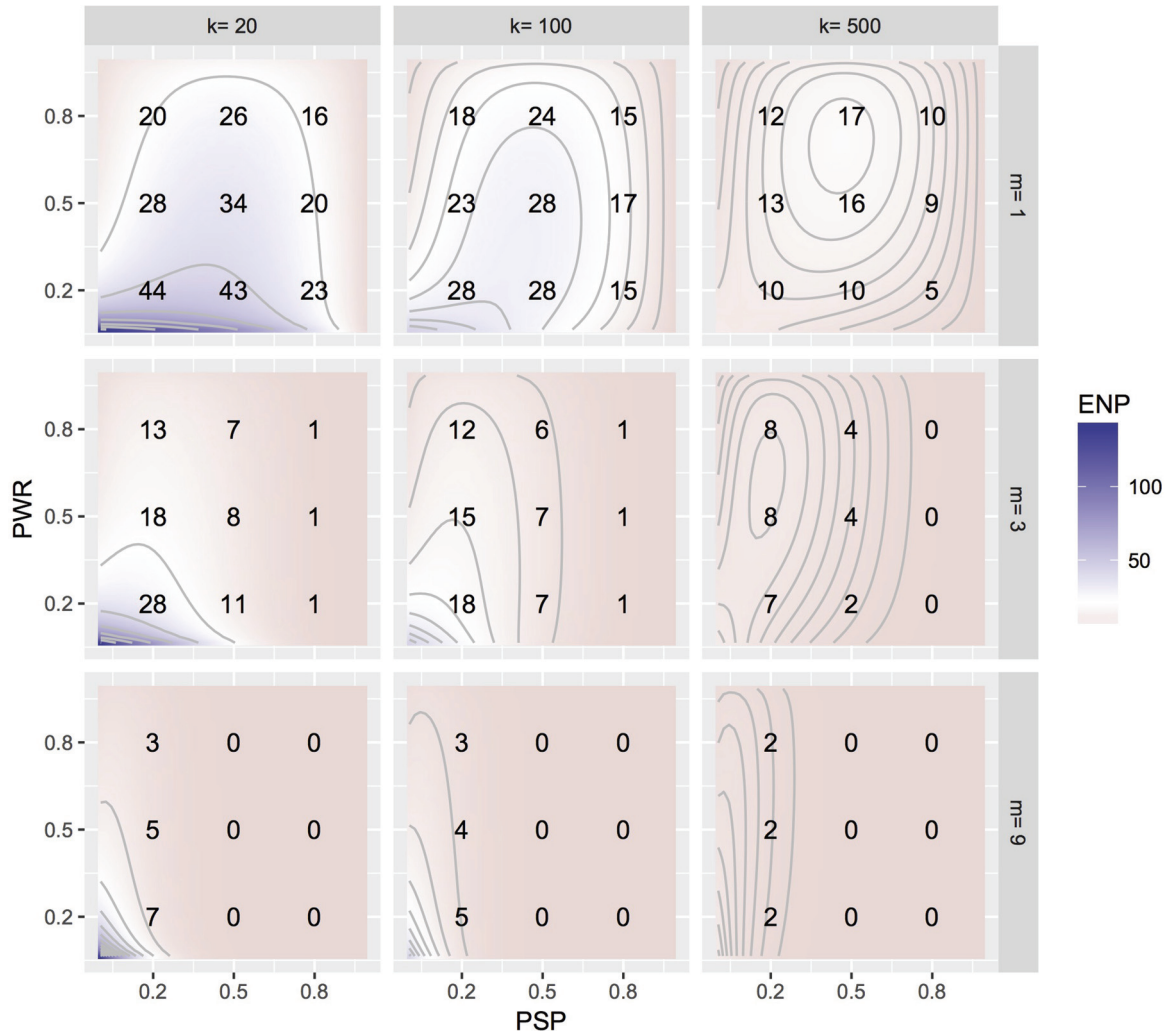


Figure 2. Plots show the number of expected publications, ENP, for different values of PSP and PWR. The printed numbers correspond to values of ENP at $psp = 0.2, 0.5,$ and 0.8; and $pwr = 0.2, 0.5,$ and 0.8. Each panel represents one ecosystem defined by $\alpha = 0.05$, $A = (1 - psp)^m$, and $B = 0$, with k and m as indicated by column and row labels, respectively. Depending on k and m, the value of (PSP, PWR) that maximizes ENP can change substantially. With a higher k, higher powered strategies will yield a greater ENP; with smaller m, optimal strategies are those with higher PSP. Assuming that researchers behave based on optimizing the use of their resources, it is interesting to observe how the"optimal strategy" changes under different scenarios.

an important attribute. Here, a true breakthrough result is defined as a TP and published study that results from a *psp* value below a threshold of 0.05. Note that we will calculate the total number of true breakthrough discoveries, *DSCV*, for each ecosystem for T units of resources. We will also note the reliability of breakthrough discoveries (D_{REL}), the proportion of positive and published studies with $psp < 0.05$ that are true.

4. *The Balance Between Exploration and Confirmation* ($\text{IQpsp}_{\text{PUB}}$). The interquartile range of the distribution of *psp* values amongst published studies provides an assessment of the much discussed balance between exploratory and confirmatory research; see Sakaluk (2016) and Kimmelman, Mogil, and Dirnagl (2014).
5. *The Median Power of Published Studies* (Mpwr_{PUB}). As mentioned earlier, higher powered research leads to fewer yet more reliable publications.
6. *The Accuracy of Published Research.* We will report the mean absolute estimated effect size in published studies, $|\hat{\mu_d}|_{\text{PUB}}$, as well as the proportion of published studies for which the estimated effect size is negative ($\hat{\mu}_{d\text{PUB}}^{\text{neg}}$). It is well known that, due to a combination of publication bias and low powered studies, published effect sizes will be inflated, see Ioannidis (2008); Lane and Dunlap (1978). In addition, we will report the exaggeration ratio (the expected Type M error), and the Type S error rate (the proportion of published studies with an error in *sign*), see Gelman and Carlin (2014).

2.5. Model Calibration

Let us begin by assessing whether or not our basic model framework is well calibrated. Are our modeling choices (specifically the choice of values to consider for parameters k and m) at all consistent with what is known about real conditions? In order to answer this important question, we considered a large number of different scenarios in which α is held fixed at 0.05, and both the fixed-cost parameter, k, and the novelty parameter, m, take various values; see Table 2.

There is a recognized lack of empirical research measuring the relationship between experimental costs and sample sizes

Table 2. For ecosystems defined by fixed $\alpha = 0.05$, $A = (1 - psp)^m$, and $B = 0$, and varying values of k, m, and δ, the table lists estimates for the metrics of interest.

δ	k	m	PR	REL	$DSCV$	D_{REL}	$Mpwr_{\text{PUB}}$	$\text{IQpsp}_{\text{PUB}}$	$[q_{25th}, q_{75th}]$
0.20	20	1	0.06	0.62	0.17	0.07	0.20	0.35	[0.19 , 0.54]
0.20	20	3	0.03	0.36	0.38	0.07	0.12	0.19	[0.07 , 0.26]
0.20	20	9	0.03	0.15	0.86	0.06	0.09	0.07	[0.02 , 0.09]
0.20	100	1	0.08	0.75	0.09	0.12	0.42	0.33	[0.25 , 0.58]
0.20	100	3	0.04	0.52	0.22	0.12	0.32	0.20	[0.09 , 0.29]
0.20	100	9	0.03	0.25	0.51	0.10	0.21	0.08	[0.03 , 0.10]
0.20	500	1	0.11	0.83	0.04	0.18	0.60	0.31	[0.29 , 0.60]
0.20	500	3	0.05	0.65	0.10	0.18	0.55	0.21	[0.12 , 0.33]
0.20	500	9	0.03	0.36	0.26	0.16	0.44	0.09	[0.03 , 0.12]
0.43	20	1	0.08	0.76	0.42	0.12	0.42	0.33	[0.25 , 0.58]
0.43	20	3	0.04	0.53	1.00	0.12	0.32	0.20	[0.09 , 0.29]
0.43	20	9	0.03	0.25	2.37	0.11	0.21	0.08	[0.03 , 0.10]
0.43	100	1	0.11	0.83	0.21	0.18	0.59	0.31	[0.29 , 0.60]
0.43	100	3	0.05	0.65	0.50	0.18	0.54	0.21	[0.12 , 0.33]
0.43	100	9	0.03	0.36	1.24	0.16	0.43	0.08	[0.03 , 0.11]
0.43	500	1	0.13	0.86	0.07	0.23	0.71	0.31	[0.30 , 0.61]
0.43	500	3	0.06	0.71	0.17	0.22	0.68	0.21	[0.13 , 0.34]
0.43	500	9	0.03	0.43	0.44	0.20	0.60	0.09	[0.03 , 0.12]

NOTES: These include the publication rate (*PR*), the reliability (*REL*), the number of true breakthrough discoveries (*DSCV*), the median power of published studies ($Mpwr_{\text{PUB}}$), and the interquartile range of the distribution of *psp* values amongst published studies ($\text{IQpsp}_{\text{PUB}}$), with corresponding 25th and 75th quartiles. Additionally, the table lists the reliability amongst the subset of published studies for which $psp < 0.05$ (D_{REL}).

(Roos 2017), and it is even more difficult to empirically access appropriate values for m. However, we were able to choose suitable values for m (= 1, 3, and 9) and k (= 20, 100, and 500) based on how our model outputs compared with empirical estimates in the meta-research literature. Specifically, values were chosen to exhibit a sufficiently wide range and so that (1) the reliability, (2) the power, and (3) the publication rate agreed, at least to a certain extent, with what has been observed empirically. Consider the following.

1. Due to the success of several recent large-scale reproducibility projects, we are able to obtain reasonable approximations for the *reliability* of published research in different scientific fields. For example, Begley and Ioannidis (2015) estimate that the reliability (i.e., TP rate) of social science experiments published in the journals *Nature* and *Science* is approximately 67%. Based on the large-scale reproducibility project of OpenScienceCollaboration (2015), Wilson and Wixted (2018) estimate that the reliability of published studies in well-respected social-psychology journals is approximately 49%, and in cognitive psychology journals, approximately 81%. For the chosen values of k and m, our model produces reliability measures ranging from 15% to 83%, for $\delta = 0.2$, and from 25% to 86%, for $\delta = 0.43$. Note that this wide range of values includes some reliability percentages that seem rather low (e.g., 15% and 25%).
2. There have also been a number of recent attempts to estimate the typical *power* of published research using meta-analytic effect sizes. For example, Button et al. (2013b) estimate that the median statistical power in published neuroscience is approximately 21% and Dumas-Mallet et al. (2017), using a similar methodology, conclude that the median statistical power in the biomedical sciences ranges from about 9% to 30% depending on the disease under study. Most recently, Lamberink et al. (2018) analyzed the literature of RCTs and determined that the median power is approximately 9% overall (and about 20% in the subset of RCTs included in significant meta-analyses). For the chosen values of k and m, our model produces a simulated published literature with median power ranging from 9% to 60%, for $\delta = 0.2$, and from 21% to 71%, for $\delta = 0.43$.
3. While the *publication rate* numbers we obtain may appear rather low, ranging from 3% to 13%, consider that Siler, Lee, and Bero (2015), in a systematic review of manuscripts submitted to three leading medical journals, observed a publication rate of 6.2%. In a review of top psychology journals, Lee and Schunn (2011) found that acceptance rates ranged between 14% and 32%. It is important to recognize that these empirical estimates are calculated based on the subset of studies that are submitted for publication. If researchers are self-selecting their "most publishable" work for submission to journals, then these estimates will substantially overestimate the true publication rate of all studies. Song, Loke, and Hooper (2014) estimate that about 85% of unpublished studies are unpublished for the simple reason that they are never submitted for publication. Our model also assumes that negative studies are never published ($B = 0$) and this (not entirely realistic) assumption also contributes to the low publication rate numbers we obtain. Senn (2012) provides substantial insights on both these related issues: the level of publication bias, and how researchers may choose to submit research based on the perceived probability of acceptance. Note that while the scenarios with higher publication rates may appear more plausible, these scenarios have very (perhaps unreasonably) high values for reliability and median power.

Regardless of the values chosen for k and m, the overall trends for our metrics of interest will be similar. See Table 2. As k increases, the reliability of published research (*REL*) will increase, while the number of true breakthrough discoveries (*DSCV*) decreases. As m increases, reliability (*REL*) decreases while the number of true breakthrough discoveries (*DSCV*) increases. This strikes us as both reasonable and realistic. When a greater emphasis is placed on novelty (i.e., when m is larger), there will be a greater number of smaller, high-risk studies published. While these publications are less reliable, they are more numerous and more likely to produce a breakthrough finding. When data are more affordable relative to overall study cost (i.e., when k is larger), there will be fewer, larger studies and as a result the published literature will be more reliable but less likely to produce a breakthrough finding.

Before moving on, let us also briefly consider how the accuracy of published effect sizes changes with k and m, see Table 3. Overall, we see that the average of the absolute published effect size, $|\hat{\mu}_d|_{\text{PUB}}$, is always much larger than δ. Furthermore, we see that a large proportion of published effect sizes are negative. This is to be expected when *psp* values are small and many published studies are FPs. For the subset of studies that are positive, published, *and true*, we see that the Type M and the Type S errors are greatest when k is small, m is large, and δ is small. This is the same configuration that produces low-powered publications. Indeed, Type M and Type S errors are known to occur more frequently when power is low (Gelman and Carlin 2014).

Table 3. For ecosystems defined by fixed $\alpha = 0.05$, $A = (1 - psp)^m$, and $B = 0$, and varying values of k, m, and δ, the table lists estimates for the metrics of interest related to the accuracy of the published effects sizes.

δ	m	k	$Mpwr_{\text{PUB}}$	$\hat{\mu_d}^{neg}_{\text{PUB}}$	$\lvert\hat{\mu_d}\rvert_{\text{PUB}}$	Type M err.	Type S err.
0.20	1	20	0.20	0.21	0.60	2.42	0.03
0.20	3	20	0.12	0.31	0.72	2.70	0.04
0.20	9	20	0.09	0.43	0.80	2.77	0.04
0.20	1	100	0.42	0.14	0.37	1.68	0.01
0.20	3	100	0.32	0.23	0.45	1.81	0.01
0.20	9	100	0.21	0.37	0.51	1.92	0.01
0.20	1	500	0.60	0.09	0.29	1.40	0.00
0.20	3	500	0.55	0.17	0.31	1.40	0.00
0.20	9	500	0.44	0.33	0.35	1.48	0.00
0.43	1	20	0.42	0.14	0.75	1.63	0.01
0.43	3	20	0.32	0.21	0.80	1.65	0.00
0.43	9	20	0.21	0.39	0.90	1.74	0.01
0.43	1	100	0.59	0.09	0.60	1.36	0.00
0.43	3	100	0.54	0.18	0.64	1.43	0.00
0.43	9	100	0.43	0.32	0.69	1.47	0.00
0.43	1	500	0.71	0.08	0.54	1.29	0.00
0.43	3	500	0.68	0.17	0.56	1.29	0.00
0.43	9	500	0.60	0.28	0.57	1.30	0.00

NOTES: These include the median power of published studies ($Mpwr_{\text{PUB}}$), the mean absolute published effect size ($\lvert\hat{\mu_d}\rvert_{\text{PUB}}$), the proportion of published effect sizes that are negative ($\hat{\mu_d}^{neg}_{\text{PUB}}$), the "Type S" error, and the "Type M" error.

Table 4. A comparison between ecosystems with $\alpha = 0.05$ and ecosystems with $\alpha = 0.005$.

δ	k	m	Change in *PR*	Change in *REL*	Change in *DSCV*	Change in D_{REL}	Change in $Mpwr_{\text{PUB}}$	Change in $\text{IQpsp}_{\text{PUB}}$
0.20	20	1	1.14	1.58	0.12	7.57	2.73	0.83
0.20	20	3	0.71	2.58	0.15	7.68	4.57	1.05
0.20	20	9	0.32	5.38	0.21	8.01	5.11	1.46
0.20	100	1	0.99	1.30	0.19	5.47	1.37	0.86
0.20	100	3	0.73	1.81	0.23	5.57	1.77	0.95
0.20	100	9	0.39	3.47	0.31	5.89	2.52	1.27
0.20	500	1	0.91	1.18	0.28	4.06	1.07	0.89
0.20	500	3	0.75	1.48	0.33	4.14	1.17	0.95
0.20	500	9	0.46	2.49	0.43	4.40	1.42	1.12
0.43	20	1	0.99	1.30	0.19	5.49	1.36	0.86
0.43	20	3	0.74	1.81	0.22	5.59	1.78	0.95
0.43	20	9	0.39	3.46	0.31	5.91	2.55	1.27
0.43	100	1	0.92	1.18	0.28	4.10	1.08	0.89
0.43	100	3	0.75	1.49	0.32	4.18	1.19	0.95
0.43	100	9	0.46	2.51	0.42	4.44	1.44	1.19
0.43	500	1	0.92	1.14	0.37	3.49	1.02	0.90
0.43	500	3	0.78	1.37	0.42	3.55	1.07	0.95
0.43	500	9	0.51	2.15	0.55	3.78	1.19	1.12

NOTES: For all ecosystems, we have $A = (1 - psp)^m$, $B = 0$, and varying values of k, m, and δ. The table lists estimates for the ratios of the publication rate (*PR*), the ratio of the reliability (*REL*), the ratio of the number of true breakthrough discoveries (*DSCV*), the ratio of the median power of published studies($Mpwr_{\text{PUB}}$), and the ratio of the interquartile range of the distribution of *psp* values amongst published studies ($\text{IQpsp}_{\text{PUB}}$).

3. The Effects of Adopting Lower Significance Thresholds

In this section, we investigate the impact of adopting lower significance thresholds. *How might the published literature be different if the* $\alpha = 0.05$ *threshold was lowered?* For positive studies, we will assume that the sample size of a study does not affect the likelihood of publication (at least not directly) and that studies with lower *psp* are more likely to be published according to the simple function introduced earlier, $A = (1 - psp)^m$. We compute the metrics of interest for 54 different ecosystems. Each ecosystem is uniquely defined with either $\delta = 0.2$ or $\delta = 0.43$, with one of three possible values for m (= 1, 3, 9), with one of three possible values for k (= 20, 100, 500), and most importantly, with one of three possible values for the α significance threshold (= 0.005, 0.020, 0.050).

3.1. Results

In Table 4, we note how the various metrics change with $\alpha = 0.005$ relative to $\alpha = 0.05$, by reporting ratios. While we focus our discussion specifically on scenarios for which $\delta = 0.20$, the results for $\delta = 0.43$ are all very similar. Figures in the supplemental material plot the complete results. Based on our results, we can make the following conclusions on the impact of adopting a lower, more stringent, significance threshold.

1. Reliability is substantially increased with a lower threshold. Based on our results, comparing $\alpha = 0.005$ to $\alpha = 0.05$, the impact on *REL* is greatest when δ is small, k is small, and m is large, see Table 4. This is due to the fact that with a lower significance threshold policy, attempted studies are typically of higher power (particularly so when k is small) and of higher prestudy probability. To understand why there would be more studies with higher power and higher prestudy probabilities, consider that with a lower α threshold, the probability of obtaining a significant result (i.e., p-value $< \alpha$) "by chance" is substantially reduced. To maximize one's expected number of publications (ENP) when $\alpha = 0.005$, it is a much better strategy to pursue higher *pwr* and higher *psp* strategies. This is particularly true when the effect size, δ, is small.
2. A disadvantage of the lower threshold is that the number of breakthrough discoveries is substantially lower, see Table 4 and Figure S3. This is due to the fact that with a lower α threshold, fewer "high-risk" (i.e., low *psp*) studies are attempted. The chance that the p-value will fall below α, when α is lowered to 0.005, is already risky enough. The D_{REL} numbers we obtain suggest that, while fewer low *psp* studies will be published, these will be much more reliable.
3. When increasing study power is less costly relative to the total cost of a study (i.e., when k is larger, or when δ is larger), the benefit of lowering the significance threshold (increased *REL*) is somewhat smaller. However, the downside (decreased *DSCV*) is *substantially* smaller, see left-panels of Figures S1 and S3. This suggests that a policy of lowering the significance threshold would perhaps be best suited in a field of research in which increasing one's sample size is less burdensome. This nuance recalls the suggestion of Ioannidis, Hozo, and Djulbegovic (2013): "Instead of trying to fit all studies to traditionally acceptable Type I and Type II errors, it may be preferable for investigators to select Type I and Type II error pairs that are optimal for the truly important outcomes and for clinically or biologically meaningful effect sizes."
4. When novelty is more of a requirement for publication (i.e., when m is larger), the benefit of lowering the significance threshold is larger and the downside smaller. This result is due to the fact that a smaller α will incentivize researchers to allocate resources in the direction toward *either* higher powered *or* higher *psp* studies (i.e., away from the South-

West corner of the plots in Figure 2). There is a choice between moving toward higher *pwr* (North) or toward higher *psp* (East), and different costs associated with each direction. This suggests that for a lower significance threshold policy to be most effective, editors should also adopt, in conjunction, stricter requirements for research novelty. To illustrate, consider three ecosystems of potential interest with their estimated *REL*, *DSCV*, and *PR* metrics (all with $\delta = 0.2$):

(1) The *baseline* defined by $\alpha = 0.05$, $m = 3$, and $k = 500$ with:
$REL = 0.650$, $DSCV = 0.105$, and $PR = 0.054$;
(2) the *alternative* defined by $\alpha = 0.005$, $m = 3$, and $k = 500$ with:
$REL = 0.963$, $DSCV = 0.034$, and $PR = 0.040$; and
(3) the *suggested* defined by $\alpha = 0.005$, $m = 9$, and $k = 500$ with:
$REL = 0.898$, $DSCV = 0.112$, and $PR = 0.015$.
Note that while the *suggested* has high REL and relatively high *DSCV*, the *PR* is substantially reduced.

5. As mentioned earlier, the balance between exploratory and confirmatory research is an important aspect of a scientific ecosystem. The results show that the interquartile range for *psp* does not change substantially with α. As such, we could conclude that even with a much lower significance threshold, there will still be a wide range of studies attempted in terms of their *psp*. However, *psp* values do tend to be substantially higher with smaller α. As such, we should expect that, with smaller α, research will move toward more confirmatory, and less exploratory studies.
6. With a lower α significance threshold, the published effect sizes are much more accurate. When $\alpha = 0.005$, the Type M error is reduced to at most 1.4 (i.e., effect size estimates are inflated by at most 40%) and Type S error is negligible; see Figures 4, 5, S7 and S8. These reductions are greatest when data are relatively expensive and when the true effect size is small (i.e. when $k = 20$ and $\delta = 0.20$).

4. The Effects of Strict *a priori* Power Calculation Requirements

In this section, we investigate the effects of requiring "sufficient" sample sizes. In practical terms, this means adopting publication policies that require studies to show *a priori* power calculations indicating that sample sizes are "sufficiently large" to achieve the desired level of statistical power, typically 80%. Whereas before, the chance of publishing a positive study in our framework depended only on *psp*, we now set the probability of publication, *A*, to also depend on power. In doing so however, we wish to

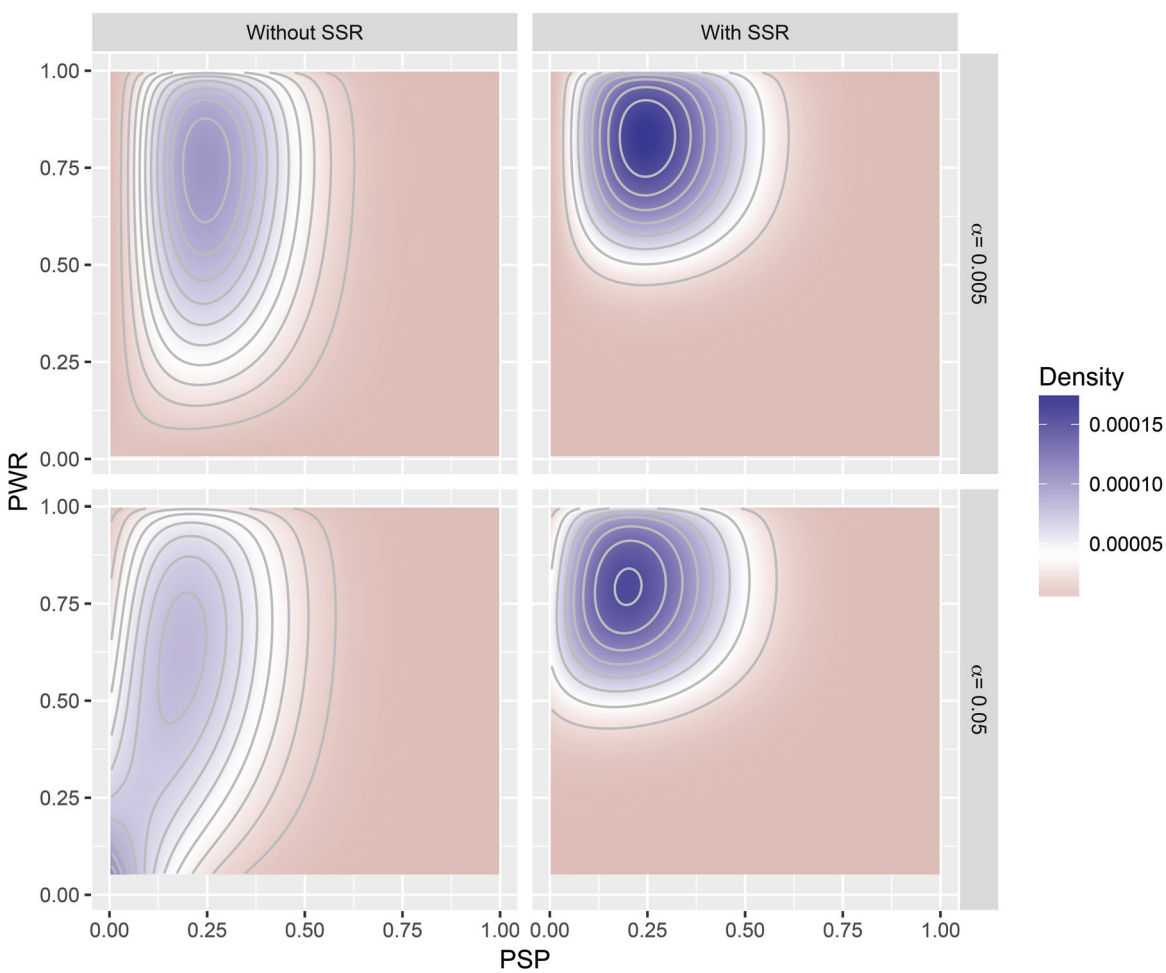


Figure 3. Heat-maps show the density of published articles, $f_{PUB}(psp, pwr)$, for the four policies of interest, with fixed $\delta = 0.2$, $k = 500$ and $m = 3$.

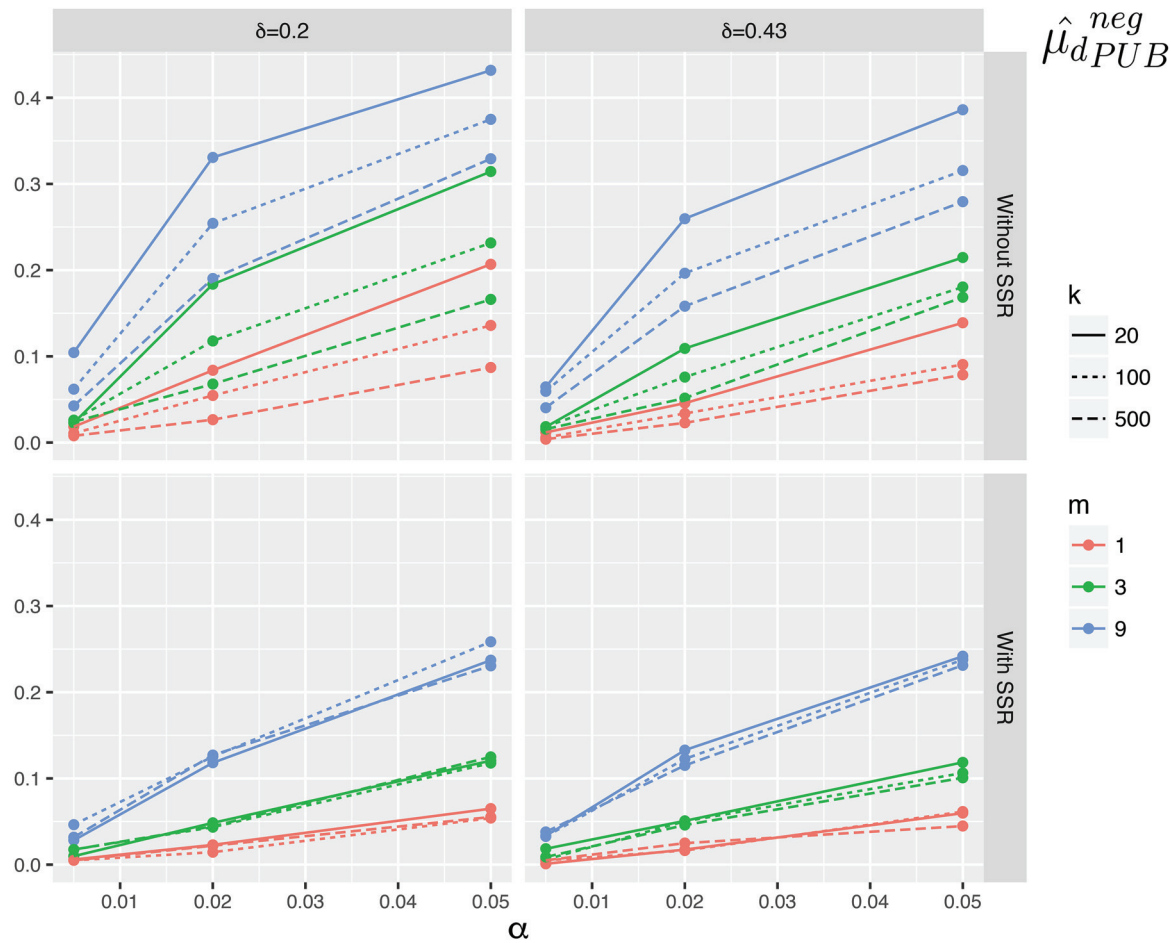


Figure 4. Values of $\hat{\mu}_{d_{\text{PUB}}}^{neg}$ for varying values of k, m and α; results for $\delta = 0.2$ and $\delta = 0.43$ on the left and right panels, respectively. Top panels shows results with no power requirement (i.e., $A = (1 - psp)^m$). Bottom panels shows results with power requirement (i.e., A defined as per equation A4 (with $c_{50} = 0.5$ and $c_{95} = 0.8$)).

acknowledge the fact that *a priori* sample size claims are often "wildly optimistic" (Bland 2009).

The "sample size samba" (Schulz and Grimes 2005)—the practice of "tweaking aspects of sample size calculation" (Hazra and Gogtay 2016) in order to obtain what is affordable—often results in a study with less than 80% power being advertised as having 80% power. Even in ideal circumstances, power can be exaggerated due to "optimism bias" (Djulbegovic et al. 2011), also known as the "illusion of power" (Vasishth and Gelman 2017), which occurs when the anticipated effect size is based on a literature filled with overestimates due to publication bias.

To take into account the problematic nature of *a priori* power calculations, our model is defined such that possessing only 50% power substantially reduces, but does not eliminate, the chance of publication. For studies which really do have 80% power and higher, there will be no notable reduction in the probability of publication. See the appendix for details on how we define A as a continuous function of both *psp* and *pwr*.

We calculated the metrics of interest for the same 54 different ecosystems as in Section 3, with our new definition of A. To contrast these ecosystems with those discussed in the previous section, we refer to these ecosystems as "with SSR" (sample size requirements).

4.1. Results

See Table 5. Based on our results, we can make the following main conclusions on the measurable consequences of adopting a journal policy requiring an *a priori* sample size justification.

1. With SSR, we observed much higher powered studies, see Figure S5. Reliability is also increased, particularly when novelty is highly prized (m is large) and effect sizes are small. This result is as expected. As soon as having a small sample size jeopardizes the probability of publication, it is in the researcher's best interest to conduct higher powered studies.
2. Requiring "sufficient sample sizes" for publication can be quite detrimental in terms of the number of breakthrough discoveries. The impact on *DSCV* is greatest when δ, k, and m are all small; see Table 5. The D_{REL} numbers show that reliability is particularly increased for low *psp* publications.
3. In conjunction with requiring larger sample sizes, it may be wise to place greater emphasis on research novelty. As in Section 3, such a combined approach could see an increase in reliability with only a limited decrease in discovery. This trade-off is most beneficial when k is small. Consider three ecosystems of potential interest (all with $\alpha = 0.05$, and $\delta = 0.2$) with their estimated *REL*, *DSCV*, and *PR* metrics:

Table 5. A comparison between ecosystems with a sample size requirement and ecosystems without.

δ	k	m	Change in PR	Change in REL	Change in $DSCV$	Change in D_{REL}	Change in $Mpwr_{\text{PUB}}$	Change in $\text{lQpsp}_{\text{PUB}}$
0.20	20	1	1.46	1.44	0.32	3.97	3.57	0.87
0.20	20	3	1.04	2.08	0.35	4.01	6.13	1.08
0.20	20	9	0.65	3.33	0.41	4.12	7.61	1.38
0.20	100	1	1.21	1.18	0.53	2.40	1.74	0.92
0.20	100	3	1.03	1.44	0.56	2.42	2.25	1.00
0.20	100	9	0.77	2.03	0.63	2.48	3.33	1.20
0.20	500	1	1.08	1.07	0.75	1.63	1.27	0.95
0.20	500	3	1.00	1.18	0.79	1.64	1.39	1.00
0.20	500	9	0.85	1.44	0.85	1.67	1.68	1.06
0.43	20	1	1.21	1.18	0.53	2.39	1.72	0.92
0.43	20	3	1.03	1.44	0.56	2.41	2.25	1.00
0.43	20	9	0.77	2.02	0.63	2.47	3.33	1.20
0.43	100	1	1.08	1.07	0.74	1.65	1.29	0.95
0.43	100	3	1.00	1.19	0.78	1.66	1.41	1.00
0.43	100	9	0.85	1.45	0.85	1.69	1.72	1.12
0.43	500	1	1.04	1.04	0.89	1.38	1.14	0.97
0.43	500	3	1.00	1.10	0.92	1.38	1.19	1.00
0.43	500	9	0.90	1.25	0.98	1.40	1.32	1.12

NOTE: A is defined by equation A4 (with $c_{50} = 0.5$ and $c_{95} = 0.8$), $B = 0, \alpha = 0.05$, and varying values of k and m.

(1) The *baseline* defined by $m = 3$, $k = 100$, and without SSR; with $REL = 0.524$, $DSCV = 0.216$, and $PR = 0.043$;

(2) the *alternative* defined by $m = 3$, $k = 100$ and with SSR; with $REL = 0.757$, $DSCV = 0.122$, and $PR = 0.044$; and

(3) the *suggested* defined by $m = 6$, $k = 100$ and with SSR; with $REL = 0.502$, $DSCV = 0.325$, and $PR = 0.022$.
Note that the while the *suggested* has both higher REL and higher $DSCV$ than the *baseline*, the PR is reduced.

4. With SSR, the Type M error is at most 1.2 (i.e., effect size estimates are inflated by at most 20%) and Type S error is negligible; see Figures 4, 5, S7 and S8.

4.2. In Tandem: The Effects of Adopting Both a Lower Significance Threshold and a Power Requirement

We are also curious as to whether lowering the significance threshold *in addition* to requiring larger sample sizes would carry any additional benefits relative to each policy innovation on its own. See Table A.1 for results, and consider two main findings:

1. For ecosystems with SSR, the distribution of published studies does not change dramatically when α is lowered. Figure 3 shows the density describing the characteristics of published studies, $f_{\text{PUB}}(psp, pwr)$, for the four policies of interest, with $k = 500$ and $m = 3$: (1) $\alpha = 0.05$, without SSR; (2) $\alpha = 0.05$, with SSR; (3) $\alpha = 0.005$, without SSR; and (4) $\alpha = 0.005$,

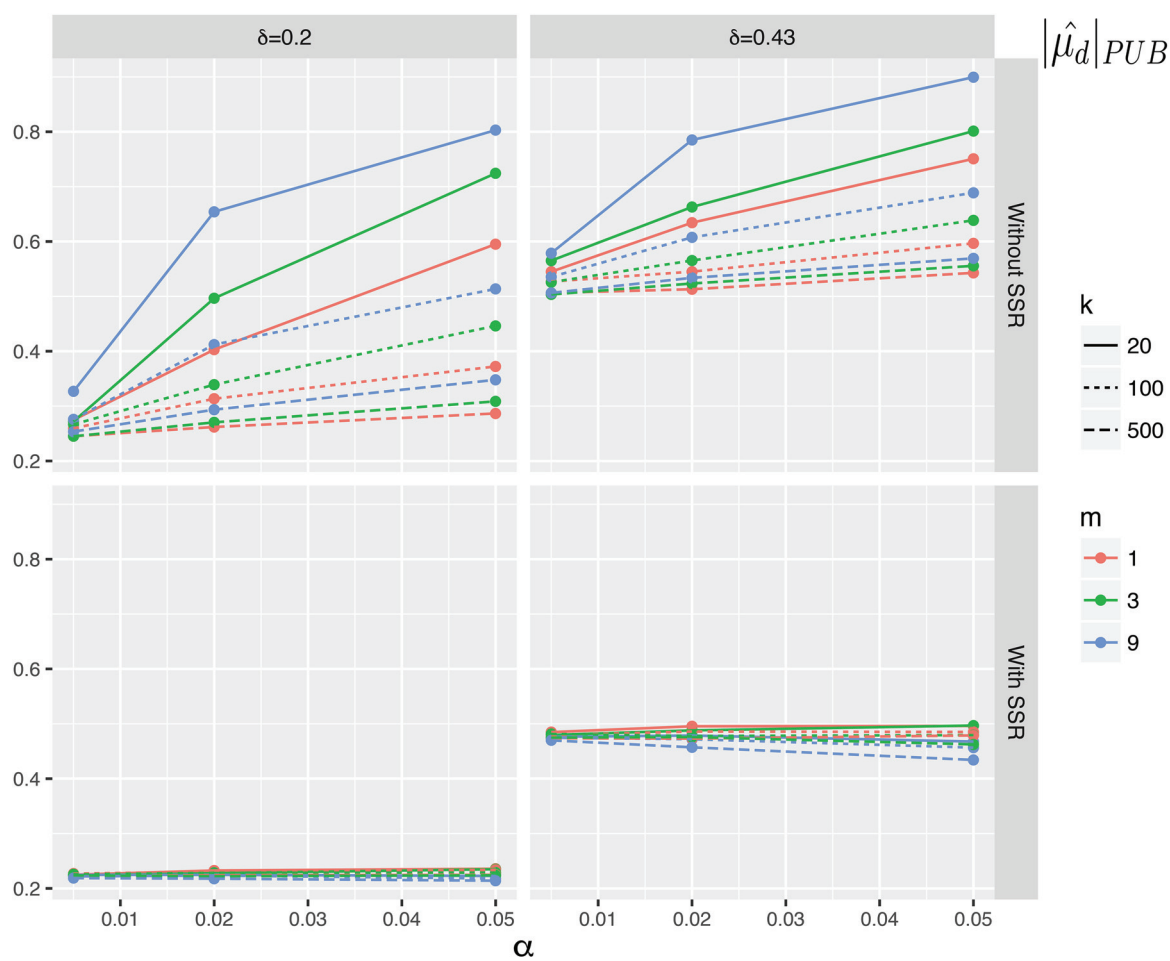


Figure 5. Values of $|\hat{\mu}_d|_{\text{PUB}}$ for varying values of k, m and α; results for $\delta = 0.2$ and $\delta = 0.43$ on the left and right panels respectively. Top panels shows results with no power requirement (i.e., $A = (1 - psp)^m$). Bottom panels shows results with power requirement (i.e., A defined as per equation A4 (with $c_{50} = 0.5$ and $c_{95} = 0.8$)).

with SSR. The difference between the densities in (2) and (4) is primarily a matter of a shift in *psp*.

2. As expected, lowering the significance threshold further increases reliability and the number of breakthrough discoveries is further decreased; see Table A.1, and Figures S1 and S3.

5. Conclusion

There remains substantial disagreement on the merits of requiring greater statistical stringency to address the reproducibility crisis. Yet all should agree that innovative publication policies can be part of the solution. Going forward, it is important to recognize that current norms for Type 1 and Type 2 error levels have been driven (almost) entirely by tradition and inertia rather than careful coherent planning and result-driven decisions (Hubbard and Bayarri 2003). Hence, improvements should be possible.

In response to Amrhein, Korner-Nievergelt, and Roth (2017) who suggest that a more stringent α threshold will lead to published science being *less* reliable, our results suggest otherwise. However, just as Amrhein, Korner-Nievergelt, and Roth (2017) contend, our results indicate that the publication rate will end up being substantially lower with a smaller α. While going from $p < 0.05$ to $p < 0.005$ may be beneficial to published science in terms of reliability, we caution that there may be a large cost in terms of fewer true breakthrough discoveries. One must ask whether a large reduction in novel discoveries is an acceptable price to pay for a large increase in their reliability? Importantly, and somewhat unexpectedly, our results suggest that this can be mitigated (to some degree) by adopting a greater emphasis on research novelty. In practice, implementing stricter requirements for research novelty, could be accomplished by greater editorial emphasis on "surprising results" as a requisite for publication. This approach, however, might be difficult to achieve unless one is willing to accept a much lower publication rate. In summary, publishing *less* may be the necessary price to pay for obtaining *more* reliable science.

Recently, some have suggested that researchers choose (and justify) an "optimal" value for α, for each unique study; see Mudge et al. (2012), Ioannidis, Hozo, and Djulbegovic (2013), and Lakens et al. (2018). Each study within a journal would thereby have a different set of criteria. This is a most interesting idea and there are persuasive arguments in favor of such an approach. Still, it is difficult to anticipate how such a policy would play out in practice and how the research incentive structure would change in response. A model, like the one presented here, but with the α threshold set to vary with *psp*, could provide useful insight.

We are also cautious about greater sample size requirements. In summary, we found that the impacts of adopting a sample size requirement policy are similar to the impacts of lowering the α significance threshold. While improving reliability, requiring studies to show "sufficient power" will severely limit novel discoveries in fields where acquiring data is expensive. Is it beneficial for editors and reviewers to consider whether a study has "sufficient power"? How much should these criteria influence publication decisions? Answers to these questions are not at all obvious. Again, using the methodology introduced, we suggest that adopting a greater emphasis on research novelty may mitigate, to a certain extent, some of the downside of adopting greater sample size requirements at the cost of lowering the overall number of published studies. Given that, as a result of publication bias, it can often be better to discard 90% of published results for meta-analytical purposes (Stanley, Jarrell, and Doucouliagos 2010), this may be an approach worth considering.

Our main recommendation is that, before adopting any (radical) policy changes, we should take a moment to carefully consider, and model how, these proposed changes might impact outcomes. The methodology we present here can be easily extended to do just this. Two scenarios of interest come immediately to mind.

First, it would be interesting to explore the impact of publication bias, i.e., the tendency of journals to reject non-significant results (Sterling, Rosenbaum, and Weinkam 1995). This could be done by allowing B to take different non-zero values. Based on simulation studies, de Winter and Happee (2013) suggest that publication bias can in fact be beneficial for the reliability of science. However, under slightly different assumptions, van Assen et al. (2014) arrive at a very different conclusion. Clearly, a better understanding of how publication bias changes a scientist's incentives is needed. If statistical significance becomes a much less important screening criterion for publication, how will published science change? Recently, a number of influential researchers have argued that to address low reliability, reviewers should "abandon statistical significance" (McShane et al. 2017). While our framework could be extended to explore this proposal (e.g., by taking $A = B$), these researchers suggest substituting statistical significance with a "more holistic view of the evidence" to inform publication decisions. Clearly, this does not lend itself to our modeling framework, or similar meta-research paradigms.

Second, it would be worthwhile to investigate the potential impact of requiring study preregistration. Coffman and Niederle (2015) use the accounting of Ioannidis (2005) to evaluate the effect of preregistration on reliability and conclude that pre-registration will have only a modest impact. However, the impact on the publication rate and on the number of breakthrough findings is still not well understood. This is particularly relevant given the current trend to adopt "result-blind peer-review" (Greve, Bröder, and Erdfelder 2013) policies including, most recently, the policy of *Registered Reports* (Chambers et al. 2015). The conditional equivalence testing publication policy proposed in our earlier work, see Campbell and Gustafson (2018), could also be considered.

Our methodology assumes above all that researchers' decisions are driven exclusively by the desire to publish. But the situation is more complex. Publication is not necessarily the end goal for a scientific study and requirements with regards to significance and power are not only encountered at the publication stage. In the planning stages, before a study even begins, ethics committees and granting agencies will often have certain minimal requirements; see Ploutz-Snyder, Fiedler, and Feiveson (2014) and Halpern, Karlawish, and Berlin (2002). And after a study is published, regulatory bodies and policy makers will also often subject the results to a different set of norms.

We also assumed a framework of independent Bernoulli trials. Of course, in practice, most studies are not conducted independently. For example, in the context of clinical trials, phase 3 studies are conducted based on the success or failure of earlier phase 2 studies (see Burt et al. (2017) who model the advantages and disadvantages of different clinical development strategies). It is difficult to anticipate the consequences of failing to incorporate such dependencies in our model. A more elaborate model, in which studies are correlated and the *psp* of subsequent studies is updated appropriately, could prove very informative.

Finally, it is important to acknowledge 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 (Contopoulos-Ioannidis, Ntzani, and Ioannidis 2003; Djulbegovic and Hozo 2007). Each policy will have its advantages and disadvantages. Our modeling exercise makes this all the more evident and forces us to carefully consider different potential trade-offs.

Code

Please note that all code to produce the results, tables, and figures in this article have been posted to a repository on the Open Science Framework, DOI 10.17605/OSF.IO/YQCVA.

Acknowledgments

We wish to gratefully acknowledge Prof. Will Welch and Prof. John Petkau for their valuable suggestions and advice. Furthermore, we wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC grant number RGPIN 183772-13).

Appendix

Here, we introduce some compact notation that will be useful for expressing distributional quantities of interest. Particularly, the probabilities comprising the distribution of a study across the eight categories are expressed as $q_{abc}(psp, pwr)$, where $a \in \{0, 1\}$ indicates the truth ($a = 0$ for null, $a = 1$ for alternative), $b \in \{0, 1\}$ indicates the statistical finding ($b = 0$ for negative, $b = 1$ for positive), and $c \in \{U, P\}$ indicates publication status. As examples, we could write $FN_{\mathrm{UN}} = q_{10U}$, or $\mathrm{TP}_{\mathrm{PUB}} = q_{11P}$. We also use a plus notation to add over subscripts, so, for instance, $q_{1+P} = \mathrm{TP}_{\mathrm{PUB}} + FN_{\mathrm{PUB}}$.

As motivated above, we consider properties that result from a scientist or group of scientists stochastically allocating T *resources* (not studies per se) according to a distribution across (psp, pwr). We denote the function $n(pwr)$ as the required sample size to obtain a power of pwr. Presuming the incentive to publish, the density of this distribution is taken proportional to $ENP(psp, pwr)$, which we express as

$$\begin{aligned} f_{\mathrm{RES}}(psp, pwr) &\propto ENP(psp, pwr) \\ &\propto \{k + n(pwr)\}^{-1} q_{++P}(psp, pwr). \end{aligned} \tag{A1}$$

Consequently, the distribution of (psp, pwr) across *attempted* studies has density

$$\begin{aligned} f_{\mathrm{ATM}}(psp, pwr) &\propto \{k + n(pwr)\}^{-1} f_{\mathrm{RES}}(psp, pwr) \\ &\propto \{k + n(pwr)\}^{-2} q_{++P}(psp, pwr). \end{aligned} \tag{A2}$$

In turn, the distribution of (psp, pwr) across *published studies* has density

$$\begin{aligned} f_{\mathrm{PUB}}(psp, pwr) &\propto f_{\mathrm{ATM}}(psp, pwr) q_{++P}(psp, pwr) \\ &\propto \{k + n(pwr)\}^{-2} \{q_{++P}(psp, pwr)\}^2. \end{aligned} \tag{A3}$$

Note particularly that $f_{\mathrm{PUB}}(psp, pwr) \propto \{f_{\mathrm{RES}}(psp, pwr)\}^2$. Hence, the distribution of (psp, pwr) across published studies is a concentrated version of the distribution describing how resources are deployed.

A1. Ecosystem Metrics: Further Details

We evaluate each ecosystem of interest on the basis of the following five metrics.

A1.1.Publication Rate and the Number of Studies Attempted/Published

If T units of resources are deployed according to $f_{\mathrm{RES}}()$, then we expect that N_{ATM} studies will be attempted, where

$$T^{-1} N_{\mathrm{ATM}} = E_{\mathrm{RES}} \left[\{k + n(\mathrm{PWR})\}^{-1} \right].$$

Similarly, N_{PUB} studies will be published, with

$$T^{-1} N_{\mathrm{PUB}} = E_{\mathrm{RES}} \left[\{k + n(\mathrm{PWR})\}^{-1} q_{++P}(\mathrm{PSP}, \mathrm{PWR}) \right].$$

The ratio $N_{\mathrm{PUB}}/N_{\mathrm{ATM}}$, which does not depend on T, is of evident interest, as the *publication rate* (*PR*) for attempted studies.

A1.1.2. The Reliability (REL)

A highly relevant metric for the scientific ecosystem is the proportion of published findings that are correct. In all the ecosystems, we consider in this article, we make the assumption that only positive results are published (i.e., $B = 0$). Therefore, we can express reliability (*REL*) simply as

$$\mathrm{REL} = \frac{E_{\mathrm{ATM}} \left\{ q_{11P}(\mathrm{PSP}, \mathrm{PWR}) \right\}}{E_{\mathrm{ATM}} \left\{ q_{+1P}(\mathrm{PSP}, \mathrm{PWR}) \right\}}.$$

More generally, in ecosystems where negative results might be published (i.e., $B(psp, pwr) \neq 0$), the reliability would equal the proportion of published articles that reach a correct conclusion, i.e.,

$$\mathrm{REL} = \frac{E_{\mathrm{ATM}} \left\{ q_{00P}(\mathrm{PSP}, \mathrm{PWR}) + q_{11P}(\mathrm{PSP}, \mathrm{PWR}) \right\}}{E_{\mathrm{ATM}} \left\{ q_{++P}(\mathrm{PSP}, \mathrm{PWR}) \right\}}.$$

A1.1.3. Breakthrough Discoveries (DSCV)

The ability of the scientific ecosystem to produce breakthrough findings is an important attribute. We quantify this in terms of spending T resource units yielding an expectation of *DSCV* breakthrough results. Here, a true breakthrough result is defined as a TP and published study that results from a *psp* value below a threshold, that is, a very surprising positive finding that gets published and also happens to be true. If we set the breakthrough threshold as $psp < 0.05$, then

$$T^{-1} DSCV = E_{\mathrm{RES}} \left\{ I_{(0,0.05)}(\mathrm{PSP}) \frac{q_{11P}(\mathrm{PSP}, \mathrm{PWR})}{k + n(\mathrm{PWR})} \right\}$$

Note that we can also write this out in terms of the density of attempted studies, f_{ATM}:

$$\begin{aligned} T^{-1}\mathrm{DSCV} &= \frac{\int h(\mathrm{PSP}, \mathrm{PWR}) f_{\mathrm{ATM}}(\mathrm{PSP}, \mathrm{PWR})(k + n(\mathrm{PWR}))}{\int f_{\mathrm{ATM}}(\mathrm{PSP}, \mathrm{PWR})(k + n(\mathrm{PWR}))} \\ &= \frac{E_{\mathrm{ATM}}[I_{(0,0.05)}(\mathrm{PSP}) q_{11P}(\mathrm{PSP}, \mathrm{PWR})]}{E_{\mathrm{ATM}}[k + n(\mathrm{PWR})]}. \end{aligned}$$

where we define the function: $h(\text{PSP}, \text{PWR}) = I_{(0,0.05)}(\text{PSP}) \frac{q_{11P}(\text{PSP},\text{PWR})}{k+n(\text{PWR})}$.

We also wish to quantify the reliability of those published studies with *psp* values below a threshold of 0.05. We have that:

$$D_{\text{REL}} = \frac{E_{\text{ATM}}\left\{I_{(0,0.05)}(\text{PSP})q_{11P}(\text{PSP}, \text{PWR})\right\}}{E_{\text{ATM}}\left\{I_{(0,0.05)}(\text{PSP})q_{+1P}(\text{PSP}, \text{PWR})\right\}}.$$

A1.1.4. The Median Power of Published Studies ($Mpwr_{\text{PUB}}$)

We already mentioned the relevance of the *psp* marginals of (A2) and (A3). In a similar vein, the marginal distributions of *pwr* under each of these distributions are readily interpreted metrics of the ecosystem. We define $Mpwr_{\text{PUB}}$ as the median of *pwr* under the $f_{\text{PUB}}()$ distribution.

A1.1.5. The Balance between Exploration and Confirmation ($\text{IQpsp}_{\text{PUB}}$)

There has been much discussion about the desired balance between researchers looking for *a priori* unlikely relationships versus confirming suspected relationships put forth by other researchers; see for example, Sakaluk (2016) and Kimmelman, Mogil, and Dirnagl (2014). The marginal distribution of *psp* arising from $f_{\text{ATM}}(psp, pwr)$ describes the balance between exploration versus confirmation for attempted studies, while the *psp* marginal from $f_{\text{PUB}}(psp, pwr)$ does the same for published studies. More specifically, we report the interquartile ranges of these marginal distributions for a given ecosystem.

A1.1.6. The Accuracy of Published Research

Consider extending $f_{\text{ATM}}(psp, pwr)$ to $f_{\text{ATM}}(psp, pwr, d, u)$, where:

$f_{\text{ATM}}(d, u|psp, pwr) = f_{\text{ATM}}(d|psp, pwr)f_{\text{ATM}}(u)$, and where

- $f_{\text{ATM}}(u)$ is the density of a standard uniform distribution; and
- $f_{\text{ATM}}(d|psp, pwr)$ is the density of a two-component mixture distribution such that:
  $Pr(d = \delta) = psp$; and $Pr(d = 0) = 1 - psp$.

In order to obtain a sample from the distribution of published effect sizes, we proceed according to the following steps.
Set $w = 1$, $q = 1$ and $b = 1$. While $w < 1000$:

1. Draw a Monte Carlo realization of $(psp_{[b]}, pwr_{[b]}, d_{[b]}, u_{[b]})$ from $f_{\text{ATM}}(psp, pwr, d, u)$.
2. Simulate a two-sample Normally distributed dataset, $data_{[b]}$, with a sample size of $n(pwr_{[b]})$, a true difference in population means of $\mu_d = d_{[b]}$, and a true variance of $\sigma^2 = 1$.
3. Run a standard t-test on $data_{[b]}$, and obtain the estimated effect size, $\hat{\delta}_{[b]}$, and the two-sided p-value, $pval_{[b]}$.
4. If $u_{[b]} < A(psp_{[b]}, pwr_{[b]})$ AND $pval_{[b]} < \alpha$, then do:
   $\hat{\delta}^{[q]}_{+1P} = \hat{\delta}_{[b]}$; and $q = q + 1$.
5. If $u_{[b]} < A(psp_{[b]}, pwr_{[b]})$ AND $pval_{[b]} < \alpha$, AND $d_{[b]} = \delta$, then do:
   $\hat{\delta}^{[w]}_{11P} = \hat{\delta}_{[b]}$; and $w = w + 1$.
6. Set $b = b + 1$.

Then we have that $\{\hat{\delta}_{+1P}\}$ is a Monte Carlo sample from the conditional distribution of effect sizes given that studies are both positive *and* published. We also have that $\{\hat{\delta}_{11P}\}$ is a Monte Carlo sample from the conditional distribution of effect sizes given that studies are positive, *and* published, *and* true.

Using these Monte Carlo samples, we can easily approximate the following metrics of interest:

- **Mean Absolute Published Effect Size ($|\hat{\mu}_d|_{\text{PUB}}$).** Approximated as the mean of $|\{\hat{\delta}_{+1P}\}|$.
- **Proportion of Published Effect Sizes that are Negative ($\hat{\mu}^{neg}_{d\text{PUB}}$).** Approximated as the proportion of $\{\hat{\delta}_{+1P}\}$ that are less than zero.
- **The "Type S" error.** Approximated as the proportion of $\{\hat{\delta}_{11P}\}$ less than 0.
- **The "Type M" error.** Approximated as the mean of the absolute effect size divided by the true effect size: $|\{\hat{\delta}_{11P}\}|/\delta$.

A2. The Probability of Publication as a Function of Both psp and pwr

In Section 3, the chance of publishing a positive study in our framework depended only on novelty via

$$A = (1 - psp)^m.$$

For Section 4, a convenient choice of function to represent a journal policy of requiring an *a priori* sample size justification would be:

$$A = \frac{(1 - psp)^m}{1 + \exp\left\{-(\log 19)\frac{pwr - c_{50}}{c_{95} - c_{50}}\right\}}. \tag{A4}$$

So A is reduced more when power is lower, with the extent of the reduction parameterized by (c_{50}, c_{95}). Specifically, c_{95} is the value for *pwr* at which the multiplicative reduction is near negligible (factor of 0.95), while c_{50} is the value for *pwr* at which the multiplicative reduction is a factor of 0.5. We conduct our experimentation using $c_{50}, c_{95} = (0.5, 0.8)$, with the following rationale. If a journal does require an a priori sample size justification, a claim of 80% power is the typical requirement. Hence, a study which really attains 80% power is not likely to suffer in its quest for publication, motivating $c_{95} = 0.8$.

A3. Additional Results

Table A.1. A comparison between ecosystems with both a sample size requirement and $\alpha = 0.005$ and ecosystems without a sample size requirement and $\alpha = 0.05$.

δ	k	m	Change in *PR*	Change in *REL*	Change in *DSCV*	Change in D_{REL}	Change in $Mpwr_{\text{PUB}}$	Change in $\text{IQpsp}_{\text{PUB}}$
0.20	20	1	1.77	1.61	0.09	11.83	3.77	0.83
0.20	20	3	1.14	2.69	0.10	12.16	6.57	1.03
0.20	20	9	0.53	6.31	0.15	13.23	8.33	1.46
0.20	100	1	1.29	1.31	0.15	6.99	1.81	0.86
0.20	100	3	0.97	1.86	0.17	7.17	2.38	0.95
0.20	100	9	0.52	3.78	0.25	7.77	3.60	1.33
0.20	500	1	1.03	1.19	0.24	4.57	1.29	0.90
0.20	500	3	0.85	1.50	0.28	4.68	1.43	0.93
0.20	500	9	0.52	2.60	0.38	5.03	1.76	1.18
0.43	20	1	1.29	1.31	0.15	6.97	1.79	0.86
0.43	20	3	0.97	1.86	0.17	7.15	2.38	0.95
0.43	20	9	0.52	3.77	0.25	7.74	3.60	1.33
0.43	100	1	1.04	1.19	0.24	4.63	1.31	0.90
0.43	100	3	0.85	1.51	0.28	4.74	1.44	0.93
0.43	100	9	0.52	2.62	0.37	5.09	1.80	1.25
0.43	500	1	0.97	1.15	0.36	3.72	1.15	0.92
0.43	500	3	0.83	1.38	0.41	3.80	1.21	0.95
0.43	500	9	0.54	2.20	0.53	4.07	1.35	1.18

NOTES: For those with a sample size requirement, A is defined by equation A4 (with $c_{50} = 0.5$ and $c_{95} = 0.8$). For those without a sample size requirement, $A = (1 - psp)^m$. For all we have $B = 0$, and varying values of k and m as indicated.

Supplemental materials

The supplementary material includes additional figures which plot various metrics of interest across the entire set of ecosystems considered.

References

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The Earth is Flat ($p < 0.05$): Significance Thresholds and the Crisis of Unreplicable Research," Technical report, PeerJ Preprints. [358,369]

Aycaguer, L. C. S., and Galbán, P. A. (2013), Explicación Del Tamaño Muestral Empleado: Una Exigencia Irracional De Las Revistas Biomédicas. *Gaceta Sanitaria 27*, 53–57. [359]

Bakker, M., van Dijk, A., and Wicherts, J. M. (2012), "The Rules of the Game Called Psychological Science," *Perspectives on Psychological Science*, 7, 543–554. [360]

Begley, C. G., and Ellis, L. M. (2012), "Drug Development: Raise Standards for Preclinical Cancer Research," *Nature*, 483, 531–533. [358]

Begley, C. G., and Ioannidis, J. P. (2015), "Reproducibility in Science: Improving the Standard for Basic and Preclinical Research," *Circulation Research*, 116, 116–126. [364]

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., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, D., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D.,. Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Holland Jones, J., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A.,. Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y. Young, C., Zinman, J. and Johnson V. E. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [358]

Bland, J. M. (2009), "The Tyranny of Power: Is There a Better Way to Calculate Sample Size? *BMJ*, 339, b3985. [367]

Borm, G. F., den Heijer, M., and Zielhuis, G. A. (2009), "Publication Bias Was Not a Good Reason to Discourage Trials With Low Power," *Journal of Clinical Epidemiology*, *62*, 47–53. [359]

Bosker, T., Mudge, J. F., and Munkittrick, K. R. (2013), "Statistical Reporting Deficiencies in Environmental Toxicology," *Environmental Toxicology and Chemistry*, *32*, 1737–1739. [359]

Brembs, B., Button, K., and Munafò, M. (2013), "Deep Impact: Unintended Consequences of Journal Rank," *Frontiers in Human Neuroscience*, 7, 1–12. [360]

Burt, T., Button, K., Thom, H., Noveck, R., and Munafò, M. R. (2017), "The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures," *Clinical and Translational Science*, 10, 470–479. [370]

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013a). "Empirical Evidence for Low Reproducibility Indicates Low Pre-study Odds," *Nature Reviews Neuroscience* 14, 877–877. [359,360]

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R.(2013b), Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience 14*(5), 365–376. [360,364]

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016), " Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436. [358]

Campbell, H., and Gustafson, P. (2018), "Conditional Equivalence Testing: An Alternative Remedy for Publication Bias," *PloS One* 13, e0195145. [369]

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., and Willmes, K. (2015), "Registered Reports: Realigning Incentives in Scientific Publishing," *Cortex*, 66, A1–A2. [369]

Charles, P., B. Giraudeau, A. Dechartres, G. Baron, and P. Ravaud (2009). Reporting of sample size calculation in randomised controlled trials. *BMJ 338*, b1732. [359]

Charlton, B. G., and Andras, P. (2006). "How Should We Rate Research?: Counting Number of Publications May Be Best Research Performance Measure," *BMJ 332*, 1214–1215. [361]

Coffman, L. C., and Niederle, M. (2015), "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible," *Journal of Economic Perspectives*, 29, 81–98. [369]

Cohen (1962), "The Statistical Power of Abnormal-Social Psychological Research: A Review," *The Journal of Abnormal and Social Psychology*, 65, 145–153. [360]

Cohen, B. A. (2017), "How Should Novelty be Valued in Science?" *eLife* 6:e28699. [360]

Contopoulos-Ioannidis, D. G., Ntzani, E., and Ioannidis, J. (2003), "Translation of Highly Promising Basic Science Research Into Clinical Applications," *The American Journal of Medicine*, 114, 477–484. [370]

Dasgupta, P., and Maskin, E. (1987), "The Simple Economics of Research Portfolios," *The Economic Journal*, 97, 581–595. [359]

de Winter, J., and Happee, R. (2013), "Why Selective Publication of Statistically Significant Results Can Be Effective," *PLoS One*, *8*, e66463. [369]

Djulbegovic, B., and Hozo, I. (2007), "When Should Potentially False Research Findings Be Considered Acceptable?" *PLoS Medcine*, 4, e26. [370]

Djulbegovic, B., Kumar, A., Magazin, A., Schroen, A. T., Soares, H., Hozo, I., Clarke, M., Sargent, D., and Schell, M. J. (2011), "Optimism Bias Leads to Inconclusive Results—An Empirical Study," *Journal of Clinical Epidemiology* 64, 583–593. [367]

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., and Munafò, M. R. (2017), "Low Statistical Power in Biomedical Science: A Review of Three Human Research Domains," *Royal Society Open Science*, 4, 160254. [359,364]

Fanelli, D. (2011), "Negative Results Are Disappearing From Most Disciplines and Countries," *Scientometrics*, 90, 891–904. [360]

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., and Thomason, N. (2006), "Impact of Criticism of Null-hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology," *Conservation Biology*, 20, 1539–1544. [359]

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (neglected) Role of Diagnosticity and a Priori Theorizing," *Perspectives on Psychological Science*, *12*, 46–61. [360]

Fire, M., and Guestrin, C. (2018), "Over-optimization of Academic Publishing Metrics: Observing Goodhart's Law in Action," *arXiv preprint arXiv:1809.07841*. [361]

Fraley, R. C., and Vazire, S. (2014), "The n-pact Factor: Evaluating the Quality of Empirical Journals With Respect to Sample Size and Statistical Power," *PloS One* 9, e109019. [360]

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015), "The economics of reproducibility in preclinical research. *PLoS Biology*, 13, e1002165. [358]

Fritz, A., Scherndl, T., and Kühberger, A. (2013), "A Comprehensive Review of Reporting Practices in Psychological Journals: Are Effect Sizes Really Enough?" *Theory & Psychology*, 23, 98–122. [359]

Gall, T., Ioannidis, J., and Maniadis, Z. (2017), "The Credibility Crisis in Research: Can Economics Tools Help?" *PLoS Biology*, 15, e2001846. [359]

Gaudart, J., Huiart, L., Milligan, P. J., Thiebaut, R., and Giorgi, R. (2014), "Reproducibility Issues in Science, is p value Really the Only Answer?" *Proceedings of the National Academy of Sciences of the United States of America* 111, E1934. [358]

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations Assessing Type s (sign) and Type m (magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [363,364]

Gervais, W. M., Jewell, J. A., Najle, M. B., and Ng, B. K. (2015), "A Powerful Nudge? Presenting Calculable Consequences of Underpowered Research Shifts Incentives Toward Adequately Powered Designs," *Social Psychological and Personality Science*, *6*, 847–854. [360]

Goodman, S., and Greenland, S. (2007), "Assessing the Unreliability of the Medical Literature: A Response to 'Why Most Published Research Findings Are False?'" Johns Hopkins University, Department of Biostatistics; Working Papers. [358]

Gornitzki, C., Larsson, A., and Fadeel, B. (2015), "Freewheelin' Scientists: Citing Bob Dylan in the Biomedical Literature," *BMJ* 351:h6505. [359]

Greenwald, A. G. (1975), "Consequences of Prejudice Against the Null Hypothesis," *Psychological Bulletin*, 82, 1–20. [359]

Greve, W., Bröder, A., and Erdfelder, E. (2013), "Result-blind Peer Reviews and Editorial Decisions: A Missing Pillar of Scientific Culture," *European Psychologist*, *18*, 286–294. [369]

Hagen, K. (2016), "Novel or Reproducible: That is the Question," *Glycobiology*, 26, 429–429. [360]

Halpern, S. D., Karlawish, J. H., and Berlin, J. A. (2002), "The Continuing Unethical Conduct of Underpowered Clinical Trials," *JAMA* 288, 358–362. [369]

Hazra, A., and Gogtay, N. (2016), "Biostatistics Series Module 5: Determining Sample Size," *Indian Journal of Dermatology 61*, 496–504. [360,367]

Higginson, A. D., and Munafò, M. R. (2016), "Current Incentives for Scientists Lead to Underpowered Studies With Erroneous Conclusions," *PLoS Biology*, *14*, e2000995. [359,360,361]

Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing," *The American Statistician*, 57, 171–178. [369]

IntHout, J., Ioannidis, J. P., and Borm, G. F. (2016), "Obtaining Evidence by a Single Well-powered Trial or Several Modestly Powered Trials," *Statistical Methods in Medical Research*, 25, 538–552. [359]

Ioannidis, J. P. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124. [358,369]

——— (2008), "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 640–648. [363]

Ioannidis, J. P., Hozo, I., and Djulbegovic, B. (2013), "Optimal Type I and Type II Error Pairs When the Available Sample Size Is Fixed," *Journal of Clinical Epidemiology*, 66, 903–910. [365,369]

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences* 110, 19313–19317. [358]

——— (2014), "Reply to Gelman, Gaudart, Pericchi: More Reasons to Revise Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 111, E1936–E1937. [358]

Kimmelman, J., Mogil, J. S., and Dirnagl, U.(2014), "Distinguishing Between Exploratory and Confirmatory Preclinical Research Will Improve Translation," *PLoS Biology*, *12*, e1001863. [363,371]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E. et al. (2018), "Justify Your Alpha," *Nature Human Behaviour* 2, 168–171. [369]

Lamberink, H. J., Otte, W. M., Sinke, M. R., Lakens, D., Glasziou, P. P., Tijdink, J. K., and Vinkers, C. H. (2018), "Statistical Power of Clinical Trials Increased While Effect Size Remained Stable: An Empirical Analysis of 136,212 Clinical Trials Between 1975 and 2014," *Journal of Clinical Epidemiology 102*, 123–128. [360,361,364]

Lane, D. M., and Dunlap, W. P.(1978), "Estimating Effect Size: Bias Resulting From the Significance Criterion in Editorial Decisions," *British Journal of Mathematical and Statistical Psychology*, 31, 107–112. [363]

Lee, C. J., and Schunn, C. D. (2011), "Social Biases and Solutions for Procedural Objectivity," *Hypatia* 26, 352–373. [364]

Leek, J. T., and Jager, L. R. (2017), "Is Most Published Research Really False?" *Annual Review of Statistics and Its Application*, 4, 109–122. [358]

Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination," *The American Statistician*, 55, 187–193. [360]

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012), "Replications in Psychology Research: How Often Do They Really Occur?" *Perspectives on Psychological Science*, 7, 537–542. [360]

Martin, G., and Clarke, R. M. (2017). Are Psychology Journals Anti-Replication? A Snapshot of Editorial Practices, *Frontiers in Psychology*, 8, 1–6. [360]

McLaughlin, A. (2011), "In Pursuit of Resistance: Pragmatic Recommendations for Doing Science Within One's Means," *European Journal for Philosophy of Science*, 1, 353–371. [359]

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2017), "Abandon Statistical Significance," *arXiv preprint arXiv:1709.07588*. [369]

Miller, J., and Ulrich, R. (2016), "Optimizing Research Payoff," *Perspectives on Psychological Science* 11, 664–691. [359]

Moonesinghe, R., Khoury, M. J., and Janssens, A. C. J. (2007), "Most Published Research Findings Are False–But a Little Replication Goes a Long Way," *PLoS Medicine* 4, e28. [360]

Mudge, J. F., Baker, L. F., Edge, C. B., and Houlahan, J. E. (2012). "Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests," *PloS One* 7, e32734. [369]

Nord, C. L., Valton, V., Wood, J., and Roiser, J. P. (2017), "Power-up: A Reanalysis of 'Power Failure' in Neuroscience using Mixture Modelling," *Journal of Neuroscience*, 3592–16. [359]

OpenScienceCollaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [358,364]

Ploutz-Snyder, R. J., Fiedler, J., and Feiveson, A. H. (2014), "Justifying Small-n Research in Scientifically Amazing Settings: Challenging the Notion That Only ?Big-n? Studies Are Worthwhile," *Journal of Applied Physiology*, 116, 1251–1252. [369]

Richard, F. D., Bond C. F. Jr, and Stokes-Zoota, J. J. (2003), "One Hundred Years of Social Psychology Quantitatively Described," *Review of General Psychology*, 7, 331–363. [361]

Roos, J. M. (2017), "Measuring the Effects of Experimental Costs on Sample Sizes." under review, available at: *http://www.jasonmtroos.com/assets/media/papers/Experimental_Costs_Sample_Sizes.pdf* [360,364]

Sakaluk, J. K. (2016), "Exploring Small, Confirming Big: An Alternative System to the New Statistics for Advancing Cumulative and Replicable Psychological Research," *Journal of Experimental Social Psychology*, 66, 47–54. [363,371]

Schulz, K. F., and Grimes, D. A. (2005), "Sample Size Calculations in Randomised Trials: Mandatory and Mystical," *The Lancet* 365, 1348–1353. [367]

Senn, S. (2012), "Misunderstanding Publication Bias: Editors Are Not Blameless After All," *F1000Research 1*. [364]

Siler, K., Lee, K., and Bero, L. (2015), "Measuring the Effectiveness of Scientific Gatekeeping," *Proceedings of the National Academy of Sciences* 112, 360–365. [364]

Smaldino, P. E., and McElreath, R. (2016), "The Natural Selection of Bad Science," *Royal Society Open Science*, 3, 160384. [360]

Song, F., Y. Loke, and Hooper, L. (2014), "Why Are Medical and Health-related Studies Not Being Published? A Systematic Review of Reasons Given by Investigators," *PLoS One 9*(10), e110418. [364]

Spiegelhalter, D. (2017), "Trust in Numbers," *Journal of the Royal Statistical Society*, Series A, 180, 948–965. [358]

Stanley, T., Jarrell, S. B., and Doucouliagos, H. (2010), "Could It Be Better to Discard 90% of the Data? A Statistical Paradox," *The American Statistician* 64, 70–77. [369]

Sterling, T. D., Rosenbaum, W. L., and Weinkam, J. J. (1995), "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa," *The American Statistician*, 49, 108–112. [369]

Szucs, D., and Ioannidis, J. P. (2016), "Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature," *bioRxiv*, 071530. [360]

van Assen, M. A., van Aert, R. C., Nuijten, M. B., and Wicherts, J. M. (2014), "Why Publishing Everything is More Effective Than Selective Publishing of Statistically Significant Results," *PLoS One*, 9, e84896. [369]

van Dijk, D., Manor, O., and Carey, L. B. (2014), "Publication Metrics and Success on the Academic Job Market," *Current Biology*, 24, R516–R517. [361]

Vasishth, S., and Gelman, A. (2017), "The Illusion of Power: How the Statistical Significance Filter Leads to Overconfident Expectations of Replicability," arXiv preprint arXiv:1702.00556. [367]

Walker, A. M. (1995), "Low Power and Striking Results—A Surprise But Not a Paradox," *Mass Medical Soc. 332*(16), 1091–1092. [359]

Wei, Y., and Chen, F. (2018), "Lowering the p Value Threshold–Reply," *JAMA* 320, 937–938. [358]

Wilson, B. M., and Wixted, J. T. (2018), "The Prior Odds of Testing a True Effect in Cognitive and Social Psychology," *Advances in Methods and Practices in Psychological Science*, 2515245918767122. [364]

Yeung, A. W. (2017), "Do Neuroscience Journals Accept Replications? A Survey of Literature," *Frontiers in Human Neuroscience*, 11, 468. [360]

Taylor & Francis
Taylor & Francis Group

The American Statistician

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

Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p*-Value Ban

Ronald D. Fricker Jr., Katherine Burke, Xiaoyan Han & William H. Woodall

To cite this article: Ronald D. Fricker Jr., Katherine Burke, Xiaoyan Han & William H. Woodall (2019) Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p*-Value Ban, The American Statistician, 73:sup1, 374-384, DOI: 10.1080/00031305.2018.1537892

To link to this article: https://doi.org/10.1080/00031305.2018.1537892

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 16434

View related articles

View Crossmark data

Citing articles: 17 View citing articles

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, 374–384: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1537892

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p*-Value Ban

Ronald D. Fricker, Jr., Katherine Burke, Xiaoyan Han, and William H. Woodall

Department of Statistics, Virginia Tech, Blacksburg, VA

ABSTRACT

In this article, we assess the 31 articles published in *Basic and Applied Social Psychology* (BASP) in 2016, which is one full year after the BASP editors banned the use of inferential statistics. We discuss how the authors collected their data, how they reported and summarized their data, and how they used their data to reach conclusions. We found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered. Readers would be largely unable to recognize this because the necessary information to do so was not readily available.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Effect size; Inference ban; NHST; Psychology; Statistical significance

1. Introduction

In February of 2015, the editors of *Basic and Applied Social Psychology* (BASP) banned the use of the "null hypothesis significance testing procedure" (NHSTP) and confidence intervals. Trafimow and Marks (2015, p. 1) said, "...prior to publication, authors will have to remove all vestiges of NHSTP (*p*-values, *t*-values, *F*-values, statements about 'significant' differences or lack thereof, and so on)" from articles published in BASP. In the 2015 editorial, in response to the question of whether *any* inferential statistical procedures were required for publication in BASP, the editors said,

> *No, because the state of the art remains uncertain. However, BASP will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem.*

While BASP stopped just short of banning all inferential methods, frequentist methods were clearly *verboten* and the editors "reserve the right to make case-by-case judgments" for Bayesian methods (Trafimow and Marks 2015). The ban is an extension of the position first laid out by Trafimow when he became editor of BASP where, in his inaugural editorial, Trafimow (2014, p. 1) said, "...BASP will no longer require inferential statistics procedures on an a priori basis." Subsequent to the ban, Trafimow is quoted as saying, "I'd rather not have any inferential statistics at all than have some that we know aren't valid" (Woolston 2015).

The BASP editors had thus chosen to only allow the use of descriptive statistics to communicate research results and, at a minimum, signaled an interest in publishing research based only on the use of descriptive statistics. After announcing the ban, BASP published an article by Valentine, Aloe, and Lau (2015) that provided "concrete guidance to researchers on ways they can explore and communicate results of their studies." This guidance consisted of the following three basic principles of data analysis: (1) use of graphical displays for data and descriptive statistics, (2) reporting measures of variation along with measures of central tendency, and (3) the use and thoughtful interpretation of effect sizes.

In our article, we consider the 31 BASP papers published in 2016; that is, one full year after the ban on hypothesis testing and confidence intervals. We do not review the 2015 papers because many seemed to contain pre-ban statistical practices, perhaps because they had been accepted for publication prior to the ban being put in place. In Section 2, we provide some additional background on the ban, including a summary of its origins. In Section 3, we describe what types of statistical practices BASP now permits. In Section 4, after reviewing the 2016 BASP papers, we present some of our findings. In Section 5, we discuss and reflect on the impact of the ban. Finally, in Section 6, we present our conclusions.

2. Criticisms of Statistical Significance Testing

Part of the BASP editors' motivation for banning statistical inference from BASP was a perceived over-reliance by psychology researchers on the use of *p*-values for judging the importance of results, where research results are often obsobscured by a stylistic convention in which authors pepper their expository text with a variety of statistics and *p*-values. For example, a typical

CONTACT Ronald D. Fricker, Jr. rf@vt.edu Department of Statistics, Virginia Tech, Blacksburg, VA 24061.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.
© 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.

psychology article often reads like the following passage from Skinner, Stevenson, and Camillus (2015, p. 61):

> *In Step 1, women ($M = 5.17, SD = 0.27$) perceived the defendant to be significantly more inattentive than men ($M = 3.95, SD = 0.25; \beta = -1.22, p < .01, \beta_{STD} = -.36$). Yet H3 was not supported: There was no significant main effect of defendant gender on perceived inattention ($\beta = 0.10, p = .73$). With regards to the sexism measures, benevolent sexism was unrelated to perceived inattention ($\beta = 0.36, p = .16$), but hostile sexism was a significant predictor of inattention ($\beta = 0.52, p = .01, \beta_{STD} = .26$). Specifically, as hostile sexism increased, participants perceived the defendant to be less attentive. Finally, as with the attribution of responsibility analysis, in Step 2 of the model ($\Delta R^2 = .01$) a Defendant Gender × Hostile Sexism interaction emerged ($\beta = -0.95, p = .02, \beta_{STD} = -.34$).*

In addition, Trafimow and Marks (2015) explicitly stated that their ban is based on the following issues:

1. Hypothesis testing does not provide the probability a particular hypothesis is true given the observed data.
2. Confidence intervals do not provide the probability the parameter is within a particular (fixed) interval.
3. Bayesian methods, which are capable of providing these types of inference, often use priors that "depend on some sort of Laplacian assumption," meaning uniform priors.

Focusing in on the role of p-values in hypothesis testing, Trafimow and Marks (2016, p. 1) wrote,

> *The probability of the finding (or one more extreme) given the null hypothesis (p) is not the same as the probability of the null hypothesis given the finding, nor does p provide a strong basis for drawing conclusions about the probability of the null hypothesis given the finding. Without a strong basis for drawing a conclusion about the probability of the null hypothesis given the finding, there is little justification for rejecting the null hypothesis simply because $p < .05$. This rejection—the heart of the null hypothesis significance testing procedure—commits the inverse inference fallacy.*

2.1. The Inverse Inference Fallacy

Cohen (1994) described the Inverse Inference Fallacy as the "misapplication of deductive syllogistic reasoning." That is, Cohen suggested hypothesis testing is interpreted as a generalization of the following deductive reasoning,

> If A is true then B cannot occur;
> However, B has occurred;
> Therefore, A is false;

that is transformed into the following probabilistic reasoning,

> If A is true then B *probably* cannot occur;
> However, B has occurred;
> Therefore, A is *probably* false;

where he then equates A with the null hypothesis and B with some form or function of the observed data. He goes on to say that the following equivalent "formulation appears at least implicitly in article after article in psychological journals and explicitly in some statistics textbooks,"

> If H_0 is true, then this result (statistical significance) would probably not occur;
> This result has occurred;
> Then H_0 is probably not true and therefore formally invalid.

Cohen (1994, p. 998) called this "the illusion of attaining improbability."

2.2. The Laplacian Assumption

The issue that BASP editors take with Bayesian methods is part of the long standing discussion over how to choose prior probabilities. In particular, the Laplacian assumption to which their 2015 editorial refers is the use of priors in the absence of any information about the probabilities of the hypotheses. There are two arguments presented. The first is that an assumed prior is based on unknown information: "...Bayesian procedures...depend on some sort of Laplacian assumption to generate numbers where none exist..." (Trafimow and Marks 2015, p. 1). The second is that uniform priors, often used to represent a state of ignorance about a parameter, may violate the "principle of indifference," by which they mean that a transformation of the parameter into an equivalent problem may no longer yield a uniform prior (Trafimow 2005). For a discussion of a Bayesian approach for use by psychologists, we recommend Wagenmakers (2007).

2.3. Other Criticisms Appearing in BASP

The BASP editors devoted an issue of the journal to mediation analysis, which seeks to identify unobservable "mediator variables" that influence the observable relationships between independent variables and a dependent variables. In that issue, Kline (2015, p. 207) listed the following five "basic criticisms of significance testing" as applied in mediation analyses in a section of his article entitled "Folly of Statistical Significance as a Decision Criteria":

1. "Outcomes of significance tests—p values—are wrong in most studies."
2. "Researchers do not understand p-values."
3. "Significance tests do not tell researchers what they want to know."
4. "Most applications of significance testing are incorrect."
5. "Respecification in SEM [structural equation modeling] based solely on significance testing merely capitalizes on chance."

The reasons for most of these criticisms will not come as a surprise to most statisticians. For example, according to Kline (2015) p-values are wrong and the applications of significance testing are incorrect in most psychology studies because "[t]rue random sampling is very rare, especially in human studies where

most samples are ad hoc (convenience)," because parametric tests are conducted on data that do not meet the assumptions of the tests, and because the wrong hypotheses are tested. He takes particular issue with "nil hypothesis testing" (as do Trafimow and Marks, 2015) because it is "scientifically far-fetched."

The criticism that (psychology) researchers do not understand p-values is also well known and was part of the justification for BASPs ban as discussed by Trafimow and Marks (2015, 2016). In his article, Kline (2015) never further expounds on the criticism that significance tests do not tell (psychology) researchers what they want to know, but presumably the rationale is similar to that discussed in Section 2; that is, by employing hypothesis testing researchers are simply testing whether mediation exists rather than determining whether or not the mediation effect is "large and precise enough to be appreciable in a particular context." Finally, the criticism that fitting a structural equation model using significance testing "merely capitalizes on chance" is mainly focused on the issue of overfitting models to data. Of course, the overfitting of models to data is not a new issue, nor is it unique to the field of psychology.

Hypothesis testing and p-value criticisms aside, most statisticians would likely agree with the main points made by Kline (2015, p. 202):

> *(1) the typical mediation study relies on an inadequate design; (2) the researcher uses a flawed analysis strategy; and (3) there is scant attention to assumptions that are required when estimating mediation. These problems stem from overgeneralizing the classical product method for estimating mediation and overreliance on statistical significance testing as a decision criterion in mediation analysis.*

3. What Does BASP Allow?

Subsequent to the ban, BASP published an article by Valentine, Aloe, and Lau (2015) entitled "Life after NHST: How to Describe Your Data Without 'p-ing' Everywhere" which we view as indicative of BASP editorial intentions. Valentine, Aloe, and Lau (2015, p. 260) said they "provide concrete guidance to researchers on ways they can explore and communicate the results of their studies" where, as mentioned in Section 1, they promote the use of descriptive statistics and the following three basic principles of data analysis:

1. Use statistics to summarize and plots to display data,
2. Report measures of variation along with measures of central tendency, and
3. Present and appropriately interpret effect sizes.

These principles are consistent with what is considered to be good statistical practice, although Valentine, Aloe, and Lau (2015) give rather cursory and overly general recommendations with no references to the broader statistical literature other than Tukey (1977). For example, the only graphical displays illustrated are bar charts, stem-and-leaf plots, and boxplots. For summary statistics, they recommend reporting the sample size, mean, median, SD, range, and coefficient of variation, which is fine as far as it goes, but limited.

A focus of Valentine, Aloe, and Lau (2015) is on the calculation and interpretation of effect sizes in the two-sample situation. They recommended measures for continuous data such as Cohen's d defined as

$$d = \frac{\bar{x}_2 - \bar{x}_1}{s_p}, \tag{1}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of samples from populations 1 and 2, respectively, and s_p is the usual pooled estimate of the standard deviation (SD). They also recommended the unstandardized difference in sample means, r and r^2, U_3,[1] "descriptive U_3," and a measure called the common language effect size (or CLES).

In summary, BASP allows basic descriptive statistics with an emphasis on observed effect sizes. Trafimow and Marks (2015) said as much in their quote given at the beginning of Section 1. Of course, a key issue is that these effect size measures are point estimates with no quantification of uncertainty, so BASP readers have no information about whether an observed effect size is consistent with random variation under the assumption of no effect being present.

4. 2016 BASP Research

The question we are interested in addressing is how the BASP editors' belief that "banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking" (Trafimow and Marks 2015, p. 2) influenced the use of data and statistics in the 31 BASP papers published in 2016. These papers consisted of:

- One editorial (Trafimow and Marks 2016);
- Three literature reviews (Kressel and Kressel 2016; Schuhr 2016; Stepanova et al. 2016);
- Ten discussion papers (Chatterjee 2016; Filho and Rettig 2016; Kuldas and Bulut 2016; Marsman, Ly, and Wagenmakers 2016; Pashler et al. 2016a, 2016b; Rose 2016; Sinha 2016; Witte and Zenker 2016a, 2016b); and,
- Seventeen research papers that used statistical methods, to a greater or lesser degree, and which are the main focus of this section.

Before proceeding, it is worth mentioning that we found some vestiges of statistical significance in this volume of BASP. The entire argument of Pashler et al. (2016b) was based on showing that some oddities in data were very unlikely under an assumption of randomness. This was demonstrated using p-values obtained using bootstrapping. Sinha (2016) and Hill (2016) reported p-values. Gatobu, Arocha, and Hoffman-Goetz (2016, p. 5) described observation-oriented modeling as an alternative to the usual methods of statistical inference. They used a "c-value" which is just a bootstrapped p-value.

[1] For two normally distributed populations with means μ_1 and μ_2 and common SD σ, Cohen (1988) defined $U_3 = \Phi\left(\frac{\mu_2-\mu_1}{\sigma}\right)$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

4.1. Assessing Cohen's d Values Reported in BASP

To begin, we considered the results of 15 experiments in seven papers in which authors used Cohen's d to assess scientific impact: Covarrubias, Herrmann, and Fryberg (2016, studies 1 & 2), Noordewier and van Dijk (2016, study 3), Hughes, Creech, and Strosser (2016, studies 2, 3, 4, & 6), Herrmann et al. (2016, studies 1 & 2), Teng, Poon, and Yang (2016, experiments 1 & 2), Rasmussen et al. (2016, studies 2 & 3), and Chen, Ding, and Li (2016, studies 1 & 3).

As shown in Equation (1), Cohen's d is the difference between the means of two samples measured in terms of their estimated assumed common SD. In the psychology literature, it has become accepted that $0.2 \leq d < 0.5$ indicates there is a "small" practical effect; $0.5 \leq d < 0.8$ indicates a "medium" practical effect; and $d \geq 0.8$ indicates a "large" practical effect (Cohen 1988, pp. 24–27).

Given the value of Cohen's d and the two sample sizes, it is straightforward to calculate the associated two-sample t-test p-value since the test statistic is just

$$t = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where n_1 and n_2 are the sample sizes. For T_ν that follows a t-distribution with ν degrees of freedom, the p-value for the two-sided test is $p = 2 \times P(T_{n_1+n_2-2} > |t|)$. We denote the total sample size by n; that is, $n = n_1+n_2$. Now, before proceeding, we wish to emphasize here that we are not advocating for p-values in this analysis but rather are using them as a convenient way (given the limited information we have available from the BASP papers) to calculate some measure of statistical significance for the reported Cohen's d values.

Figure 1 is a plot of Cohen's d values from the seven papers versus their p-values, where most of the p-values are approximate because for 59 of the 64 reported values only the total sample size n was given and so we had to assume n_1 and n_2 were equal. Of the 64 published d values, 32 (50%) were not statistically significant, if statistical significance is defined as having a p-value less than or equal to 0.05. As the figure shows, for this particular set of results, nearly all of the d values larger than Cohen's minimum medium effect size were statistically significant and all of those below Cohen's small effect size limit were not. Those are the more straightforward cases.

However, for those d values that correspond to small effect sizes shown in the light blue shaded area in the plot (29 of 64 or 45.3% of the d values), there is no way to decide whether a particular observed d value is consistent with random fluctuations in the data under the assumption of equal population means without some measure of statistical significance. Furthermore, for those Cohen's d values that fell in the small effect size range, more than two-thirds (20 out of 29 or 68.9%) are not statistically significant if one uses the admittedly arbitrary cutoff of $p < 0.05$. We can debate whether the p-value is the best way to separate signal from noise, including the question of whether to use a simple dichotomization rule to decide, but in the absence of any information other than observed effect size, it surely seems that subsequent researchers will be at a disadvantage trying to interpret the results.

This issue is further compounded by authors who sometimes seem to over-interpret the observed effect sizes relative to someone who uses the arbitrary $p < 0.05$ cutoff. For example, for a d value of 0.44 Teng, Poon, and Yang (2016, p. 273) conclude that "...those in the high group identification condition rated the protest as less righteous than participants in the low group identification condition...," yet the associated p-value is 0.065. In another example, Herrmann et al. (2016, p. 262) conclude that "A d of .24 indicates that participants in the intervention group on average earned approximately one fourth of a SD higher grade than participants in the control group, which is considered a small effect size." This sentence is, of course, correct because it only refers to the two samples, but the associated p-value of 0.064 would make someone using the $p < 0.05$ threshold careful in inferring anything more from the observed difference. Yet, Herrmann et al. (2016, p. 263) later go on to conclude that "These findings provide further support for the notion that brief psychological interventions can yield significant effects in educational contexts."

Now, to be fair, most of the initial descriptions of the observed Cohen's d effect sizes seem consistent with a p-value dichotomization at 0.05, but that could be due to authors who first conducted a standard statistical analysis and then removed it from their papers prior to publication. If that is the case, then BASPs ban on inference makes less sense as all it does is deprive readers of potentially useful information. However, as in the Herrmann et al. (2016, p. 263) example, not infrequently the specific results are appropriately presented, but then in the title, abstract, or elsewhere in the article, conclusions are drawn that seem stronger than what the actual results appear to support. For example, studies 1 and 2 of Covarrubias, Herrmann, and Fryberg (2016) reported 19 Cohen's d values, where each observed effect size was described in terms of the study participants without reference to a larger population, which is perfectly fine in that context. However, in spite of the fact that some of the results in study 1 were inconsistent, and one of the significant d values *contradicted* the original hypothesis until a post hoc reassignment of some of the results achieved a significant result in the desired direction, the final paragraph of the article is a broad generalization, concluding:

> *The positive influence of affirmations that incorporate families for Latino students signifies the importance of integrating these close others into the school environment. For example, teachers can empower Latino students by including important others, such as family members, in their school experiences. Given the positive effects on performance in our studies, teachers could utilize number or word problems that are culturally relevant (e.g., ratio of ingredients for your family's favorite meal, distance to travel to see your relatives). By framing affirmations to match students' ideas about what is a 'right' self, teachers can create cultural matches that foster a learning environment that is welcoming, familiar, and safe for students who have been historically underserved in education* (Covarrubias, Herrmann, and Fryberg 2016, pp. 54–55).

In the authors' defense, they do describe their post hoc analysis. However, this does suggest that some sort of caveat is warranted in reporting the results, where it seems that in the absence of a measure of statistical significance the authors

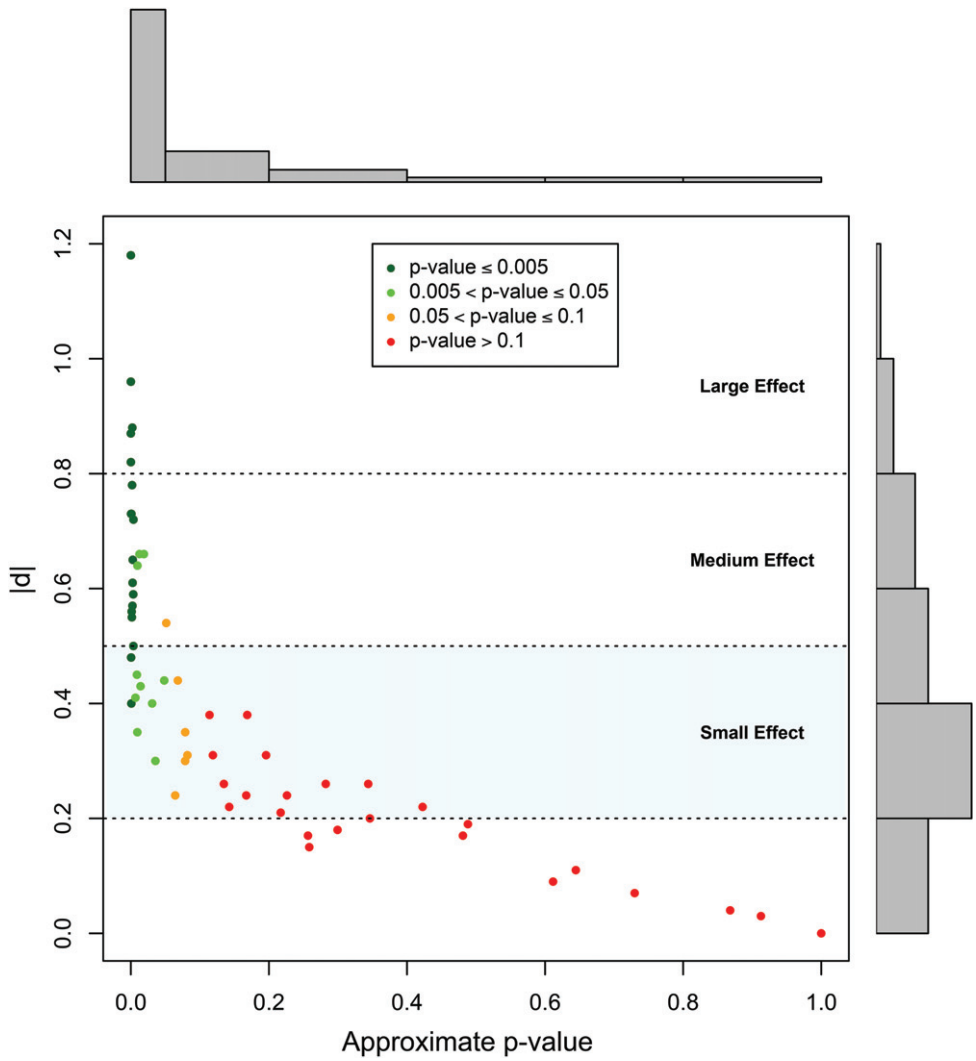


Figure 1. Cohen's *d* values plotted against the approximate *p*-values from 15 experiments in seven BASP papers that used Cohen's *d* to assess scientific impact. Four points with extremely large *d* values ranging from 2.85 to 3.71 are not shown.

focused only on the effect size measures and this led to a potential overstatement of Study 1 results.

4.2. Assessing Some Reported Regression Coefficients

Within the 31 papers, two papers contained regression results that also included the standard errors (SEs) of the estimated regression coefficients: Teng, Poon, and Yang (2016) and Haas (2016). Using this information, we were able to calculate approximate 95% percent confidence intervals and thus assess some inferential aspects of the reported results.

We begin with Teng, Poon, and Yang (2016) who conducted two studies intended to evaluate how reading about social protests affects an individual's perception of social justice as a function of whether he or she identifies with the protestors. The data were collected via surveys of students in a Chinese university, with 128 students in the first experiment and 142 in the second.

For the 20 estimated regression coefficients reported in the article, Figure 2 displays 95% confidence intervals color coded as to whether the interval contains zero (red) or not (green) along with the affiliated text describing each result. The first nine estimated coefficients are from the first experiment and the rest from the second experiment.

There are a few observations evident from this plot. The first is that all of the estimated coefficients are quite small (the largest in an absolute sense is -0.6) and the majority (13 of 20) are statistically insignificant at an approximate 5% significance level. Secondly, at least five of the insignificant results are interpreted as if there were effects (see the bold text for lines 1, 10, 12, 15, and 18). This suggests that, in the absence of interval estimates, the authors tended to over-interpret their results (compared to the use of the p-value <0.05 threshold) by asserting the existence of effects that are not supported if one considers the variation in the data.

A particularly egregious example is plotted in the article. The authors say, "Moreover, consistent with our predictions, there was evidence for the existence of an interaction between protest exposure and group identification ($b = 0.10, SE = 0.08$; see Figure 2)." The interaction plot is given in Figure 3 where we note that in the published figure the range of the vertical axis is (3.0, 3.8) which visually amplified what would be considered a statistically insignificant interaction. Furthermore, the title

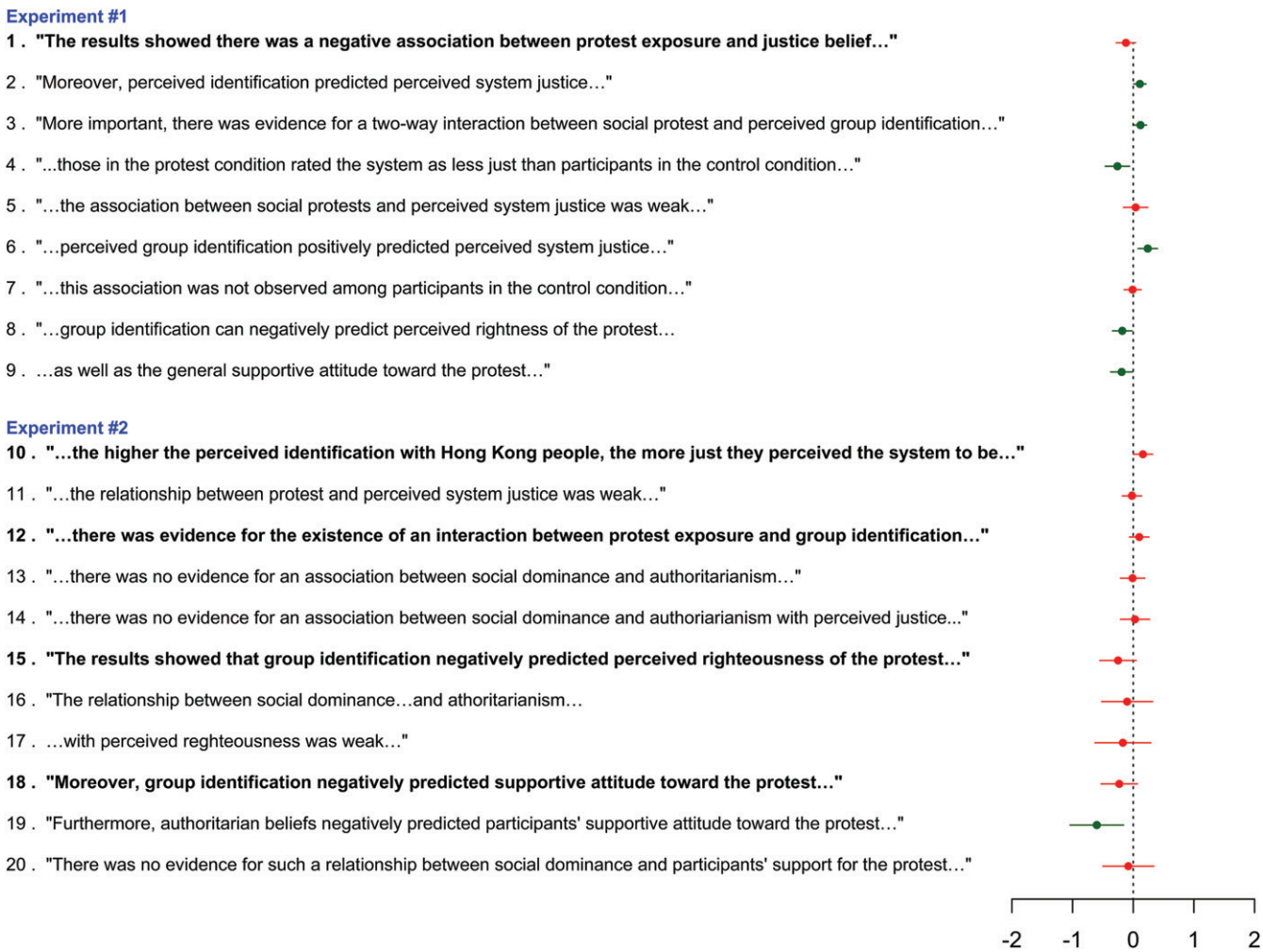


Figure 2. Approximate 95% confidence intervals for estimated regression coefficients in Teng, Poon, and Yang (2016, experiments 1 & 2).

of their article explicitly states without any qualification that the interaction effect exists: "Does Social Protest Shake People's Justice Beliefs? *It Depends on the Level of Group Identification*" (our emphasis). They also stress this in the conclusion section of their article, "We found that protest exposure and group identification interacted to predict participants' justice belief. Specifically, the more participants perceived identification with the protesters' group, the less they thought the protest was right, and thus less supportive of the protest" (Teng, Poon, and Yang 2016, p. 274).

We now turn to Haas (2016), the goal of which was to examine the impact of uncertainty and threat on support for political compromise. For study 2 of that article, data were collected on a sample of 210 subjects that consisted of a combination of 152 individuals recruited via Amazon's Mechanical Turk website (which we will discuss in more detail in Section 5) and 58 undergraduates. The participants were randomly assigned to one of four conditions in a 2×2 design: threat versus control by uncertain versus certain.

Figure 4 shows plots of approximate 95% confidence intervals for the coefficients from three linear regression models, each with successively more terms, calculated from Table 1 of Haas (2016, p. 143). Assuming the regression models were appropriately fit (something we cannot check), the plots clearly show that only the intercept, ideology extremity, and perhaps political

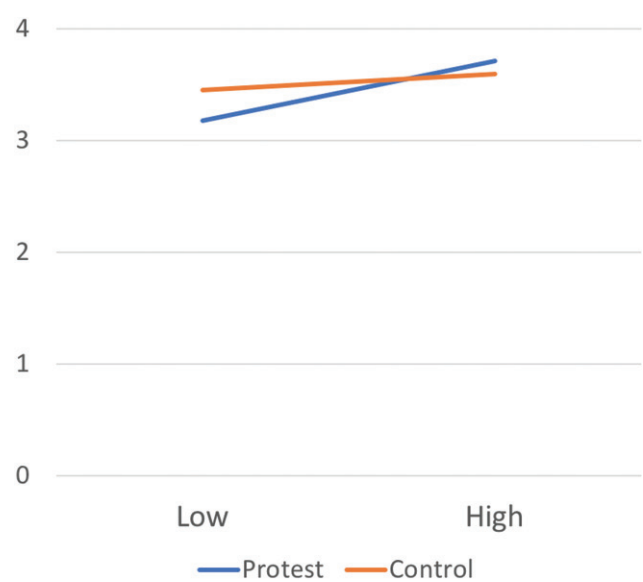


Figure 3. Interaction plot using data from Teng, Poon, and Yang (2016, p. 273). The original figure truncated range of the *y*-axis which visually magnified the differences in the slopes of the lines.

ideology are statistically significant at the $p < 0.05$ level. That is, the uncertainty and threat terms, as well as all the associated interaction terms, are not significant (where we discount the

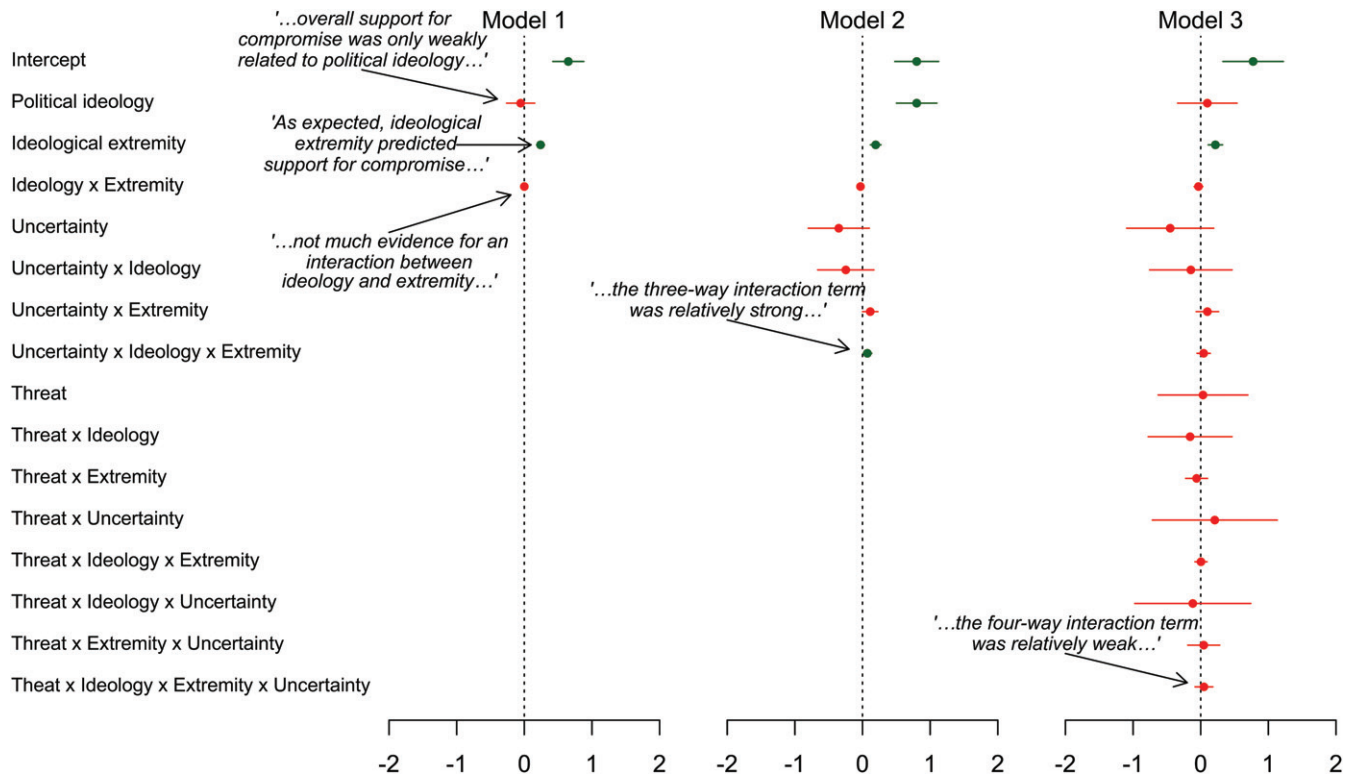


Figure 4. Approximate 95% confidence intervals for estimated regression coefficients in Haas (2016, study 2).

negligible three way interaction in Model 2). Yet, the author then devotes an entire page to interpreting statistically insignificant differences between a variety of pairwise sample mean comparisons. For example: "Examining mean support for compromise by condition shows that in the full sample, uncertainty *increased* [author's emphasis] support for compromise ($M = 1.28, SD = 1.32$) relative to certainty ($M = 1.44, SD = 1.40$) in the control condition" (Haas 2016, p. 144). However, assuming the study participants were evenly divided between each condition, the p-value for a two sample t-test is about 0.4, showing that the results are consistent with the assumption of equal population means. Yet, the author goes on to state conclusions such as, "Consistent with the results of Study 1, uncertainty was more likely to increase support for compromise when participants were not also feeling threatened" (Haas 2016, p. 145).

In summary, in this section we began by noting that the 2016 papers were not entirely free of inferential methods. Interestingly, Gatobu, Arocha, and Hoffman-Goetz (2016) reinvented the bootstrap p-value, perhaps to fill the vacuum left by the ban. Mainly, however, we reviewed results in BASP papers related to Cohen's d or regression analyses when the results included standard errors. In so doing, note that we did not cherry pick the most egregious examples to present; rather, we simply selected the papers that used statistics for which it was easy to do some basic inferential checks. When researchers only employ descriptive statistics we found that they are likely to overinterpret and/or overstate their results compared to a researcher who uses hypothesis testing with the $p < 0.05$ threshold.

5. Discussion

In reading the psychology literature, we found what we consider to be exemplary guidance for the rigorous application of statistical methods in psychology research. Specifically, the recommendations of the American Psychological Association's Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference 1999) are excellent, where we note that there were statisticians on the task force (Donald Rubin, Frederick Mosteller, and John Tukey), and where the recommendations fully support the appropriate use of descriptive and inferential statistical methods in psychological research. In fact, the Task Force wrote:

> *Some had hoped that this task force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought that there were enough counterexamples (e.g., Abelson 1997) to justify forbearance.* (Wilkinson and the Task Force on Statistical Inference 1999, pp. 602–603).

Similarly, the *Publication Manual of the American Psychological Association* (APA) provides very relevant and appropriate advice for how to appropriately convey research results, including the following:

> *APA stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results. The degree to which any journal emphasizes (or de-emphasizes) NHST is a decision of the individual editors. However, complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are*

> *the minimum expectations for all APA journals* (American Psychological Association 2009, p. 33).

However, the BASP editors decided to abandon this approach, where we note that BASP is not an APA journal.

Also in reading the literature, we noted that there is a very vocal subset of psychology researchers (in addition to the BASP editors) who believe that inferential statistical methods generally accepted in many other fields are not serving their research needs and goals. In particular, there has been severe criticism in some corners of psychology, as well as in the BASP editorials, against "nil-hypothesis testing." Yet, a point that seems to have been missed in these discussions is one of statistics' seminal contributions to the practice of science: tools that facilitate insight into whether an observed phenomenon could be a real effect or just noise in the data. As Benjamini (2016) said of the p-value specifically, "In some sense it offers a first line of defense against being fooled by randomness, separating signal from noise, because the models it requires are simpler than any other statistical tool needs."

Lakens (2016) who did a brief review of some 2015 post-NHST ban BASP papers expressed a similar sentiment, albeit a bit more colorfully:

> *In their latest editorial, Trafimow and Marks [2016] hit down some arguments you could, after a decent bottle of liquor, interpret as straw men against their ban of p-values. They don't, and have never, discussed the only thing p-values are meant to do: control error rates. Instead, they seem happy to publish articles where some . . . authors get all the leeway they need to adamantly claim effects are observed, even though these effects look a lot like noise.*

However, this point is not lost on all psychology researchers. More than 20 years ago, Abelson (1997, p. 14) wrote, "Null hypothesis tests are cogent in scrutinizing surprising results that critics doubt. They address the alternative argument that some familiar chance mechanism could have produced the result." In a similar vein, 30 years ago, Cohen (1988, p. 16) wrote, "Research reports in the literature are frequently flawed by conclusions that state or imply that the null hypothesis is true. For example, following the finding that the difference between two sample means is not statistically significant, instead of properly concluding from this failure to reject the null hypothesis that the data do not warrant the conclusion that the population means differ, the writer concludes, at least implicitly, that there is *no* difference."

Our sense from reading these papers is that psychology research is much less focused on the question of whether an observed effect can be attributed to random variation in the data as it is on proving a particular hypothesis that explains the existence of a presumed effect. We see both as important, but failing to first skeptically assess whether an observed effect could be consistent with random variation may result in an excessive number of false positives in research results. Indeed, the scientific method demands skepticism of any observed results, where one should first want to rule out the simplest explanation that an observed result is consistent with random variation in the data before then seeking to find another explanation.

In an article published in *BASP*, Kline (2015, p. 208), citing a study by Haller and Krauss (2002), perhaps provides some insight into why some misconceptions persist in psychology when he writes ". . .about 80%–90% of psychology professors endorse multiple false beliefs about statistical significance, no better than psychology undergraduate students in introductory statistics courses." Gigerenzer (2004, p. 588) also cites Haller and Krauss (2002) and goes on to say, "Textbooks and curricula in psychology almost never teach the statistical toolbox, which contains tools such as descriptive statistics, Tukey's exploratory methods, Bayesian statistics, Neyman–Pearson decision theory and Wald's sequential analysis." Instead, he says that psychology statistics textbooks focus exclusively on null hypothesis testing, where in referring to a "distinguished statistical textbook author" Gigerenzer says "He was not a statistician; otherwise, his text would likely not have been used in a psychology class" (Gigerenzer 2004, p. 587).

To illustrate some misperceptions, one need look no further than Valentine, Aloe, and Lau (2015). This article was published in *BASP* after the p-value ban, ostensibly to help psychology researchers improve research practices. Yet as an example of questionable guidance, Valentine, Aloe, and Lau (2015, p. 262) used questionnaire data collected on a 5-point Likert scale to "illustrate how to explore and present data for continuous outcomes." We also noted multiple BASP papers that contained linear regression models using dependent variables based on Likert scale data (see, e.g., Haas 2016; Noordewier and van Dijk 2016; Skinner, Stevenson, and Camillus 2015), suggesting an issue with what it means to have continuous data and/or how to appropriately model using ordinal data.

Returning to the Inverse Inference Fallacy in Section 2.1, if psychologists are using this type of logic to understand hypothesis testing, then it is perhaps not surprising that some have difficulty with p-values. The Inverse Inference Fallacy seems, at least to us, like a strawman "derivation" of hypothesis testing, where it:

- Focuses on a dichotomized decision (the result is "statistical significance"),
- Overstates the resulting conclusion (H_0 is "formally invalid"), and,
- Suggests that $P(R|H_0) = P(H_0|R)$, where R is some result derived from the observed data, say a p-value, because the same word "probably" is used in both the first and third lines of Cohen's "equivalent formulation."

To the non-probabilistically inclined, the last point may stem from a general difficulty people have with understanding and interpreting conditional probabilities, where other common mistakes include thinking that $P(R|H_0) = P(H_0 \cap R)$ and that p-value $= P(H_0|R)$. The latter misinterpretation can be particularly troublesome because p-values can be dramatically smaller than $P(H_0|R)$, the probability that the null hypothesis is true given the observed data (Berger and Sellke 1987).

In addition, in our opinion, one of the problems with tying hypothesis testing to deductive syllogistic reasoning is that it divorces it from the scientific method applied to experiments. That is, rather than being part of an experimental process that seeks to better understand what is more or less likely to be

true, deductive syllogistic reasoning makes hypothesis testing seem like a logic exercise where the goal is to prove one of the hypotheses definitively true and the other definitively false.

In our review, we also found that some authors fail to clearly distinguish between sample and population quantities or between estimates and the quantity being estimated. For example, Valentine, Aloe, and Lau (2015, p. 263, eq. 1) define the coefficient of variation as $\mathrm{CV} = \sigma/\mu$ and they explicitly state "σ is the population SD and μ is the population mean," which is fine. However, they then go on to illustrate with an example using data, where it then seems more appropriate to think about the data as a sample from some larger population so that they were estimating the coefficient of variation with $\widehat{\mathrm{CV}} = s/\bar{x}$. This points to a deeper issue where, if the distinction between sample and population is not clear then the purpose of inference must also be unclear. As discussed by Wagenmakers et al. (2011), it is also important to distinguish between exploratory and confirmatory analyses.

In reading the 2016 BASP papers, in a practice that is common throughout psychology research, nearly all of them used convenience samples, where they are almost exclusively comprised of undergraduate students and individuals recruited using Amazon's Mechanical Turk. Specifically, within the 17 BASP 2016 research papers, there were a total of 47 studies, of which only four studies in three papers involved sampling from some sort of general population (Dolinski and Grzyb 2016; Gatobu, Arocha, and Hoffman-Goetz 2016; Montoya and Pittinsky 2016). For the other 43 studies:

- Nineteen used Mechanical Turk in whole or part;
- Twenty-three used college students, mainly undergraduates, in whole or part (with potentially two more in Bhogal, Galbraith, and Manktelow (2016) but not enough information was given to be sure); and,
- Two used preuniversity children/students.

For those unfamiliar with it, Mechanical Turk is an Internet-based crowdsourcing platform within which one can post tasks that people will do for pay, often in very small amounts. As the Mechanical Turk website says, "Amazon Mechanical Turk (MTurk) operates a marketplace for work that requires human intelligence. The MTurk web service enables companies to programmatically access this marketplace and a diverse, on-demand workforce" (*www.mturk.com*). Psychologists now often use it to conduct a variety of experiments, though it is important to note that those doing the tasks choose to do them; there is no way to conduct probability-based sampling nor any feature to sample from some larger targeted population.

Of course, the fact that nearly all the studies in the BASP research papers are based on convenience samples raises basic questions about whether inferential statistics are even appropriate to use on this data. For example, in a conclusion that probably won't surprise statisticians, Peterson and Merunka (2014, p. 1035) wrote:

> *In a study of business-related ethicality, analysis of data from four dozen convenience samples of undergraduate business students revealed significant differences in means, variances, intercorrelations, and path parameters across the samples. Depending on the particular convenience sample used, relationships between variables and constructs were positive or negative and statistically significant or insignificant.*

Such a result raises important questions about whether many of the experiments in the papers we reviewed are reproducible using other convenience samples.

6. Conclusions

To statisticians, *BASPs* banning of the use of inference may seem baffling. Although there are limitations of statistical inference, we ascribe to the philosophy of the renowned British scientist Francis Galton (1894, pp. 62–63):

> *Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man.*

Here we are 125 years later and it seems that it remains important to emphasize Galton's key caveats that statistical methods "are not brutalized" and that they are "warily interpreted." Furthermore, after this review, we also feel it is important to re-emphasize Spiegelhalter's comments on the ASA Statement on *p*-Values (Wasserstein and Lazar 2016) in "Too familiar to ditch:"

> *. . .many point out that the problem lies not so much with p-values in themselves as with the willingness of researchers to lurch casually from descriptions of data taken from poorly designed studies, to confident generalisable inferences* (Matthews, Wasserstein, and Spiegelhalter 2017, p. 41).

In the final accounting, it is the responsibility of the individual researcher to do good science in his or her field. All the statistics community can do is provide the best methods and tools possible; it is up the researcher to properly apply them ("delicately handled") and appropriately decipher the results ("warily interpreted"). If a drunken driver crashes his car into the tree, it is not the car's fault (at least not today and until such time as we have fully automated cars). Similarly, there is only so much the statistics profession can do to prevent unfair criticism or abuse of statistical methods and tools. However, when appropriate, we should speak up.

For example, the American Statistical Association's *Ethical Guidelines for Statistical Practice* requires individual statisticians to be "candid about any known or suspected limitations, defects, or biases in the data that may impact the integrity or reliability of the statistical analysis" (American Statistical Association 2016). We should be just as candid as a profession about other organizations that institutionalize bad statistical practices. Of course, we do not want to be the statistical police, but when a journal

like *BASP* tries to blame statistics for poor scientific practices, we should push back against arguments that lack merit.

That said, we have to take some responsibility for oversimplifying hypothesis testing in our introductory classes with simple dichotomized accept–reject rules and the language of "rejecting the null hypothesis" and "proving the alternative hypothesis" and the like. Furthermore, as a discipline, we would do better by our students to make teaching the scientific method part of our standard curriculum, as well as bringing more experimentation into the classroom, so that students begin to understand what it is like to try to learn through experimentation and particularly to give them actual experience trying to learn from data. Specifically, many of our introductory classes have become so focused on methods in the absence of their scientific application that students do not connect the two. We need to stop oversimplifying or assuming away the actual problems and teach our students how to apply statistics to real problems.

In addition, we need to pay more attention to how our tools are used and, to the extent possible, design them to be less prone to errors of human intuition. For example, we know that p-values confuse some practitioners (see, e.g., Greenland et al. 2016) and, while there are legitimate concerns and philosophical divides about their use, frequentist-based hypothesis testing will be with us for a very long time. Thus, we should take a pragmatic view about "mistake proofing" things like p-values. For example, we agree with Greenland (2017) that transforming the p-value onto a "surprisal scale" of, say, $-\log_2(p)$, has the potential to align intuitive interpretations more closely with correct use and thereby help eliminate some of the mistakes so common with the use of p-values.

Regardless of whether one agrees with p-values and the hypothesis testing paradigm, psychology researchers would also do well to routinely report some measure of the uncertainty in their effect size point estimates. A good start would be to follow Cohen's recommendation from 25 years ago:

> *. . .my third recommendation is that, as researchers, we routinely report effect sizes in the form of confidence limits. "Everyone knows" that confidence intervals contain all the information to be found in significance tests and much more. . . . Yet they are rarely found in the literature. I suspect that the main reason they are not reported is that they are so embarrassingly large!* (Cohen 1994, p. 1002).

In summary, in this article we presented our findings from reviewing the papers published in *BASP* in 2016 a full year after the editors banned the publication of inferential statistical methods. We found multiple instances of results seemingly being overstated beyond what the data would support if p-values (or some other form of statistical inference) had been used. Thus, the ban seems to be allowing authors to make less substantiated claims, the net result of which could be misleading to readers and could well result in a body of literature within which there are a greater number of irreproducible effects.

The Open Science Collaboration (2015) attempted to replicate 100 psychology studies published in three psychology journals. They found that "[r]eplication effects were half the magnitude of original effects" and while "[n]inety-seven percent of original studies had statistically significant results," only "[t]hirty-six percent of replications had statistically significant results." In our opinion, the practices we have observed in the papers published in BASP post-ban will not help to solve this problem; in fact, we believe they will make it worse.

Acknowledgments

The authors gratefully acknowledge an associate editor and two reviewers for their thoughtful and very helpful comments on earlier drafts of this article. They were instrumental in helping us substantially improve the article.

References

Abelson, R. P. (1997), "On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test," *Psychological Science*, 23, 12–15. [380,381]

American Psychological Association (2009), *Publication Manual of the American Psychological Association* (6th ed.), Washington, DC: American Psychological Association. [381]

American Statistical Association (2016), "Ethical Guidelines for Statistical Practice," American Statistical Association, available at *www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx*. [382]

Benjamini, Y. (2016), "It's Not the p-Values' Fault," American Statistical Association, available at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top*. [381]

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [381]

Bhogal, M. S., Galbraith, N., and Manktelow, K. (2016), "Physical Attractiveness and Altruism in Two Modified Dictator Games," *Basic and Applied Social Psychology*, 38, 212–222. [382]

Chatterjee, P. (2016), "Response to Pasher et al. (2016)," *Basic and Applied Social Psychology*, 38, 19–29. [376]

Chen, N., Ding, G., and Li, W. (2016), "Do Negative Role Models Increase Entrepreneurial Intentions? The Moderating Role of Self-Esteem," *Basic and Applied Social Psychology*, 38, 337–350. [377]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Erlbaum. [376,377,381]

——— (1994), "The Earth Is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [375,383]

Covarrubias, R., Herrmann, S. D., and Fryberg, S. A. (2016), "Affirming the Interdependent Self: Implications for Latino Student Performance," *Basic and Applied Social Psychology*, 38, 47–57. [377]

Dolinski, D., and Grzyb, T. (2016), "One Serious Shock Versus Gradated Series of Shocks: Does 'Multiple Feet-in-the-Door' Explain Obedience in Milgram Studies?," *Basic and Applied Social Psychology*, 38, 276–283. [382]

Filho, E., and Rettig, J. (2016), "Intergroup Conflict Management Strategies From a Nobel Peace Laureate: The Case of José Ramos-Horta," *Basic and Applied Social Psychology*, 38, 351–361. [376]

Galton, F. (1894), *Natural Inheritance*, London: McMillan and Company. [382]

Gatobu, S. K., Arocha, J. F., and Hoffman-Goetz, L. (2016), "Numeracy, Health Numeracy, and Older Immigrants' Primary Language: An Observation-Oriented Exploration," *Basic and Applied Social Psychology*, 38, 185–199. [376,380,382]

Gigerenzer, G. (2004), "Mindless Statistics," *The Journal of Socio-Economics*, 33, 587–606. [381]

Greenland, S. (2017), "Cognitive Biases Built Into Testing, and a Surprising Solution: The Information Content of a P-Value as Measured by Surprisal," American Statistical Association, Presentation at the ASA Symposium on Statistical Inference on October 12, 2018, Bethesda, MD. [383]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, p-Values, Confidence

Intervals, and Power: A Guide to Misinterpretations," *The American Statistician* Online Supplement, 70, 1–20. [383]

Haas, I. J. (2016), "The Impact of Uncertainty, Threat, and Political Identity on Support for Political Compromise," *Basic and Applied Social Psychology*, 38, 137–152. [378,379,380,381]

Haller, H., and Krauss, S. (2002), "Misinterpretations of Significance: A Problem Students Share With Their Teachers?," Institute for Science Education, available at *http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf*. [381]

Herrmann, S. D., Adelman, R. M., Bodford, J. E., Graudejus, O., Okun, M. A., and Kwan, V. S. Y. (2016), "The Effects of a Female Role Model on Academic Performance and Persistence of Women in STEM Courses," *Basic and Applied Social Psychology*, 38, 258–268. [377]

Hill, C. A. (2016), "Implicit and Explicit Sexual Motives as Related, But Distinct Characteristics," *Basic and Applied Social Psychology*, 38, 59–88. [376]

Hughes, J., Creech, J. L., and Strosser, G. L. (2016), "Attributions About Morally Unreliable Characters: Relationship Closeness Affects Moral Judgments," *Basic and Applied Social Psychology*, 38, 173–184. [377]

Kline, R. B. (2015), "The Mediation Myth," *Basic and Applied Social Psychology*, 37, 202–213. [375,376,381]

Kressel, N. J., and Kressel, S. W. (2016), "Trends in the Psychological Study of Contemporary Antisemitism: Conceptual Issues and Empirical Evidence," *Basic and Applied Social Psychology*, 38, 111–126. [376]

Kuldas, S., and Bulut, S. (2016), "What Is Called Self in Educational Research and Practice?," *Basic and Applied Social Psychology*, 38, 200–211. [376]

Lakens, D. (2016), "So You Banned *p*-Values, How's That Working Out for You?," Eindhoven University of Technology, available at *http://daniellakens.blogspot.com/2016/02/so-you-banned-p-values-hows-that.html*. [381]

Marsman, M., Ly, A., and Wagenmakers, E.-J. (2016), "Four Requirements for an Acceptable Research Program," *Basic and Applied Social Psychology*, 38, 308–312. [376]

Matthews, R., Wasserstein, R. L., and Spiegelhalter, D. (2017), "The ASA's *p*-Value Statement, One Year on," *Significance*, 14, 38–41. [382]

Montoya, R. M., and Pittinsky, T. L. (2016), "Bounded Rationality's Account for the Influence of Group Identification on Ingroup Favoritism: A Field Investigation Using Jewish and Arab Populations in Israel," *Basic and Applied Social Psychology*, 38, 127–136. [382]

Noordewier, M. K., and van Dijk, E. (2016), "Interest in Complex Novelty," *Basic and Applied Social Psychology*, 38, 98–110. [377,381]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716-1–aac4716-8. [383]

Pashler, H., Rohrer, D., Abramson, I., Wolfson, T., and Harris, C. R. (2016a), "Response to comments by Chatterjee, Rose and Sinha," *Basic and Applied Social Psychology*, 38, 41–46. [376]

——— (2016b), "A Social Priming Data Set With Troubling Oddities," *Basic and Applied Social Psychology*, 38, 3–18. [376]

Peterson, R., and Merunka, D. (2014), "Convenience Samples of College Students and Research Reproducibility," *Journal of Business Research*, 67, 1035–1041. [382]

Rasmussen, K. R., Alibhai, A., Boon, S. D., and Ellard, J. H. (2016), "Trust as an Explanation for Relational Differences in Revenge," *Basic and Applied Social Psychology*, 38, 284–300. [377]

Rose, R. L. (2016), "Cautious Thoughts on 'A Social Priming Data Set With Troubling Oddities'," *Basic and Applied Social Psychology*, 38, 30–32. [376]

Schuhr, A. (2016), "Seizing Control: Estimating Multiple Decision Processes and the Investigation of Self-Control," *Basic and Applied Social Psychology*, 38, 241–257. [376]

Sinha, J. (2016), "Selective Literature Review and Selective Data Analyses: Implications for the (Re)analysis of Public Access Research Data," *Basic and Applied Social Psychology*, 38, 33–40. [376]

Skinner, A. L., Stevenson, M. C., and Camillus, J. C. (2015), "Ambivalent Sexism in Context: Hostile and Benevolent Sexism Moderate Bias Against Female Drivers," *Basic and Applied Social Psychology*, 37, 56–67. [375,381]

Stepanova, E. V., Strube, M. J., Clote, L. E., and Limes, D. (2016), "Pictorial Race Activation in Priming Measures," *Basic and Applied Social Psychology*, 38, 223–239. [376]

Teng, F., Poon, K.-T., and Yang, Y. (2016), "Does Social Protest Shake People's Justice Beliefs? It Depends on the Level of Group Identification," *Basic and Applied Social Psychology*, 38, 269–275. [377,378,379]

Trafimow, D. (2005), "The Ubiquitous Laplacian Assumption: Reply to Lee and Wagenmakers (2005)," *Psychological Review*, 112, 669–674. [375]

——— (2014), "Editorial," *Basic and Applied Social Psychology*, 36, 1–2. [374]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [374,375,376]

——— (2016), "Editorial," *Basic and Applied Social Psychology*, 38, 1–2. [375,376]

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Pearson. [376]

Valentine, J. C., Aloe, A. M., and Lau, T. S. (2015), "Life After NHST: How to Describe Your Data Without '*p*-ing' Everywhere," *Basic and Applied Social Psychology*, 37, 260–273. [374,376,381,382]

Wagenmakers, E.-J. (2007), "A Practical Solution to the Pervasive Problems of *p* Values," *Psychometric Bulletin and Review*, 14, 779–804. [375]

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. (2011), "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)," *Journal of Personality and Social Psychology*, 100, 426–432. [382]

Wasserstein, R. L., and Lazar, N. A. (2016), "Editorial: The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [382]

Wilkinson, L., and the Task Force on Statistical Inference (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist*, 54, 594–604. [380]

Witte, E. H., and Zenker, F. (2016a), "Beyond Schools: Reply to Marsman, Ly, and Wagenmakers," *Basic and Applied Social Psychology*, 38, 313–317. [376]

——— (2016b), "Reconstructing Recent Work on Macrosocial Stress as a Research Program," *Basic and Applied Social Psychology*, 38, 301–307. [376]

Woolston, C. (2015), "Psychology Journal Bans *P* Values," *Nature*, 519, 9. [374]

The American Statistician

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

Content Audit for *p*-value Principles in Introductory Statistics

Karsten Maurer, Lynette Hudiburgh, Lisa Werwinski & John Bailer

To cite this article: Karsten Maurer, Lynette Hudiburgh, Lisa Werwinski & John Bailer (2019) Content Audit for *p*-value Principles in Introductory Statistics, The American Statistician, 73:sup1, 385-391, DOI: 10.1080/00031305.2018.1537890

To link to this article: https://doi.org/10.1080/00031305.2018.1537890

© 2019 The Author(s).

Published online: 20 Mar 2019.

Submit your article to this journal

Article views: 4549

View related articles

View Crossmark data

Citing articles: 3 View citing articles

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, 385–391: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1537890

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Content Audit for *p*-value Principles in Introductory Statistics

Karsten Maurer, Lynette Hudiburgh, Lisa Werwinski, and John Bailer

Department of Statistics, Miami University, Oxford, OH

ABSTRACT

Longstanding concerns with the role and interpretation of *p*-values in statistical practice prompted the American Statistical Association (ASA) to make a statement on *p*-values. The ASA statement spurred a flurry of responses and discussions by statisticians, with many wondering about the steps necessary to expand the adoption of these principles. Introductory statistics classrooms are key locations to introduce and emphasize the nuance related to *p*-values; in part because they engrain appropriate analysis choices at the earliest stages of statistics education, and also because they reach the broadest group of students. We propose a framework for statistics departments to conduct a content audit for *p*-value principles in their introductory curriculum. We then discuss the process and results from applying this course audit framework within our own statistics department. We also recommend meeting with client departments as a complement to the course audit. Discussions about analyses and practices common to particular fields can help to evaluate if our service courses are meeting the needs of client departments and to identify what is needed in our introductory courses to combat the misunderstanding and future misuse of *p*-values.

ARTICLE HISTORY
Received February 2018
Revised September 2018

KEYWORDS
Audit; Curriculum; Education; Service courses

1. Introduction

The publication of this journal's special issue reflects a growing consensus: *p*-values are often misused, and that misuse often leads to bad science. Many argue that the main challenges are to understand the logic of testing scientific hypotheses, and to get away from mechanical rules like $p < 0.05$ as a substitute for contextual reasoning. The logic of testing, we argue, is best taught in a first statistics course. However, research in statistics education makes clear that this logic is far harder to teach and to learn than is the simple $p < 0.05$. This poses a particular challenge for those who teach introductory statistics courses. In this article, we propose a process for auditing the coverage of *p*-value principles in an introductory statistics course.

The misuse of *p*-values is frequent, well-documented, and potentially leads to bad science; most notably, the "reproducibility crisis." Ioannidis sounded an alarm with his paper, "Why most published research findings are false" (Ioannidis 2005). A decade later, the Open Science Collaboration (2015) repeated 100 experiments taken from the psychology literature. Of these hundred, only 39 produced results that replicated the original findings. In the health sciences, Greenland et al. (2016) offered readers a catalogue of misinterpretations of *p*-values. As noted by Berry (2016), "(o)ur collective credibility in the science community is at risk." These and other articles led the American Statistical Association (ASA) to issue a statement on proper use of *p*-values (Wasserstein and Lazar 2016). The ASA statement spurred a flurry of responses and discussions by statisticians, along with the 2017 ASA Symposium on Statistical Inference. We now look toward the next steps necessary to expand the adoption of these principles.

Although there is a growing consensus about the nature of the problem, there is little consensus on simple remedies like banning *p*-values altogether or reducing the threshold for "significance" to $p < 0.005$. Most agree that the heart of the problem is reliance on mechanical rules like $p < 0.05$. Such rules cannot substitute for the logic of hypothesis testing applied in the scientific context. Unfortunately, decades of research have shown that this logic is not easy to learn (Falk and Greenbaum 1995; Williams 1999; Batanero 2000; Garfield and Ben-Zvi 2003; Harradine, Batanero, and Rossman 2011). DelMas et al. (2007) gave a multiple choice test to students who had completed an introductory statistics course and found that only 54.5% could identify the correct interpretation for *p*-values; only 58.6% could identify incorrect interpretations. Rossman (2008) cited the work of Nickerson (2004) in cognitive psychology to argue that one explanation "surely rests in all of the research that has shown how difficult probabilistic reasoning is for people."

Stangl (2016) argued that we have a responsibility in statistics education to preempt and end the perpetual misuse of *p*-values. Cobb (2016) responded to the ASA statement by saying, "(w)hat ASA has done here should spur a reshaping of the way we teach—both *p*-values in particular and statistics generally." There are many reasons to focus on the introductory course in statistics. For many students it is their first encounter with the logic of statistical inference. For most of those students it is also their last formal encounter with that logic in an academic setting. Moreover, courses that introduce statistical thinking and methods reach a very large percentage of the students who will become practicing scientists and evidence-based decision makers.

CONTACT Karsten Maurer *maurerkt@miamioh.edu* Miami University, 100 Bishop Circle, 311 Upham Hall, Oxford, OH 45046.

© 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.

The ASA-sponsored report, *Guidelines for Assessment and Instruction in Statistics Education* (GAISE), outlines goals and methods for teaching introductory statistics (Carver et al. 2016). The GAISE report sets as a goal "(s)tudents should demonstrate an understanding of, and be able to use, basic ideas of statistical inference, both hypothesis tests and interval estimation, in a variety of settings." Millar (2016) argued that "(s)tudents of other disciplines will be in our service courses, and while we should not advocate for hypothesis tests as the monolithic statistical inference method, they do need to know what it is and what its shortcomings are because they will encounter it." Goodman (2016) also pointed out the responsibility our courses have to future scientists, saying "(t)he fact that statisticians do not all accept at face value what most scientists are routinely taught as uncontroversial truisms will be a shock to many. But if we are to move science forward, we must speak to scientists." Berry (2016) was more direct: "We must communicate better even if we have to scream from the rooftops." But as statisticians, before we scream, we should gather data.

In the remainder of this article we describe and illustrate the use of a rubric that can be used to assess how well a course addresses the ASAs principles for sound use of p-values and the use of focus group discussions between teachers of statistics and their colleagues in the sciences. We hope these will serve as tools to help frame systematic conversations within and across departmental lines. In what follows, Section 2 describes the rubric and focus group questions. Section 3 illustrates an example of conducting the course audit with these methods. Section 4 concludes with a discussion.

2. Methods

For a department to perform a comprehensive audit of p-values concepts in their introductory statistics courses, we propose a framework that elicits both intra-departmental and inter-departmental feedback on the current curriculum. The combination of introspection and external suggestions can then drive targeted curricular adaptation to better cover the challenging topics related to p-values. The key idea for this assessment is that each of the six principles articulated by the ASA statement provide a target that can be directly evaluated with a common rubric as we discuss next.

2.1. Intra-Departmental Evaluation

For a department with several instructors involved in teaching introductory statistics courses, it may be helpful to use a common rubric as a unified starting point for identifying strengths and/or weaknesses in teaching about p-values. We propose the rubric found in Table 1. The rubric is structured to be applied to each of the six principles from the ASA statement:

1. *p-values can indicate how incompatible the data are with a specified statistical model.*
2. *p-values do not measure the probability that the studied hypothesis is true, or the probability that data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

The rubric evaluates how each principle is formally introduced by the instructor, reinforced with classroom activities, assessed for comprehension and is supported with appropriate supplementary learning resources. These broad categories of course components are used because they are likely to be present, in some form, in most introductory courses.

Rubric instructions for evaluating ASA p-value principle:
Repeat the following steps for each principle $i = 1, \ldots, 6$

1. Reread the ith ASA p-value principle.
2. Reflect on how completely, correctly and consistently each of the four broad curricular components (instruction, class activity, assessment and support materials) reflect the ith ASA p-value principle.
3. Score the corresponding item based on the statement in Table 1 that you feel most accurately describes how substantially the curricular component reflects the ith ASA p-value principle. Enter this in the curriculum audit report card in Table 2.
4. Sum the score from the four curricular components. The sum will range from to 4. Treat this sum as a GPA value and assign a letter grade.

$$(1.0 = \text{D}, 1.3 = \text{D+}, 1.7 = \text{C−}, 2.0 = \text{C}, 2.3 = \text{C+},\\ 2.7 = \text{B−}, 3.0 = \text{B}, 3.3 = \text{B+}, 3.7 = \text{A−}, 4.0 = \text{A})$$

With such a general grading guide, the scores and letter grades are clearly subjective in nature but may provide an informative shorthand familiar to most educators. It is important to highlight the rubric is primarily designed as a support tool to help frame discussion and help target areas for improvement in an internally administered content audit. Creating an overall score is secondary, as it is *not* a nationally normed instrument for comparing quality of curriculum across all universities. The rubric can be completed collectively or by individual instructors and then used to drive a discussion on steps to improve the curriculum coverage of the ASA principles. Alternative rubrics could be devised by moving to simple dichotomous responses for the inclusion or exclusion of components covering particular ASA principles, or by including weights for components that a department finds more essential, or by breaking from the letter-grade theme altogether. In the end, we employed a relatively interpretable scale to provide a basis for reflection and discussion. If several department members identify a deficiency in teaching one of the principles from the ASA statement, a collective curricular remediation can be planned. In the case that a principle is covered well by one instructor but not by another, class materials and teaching advice can be shared to help patch the gap. A result of this audit conducted by a subgroup of a department might provide the basis for discussion at a departmental retreat or meeting.

Table 1. Curricular component rubric.

The curricular component does not reflect the ASA principle.	The curricular component barely reflects the ASA principle.	The curricular component mostly reflects the ASA principle.	The curricular component fully reflects the ASA principle.
Score = 0.	Score = 0.3	Score = 0.7	Score = 1.0

Table 2. Curriculum audit report card.

		ASA *p*-value principles					
		1	2	3	4	5	6
Curricular component	*Instruction:* lecture, discussion, video lecture, etc.						
	Activity: lab, worksheet, case studies, etc.						
	Assessment: homework, quiz, exam, etc.						
	Support Materials: readings, tutorials, apps, etc.						
	Totals						
	Overall *p*-value principle grades						

We conducted the content audit for *p*-value principles within our department at Miami University, a mid-sized public Midwestern university. Multiple introductory statistics courses are offered through our department, each geared toward a different subpopulation of students. We applied the rubric to our algebra-based introductory statistics service course for undergraduate students, STA 261. This course serves an assortment of majors. The course begins with examples that challenge students to begin inferential thinking by using simulation-based methods to evaluate likelihood of observed data under assumed conditions. The curriculum then proceeds through a unit on probability and sampling distributions before introducing *p*-values in a probability-based framework.

This class is taught to 600 students each semester using a hybrid model with online introduction of concepts, just-in-time teaching of problematic concepts in a large lecture meeting, and a smaller lab section where statistical principles are explored. In a typical semester, two large lectures are taught by a continuing lecturer and four other large lectures are taught by term-limited faculty, namely Instructors or Visiting Assistant Professors. Graduate students facilitate lab sections. Course materials and labs are all centrally constructed by the course coordinator. In our case, the rubric was completed collectively by members of the author group and the results can be found in Section 3.1.

2.2. Inter-Departmental Evaluation

Next, we recommend running a small focus group discussion with analytically savvy members from client departments to gain an interdisciplinary perspective on the statistics curriculum at your institution. A great starting place is to identify departments with large numbers of undergraduate majors who are required take your introductory service course in statistics. These will often be psychology, political science and biology departments, but this will vary from campus to campus. The goals would be to advocate for good analytical practice in the scientific community, in this case pertaining to *p*-values, and to ask for candid feedback on how the statistics service classes meet the needs of their respective fields based on their general observations. Our focus group included faculty members from psychology, biology, geology, and kinesiology.

Along with an invitation to the meeting, we suggest sending a brief description with a web-link to the ASA *p*-value statement and a short list of questions you plan to discuss. We encourage you to consider a set of questions that can be used as discussion prompts. For example, when we conducted this focus group, we asked a number of questions to prompt additional discussion including:

- Teaching in our introductory statistics service courses:
  - Do your students show understanding of *p*-values and hypothesis testing?
  - What methods should our introductory statistics course include?
- Teaching in our advanced statistics service courses:
  - What do our advanced service courses do well to prepare your students?
  - Do your advanced students show understanding of *p*-values and hypothesis testing?
  - What is missing and what would you like to see us address in more detail?
- Why is a fundamental understanding of statistics important in your field?
  - What attracted and engaged you personally with statistics?

3. Results

In Fall 2017, we conducted a content audit for the coverage of *p*-values in our algebra-based introductory statistics course and held a focus group meeting with representatives from client departments. The results of the content audit and focus group are summarized in Sections 3.1 and 3.2, respectively.

Table 3. Completed curriculum audit report card.

		ASA *p*-value principles					
		1	2	3	4	5	6
Curricular Component	*Instruction:* lecture, discussion, video lecture, etc.	0.3	0.7	0.7	0.3	0.7	0.3
	Activity: lab, worksheet, case studies, etc.	0.3	1.0	0.7	0.7	1.0	0.3
	Assessment: homework, quiz, exam, etc.	0.3	1.0	0.7	0.7	1.0	0.7
	Support Materials: readings, tutorials, apps, etc.	0.3	1.0	0.7	0.7	1.0	0.7
	Totals	1.2	3.7	2.8	2.4	3.7	2.0
	Grades	D	A−	B−	C+	A−	C

Table 4. Results of our content audit with a short explanation specific to each principle.

1. *p*-values can indicate how incompatible the data are with a specified statistical model.
Grade: D
Reason: Hypothesis tests (HT) are presented and assessed with statements of assumptions, but these assumptions are often made without consideration of why these assumptions are important. May not be until the first regression / modeling class before intuition about the importance of assumptions on validity of HT. Even then, the *p*-value is more often emphasized as linked to a parameter vs. a model.

2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that data were produced by random chance alone.
Grade: A−
Reason: Conditional probability is strongly emphasized and *p*-values are consistently discussed and assessed in the context of a particular hypothesis being true. Minor problems remain where we may still see the second interpretation sneak into instructor or graduate assistant descriptions.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
Grade: B−
Reason: We regularly and thoroughly challenge the bright line rules ($p < 0.05$ significant) with examples, readings and discussions; for instance questioning what the difference between $p = 0.049$ and $p = 0.051$ really means. However, context factors—design of study, quality of measurements, and data analysis assumptions—are often given little emphasis or done as a topic of discussion separate from the *p*-value calculation and interpretation.

4. Proper inference requires full reporting and transparency.
Grade: C+
Reason: Our curricular components currently discuss *p*-values in the context of a single study. The class does a lab exercise on data cleaning and ethical / transparent data handling procedures, but largely omits the discussion about transparent reporting of analysis choices. Topics like multiplicity of testing and "p-hacking" are rarely discussed in our introductory classes, which do not tend to emerge until a second course in statistics.

5. *p*-value does not measure the size of an effect or the importance of a result.
Grade: A−
Reason: Exercises and class activities address practical significance vs. statistical significance. Emphasize interval estimation and reporting effect sizes across curricular components. Need to be careful that we do not implicitly suggest that larger *p*-values imply "lack of importance or even lack of effect" or that smaller *p*-values necessarily imply "larger or more important effects."

6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.
Grade: C
Reason: We do not suggest that data analysis ends with the construction of a *p*-value; but rarely do we discuss or evaluate other evidence in support of our models/hypotheses.

3.1. Application of Course Content Audit Rubric

Table 3 contains the rubric that was collectively completed by members of the author team to evaluate the curriculum of the introductory course discussed in Section 2.1. The audit for our algebra-based introductory course found that we avoided improper probabilistic interpretations and clearly articulate the difference between statistical and practical significance as highlighted in principles two and five, respectively. However, the results of the audit suggested that we had ample room to improve our coverage of some principles. As a result, we developed a set of action items to address these areas' curricular weaknesses. A natural follow up to the assessment is to document what features were highlighted in this evaluation as a justification of the assigned grade. Table 4 provides this reflection.

The most urgent action item was to replace the rote procedural approach of running down the checklist of model assumptions taught to accompany each hypothesis test, and instead begin to frame hypothesis tests more comprehensively with respect to clearly specified statistical models. For example, we can discuss a hypothesis test for a population proportion of success equal to 0.5, where a binomial exact test runs from a simple and familiar binomial model. Here, we can note that the *p*-value calculated could reflect the true proportion differing from 0.5 or that the observations were not from a binomial experiment. Another action item from the audit is to include a case study to deconstruct a real-world analysis involving hypothesis tests, where we can highlight study design, dangers of "p-hacking," effect sizes, integration of additional evidence, and resulting policy decisions. The last action item is to expand our discussions of reproducibility beyond documenting data handling to include all stages of a study, from design through analysis. The Reproducibility Project led by the Center for Open Science can provide a good launching point for this discussion (Weir 2015).

We have shared results of the audit of this introductory course curriculum with our colleagues at a recent departmental retreat. The discussion encouraged change to modify content and exercises to better address the ASA *p*-value principles throughout our curriculum. We also asked our colleagues to consider how we can similarly focus the curriculum of our upper-level classes to reinforce areas that are challenging to impart with students in an introductory course.

Table 5. Highlights from focus group meeting with colleagues from client departments.

Question prompts	Summary of focus group discussion responses
Teaching in our *introductory* statistics service courses: - Do your students show understanding of *p*-values and hypothesis testing? - What methods should an introductory statistics course include?	*p-value understanding:* Students do not adequately understand *p*-values after intro courses • Need more theory and intuition about *p*-values, less procedural mechanics and thresholds, show impacts of assumptions *Desired coverage:* Want students to be good data consumers—need "soft touch" in intro • Considering alternative sources of uncertainty • Students do not seem to understand discrete probability distributions • More fundamental probability, including conditional probability • More exposure and practice with software and packages • Emphasize effect sizes and interval estimates • May benefit from resampling / randomization—illustrate null model • Open question: How can we get more advanced methods from intro?
Teaching in our *advanced* statistics service courses: - What does it do well to prepare your students? - Do your students show understanding of *p*-values and hypothesis testing? - What is missing or needs more detail?	*Coverage of Methods:* *p*-value good starting point, necessary for reading literature • Many students do not understand *p*-values even after a second course • Becoming practitioners, need to know principles and adv. methods *What is needed:* More modeling vs. HT—focus on fit (R^2) and selection vs. tests • Behind in teaching Bayesian methods *Challenges to extending beyond HT:* • May go against adviser's training and level of expertise • Books/literature still have heavy focus on HT • "$p < 0.05$ will continue to be taught b/c it is so much easier to teach" • "What do we reinforce? We implicitly support this bright line" • Grad students too focused on application, ignoring important nuance • Simplicity of "yes/no" type of outcome is attractive
Why is a fundamental understanding of statistics important in your field? What attracted and engaged you personally with statistics?	*Big issues for stats in STEM education:* • Too much focus on "what of statistics", need more "why of statistics" • Motivation gap for learning statistics vs. science specialty • Fighting student expectations (and sometimes their adviser's too) • Challenging idea: something could be wrong in a scientific model *Individual responses for personal experience with learning statistics:* • Tools for addressing the idea of significance / expressing confidence • Not interested as an undergraduate—took off as a graduate student when encountered Markov models to explain behavior "Hated undergraduate stats class but in graduate school it made sense and I love it now" • Interest and passion emerged in grad school but stat classes not directly relevant to their field—90% of statistics background was learned as a researcher. Need to start motivating the learning earlier.

3.2. Summary of Focus Group Discussion

Looking externally for feedback, our focus group was comprised of five colleagues from the fields of biology, kinesiology, psychology, and environmental science. During the ninety minute session with these representatives of our client departments, we asked the questions listed in Section 2.2. The questions were posed as prompts to discussion periods where a member of the statistics faculty primarily acted as a facilitator: taking notes, asking for additional details and prompting for points of clarification. The discussion yielded fruitful conversation; the detailed summary is found in Table 5.

The focus group provided strong insights for us to consider when we engage students from many scientific disciplines in our service courses. We are not proposing that we completely restructure our course based this input, only that we acknowledge what is viewed as important by client departments when evaluating service course curricula. A major theme that came out of the conversation was that hypothesis testing and *p*-values are used widely in their fields, so students need a better understanding of the ASA principles. The focus group would also like to see a more thorough coverage of additional concepts such as the fundamentals of probability, model fitting, effect size estimation, and Bayesian methods in our curricula for future STEM students. Lastly, the personal feelings toward statistics by the faculty members from other disciplines were illuminating and encouraging. Several of them divulged an animosity or an ambivalence toward statistics in their undergraduate educations, but later came to recognize the value in their careers during graduate study. This revealed the urgent need to provide stronger motivation for STEM students in an undergraduate statistics course. We should strive to motivate *why* the statistical methods they are learning are necessary within their respective fields, and demonstrate the value that statistics provides within science. A final benefit of this meeting with representatives from other departments is that this strengthens the connections between statistics departments and client departments.

4. Discussion

We encourage statisticians in academia to evaluate the coverage of *p*-values and hypothesis testing in their curriculum, in light of the ASA *p*-value statement. Within our statistics department, we conducted an audit of our algebra-based introductory course using our broad rubric to evaluate how effective the curriculum is at conveying the six principles from the ASA *p*-value statement. We then looked outside the department, through a focus group conversation with colleagues across several scientific disciplines. Through these activities, we have formed a better understanding of the coverage of *p*-values and hypothesis tests within our introductory curriculum and have developed a set of action items to remediate areas where we can improve student learning outcomes. Following the research of Dewey (1933), we feel that the critical reflection on the state of our inference curriculum for our introductory service courses will make us more aware of how we approach these concepts in our other courses. We propose the framework that we used in our reflective process as a basic template that can be adapted and implemented at other universities.

Several challenges remain for fully improving the quality of *p*-value understanding in introductory courses. At many institutions, introductory statistics courses are staffed by temporary instructors or visiting faculty, who may lack experience in the teaching or study of statistics. In addition, lab sections are often facilitated by graduate assistants, many of whom are new to statistics. Continuing research methodology courses in other departments are also often taught by non-statistician methodological instructors who may hold misconceptions about significance and *p*-value interpretations (Harradine, Batanero, and Rossman 2011). This suggests that instructor preparation is a necessary component for improving the teaching of *p*-value concepts. We need to have a continuing process of training the instructors for our introductory classes. A great set of resources for staying current with instructional practices can be found through the ASA Section on Statistics Education (*community.amstat.org/statisticaleducationsection*), the consortium for the Advancement of Undergraduate Statistics Education (*causeweb.org*) and the Statement on Qualifications for Teaching an Introductory Statistics Course by the ASA/MAA Joint Committee on Undergraduate Statistics (JCUS) (2014).

Another challenge for introductory statistics is that the ASA statement encourages that *p*-values be framed within a wider range of methods. We cannot expect students new to statistics to appreciate the value of robust analyses and reporting without some real exposure to those methods. For this, we need to find space in the curriculum—a scarce commodity in any statistics course—to cover new analysis methods, such as model selection, bootstrapping, generalized linear models, or Bayesian methods. Including a broader context of *p*-values and/or reinforcing the introductory curriculum will almost certainly come at the cost of another topic, and clearly not all of these topics can be added to an introductory statistics class. While the particular weight given in the curriculum is subjective, it should be heavily considered when striking the balance, given the general consensus in statistics education that a proper understanding of inference is a foundational learning objective in introductory statistics (Cobb 2007; Garfield and Ben-Zvi 2008; Rossman 2008; Carver et al. 2016). Realistically, we concede that a truly robust understanding of statistical analyses is developed through a continued statistical education beyond the introductory course. While this is a given for the future statisticians in our classrooms, our discussions with colleagues from the focus group made it clear that we need to provide consistent motivation, starting in the earliest service courses, on how statistical foundations—such as a correct understanding of *p*-values—are valuable across all scientific fields. We agree with Harradine, Batanero, and Rossman (2011) who argued that with students understanding of inference, "the underpinning ideas need to be developed over years, not weeks."

As statistics educators, we need to recognize the potential of our service courses to inform the next wave of scientists and scholars about the correct interpretation and fundamental use of *p*-values. Statisticians should also strive to reinforce the value that statistical analyses bring to general scientific inquiry by actively engaging in conversation with our peers in other scientific professions.

Acknowledgments

The authors thank our collaborators—Drs. Tom Crist, Joe Johnson, Jonathan Levy, Hank Stevens, and Rose Marie Ward—for representing their respective scientific disciplines and providing insights in the curricular focus group discussed. We also thank the anonymous referees and associate editor for the helpful suggestions they provided on the first version of this article.

References

ASA/MAA Joint Committee on Undergraduate Statistics (JCUS) (2014), "Qualifications for Teaching an Introductory Statistics Course.", American Statistical Association and Mathematical Association of America Joint Committee on Undergraduate Statistics, available at *amstat.org/asa/files/pdfs/EDU-TeachingIntroStats-Qualifications.pdf*.

Batanero, C. (2000), "Controversies Around the Role of Statistical Tests in Experimental Research," *Mathematical Thinking and Learning*, 2, 75–97. [385]

Berry, D. A. (2016), "P-Values Are Not What They're Cracked Up to Be, Online Discussion: ASA Statement on Statistical Significance and P-values," *The American Statistician*, 70, 1–2. [385,386]

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G.H., Velleman, P., Witmer, J., and Wood, B. (2016), "Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016," American Statistical Association, available at *amstat.org/education/gaise*. [386,390]

Cobb, G. W. (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum?" *Technology Innovations in Statistics Education*, 1, 1–16, available at *escholarship.org/uc/item/6hb3k0nz* [390]

——— "ASA Statement on P-Values: Two Consequences We Can Hope For," Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values," *The American Statistician*, 70, 1. [385]

DelMas, G., Joan, G., Ooms, A., and Chance, B. (2007), "Assessing Students' Conceptual Understanding After a First Course in Statistics," *Statistics Education Research Journal*, 6, 28–58. [385]

Dewey, J. (1933), *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*, New York: D.C. Heath. [390]

Falk, R., and Greenbaum, C. W. (1995), "Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception," *Theory & Psychology*, 5, 75–98. [385]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, P-values, Confidence

Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [385]

Goodman, S. N. (2016), "The Next Questions: Who, What, When, Where, and Why? Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values." *The American Statistician*, 70, 1–2. [386]

Garfield, J., and Ben-Zvi, D. (2008), *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*, Berlin: Springer Science & Business Media. [390]

Harradine, A., Batanero, C., and Rossman, A. (2011), *Students and Teachers Knowledge of Sampling and Inference in Teaching Statistics in School Mathematics—Challenges for Teaching and Teacher Education. A Joint ICMI/IASE Study: The 18th ICMI Study*, New York: Springer. [385,390]

Ioannidis, J. P. (2005), "Why Most Published Research Findings are False," *PLoS Medicine*, 2, e124. [385]

Millar, M. (2016), "ASA Statement on P-values: Some Implications for Education. Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values," *The American Statistician*, 70, 1. [386]

Nickerson, R. S. (2004), *Cognition and Chance: The Psychology of Probabilistic Reasoning*, New York, NY: Psychology Press. [385]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [385]

Rossman, A. J. (2008), "Reasoning about Informal Statistical Inference: One Statistician's View," *Statistics Education Research Journal*, 7, 5–19. [385,390]

Stangl, D. (2016), "Comment. Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values," *The American Statistician*, 70, 1. [385]

Wasserstein, R. L., and Lazar, N. A. (2016), "ASA Statement on P-values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [385]

Weir, K. (2015), "A Reproducibility Crisis? The Headlines Were Hard to Miss: Psychology, They Proclaimed, is in Crisis," *Monitor on Psychology*, 46, 39. [388]

Williams, A. M. (1999), "Novice Students' Conceptual Knowledge of Statistical Hypothesis Testing," in *Making the Difference: Proceedings of the Twenty-Second Annual Conference of the Mathematics Education Research Group of Australasia*, 554–560. [385]

The American Statistician

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

Beyond Calculations: A Course in Statistical Thinking

E. Ashley Steel, Martin Liermann & Peter Guttorp

To cite this article: E. Ashley Steel, Martin Liermann & Peter Guttorp (2019) Beyond Calculations: A Course in Statistical Thinking, The American Statistician, 73:sup1, 392-401, DOI: 10.1080/00031305.2018.1505657

To link to this article: https://doi.org/10.1080/00031305.2018.1505657

© 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: 17059

View related articles

View Crossmark data

Citing articles: 8 View citing articles

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, 392–401: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1505657

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Beyond Calculations: A Course in Statistical Thinking

E. Ashley Steel[a], Martin Liermann[b], and Peter Guttorp[c,d]

[a]PNW Research Lab, USDA Science Center, Seattle, WA; [b]Northwest Fisheries Science Center, Seattle, WA; [c]Norwegian Computing Center, Oslo, Norway; [d]Professor Emeritus, University of Washington, Seattle, WA

ABSTRACT

Statisticians are in general agreement that there are flaws in how science is currently practiced; there is less agreement in how to make repairs. Our prescription for a Post-$p < 0.05$ Era is to develop and teach courses that expand our view of what constitutes the domain of statistics and thereby bridge undergraduate statistics coursework and the graduate student experience of applying statistics in research. Such courses can speed up the process of gaining statistical wisdom by giving students insight into the human propensity to make statistical errors, the meaning of a single test within a research project, ways in which *p*-values work and don't work as expected, the role of statistics in the lifecycle of science, and best practices for statistical communication. The course we have developed follows the story of how we use data to understand the world, leveraging simulation-based approaches to perform customized analyses and evaluate the behavior of statistical procedures. We provide ideas for expanding beyond the traditional classroom, two example activities, and a course syllabus as well as the set of statistical best practices for creating and consuming scientific information that we develop during the course.

ARTICLE HISTORY
Received January 2018
Revised June 2018

KEYWORDS
Probabilistic thinking; *p*-Values; Simulation-based testing; Statistical intuition Statistics education

1. A Prescription for a Post-$p < 0.05$ Era: Teach Statistical Wisdom

Statisticians and others are in general agreement that the current system of curating scientific knowledge is flawed (Sterne and Smith 2001; Leek and Peng 2015; Wasserstein and Lazar 2016). In fact, 90% of 1576 researchers surveyed agreed that there is a crisis in reproducibility in science with more than half agreeing that it is a major crisis (Baker 2016). There are concerns about scientists' inability to replicate published results and the cost to society of an ineffective scientific process. To identify just a sample of the issues underlying these sentiments: questions are poorly formulated, too many tests are conducted, *p*-values are overvalued or misinterpreted altogether, and the peer-review process is overwhelmed. There is a pressing need to move statistical analysis and evidence-based decision making beyond the current system of overtesting and underthinking.

There is less agreement among statisticians on exactly how to repair our scientific machinery. Some promote alternative systems of assessing statistical significance, for example, $p < 0.005$, (Benjamin et al. 2018) or of evaluating information, for example, Bayesian methods (e.g., Johnson 2013). Some statisticians promote changes to the publication process (e.g. Sterne and Smith 2001) such as preregistration, preacceptance of study designs, and publication of negative results. Still others promote shifts in the way we evaluate scientists themselves with emphasis on quality of scientific work rather than quantity of high-ranking publications (e.g., Martinson 2017).

We follow on a proposal from Brown and Kass (2009) to expand our view of what constitutes statistical training. In fact, we believe that the domain of statistics includes not only calculations but also the beginning of the scientific process, for example, asking a clear question and tying it correctly to analytical methods, and the end of the scientific process, for example, communicating results to the general public, as well as all the steps in-between. If the problem is a flawed system of curating scientific knowledge, in which *p*-values have played a leading role, our prescription for a Post-$p < 0.05$ Era is to equip students with a clear understanding of the opportunities and pitfalls in the application of statistics within that entire system.

We propose that there are opportunities to build a more efficient and effective knowledge supply chain. The wisdom necessary to identify and avoid common statistical mistakes or to understand flaws in the scientific process is generally arrived at through years, if not decades, of experience with data analysis, repeated exposure to common statistical errors and, long-term observation of the publication process. Instead of waiting for this wisdom to be accumulated, we could engage purposefully in teaching this statistical understanding to graduating statisticians and new researchers. In the previously mentioned poll on the reproducibility crisis, the number one factor needed for boosting reproducibility in science, cited by just over 50% of those surveyed, was "better understanding of statistics." (Baker 2016).

In this article, we begin by outlining the history of efforts to improve teaching of statistical thinking and by describing the current atmosphere of opportunity for making an impact. We then explain the structure and unique features of our course for advanced undergraduate statistics majors and early graduate students from any discipline. Our course is explicitly designed

CONTACT E. Ashley Steel easteel.2018@gmail.com PNW Research Lab, USDA Science Center, 400 N 34th Street, Suite 201, Seattle, WA 98103.

© 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.

to teach students why statistical thinking can be challenging; what common errors are and how they can be avoided; how a single statistical analysis fits into the larger supply-chain of knowledge; how errors can be induced from mismatched data summaries and research questions; what *p*-values are and how they are often misused; and what the responsibilities of scientists, including statisticians, are in communication of research results and uncertainty. We include two example activities and a detailed syllabus to promote the development of a shared library of resources for teaching statistical thinking beyond first-year courses. We conclude with a short synthesis of lessons learned in the first few years of delivering our course, and a call for expanding the domain of statistics to other courses and across communication platforms.

2. The Value of Teaching Statistical Thinking: Now is an Opportunity

Our ideas build on the efforts of many to improve statistical education. A conceptual framework for K-12 statistics education aimed at enhancing existing mathematics standards was produced in 2007 (Franklin et al. 2007). One of the stated goals of this endeavor was that "every high-school graduate should be able to use sound statistical reasoning to intelligently cope with the requirements of citizenship, employment, and family and to be prepared for a healthy, happy, and productive life." Similarly, a tremendous effort has gone into researching and providing guidance for college-level introductory statistics courses. The GAISE College Report lists "teach statistical thinking" as the first of six recommendations for introductory courses (GAISE College Report ASA Revision Committee 2016). And, the American Institute of Biological Sciences (AIBS) 2015 report on bioinformatics workforce needs recommended training that includes the concept of the data lifecycle and project-based learning (AIBS 2015). There is still much room for innovation, particularly in integrating statistics and science. Statistics is "vastly underappreciated because too few statisticians, relatively speaking, have effectively conveyed the excitement of statistics, as a way of scientific thinking" (Meng 2009). Our course focuses on statistical thinking and, where possible, statistical wisdom across the full domain of statistics from asking questions to interpreting and communicating results. In so doing, it provides a bridge between undergraduate statistics coursework and the real-world application of statistical methods.

Teaching statistical thinking is not a new proposal. Twenty years ago, Moore (1998) envisioned statistics as belonging to the liberal arts because our discipline offers tools and ideas for reasoning and a framework for evaluating evidence. Drawing conclusions from observations and data, understanding variation, and reasoning about chance are essential skills for today's world. "We [statistics] offer broad and flexible modes of reasoning that make smart people smarter in every aspect of life and work" (Moore 1998). He noted that the way we introduce the discipline of statistics to the hundreds of thousands of students who are subjected to a first-year statistics course every year is both an opportunity to communicate the value of statistical thinking and a public responsibility. We would add that similar opportunities and responsibilities exist at the transition from undergraduate study to graduate research and beyond.

In 1992, Cobb summarized three recommendations for curricular change in statistics as identified by a Statistics Focus Group on curricular action (Cobb 1992): (1) Emphasize statistical thinking; (2) more data and concepts: less theory, fewer recipes; and (3) foster active learning. Under "Emphasize statistical thinking," the focus group included topics and principles for classroom learning such as the need for real data, consideration of how data are produced, the omnipresence of variability, and the quantification and explanation of variability. We have observed that, although the vast majority of statisticians agree with these ideals, many sequences of coursework for statistics majors fail in at least some of these and related elements. Graduate students outside of statistics are rarely exposed to these ideas at all and certainly not before they design their first research project. It remains unusual that these topics are covered explicitly, with the same depth or rigor of the technical content for which our discipline is, perhaps incorrectly, better known.

The ASA Ethical Guidelines describe statistician's responsibilities to the public, research subjects, colleagues, other statisticians, employers, and clients (ASA 2016). Strong statistical thinking skills and an expanded concept of the domain of statistics can help early-, mid-, and late-career statisticians meet these ethical standards that include, for example, avoidance of running multiple tests. By virtue of our experience, statisticians also have an ethical obligation to students of all ages and disciplines, an obligation to share and teach skills for critically evaluating claims of new knowledge, making strong decisions in the face of uncertain information, and avoiding egregious errors in probabilistic thinking.

As a pointed example of our current failures, and at the risk of drawing false conclusions from anecdotes, here is a short story. One of us recently suggested to a student applying for medical school that a good course in statistics was essential for physicians. She replied that she had taken AP Statistics in high school, that it was horrid, and that she had therefore met her statistics coursework requirements forever. She had no intention of ever taking another statistics course. Vaguely horrified, we asked "How will you know whether to take newly published research seriously and how will you know when to change your ideas about the best treatments? And how will you evaluate ..?" "Oh I know all that" she interrupted nonchalantly, "just check the *p*-value." This student might be making health care decisions for you in 10 years.

The enthusiasm and energy surrounding this special issue and the ASA's Symposium on Statistical Inference, Bethesda, MD October 2017, together with the following trends brings unprecedented opportunity for change. The recent growth in the number of undergraduate statistics majors makes this an ideal moment in history to experiment with new course offerings. From 2003 to 2011, there was a 78% increase in bachelor's degrees in statistics, biostatistics, and related fields (Pierson 2013). We would expect the trend to continue as the market demand for quantitative skills continues to increase. The Bureau of Labor Statistics (2017) projected that the number of statisticians would increase by 27% between 2012 and 2022, while the general job market was only expected to grow by 11% during that same time. There is, of course, expected to be a large number of jobs in the general area of Big Data; Lund et al. (2013) expect openings for nearly 200,000 people in this field.

Table 1. Core Topics for an Advanced Course in Statistical Thinking.

Topics	Detailed Understanding
Errors in thinking	How the wiring of the human brain leads to incorrect conclusions from data.
Clear questions	Quantifying the metric of interest in scientific questions.
Simulations	Using simulations to understand the behavior and power of statistical procedures.
Estimation	The advantages of estimation over testing for many common situations.
The p-value	Understanding what it is, what it tells us, and what it does not tell us.
Repeated experiments	Why results from a single study are inconclusive and how the scientific process is iterative.
The science process	Pitfalls and opportunities of the project life-cycle and peer-review process.
Study design	Principles of good study design and how poor design can lead to incorrect conclusions.
Sample size	Issues with small samples and with big data. How these both can lead to incorrect conclusions.
Communication	Principles of scientific communication and how they apply to statistics.

The analytics area, be it in business, sports, or criminology, is seeing vastly increasing demand. The workforce a decade from now will employ a large number of statisticians all of whom must be able to both calculate to understand the context of their analysis. As well, students now enter college having completed math curricula with more probability theory and early statistical thinking skills than in the past, providing an "opportunity to build on a broader foundation of prior knowledge that leaves room to delve deeper and farther than ever before possible" (GAISE College Report ASA Revision Committee 2016).

3. Structure of Our Course in Statistical Thinking

The course, focused on a set of core topics that move from how the human brain works to statistical communication (Table 1), is intended for senior statistics majors, though some sophomores and juniors have enrolled, and anyone else interested in conducting research (usually graduate students but we have had some research-minded undergraduates from other disciplines). In terms of their understanding of the role of statistics in science, statistics senior undergraduates and, say, first-year ecology graduate students are not so far apart. There are no mandatory prerequisite courses, but some knowledge of standard statistical procedures, such as *t*-tests or linear regression, is strongly recommended. We aim to reach students who think they understand something about statistics and who are ready to think hard about what they have yet to learn. We were proud when one student described our course as having a "completely different framing than any other course: how to understand data, how to filter out nonsense, how to navigate quantitative data."

We interweave lectures, in-class activities such as analog simulations, for example, drawing marbles from a bag, discussion groups, and computer labs. A key feature of the course structure is that all in-class work is done in assigned groups, which are designed to mix students from multiple disciplines, with different perspectives, and with a range of experiences. We have observed that the broader the range of backgrounds, the more effective the group dynamics. The application of statistical tools, after all, requires both mathematical and non mathematical knowledge (Kuzmak 2016). The students with domain-related knowledge and experience are essential for helping the statistics majors interpret results and engage in the exercises. Interestingly, the number of statistics courses previously taken does not seem to predict course performance. Students with lots of procedural knowledge are not necessarily those with the interest and ability to think about how those procedures can inform our understanding of the world around us. The limiting factor for many students is confidence in R programming skills. We design the groups to distribute the most experienced R programmers across groups. Sample code is always provided and we have added a lab in the second week of the course specifically designed to teach the bare bones of running simulations in R. We also ask that the least experienced R programmer does the typing which keeps all students engaged, either typing or coaching.

The course is divided into three sections (see Table 2) which form a story arc from how humans make observations to making inference from observations to communicating what has been learned. In the first section, Making Sense of Data, we focus on the workings of the human brain and on sampling distributions. We provide students with many examples to demonstrate how easily we, as humans, are fooled by pattern and coincidence (Tversky and Kahnemann 1974; Diaconis and Mosteller 1989). And we argue that, by heightened awareness of these common errors, we can avoid them. We also discuss big ideas like confidence intervals, introduce simulations as a way of learning and testing, and give students the quick overview of R mentioned above. The second section, Using Data for Scientific Understanding, is all about making inference. Students gain experience through readings and exercises that increase their understanding of *p*-values (as described in detail below) and promote the idea that estimation is often a better approach than testing. We help them identify when statistical testing is appropriate and enable them to design their own randomization tests to ask and answer exactly the question of interest. We also discuss inference on a macroscale. How does science proceed from an experiment in a lab to common understanding or accepted truth? How does the peer-review process influence the application of statistics? In the third section, Using Science in Society, we focus on interpreting scientific claims (Sutherland et al. 2013), communication of risk, communication with and by the media, applications of statistical insight in controversial topics such as climate change, and the ability (or lack of ability) for scientific fact to change people's minds about controversial topics. Student understanding of the big ideas in the course doesn't happen all at once but through spiraling back to ideas over and over again from different perspectives.

During the first half of the course, students are assigned both readings (or similar activities) and R homework each week. Assigned readings and activities range from original statistics papers to science articles, magazine articles, blog posts, spurious correlation and *p*-hacking simulators, and videos. Many of the key readings are cited in this article. R homework generally follows directly from the in-class labs. During the second half

Table 2. Condensed syllabus for our 10-week course in statistical thinking.

Week	Main Topic	Example Lecture Modules
SECTION I: MAKING SENSE OF DATA		
1	How the human brain works, scientific reasoning, & simulation thinking	Human judgement under uncertainty and common errors in statistical thinking. Interpreting patterns visually then using physical and computer simulations to improve on that interpretation (Example 1: Explain this … or that!).
2	Coincidences, randomness, & measurement (and an introduction to R)	Frequency of coincidences and four main reasons for coincidence (with group discussion). Patterns emerge from randomness. Basics of using R statistical software—particularly for simulations.
3	Populations, samples, means, confidence Intervals, & over-confidence	The difference between a sample and a population, sampling distributions. Understanding confidence intervals, credible intervals, and prediction intervals. Over-confidence in the human brain and its relationship to statistical thinking. Hands-on simulation of sampling with $n = 3$ versus $n = 10$ to understand that the sample mean is random; how confidence intervals of the mean arise; and the role of sample size (Figure 1). Bootstrapping confidence intervals in R.
SECTION II: USING DATA FOR SCIENTIFIC UNDERSTANDING		
4	Comparing two populations	Multiple ways to compare two populations using samples. What is the metric of interest? It is not always the mean. Effect sizes and estimation. Hands-on permutation tests using paper data. Permutation testing in R. Effect sizes in R.
5	*P*-values	A history of *p*-values. *P*-hacking: what is it and why does it matter? Peer review process and how it influences abuse of *p*-values and significance testing. Students conduct simulation of a research study famous for *p*-hacking to discover how it works and why is it so dangerous (Example 2: The Science of Bad Science). Simulation experiments on p-values.
6	The life cycle of science	Statistical power. Using simulations to evaluate statistical power in R. How ideas are curated in science and the crisis of replication. File drawer effect (Rosenthal 1979) and problems with *p*-values. Garden of forking paths (Gelman 2013) and problems with multiple testing. Peer review in the context of the crisis of replication. Proposed solutions to the crisis of replication (with group discussion).
SECTION III: USING SCIENCE IN SOCIETY		
7	Study design & Big data	Students collect observational data on cell phone use while driving as pilot for a well-designed study. Key principles of study design. A brief history of data. What is "Big Data"? Perils and opportunities of big data. Student lab using Google trends to enrich group project.
8	Statistical & scientific communication	Communication of risk. Communicating science and statistics with the media. The role of the press release. How twitter and social media influence science communication. Students work in groups to evaluate a selection of science media reports. Science and statistical communication best practices.
9	Application: Climate change research	How does the Intergovernmental Panel on Climate Change (IPCC) work and what is the role of statistics? Communication of risk and scientific evidence in controversial topics. Estimating and ranking global mean temperature in R.
10	Student project presentations	Short summary presentation on best statistical practices (Table 3).

of the course, students have very little programming homework and, instead, work in groups on a final project to apply their new understanding and communication skills in conducting statistical testing, drawing conclusions from data, and communicating findings honestly yet with enthusiasm.

Final projects are delivered as oral presentations, newspaper or magazine articles, and social media posts. These group projects not only include real data but they apply to real-world situations for which one truly seeks and cares about the answer (Kuzmak 2016). The data are messy; the questions are poorly specified; graphs must be designed to answer a question; and *t*-tests are not allowed. Although each group project is designed to cover the full spectrum of activities involved in using data to answer a question, each project emphasizes a particular element of the course from heuristics that lead to flawed thinking to the perils of multiple tests to the challenges of communicating exciting results with honesty. Past group projects have included conducting a survey similar to that of Tversky and Kahneman (1974) to reproduce and/or extend their results, for example by comparing performance between statistics and nonstatistics majors; exploring the first four years of data from a large Federal research project on stream restoration in Alaska to provide the researchers with preliminary information relevant to their stated research questions; re-evaluating an article in Grist claiming that large snowstorms have become more frequent on the East Coast using the data reported in the article and publicly accessible data for several cities; and graphing downscaled climate hindcast and forecast data for a watershed south of Bangkok, Thailand, to provide insight about likely future changes in climate that would be relevant to local farmers. In the final week of the course, group project presentations are ordered so as to provide a fairly organized review of the key messages and ideas in the course. For example, the above projects would be presented in the order described so as to cover typical human errors, matching analysis to research questions, how the media communicates scientific results, and communicating to particular stakeholder groups. To avoid what might be seen as a cynical and depressing view of all the opportunities to make mistakes, we also accumulate a set of best statistical practices during the course. On the final day, we review and distribute this succinct list of reminders for conducting and consuming scientific analyses (see Table 3).

After taking our course, we expect students to have gained (or sharpened) the following understandings and skills: (1) the ability to identify and prevent common errors in thinking that lead to erroneous conclusions; (2) the ability to match a research question to a statistical procedure, including the metric of interest; (3) the ability to create a bootstrap confidence interval for a quantity of interest; (4) the ability to customize a

Table 3. Best statistical practices: for implementation by undergraduates and early graduate students

Plot your data—early and often.
Understand your dataset as one of many possible sets of data that could have been observed.
Understand the context of your dataset—what is the background science and how were measurements taken.
Be thoughtful in choosing summary metrics.
Decide early which parts of your analysis are exploratory versus confirmatory and pre-register your hypotheses in your own mind.
If you are going to use *p*-values, which can be useful summaries when testing hypotheses, follow these principles: • Report estimates and confidence intervals; • Report the number of tests you conduct (formal and informal); • Interpret the *p*-value in light of your sample size (and power); • Don't use *p*-values to claim that the null hypothesis of no difference is true; • Consider the *p*-value as one source of support for your conclusion not the conclusion itself.
Compute (and display) effect sizes and confidence intervals as an alternative to or in addition to statistical testing.
Consider creating customized, simulation-based statistical tests for answering your specific question with your particular dataset.
Use simulations to understand the performance of your statistical plan on datasets like yours and to test various assumptions.
Read with skepticism, remembering that pattern can easily occur by chance (especially with small samples), and that unexpected results based on small sample sizes are often wrong.

permutation test for a comparison of interest; (5) the ability to use simulations to conduct and evaluate statistical analyses; (6) an understanding of how knowledge is curated and the role that statistics plays in that process; and (7) skills for communicating scientific and probabilistic information honestly and clearly. We summarize our teaching philosophy as "building efficient skepticism to improve the application of statistics in scientific research." By efficient, we intend to imply both that skepticism should be a natural part of consuming scientific information and also that it need not prevent the forward momentum of ideas.

Our course is called "So You Think You Can Do Statistics?", a pun on the TV show "So You Think You Can Dance?", and has been taught three times as a 3-credit, 1-quarter class at the University of Washington, cross-listed in Statistics and College of the Environment. Student reviews have been exceptional. "The most interesting stats class I've ever taken," "It teaches things that weren't mentioned in a lot of theory classes," "requires us to think about the process of doing statistics," "mostly it really makes sense when you think about it, but you would never think about it on your own," and "permutation tests are a Swiss-army knife for designing statistical tests." In 2017, 72% of students ranked the course content as excellent and there was a request to increase maximum enrollment for 2018.

By the very principles we teach, we can make no assessment of the "significance" of our course. Early indications, however, are positive. As a part of our course evaluation process, students read and evaluated similar articles before and after the course. Before our class, no students mentioned that they considered the sample size or that they wished for confidence intervals in their evaluations of these articles but, after the course, students focused on these ideas. After the course, students were also concerned about the influence of base rates, cherry-picking results, and decisions about what was measured on the article's final conclusions. When asked anonymously what they would take away from the course, undergraduate students wrote "Those fallacies when interpreting those scientific news"; "Awareness of common statistical pitfalls and bootstrapping. Specifically knowing how easy it is to manipulate information to support a desired outcome"; and "All of the things that we need to look out for, and then just about how *p*-values aren't the end all be all of statistics like they're made out to be." Graduate students from 2018 described their learning informally as "I learned that even published scientific studies should be taken with a grain of salt" and "I feel that I will take a much more cautious approach to analyses I will conduct in my own research, particularly when it comes to choosing a particular analysis approach ahead of time and sticking to it." In the formal evaluation, one graduate student wrote "Since beginning this class, I am definitely more critical of every paper, article, or news story that I read. I always check sample size, and then look to see if they report confidence intervals or uncertainty. I have started to mistrust *p*-values and have noticed the dependence on them in literature. I have tried to decipher how much is communicated to the general public versus the research community and how data is misconstrued." Whether this same wise skepticism is applied and retained in the future remains to be seen.

4. Unusual Features of Our Course in Statistical Thinking

There are similarities between our approach and the approaches of others, including the GAISE recommendations for first year courses in statistics (GAISE College Report ASA Revision Committee 2016). Like others, we also hope our course is fun and relevant to the daily life of students (Meng 2009), we encourage effective graphing, and we discourage overreliance on statistical testing.

Our course is, however, different from traditional statistics courses in several ways. First, our course is *predicated on the idea that the biggest problems with statistics as currently practiced are not in the details of the techniques* (e.g., nonnormality, nonconstant variance, unaccounted for nonlinearity) *but in the big ideas* surrounding application of statistical tools (e.g., how does what you have measured relate to the subject of interest (*sensu* Mallows 1998), the likelihood of observing strong patterns by chance alone, the tendency to oversell novel results). Our course therefore presents a body of knowledge not commonly taught as a cohesive whole. As a result, and secondly, *the course is inherently multidisciplinary*, including topics rarely covered in Statistics Departments such as human psychology, communication, and peer review. Third, *our course is intended as a bridging course*. It serves as a bridge between statistical knowledge and successful application of statistics in science. We have found that the sharing of experience across advanced undergraduates and early graduate students enriches the learning outcomes for both. Fourth, *our course focuses on what can go wrong* in the application of statistics. Statistical tools are often taught within a framework in which they are useful. We spend nearly equal

time focusing on examples of when, how, and why application of statistical tools go wrong. Fifth, *our course considers how a single analysis fits into the lifecycle of science*. For example, we explicitly consider the file drawer effect (Rosenthal 1979) and the concept of the garden of forking paths (Gelman and Loken 2013), as well as the filtering that occurs during the peer-review process. Sixth, *we conduct experiments on p-values and explicitly discuss the replication crisis in science*. Seventh, we devote nearly 20% of the course to *statistical communication* including activities such evaluating press releases, watching videos of prominent statisticians explaining the demands of statistical communication, and considering how social media magnifies the pressure and opportunity for scientists to overstate their claims. And, finally, *we confront the messiness of data analysis head-on*. Subjective decisions are required at multiple steps along the way and answers are not always conclusive.

A very similar course could be taught as a seminar in "How science happens" to graduate students or as a capstone course for statistics majors. Statistical capstone courses have been designed in a variety of formats and often focus on large-scale data analysis projects (Martonosi and Williams 2016). One common format for a capstone course is an in-depth analysis of real-world data embedded in an advanced methodology course. As a capstone, our course would differ in its lack of emphasis on advanced methods and in the focus of our final projects on specifying the question of interest, identifying best metric to answer the question, creating graphics that tell a story, and communicating, even in the form of tweets and Facebook posts, the scientific conclusion.

Three features of our course demand a more in-depth explanation.

Simulations. Simulations can form a foundation for K-12 courses (Franklin et al. 2007), introductory college-level statistics courses (e.g., Lock et al. 2013 and Tintle et al. 2015), and advanced courses such as "Advanced Resampling Methods." Simulations give us a multipurpose tool to build intuition and to create customized estimation and testing approaches that would otherwise be difficult. We use simulations in four ways: sampling from a population and observing the sampling distribution of some quantity; sampling with replacement to estimate a confidence interval; sampling without replacement to conduct permutation tests; creating multiple datasets from known distributions to evaluate the performance of statistical procedures.

It is somewhat unusual that we pair analog and digital simulations in order to build a deep, perhaps gestalt, sense of how statistical processes work. In the first week (See Example Exercise One), students draw marbles from a bag to simulate, effectively, balls in urns and experience differences in the probability of extreme events with changes in sample size. We then simulate that same experiment digitally in R (R Core Team, 2017). We bring a sampling distribution to life by calculating the mean weights of Faux Fish (data printed on fish-shaped bits of paper) in a pond (an envelope) from a sample of $N = 3$ Faux Fish and then a sample of $N = 10$ Faux Fish. Each mean value is written on a small sticky note and used to build a histogram of the results from all groups at the front of the room (Figure 1).

The Faux Fish exercise is taken from a curriculum aimed at middle school students (Kelsey and Steel 2002) but it is also effective and powerful for advanced students and even

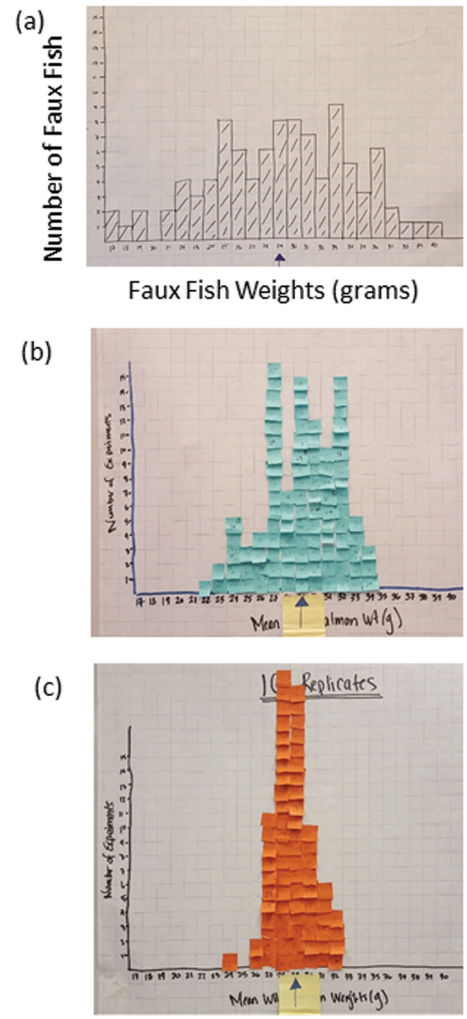


Figure 1. Visualizing sampling distributions using analog simulations: (a) histogram of weights of all Faux Fish in a pond; (b) sampling distribution of the mean of samples of size $n = 3$; (c) sampling distribution of the mean of samples of size $n = 10$. The arrows correspond to the true mean from (a). Photos by E. A. Steel.

practicing professionals. There is something magical about watching a sampling distribution materialize, just as expected, as a result of random chance. We then simulate that exact same experiment digitally. Students also conduct a permutation test by physically cutting off data labels and permuting the data values, building a null model at the front of the room by graphing class results in real time. Of course, permutation tests are then conducted with a much higher sample size in R. With this foundation in simulation thinking, students go on to use simulations to evaluate the performance of p-values under a variety of situations (see Example Exercise Two), compute statistical power, and compare the performance of testing procedures under various conditions.

P-values. Because we are proposing our course, and other courses like it, as one part of the prescription for a post $p < 0.05$ world, it is important to describe what and how we teach students specifically about p-values. In week four, we introduce the idea of a p-value without naming it. When we permute data after cutting off the labels, as described above, we build a histogram of the distribution of the difference between group means under a null hypothesis of "no difference between the group means." We then ask the students, "what percent of the

time did we see a difference between group means as extreme or more extreme than the difference we actually observed in the original data?" And, we have a short discussion about what we conclude given the observed data and the histogram of the null distribution still without naming the idea as a *p*-value. Although many students are familiar with *p*-values, this way of approaching it is new to most of them.

We then begin week five with a 20-min lecture on the history of *p*-values, starting with a definition from Karl Pearson (Pearson 1900) and eventually comparing and contrasting Ronald Fisher's testing approach (Fisher 1925) with that of Jerzy Neyman and Egon Pearson (Neyman and Pearson 1933). We remind students that the *p*-value is a random variable and explain that, under the null hypothesis, the *p*-value has a uniform distribution. We spend time considering how likely a probability of 0.05 really is. Most students have played games of chance in which rolling two die and getting two sixes is an exciting but relatively common outcome. A *p*-value of 0.05 is nearly twice as likely! Students then enjoy The Chocolate Lab (see Example Exercise Two) which should reinforce a growing understanding about how *p*-values can be misused and misunderstood. For homework, students conduct one of five simulation experiments on *p*-values. The experiments build intuition about how *p*-values behave when the null hypothesis is true; distinguish between statistical and practical significance; determine how effect size and sample size influence the probability of rejecting; estimate the probability that the null hypothesis is true when you have and when you have not rejected; and demonstrate that when you reject with low power, the observed difference in means is likely much higher than the true difference in means. Students report out the results of their experiments in the following class so that all students can learn from all five experiments. Finally, we talk extensively about how to communicate to scientists, journalists, or user groups such as farmers, the results of statistical analyses when *p*-values are small and when they are not.

Limiting our course to two-sample comparisons. There is surprisingly little time in a semester course. Teaching multiple techniques or covering a range of complex applications does not leave enough time for big picture ideas on which we want to focus our efforts; therefore, we have made thoughtful tradeoffs and exclusions. For example, we limit our formal lectures and exercises to the comparison of samples from two populations using graphs, estimation, and permutation tests. This allows us to build on knowledge of *t*-tests, which almost all our students have been introduced to in past coursework, create manageable simulations, and explore parallels across examples. Most of the skills and big ideas are relevant beyond applications that compare two samples, for example identifying the relevant measure to compare, bootstrapping a confidence interval for an effect size, or limiting the number of tests. Student projects generally involve comparisons between many groups and the students have had no problem extending the ideas to this expanded situation. Occasionally, we do touch on linear regression because many students have been exposed to it in past coursework. For example, students simulate the distribution of *r*-squared values for linear regressions when there is no effect, and students who are particularly comfortable with linear regression sometimes bring it to their group projects. In this case, they might create a scatterplot, fit a regression line, and bootstrap a confidence interval for the regression coefficient. Beyond permutation tests, we do not teach any new statistical testing procedures. Although our focus on two-sample situations could be perceived as a limitation of our course, we prefer to think of it as a feature of our course. There are lots of courses that teach statistical tools. Our course teaches students about issues and opportunities that arise when applying a single, simple tool, and by extension many other statistical tools, to generate new knowledge.

5. Example Activities to Promote Statistical Thinking

We challenge students to think deeply about their own cognitive biases, issues that arise in the application of statistical testing, types of statistical mistakes that are most common, and the cumulative effect of those mistakes. Some of our exercises include a surprise so that students are forced to face their own propensity for error. We describe two such exercises here as they may be particularly useful for other courses or workshops. Other course exercises focus on activities such as writing R code to create bootstrap confidence intervals, permutation tests, and power analyses or even collecting small datasets as an opportunity to consider the effect of study design on study conclusions.

Example Activity 1: Interpret this … and that! The perils of small samples and how strong patterns arise by chance alone.

We begin this activity on the second day of the course. Students sit in mixed groups, statistics students with students from other disciplines, undergraduates and graduate students, R programmers and the computer-phobic. Each group is given a paper to direct their discussion. Half of the groups are given the following statement "Kidney Cancer across America: A study of new diagnoses of kidney cancer in the 3141 counties of the United States reveals a remarkable pattern. The counties in which the incidence of kidney cancer are *lowest* are mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, South, and West. Think about these findings and identify three likely explanations." Unknown to the students, half of the groups are given the opposite statement in which kidney cancer are *highest* in rural counties. After the students generate ideas in their small groups, we open a discussion by making a list on the white board of "likely explanations for these observations." The students generate explanations ranging from the magic of a more natural environment to the perils of agricultural chemicals and from access to smaller hospitals to lack of access to high-end hospitals. Because the students are unaware that groups were given different information, discomfort grows. The list of answers stops making sense; time for the class runs out; and we remind students to complete the assigned reading.

Assigned reading after this exercise includes "The most dangerous equation" (Wainer 2007), "The most dangerous hospital or the most dangerous equation?" (Tu and Gilthorpe 2007), a press release for a $25M grant to create smaller schools across Oregon State (Bill and Melinda Gates Foundation Press Room 2003), and a news article from the Washington Post about the fallacies of the Gates Foundation small schools initiative and the dollars potentially wasted (Strauss 2014). In the first reading, Wainer explains the kidney cancer example with a beautiful map that shows both the highest and the lowest kidney cancer rates in rural counties. Of course, where you have fewer residents you

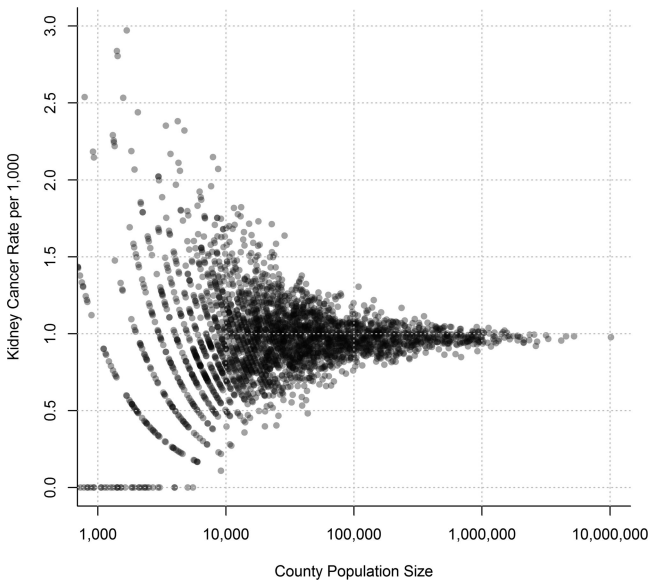


Figure 2. Simulated county cancer rates as a function of county population size using a fixed kidney cancer rate. Results demonstrate that, although the cancer rate in these simulations is the same in every county, by chance we will observed the highest and lowest county cancer rates in the counties with the fewest people.

have smaller sample sizes; you are, by chance alone, likely to see more extreme results.

When the students return to class, they are provided with a bag of marbles in which there are 40 red marbles and 40 clear marbles, our "balls in urns." Students take samples of size $N = 7$ marbles and record how many times they get an extreme event, defined as all the marbles are the same color. There are few extreme events. The exercise is repeated for samples of size $N = 4$ marbles and there are, as expected, many more extreme events. Class results are combined and we calculate the probability of the extreme event when you draw 4 marbles and the probability of the extreme event when you draw 7 marbles. What's the answer? For $n = 4$, $p = 0.125$. Not really all that unlikely. For $n = 7$, $\mathrm{p} = 0.0156$. Rare!

We then turn the discussion back to the kidney cancer rates and the effect of sample size on the probability of extreme events which, in that case, were high or low kidney cancer rates. Students learn that all groups were actually given correct information. We discuss the ease with which the human mind is fooled by pattern and refer back to several of the System 1 shortcuts described by Tversky and Kahnemann (1974) which students had read earlier. We also comment on how willingly and confidently each group had created a story to explain an observed pattern. We then simulate the pattern in R by assuming a constant disease rate and applying it to county population data from the U.S. Census Bureau (https://www.census.gov/data/tables/2016/demo/popest/counties-total.html) (Figure 2). We demonstrate to the students that, by chance alone, we nearly always see the highest and lowest disease rates in the smallest counties.

Example Activity 2: The science of bad science. How you too can fool the world into thinking almost anything.

We welcome the students to class with local chocolate, "Good news. This is healthy so enjoy!" The students, if they have done the reading, should be in agreement. We had asked them to read "Chocolate with High Cocoa Content as a Weight-Loss Accelerator" (Bohannon et al. 2015). We put up a nice slide full of happy headlines reporting on the joy of chocolate as a new weight loss tool and we pull out the bare bones of the methods for the students.

The next step is to create a null model of the experiment. What would these results be expected to look like by chance alone? Students are given guidance and snippets of R code to simulate a simplified version of this study in which the researchers compared 18 responses of a small number of participants to a diet that included chocolate versus a diet that did not. Using this null model, students simulate the frequency of a particular significant effect (e.g., weight change), the frequency of any significant effect, the frequency of a significant effect when one outlier is removed (as was done in the study), the distribution of simulated effect sizes for those datasets in which the null hypothesis was rejected (as compared to the known, simulated effect size), and the number of study subjects that might truly be necessary to detect such a small difference. Note that statistical power is not formally introduced until the following week.

We then read the article "I fooled millions into thinking chocolate helps weight loss. Here's how," a blog post by lead author, John Bohannon (Bohannon 2015). "The study was 100 percent authentic. My colleagues and I recruited actual human subjects in Germany. We ran an actual clinical trial, with subjects randomly assigned to different diet regimes. And the statistically significant benefits of chocolate that we reported are based on the actual data … It was terrible science. The results are meaningless, and the health claims that the media blasted out to millions of people around the world are utterly unfounded. Here's how we did it …"

Ensuing class discussions are fun. A few of the more advanced students were a little skeptical of the whole process and knew intuitively where this was going but few had understood the details of how it works, from the perils of the statistical process to entrancing the media. We explain that any one particular statistical test worked correctly and the odds of observing any one particular outcome, for example, a loss of weight, was just 0.05. But the overall odds of finding something "surprising" is very high when there are multiple tests and higher still when you make decisions such as systematically deleting a type of outlier.

This exercise, coming at the center of the course, reflects back on earlier exercises in which students competed in teams to identify neat patterns in data (which turned out to be randomly generated) and were quickly willing to oversell their "findings" and it opens up deep topics about the lifecycle of science, how the peer review process really works, issues embedded in working with the media, and the importance of clear scientific communication. As well, the exercise builds intuition about the value of a null model and it arms students with simulation skills (and sample code) for evaluating the behavior of various testing procedures as part of their final project.

6. Conclusion

A course in statistical thinking that considers the full domain of statistics is fundamentally different from a traditional statistics course. First, the errors of concern in statistics courses are often errors in calculation, which are only one of the many kinds of statistical mistakes that can lead to poor science and bad

decision-making. A large proportion of statistical mistakes result from incorrect logic or interpretation despite correct numerical calculations (Steel et al. 2013; Leek and Peng 2015). Second, the course must weave in content not traditionally taught in statistics departments. Few courses train students to understand the role of statistics in the scientific process. We argue that, if more scientists and statisticians had a clearer understanding of these interactions, there would be fewer flaws in how the scientific method is applied and we would be taking a big step in moving beyond the "bright line rule" of $p < 0.05$.

To think statistically is to think in distributions and probabilities and also to understand the role of statistical analysis within the greater machinery of generating scientific knowledge. A course in statistical thinking can use simulations to understand the role of chance in creating patterns, the importance of randomness in making inference from a sample to a population, the multiple ways of quantifying uncertainty, and the host of potential errors that can occur when conducting statistical tests. Such a course also needs to speed up the process of building statistical wisdom about how things can go wrong in the integration of mathematical and nonmathematical knowledge. Understanding how human fallacies make particular statistical errors very likely can be a foundation of such wisdom. Activities that surprise students or lead them to question their understanding of statistics provide benchmark lessons that students are likely to refer to as practicing statisticians and scientists. Teaching statistical thinking and teaching across the full domain of statistical content is challenging. It requires that we share success stories, great teaching examples, teaching resources, and teaching mistakes. A shared library of resources, for example, would be a great help.

In teaching the course, we have learned a good deal about how to develop deeper statistical skills. First, students need to understand how pervasive problems in statistical intuition are and they need to be convinced that they too make predictable errors. We achieve this by a precourse survey that includes a few questions even trained statisticians often get wrong and a few carefully structured trick examples as described above. Second, students need a gut-level understanding of statistical topics, and this gut-level understanding can best be achieved through simulation. In order to understand sampling distributions, they need to take samples in many ways and observe the outcomes. In order to group how p-values do and don't work, students need to experience for themselves significant results that are generated from noise. Third, students need both positive and negative examples. In our first year, we had fun presenting dozens of flawed papers, only to realize that we had left students without confidence in the best practices we were promoting. Our short list of best practices (Table 3) has helped students envision how to make improvements in the system. Fourth, students need real datasets of which to ask real-world questions (Willet and Singer 1992; Neuman et al. 2013). Only through this process can students struggle with all the steps in the scientific lifecycle, from how to refine a question and how to best summarize data in a way that answers it, to how to make an informative graph and when to stop searching for patterns. Fifth, we limit lecturing through increased use of group discussions and hands-on activities. Thinking skills need to be practiced, not just preached.

Courses dedicated to statistical thinking are essential and fun, but there are many other opportunities to inject statistical thinking and wisdom into the education and minds of future lawyers, politicians, parents, investors, and even data scientists. There are engaging statistics courses that bring statistical thinking into daily life, such as Stat 105 at Harvard University, "Real Life Statistics: Your Chance for Happiness (or Misery)", that arranges course content into real-life applications (Meng 2009). As well, there are myriad opportunities for speaking on these topics beyond traditional statistics curriculum or statistics departments. Big ideas in statistical thinking can also be presented as single lectures in topical undergraduate or graduate courses, as guest presentations in your local high school, and even anecdotally in talks presented at professional societies.

Teaching statistical thinking and explaining the integration of statistics in the scientific process is an essential part of repairing our scientific machinery. It doesn't stop with one new course. If we had the opportunity to teach for an additional quarter, we could design and add activities that add similar depth to linear regression and we would dedicate more time to principles of sampling design (what works and what can go wrong), the art of reviewing science papers (formally and also informally), and science journalism. Workshops for a range of practicing professionals and even for the general public would also be incredibly valuable: "Five statistical red flags: A workshop for managers who rely on the best available science to make decisions" or "Better living through statistics: A whirlwind tour of how big ideas in the quantitative world can help you make better decisions on everything from whether to buy travel insurance to whether to get a mammogram." Through increased promotion of these ideas beyond academia, future physicians might be better equipped to evaluate new research, future judges might be less susceptible to the negative effect fallacy (Enos et al. 2017; Fowler 2017), and future charitable foundation leaders might be more skeptical of interesting patterns (Bill and Melinda Gates Foundation Press Room 2003). Formally and informally, teaching statistical thinking is, perhaps, our biggest, best, and easiest opportunity to make a positive impact.

References

American Institute of Biological Sciences (AIBS) Council of Member Society and Organizations Meeting (2015), *Leadership in Biology an AIBS Initiative: Addressing Biological Informatics Workforce Needs*, available at *https://www.aibs.org/public-policy/resources/AIBS_2015_Council_Report.pdf*. [393]

American Statistical Association (ASA) (2016), *Ethical Guidelines for Statistical Practice*, available at *http://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf*. [393]

Baker, M. (2016), "Is There a Reproducibility Crisis in Science?" *Nature*, 533, 452–454. [392]

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., Cesarini D. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [392]

Bill and Melinda Gates Foundation Press Room (2003), *Stronger, Smaller High Schools to Help Thousands of Oregon Students*, available at *https://www.gatesfoundation.org/Media-Center/Press-Releases/2003/04/Oregon-Small-Schools-Initiative*. [398,400]

Brown, E. N., and Kass R. (2009), "What is Statistics?" *The American Statistician*, 63, 105–110. [392]

Bohannon, J., Koch, D., Homm, P., and Driehaus, A. (2015), "Chocolate with High Cocoa Content as a Weight-Loss Accelerator," *Global Journal of Medical Research*, 15, 9–14. [399]

Bohannon, J. (2015), *I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How*, available at *https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800*. [399]

Bureau of Labor Statistics (2017), *Occupational Outlook Handbook, Mathematicians and Statisticians*, available at *http://www.bls.gov/ooh/math/statisticians.htm*. [393]

Cobb, G. W. (1992), "Report of the Joint ASA/MAA Committee on Undergraduate Statistics, in *Proceedings of the Section on Statistical Education*, Alexandria, VA: American Statistical Association, pp. 281–283. [393]

Diaconis, P., and Mosteller, F. (1989), "Methods for Studying Coincidences," *Journal of the American Statistical Association*, 84, 853–861. [394]

Enos R. D., Fowler, A., and Havasy, C. S. (2017), "The Negative Effect Fallacy: A Case Study of Incorrect Statistical Reasoning by Federal Courts," *Journal of Empirical Legal Studies,* 14, 618–547. [400]

Fisher, R. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd. [398]

Fowler, A. (2017), "Chief Justice Roberts and Other Judges Have a Hard Time with Statistics. That's a Real Problem," *The Washington Post*, available at *https://www.washingtonpost.com/news/monkey-cage/wp/2017/10/31/chief-justice-roberts-and-other-judges-have-a-hard-time-with-statistics-thats-a-real-problem/?utm_term=.28de5b8f0a3e*. [400]

Franklin, C., Kader, G., Mewborn, D., Moreno. J., Peck, R., Perry, M., and Scheaffer, R. (2007), *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework*, available at *www.amstat.org/education/gaise*. [393,397]

GAISE College Report ASA Revision Committee (2016), *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*, available at *http://www.amstat.org/education/gaise*. [393,396]

Gelman, A., and Loken, E. (2013), "The Garden of Forking Paths: Why Multiple Comparisons Can Be A Problem, Even When There Is No 'Fishing Expedition' Or 'P-Hacking' and the Research Hypothesis was Posited Ahead of Time," Department of Statistics, *Columbia University*, available at *http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf*. [397]

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110, 19313–19317. [392]

Kelsey, K., and Steel, A. (2002), *The Truth About Science: A Curriculum for Developing Young Scientists*, Arlington, VA: National Science Teachers Association Press, 226 pp. [397]

Kuzmak, S. (2016), Mapping Knowledge for Probability and Statistics Application: Mathematical and Non-Mathematical, in *Promoting Understanding Of Statistics About Society*, *Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE)*, July 2016, Berlin, Germany, ed. J. Engel, available at *https://iase-web.org/documents/papers/rt2016/Kuzmak.pdf*. [394,395]

Leek, J. T., and Peng, R. D. (2015), "Statistics: P Values Are Just The Tip Of The Iceberg," *Nature*, 520, 612. [392,400]

Lock R. H., Lock, P. F., Morgan, K. L., Lock, E. F., and Lock D. F. (2013), *Statistics: Unlocking the Power of Data*, New York: Wiley. [397]

Lund, S., Manyika, J., Nyquist, S., Mendonca, and Ramaswamy, S. L. (2013), *Game Changers: Five Opportunities for US Growth and Renewal*, available at *http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Growth/US%20game%20changers/MGI_Game_changers_US_growth_and_renewal_Full_report.ashx*. [393]

Mallows, C. (1998), "The Zeroth Problem," *The American Statistician*, 52, 1–9. [396]

Martinson B. C. (2017), "Give Researchers A Lifetime Word Limit," *Nature News*, 550, 303. [392]

Martonosi, S. E., and Williams, T. D. (2016), "A Survey of Statistical Capstone Courses," *Journal of Statistics Education*, 24, 127–135. [397]

Meng X. L. (2009), "Statistics: Your Chance For Happiness (Or Misery)," *The Harvard University Undergraduate Research Journal*, 2, 21–7. [393,396,400]

Moore, D. S. (1998), "Statistics Among The Liberal Arts," *Journal of the American Statistical Association*, 93, 1253–1259. [393]

Neumann, D. L., Hood, M., and Neumann, M. M. (2013), "Using Real-Life Data When Teaching Statistics: Student Perceptions of this Strategy in an Introductory Statistics Course," *Statistics Education Research Journal*, 12, 59–70. [400]

Neyman, J., and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Biometrika*, 20, 175–240. [398]

Pearson, K. (1900), "On The Criterion That A Given System Of Deviations From The Probable In The Case Of A Correlated System Of Variables Is Such That It Can Be Reasonably Supposed To Have Arisen From Random Sampling," *Philosophical Magazine Series 5*, 50, 157–175. [398]

Pierson, S. (2013), "Growing Numbers of Stats Degrees," *AmStat News May*, available at *http://magazine.amstat.org/blog/2013/05/01/stats-degrees/*. [393]

R Core team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [397]

Rosenthal, R. (1979), "The File Drawer Problem And Tolerance For Null Results," *Psychological Bulletin*, 86, 638–641. [395,397]

Steel, E. A., Kennedy, M. C., Cunningham, P. G., and Stanovick, J. S. (2013), "Applied statistics in ecology: common pitfalls and simple solutions," *Ecosphere*, 4, 1–15. [400]

Sterne, J. A. C., and Smith, G. D. (2001), "Sifting the Evidence—What's Wrong With Significance Tests?" *Physical Therapy*, 81, 1464–1469. [392]

Strauss, V. (2014), "How Much Bill Gate's Disappointing Small-Schools Effort Really Cost," *The Washington Post*, available at *https://www.washingtonpost.com/news/answer-sheet/wp/2014/06/09/how-much-bill-gatess-disappointing-small-schools-effort-really-cost/*. [398]

Sutherland, W. J., Spiegelhalter, D., and Burgman, M. A. (2013), "Twenty Tips for Interpreting Scientific Claims," *Nature*, 503, 335–337. [394]

Tintle N., Chance B. L., Cobb G. W., Rossman A. J., Roy S., Swanson T., and VanderStoep J. (2015), *Introduction to Statistical Investigations*, New York: Wiley. [397]

Tu, Y. K., and Gilthorpe, M. S. (2007), "The Most Dangerous Hospital or the Most Dangerous Equation?" *BMC Health Services Research*, 7, 185. [398]

Tversky, A., and Kahneman, D. (1974), "Judgment Under Uncertainty: Heuristics And Biases," *Science*, 185, 1124–1131. [394,395,399]

Wainer, H. (2007), "The Most Dangerous Equation," *American Scientist*, 95, 249–256. [398]

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 [392]

Willet, J. B.. and Singer, J. D. (1992), "Providing a Statistical "Model": Teaching Applied Statistics using Real-World Data," in *Statistics for the Twenty-first Century*, eds. E. Gordon, and S. Gordon, Washington, DC: Mathematical Association of America, pp. 83–98. [400]