Rejecting Statistical Significance Tests: Defanging the Arguments

D. G. Mayo
Philosophy Department, Virginia Tech, 235 Major Williams Hall, Blacksburg VA 24060

Abstract
I critically analyze three groups of arguments for rejecting statistical significance tests (don't say 'significance', don't use P-value thresholds), as espoused in the 2019 Editorial of *The American Statistician* (Wasserstein, Schirm and Lazar 2019). The strongest argument supposes that banning P-value thresholds would diminish P-hacking and data dredging. I argue that it is the opposite. In a world without thresholds, it would be harder to hold accountable those who fail to meet a predesignated threshold by dint of data dredging. Forgoing predesignated thresholds obstructs error control. If an account cannot say about any outcomes that they will not count as evidence for a claim—if all thresholds are abandoned—then there is no a test of that claim. Giving up on tests means forgoing statistical falsification. The second group of arguments constitutes a series of strawperson fallacies in which statistical significance tests are too readily identified with classic abuses of tests. The logical principle of charity is violated. The third group rests on implicit arguments. The first in this group presupposes, without argument, a different philosophy of statistics from the one underlying statistical significance tests; the second group—appeals to popularity and fear—only exacerbate the 'perverse' incentives underlying today's replication crisis.

Key Words: Fisher, Neyman and Pearson, replication crisis, statistical significance tests, strawperson fallacy, psychological appeals, 2016 ASA Statement on P-values

1. Introduction and Background

Today's crisis of replication gives a new urgency to critically appraising proposed statistical reforms intended to ameliorate the situation. Many are welcome, such as preregistration, testing by replication, and encouraging a move away from cookbook uses of statistical methods. Others are radical

and might inadvertently obstruct practices known to improve on replication. The problem is one of *evidence policy*, that is, it concerns policies regarding evidence and inference. Problems of evidence policy call for a mix of statistical and philosophical considerations, and while I am not a statistician but a philosopher of science, logic, and statistics, I hope to add some useful reflections on the problem that confronts us today.

In 2016 the American Statistical Association (ASA) issued a statement on P-values, intended to highlight classic misinterpretations and abuses.

> The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. …. much confusion and even doubt about the validity of science is arising. (Wasserstein and Lazar 2016, p. 129)

The statement itself grew out of meetings and discussions with over two dozen others, and was specifically approved by the ASA board. The six principles it offers are largely rehearsals of fallacious interpretations to avoid. In a nutshell: P-values are not direct measures of posterior probabilities, population effect sizes, or substantive importance, and can be invalidated by biasing selection effects (e.g., cherry picking, P-hacking, multiple testing). The one positive principle is the first: "*P*-values can indicate how incompatible the data are with a specified statistical model" (ibid., p. 131).

The authors of the editorial that introduces the 2016 ASA Statement, Wasserstein and Lazar, assure us that "Nothing in the ASA statement is new" (p. 130). It is merely a "statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value" (p. 131). Thus, it came as a surprise, at least to this outsider's ears, to hear the authors of the 2016 Statement, along with a third co-author (Schirm), declare in March 2019 that: "The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of 'statistical significance' be abandoned" (Wasserstein, Schirm and Lazar 2019, p. 2, hereafter, WSL 2019).

The 2019 Editorial announces: "We take that step here….[I]t is time to stop using the term 'statistically significant' entirely. …[S]tatistically significant –don't say it and don't use it" (WSL 2019, p. 2). Not just outsiders to statistics were surprised. To insiders as well, the 2019 Editorial was sufficiently perplexing for the then ASA President, Karen Kafadar, to call for a New ASA Task Force on Significance Tests and Replicability.

> Many of you have written of instances in which authors and journal editors—and even some ASA members—have mistakenly assumed this

> editorial represented ASA policy. The mistake is understandable: The editorial was co-authored by an official of the ASA.
>
> … To address these issues, I hope to establish a working group that will prepare a thoughtful and concise piece … without leaving the impression that *p*-values and hypothesis tests…have no role in 'good statistical practice'. (K. Kafadar, President's Corner, 2019, p. 4)

This was a key impetus for the JSM panel discussion from which the current paper derives ("P-values and 'Statistical Significance': Deconstructing the Arguments"). Kafadar deserves enormous credit for creating the new task force.[1] Although the new task force's report, submitted shortly before the JSM 2020 meeting, has not been disclosed, Kadar's presentation noted that one of its recommendations is that there be a "disclaimer on all publications, articles, editorials, ... authored by ASA Staff".[2] In this case, a disclaimer would have noted that the 2019 Editorial is not ASA policy. Still, given that its authors include ASA officials, it has a great deal of impact.

We should indeed move away from unthinking and rigid uses of thresholds—not just with significance levels, but also with confidence levels and other quantities. No single statistical quantity from any school, by itself, is an adequate measure of evidence, for any of the many disparate meanings of "evidence" one might adduce. Thus, it is no special indictment of P-values that they fail to supply such a measure. We agree as well that the actual P-value should be reported, as all the founders of tests recommended (see Mayo 2018, Excursion 3 Tour II). But the 2019 Editorial goes much further. In its view: Prespecified P-value thresholds should not be used at all in interpreting results. In other words, the position advanced by the 2019 Editorial, "reject statistical significance", is not just a word ban but a gatekeeper ban. For example, in order to comply with its recommendations, the FDA would have to end its "long established drug review procedures that involve comparing *p*-values to significance thresholds for Phase III drug trials" as the authors admit (p. 10).

Kafadar is right to see the 2019 Editorial as challenging the overall use of hypothesis tests, even though it is not banning P-values. Although P-values can be used as descriptive measures, rather than as tests, when we

---

[1] Linda Young, (Co-Chair), Xuming He, (Co-Chair) Yoav Benjamini, Dick De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry Graubard, Xiao-Li Meng, Vijay Nair, Nancy Reid, Stephen Stigler, Stephen Vardeman, Chris Wikle, Tommy Wright, Karen Kafadar, Ex-officio. (Kafadar 2020)

[2] Kafadar, K., "P-values: Assumptions, Replicability, 'Significance'," slides given in the Contributed Panel: *P-Values and "Statistical Significance": Deconstructing the Arguments* at the (virtual) JSM 2020. (August 6, 2020).

wish to employ them as tests, we require thresholds. Ideally there are several P-value benchmarks, but even that is foreclosed if we take seriously their view: "[T]he problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups…" (WSL 2019, p. 2).

The March 2019 Editorial (WSL 2019) also includes a detailed introduction to a special issue of *The American Statistician* ("Moving to a World beyond $p < 0.05$"). The position that I will discuss, reject statistical significance, ("don't say 'significance', don't use P-value thresholds"), is outlined largely in the first two sections of the 2019 Editorial. What are the arguments given for the leap from the reasonable principles of the 2016 ASA Statement to the dramatic "reject statistical significance" position? Do they stand up to principles for good argumentation?

2. Statistical Significance Tests

Statistical significance tests are a small part of what must be understood as a piecemeal approach, providing "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, p. 1033). These may be called *error probabilities.* The one piece addressed by statistical significance tests concerns mistaking an observed effect, difference, or association that is due to ordinary or random variability as a genuine or systematic effect. Any methods proposed as substitutes must show they can perform this task. Accounts that employ error probabilities to control and assess the capability of a method to avoid error, I call *error statistical*. This umbrella includes simple Fisherian tests, and Neyman-Pearson (N-P) formulations of hypotheses tests.

> [The significance test arises] to test the conformity of the particular data under analysis with [a statistical hypothesis] H_0 in some respect to be specified. To do this we find a function $d = d(y)$ of the data, to be called the test statistic, such that
>
> - the larger the value of d the more inconsistent are the data with H_0;
>
> …[We define the] p-value corresponding to any d as
>
> $$p = p(d) = P(D \geq d; H_0).$$
>
> (Mayo and Cox 2006, p. 81, substituting d for t)

I recommend this reading, in relation to a given test T: A P-value is the probability test T would have given rise to a result more incompatible with

H_0 than d is, were the results due to background or chance variability (as described in H_0). It is a counterfactual claim. The probability is accorded to the overall method of testing. It is important to see that in computing the P-value under the assumption H_0, hypothesis H_0 serves only as an *implicationary* assumption—it is assumed solely for drawing out the probabilistic implications for purposes of testing. This is no different than ordinary tests, even if they are non-probabilistic. Clearly, if even larger differences than d are frequently brought about by chance variability alone (P-value is not small), the data are not evidence of incompatibility with H_0. Requiring a small P-value before inferring an indication of a genuine incompatibility or discrepancy from H_0 controls the probability of a *Type I error*: erroneously finding evidence against H_0.

> [The p value] is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0. (Cox & Hinkley 1974, p. 66)

But the justification for testing reasoning is not merely a concern to control errors in the long-run use of tests. It is because of what the evaluation means for the test at hand. If the P-value is low, then there's a high probability that a less extreme value of D would have occurred, were H_0 in fact adequate, 1 – P. Since the test very probably would have produced a result more compatible with H_0, were we dealing with chance variability alone, a low P-value indicates incompatibility with H_0.

Trouble only begins if one moves from such an indication to inferring evidence of a substantive scientific hypothesis H^* which might explain the effect. The probability of inferring H^* erroneously is not bound by the small P-value, even where underlying statistical assumptions hold. N-P tests are explicit in avoiding such fallacies by considering the alternative statistical hypothesis H_1 where H_0 and H_1 together exhaust the possibilities for the test. N-P tests are specified to also ensure control of the probability of a *Type II error*: erroneously failing to find evidence against the null hypothesis. Equivalently, N-P tests are specified to have reasonably high *power* to detect alternatives of interest. Notice that the concept of power turns on there being a threshold value for test statistic D beyond which we infer there is evidence against H_0. N-P called H_0 the *test hypothesis*, rather than Fisher's *null hypothesis*, a more suitable term, less open to misinterpretation.

We should not confuse prespecifying minimal thresholds in each test, with fixing a value to habitually use, especially without tying it to theoretical and empirical background. N-P tests called for the practitioner to balance error probabilities according to context, not rigidly fix a value like 0.05.

As Erich Lehmann, a Neyman student and leading spokesperson on N-P statistics, observes:

> Both Neyman–Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. (Lehmann 1993, p. 1248)

David Cox (2019) observes that, in practice, the founders altered their stances:

> In his later, more applied work, Neyman … used *p*-values flexibly, whereas Fisher paradoxically, in some at least of his work, used 5% significance rather rigidly, although he recognized the arbitrariness of that specific choice. (p. 6)

Moreover, we should move away from being hamstrung by what the founders thought, or by what some people assume they thought. We should reformulate and reinterpret statistical tests to grapple with the replication problems we now face.

3. Does Abandoning Significance Tests Block Biased Selection?

The sources of irreplication are not mysterious: in many fields, latitude in collecting and interpreting data makes it too easy to dredge up impressive looking findings even when spurious. The low P-value initially found is not found when an independent group seeks to replicate the results. Significance testers have an argument to block data dredging–it wrecks the error probability guarantees of tests. Aware of the problem, Neyman and Pearson insisted that the criterion used to test a statistical hypothesis be predesignated.

> To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the hypothesis [H_0] is true. (Pearson and Chandra Sekar 1936, 127)

3.1 Data Dredging Is Not Blocked

It is important to see that even agreement on sources of poor replication may lead to opposing standpoints on the importance of P-value thresholds in interpreting results. This helps us to understand a key source of disagreement about whether to remove the use of P-value thresholds. To be fair, perhaps the strongest argument is the supposition that without a P-value threshold, the eager researcher would lose the (perverse) incentive to data dredge, multiple test and P-hack when confronted with a large,

statistically insignificant P-value. Even without the word 'significance', eager researchers could not present the large (insignificant) P-value as indicating a genuine effect—and they will still want to show this. For to do so would be to say something nonsensical. It would be to say:

> Even though more extreme results than ours would frequently occur by random variability alone, I maintain our data provide evidence they are not due to chance variability.

In short, researchers would still need to report a reasonably small P-value, to claim an effect. But this is to use a threshold. Any eager researchers incentivized to data dredge before, would be that much more incentivized to dredge in a world without statistical significance level thresholds (although perhaps not quite as far). That is because, in a world without thresholds, it would be hard for a critic to hold them accountable for reporting a *nominally* small P-value attained through ransacking the data, outcome-switching and the like. (See Mayo 2019). According to the 2019 Editorial, "whether a *p*-value passes any arbitrary threshold should not be considered at all" in interpreting data (WSL 2019, p. 2).

John Ioannidis is right to charge that "fields that obstinately resist refutation can hide behind the abolition of statistical significance but risk becoming self-ostracized from the remit of science" (2019, p. 2068). However, if the ASA executive director gives a green light to rejecting statistical significance, it might be easy to escape opprobrium.

> By removing the prespecified significance level, typically 5%, interpretation could become completely arbitrary. It will also not stop data-dredging, selective reporting, or the numerous other ways in which data analytic strategies can result in grossly misleading conclusions. (Cook et al. 2019, p. 224)

Already we see the 2016 ASA Statement used as grounds to free researchers from culpability for failing to report or adjust for data dredging and multiple testing. One case even reached the Supreme Court of the United States. In 2009, Scott Harkonen was found guilty of issuing a misleading press report on results of a drug advanced by the company of which he was CEO. Downplaying the high P-value on the primary endpoint (and 10 secondary endpoints), he reported statistically significant drug benefits had been shown, without mentioning this referred only to a subgroup he identified from ransacking the unblinded data. Nevertheless, Harkonen and his defenders argued that "the conclusions from the ASA Principles are the opposite of the government's" conclusion that his construal of the data was misleading (*Harkonen v. United States*, 2018, p. 16). The theory on which the client's guilt rests—statistical significance

tests—is declared to have been "shown false" by the 2016 Statement. (See Mayo 2020.)

3.2 No Threshold, No Error Control

The 2016 ASA statement warned (Principle 4) that data dredging "renders the reported *p*-values essentially uninterpretable":

> Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. … Valid scientific conclusions based on *p*-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including *p*-values) were selected for reporting. (pp. 131-32)

Two other contributions to our panel (by S. Young, and Y. Ritov) focus on how P-values are invalidated by multiple testing.

However, the same P-hacked hypothesis can occur in Bayes factors, likelihood ratios, and a number of alternative methods. While the 2019 Editorial mentions other "don'ts" from the 2016 Statement, albeit in much stronger forms, there is no mention of Principle 4. As Yoav Benjamini (2016) emphasizes, selection effects are a problem affecting all statistical methods, especially in today's uses of Big Data.

One paper within the special issue introduced by the 2019 Editorial takes this up:

> 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!' (Hurlbert et al. 2019, p. 354)

If they are correct, then losing thresholds is to lose the intrinsic property enjoyed by statistical significance tests. It is scarcely a happy answer for those seeking to discriminate real from spurious effects.

The central reason that researchers look to controlled trials of treatments for Covid-19 is to sustain error control. Fisher emphasizes how statistical significance is tied to randomized controlled trials:

> the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged. (Fisher 1935, 21)

Cook et al., in the journal *Clinical Trials*, respond to the recommendation to reject statistical significance tests warning that "it would be a mistake to allow the tail to wag the dog by being overly influenced by flawed statistical inferences that commonly occur in less carefully planned settings", in contrast to the "protection of scientific validity provided by the randomisation of the interventions being compared" (p. 223). The authors of the 2019 Editorial do not restrict their recommendations, but call for rejecting statistical significance tests across all science.

The New England Journal of Medicine (NEJM) refuses the 2019 Editorial's call to reject statistical significance, specifically emphasizing that a central premise on which their revisions are based is "the use of statistical thresholds for claiming an effect or association should be limited to analyses for which the analysis plan outlined a method for controlling type I error" (Harrington et al. 2019, p. 286).[3]

> A well-designed randomized or observational study will have a primary hypothesis and a prespecified method of analysis, and the significance level from that analysis is a reliable indicator of the extent to which the observed data contradict a null hypothesis of no association between an intervention or an exposure and a response. Clinicians and regulatory agencies must make decisions about which treatment to use or to allow to be marketed, and P values interpreted by reliably calculated thresholds subjected to appropriate adjustments [for multiple trials] have a role in those decisions. (Harrington et al. 2019, p. 286)

3.3 No Thresholds, No Tests

A common fallacy is to suppose that because we have a continuum, and cannot point to a value where there is a conversion from one state to another, we cannot distinguish points at the extremes. It may be called the *fallacy of the beard*. (There is no one point at which a man goes from having, to not having, a beard.) But we *can* distinguish results readily produced by random variability from cases where there is evidence of incompatibility with the chance variability hypothesis.

Kafadar's JSM presentation listed numerous thresholds—bone density, blood pressure, prostate specific antigen (PSA), asking. "Is anyone complaining about these thresholds?" The use of thresholds for the categories of Covid-19 risk faced by U.S. counties—green, yellow, orange, red, based on the number of new daily cases, provide broad guidance for control efforts. It would be a fallacy to claim that no useful distinctions can be made because there is no substantial difference

[3] That the NEJM was asked to revise their guidelines taking into account both the 2016 ASA Statement and the 2019 Editorial underscores the need for a disclaimer.

between, say, the highest number of cases per 100,000 in yellow (9) and the lowest number in orange (10).

Neyman and Pearson originally had a space of outcomes to be construed as an undecided range, and there is nothing to stop us from viewing tests that way. Yet the 2019 Editorial, recall, rejects "any number of" categories. Taken strictly, this would preclude distinguishing the interpretation of results at their particular P-values. This is after all, to classify results according to the P-value reached.

Confidence interval (CI) estimates are often advanced as replacements for statistical significance tests, yet its advocates standardly use 95% confidence levels. An objection to taking a difference that reaches P-value 0.025 as evidence of a discrepancy from the null hypothesis, would also be an objection to taking it as evidence the parameter exceeds the lower 0.025 CI bound. They are identical, insofar as CIs retain their duality with tests (likewise for the upper limit). A better alternative would be to report several intervals at different levels.[4]

Nor could Bayes factor thresholds be used, as they often are, to test a null against an alternative. It is not clear how any statistical tests survive. If you cannot say about any results, ahead of time, they will not be allowed to count in favor of a claim, then you do not have a test of it. No tests, no falsification. We are not told what happens to the use of significance tests to check if statistical model assumptions hold approximately, or not–essential across methodologies. As George Box, a Bayesian, remarks, "diagnostic checks and tests of fit … require frequentist theory significance tests for their formal justification" (1983, p. 57). What's the point of insisting on replications if at no point can you say, the effect has failed to replicate?

4. 'Statistical Significance' is Meaningless, and Other Strawperson Fallacies

A second class of arguments points to misinterpretations and abuses of statistical significance, even without data dredging. Appraising such arguments requires being on the lookout for *strawperson fallacies*.

A strawperson fallacy argues against a view by distorting or exaggerating it in order to make it easy to knock down. The pattern is this:

[4] This is done in confidence distributions (Xie and Singh 2013), and in what I call a severity assessment. With the latter, the evidential warrant associated with different points in any interval are distinguished (Mayo 1996, 2018; Mayo and Cox 2006; Mayo and Spanos 2006).

> *Strawperson Fallacy*: The view or method I wish to reject (in this case statistical significance tests), is tantamount to something clearly problematic (fallacious interpretations of data) and must be avoided. Therefore the method I wish to reject (statistical significance tests) must be avoided.[5]

To reject statistical significance tests because they can be used very badly is itself a very bad argument.

4.1 Statistical Significance Gets Its Full Meaning in Context

The 2019 Editorial declares:

> Regardless of whether it was ever useful, a declaration of 'statistical significance' has today become meaningless.… Statistical significance was never meant to imply scientific importance. (p. 2)

Granted statistical significance should be distinguished from scientific significance. Placing "statistical" before "significance" is intended to have the diminutive effect of avoiding just such a fallacy. It is saying merely that the observed effect or difference is not readily explained by random or chance variability. That is a meaningful assertion.

As Karen Kafadar, in her JSM presentation, puts it: "p-values do not tell the whole story. But they do tell us *something". They are not meaningless.* She gives an example from Dr. Fauci's recent assessment of remdesivir in battling Covid-19:

> Dr. Fauci said the NIAID trial, called the Adaptive COVID-19 Treatment Trial, (ACTT) showed a statistically significant difference in the primary endpoint, time to recovery, between patients randomized to remdesivir and those in the placebo group. (Wehrwein 2020)

Although "there were fewer deaths in the remdesivir group, [and the P-value was small] the result did not reach statistical significance, Dr. Fauci said. Deaths were not a primary measure in the trial" (Kolata et al. 2020).

Moreover, Fauci acknowledged the effect size indicated was modest:

> 'Although a 31% improvement doesn't seem like a knockout 100%, it is a very important proof of concept because what it has proven is that a drug can block this virus,' Fauci told reporters. (Wehrwein 2020).

[5] This is an example of an *informal* fallacy: the form itself is deductively valid, but its premises are unsound.

The drug indicates a genuine but modest improvement in time to recovery (for a given group of patients), but we should be extremely cautious in taking the observed survival benefit as genuinely due to remdesivir. The trial did not provide evidence the observed survival benefit was genuinely due to remdesivir and not ordinary variability. It is wrong to suppose nuanced interpretations are not grasped or are so unusual as to render "statistical significance" meaningless. Whether one agrees with Fauci's decision to make remdesivir a standard of care—an act that goes beyond the statistical inference—the assessment was not meaningless. His expectation is to try to combine remdesivir with other anti-virals to hopefully achieve larger benefits.

R.A. Fisher was clear that we are not interested in isolated results:

> [W]e need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher 1935, p. 14)

If such statistically significant effects are produced reliably, as Fisher required, they indicate a genuine effect. This is the essence of statistical falsification in science.

Throughout the full 2019 Editorial, there are plenty of useful points about the importance of context in linking formal tools to scientific claims. These points are right-headed, but they do not rescue their arguments to reject statistical significance tests. Statistical significance tests are always intended to serve as a small piece of full-bodied inquiries. The formal notions are largely intended to clarify the properties of the tools, not as rigid rules for subsequent interpretation. "The interpretation to be attached to accepting or rejecting a hypothesis is strongly context-dependent" (Cox 2006, p. 36). Cox gives a rich taxonomy of null hypotheses that recognizes how significance tests can function as part of complex and context-dependent inquiries (see Cox 1977, Cox 2019).

Neyman and Pearson emphasized that tests should be used with "discretion and understanding" (1928, p. 58). Even in their earliest papers they say:

> it is doubtful whether the knowledge that P_z [the P-value associated with test statistic z] was really 0.03 (or 0.06) rather than 0.05, . . . would in fact ever modify our judgment . . . regarding the origin of a single sample. (Neyman and Pearson 1928, p. 27)

Consider how the discovery of the Higgs particle in 2012 moved in stages. The existence of the particle was shown by statistical significance tests (a 5 sigma effect), followed by inquiries into its properties via confidence intervals. Physicists then moved to other rounds of statistical significance tests at more substantive levels. Here, departures from null hypotheses represent ways to develop physics "beyond the standard model" (BSM). Statistical insignificance plays an important role in denying that various observed BSM anomalies are real. These serve to rule out avenues for development of BSM theories, even though physicists presume that such theories will be needed. (See Mayo 2018, Excursion 3 Tour III.)

4.2 The Revised Principles Assume Strawpersons

The 2019 Editorial opens with the suggestion that it is merely reviewing some of the 2016 principles for the uninitiated reader. However, in the service of supporting their stronger position to reject statistical significance, the principles they consider get a stronger construal.[6] Each is open to the strawperson charge. In every case, we see the same pattern. For example, 'Statistical significance *can be* used thoughtlessly' becomes the premise:

> A declaration of statistical significance is the antithesis of thoughtfulness. (WSL, p. 4)

From this they conclude we should reject statistical significance.

Here's another example. While a P-value does not quantify the indicated population effect size, it is incorrect to allege that we are not to infer anything of scientific importance based on statistical significance. Yet the 2019 Editorial declares:

> Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof). (WSL, p. 1)

Granted as well, a small P-value does not entail a substantively large incompatibility. Suddenly, it cannot even be taken to indicate the mere presence of a discrepancy from H_0.

> No *p*-value can reveal the …presence… of an association or effect. (WSL, p. 2)

Whether a word other than significance would serve better can be debated. For example, we might say that the results are statistically distinguishable

[6] I assumed these stronger variants were inadvertent at first, and I delineated ways to reformate them. These reformulations were not accepted by the authors.

from, or statistically inconsistent with, random error. But this does not rescue the above charges from being strawpersons.

The stronger stipulations in the 2019 Editorial also conflict with the one positive principle from the 2016 ASA Statement:

> **1. *P*-values can indicate how incompatible the data are with a specified statistical model.**
>
> … Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. … This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions." (p. 131)

However, an indication of how incompatible data are with a claim of the absence of a relationship *would* be an indication of the *presence* of the relationship. Likewise providing evidence against a claim of no difference between two groups *would* often be of scientific or practical importance. So, the 2019 Editorial is at odds with the first principle of the 2016 ASA Statement.

4.3 Fallacies about Fallacies of Statistically Insignificant Results

To herald the 2019 Editorial, and the special issue of *TAS*, the journal *Nature* requisitioned a commentary from Amrhein, Greenland and McShane (2019). The premise for their argument for "retiring" the concept of statistical significance is that

> a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment …). (Amrhein et al. 2019, p. 305)

The fact that it is possible to fallaciously take a statistically nonsignificant difference as *proving* the truth of a zero-effect null hypothesis, is a strawperson argument against statistical significance. Moreover, obliterating thresholds would remove the very standards we need to call out the fallacies. A rule that allowed inferring, from a statistically insignificant result, that H_0 is proved, or even well warranted, would have extremely high Type II error probabilities. The fact that these authors deal with a point null hypothesis makes it even worse.

Even where Neyman-Pearson, in their search for optimality, formulate tests as a binary classification: "reject *H*" and "do not reject *H*," Neyman made clear that the meaning of "'do not reject *H*' is 'no evidence against *H* is found'" (Neyman 1976, p. 749). He developed power, and power analysis, to block the very fallacy of non-significance considered by

Amrhein et al. (2019). The authors of an article in *Clinical Trials*, in response to Amrheim et al., emphasize this. "[I]t is important to recognise that an appropriately designed and powered clinical trial enables the investigators to potentially conclude there is 'no meaningful effect' for the principal analysis" (Cook et al. 2019, p. 224).

If the test very probably would have resulted in a statistically significant result, were a meaningful effect to exist, and yet it failed to do so, then there is an indication it is absent. A more data-dependent way of interpreting insignificant P-values is to consider the P-value distribution under various discrepancies. Consider testing a Normal mean H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$. If the test very probably would have resulted in a more impressive (smaller) P-value than observed, if $\mu = \mu_1$ (where $\mu_1 = \mu_0 + \gamma$), then the data are evidence that $\mu < \mu_1$[7] This also matches inferring that μ is less than the upper bound of the corresponding upper confidence bound, at the associated confidence level.

A final strawperson is in the warning of "the seductive certainty falsely promised by statistical significance" (WSL 2019, p. 3). The implied argument is: Statistical significance tests promise certainty, any method that promises certainty should be rejected, so statistical significance tests should be rejected. But statistical significance tests promise no such thing. This charge is especially egregious given that all error statistical inferences are qualified with error probabilities. Many other approaches simply state inferences without such a qualification.

An important principle in logical argumentation—the *principle of charity*—stipulates that an arguer not give a false or extreme (straw) reading to a view under analysis, so long as there is a plausible alternative reading available. It is not a matter of being kind, it is that to violate this principle results in a criticism being fallacious—arguing against a strawperson.

5. Two Subliminal Appeals: To Philosophy of Statistics and To Popularity

Arguments need not be explicit to be convincing. There are two types of implicit appeals that underlie the call to end statistical significance tests: the first implicitly appeals to a philosophy of statistics that differs from the one underlying statistical significance tests; the second relies on implicit psychological appeals.

[7] This is an application of a general principle put forward in Mayo and Cox 2006 to capture both Fisherian and N-P tests. We dub it the Frequentist Principle of Evidence (FEV). (See also Mayo 2018, p. 149.)

5.1 How Believable vs How Well Tested

The first reflects long-standing philosophical controversies about the very role of probability in statistical inference: Should probability enter to control the probability of serious misinterpretations of data? Or to give a comparison of degrees of belief or support about claims? Disagreements between frequentists and Bayesians have been so contentious that everyone wants to believe we are long past them. Yet these battles still simmer below the surface of criticisms of statistical significance tests, and they must be unearthed to properly appraise them. Notably, if it is assumed that statistical inference should take the form of a degree of belief in statistical hypotheses, then it might appear that the P-value has to be misinterpreted to be relevant.

Given that most practitioners find the need to use an eclectic toolbox in statistics, it is important to avoid expecting an agreement on numbers from methods evaluating different things. Hence, it is incorrect to claim a P-value is "invalid" for not matching a posterior probability or a Bayes factor based on one or another prior distribution (whether subjective, empirical, or one of the many conventional measures). Statistical significance tests are designed to avoid reliance on Bayesian priors—around which there continues to be radical disagreement—unless the parameter itself is a random variable with a frequency distribution.

Andrew Gelman holds a hybrid "falsificationist Bayesian" view:

> a philosophy that openly deviates from both objectivist and subjectivist Bayesianism, integrating Bayesian methodology with an interpretation of probability that can be seen as frequentist in a wide sense and with an error statistical approach to testing assumptions. (Gelman and Hennig 2017, p. 991)

The falsification part calls for error statistical testing. How believable a claim is differs from how well it has been tested. But there's an important difference. A claim can be probable or even known to be true while very poorly tested by the data at hand. We do not want to lose that distinction.

The bottom line is this: Regardless of your philosophy of statistics, it will not do to declare by fiat that science should reject the falsification or testing view.

> The proposals for abandoning p-values altogether often suggest adopting the exclusive use of Bayesian methods. For these proposals to be convincing, it is essential their presumed superior attributes be demonstrated without sacrificing the clear merits of the traditional framework. (Cook et al. 2019, p. 223)

5.2 Psychological Appeals

The second type of implicit argument, fallacious appeals to popularity and bandwagon effects are called psychological fallacies for good reason. They provide persuasive appeals to go along with a position, despite not being warranted by sound arguments. That a position is popular or endorsed by a powerful group is not to give an argument warranting the position. But appealing to popularity gives a *prudential* reason to go along. It is risky to stand in opposition to journal and administrative leaders at the ASA. There is also an appeal to fear, with the result that many will fear using statistical significance tests altogether. Why risk using a method that is persecuted with such zeal?

It is generally agreed that a large part of the blame for lack of replication in many fields may be traced to biases encouraged by the reward structure. On this "perverse incentives" hypothesis, the pressure to publish, to advance one's career, is so great as to seduce even researchers aware of the pitfalls of capitalizing on selection biases. That mindset makes for a highly susceptible group. When those with professional power use questionable arguments, it only reinforces any existing tendencies practitioners have to use questionable methods in their own work. Thus, the very process being used to advance a position purporting to improve on replication will actually inculcate the bad habits that lead to irreplication.

The authors of the 2019 Editorial admit there is no agreement on statistical inference: "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" (WSL p. 2). This makes it all the more curious that the 2019 Editorial comes out stridently with a highly uncharitable view of statistical significance tests, rather than see the ASA as a forum that nurtures vigorous debate of all of the methods used by ASA members. Sharing the recommendations of the new ASA Task Force on Significance Tests and Replicability will be important.

6. Conclusion

Statistical significance tests have an important role in distinguishing genuine from spurious effects. They have the intrinsic features for this task, if used correctly. They shouldn't be replaced by tools that have not been shown to have these features.

It is mistaken to suppose that banning P-value thresholds would diminish P-hacking—just the opposite. In a world without thresholds, we would be hamstrung from highlighting, critically, P-values that breach (as opposed

to uphold) preset thresholds. It would make it harder to hold accountable those who fail to meet a predesignated threshold by dint of P-hacking.

To argue we should not use them because they may be used badly is itself a bad argument, guilty of the strawperson fallacy. Moreover, the premises of those arguments are in tension with Principle 1 of the 2016 ASA Statement. To press the rejection of statistical significance tests in the 2019 Editorial, we also saw, is in tension with Principle 4 on avoiding data-dredging. Finally, we considered two implicit fallacious arguments. The first assumes a different philosophy of statistics from the one underlying statistical falsification; the second—appeal to popularity—only exacerbates the perverse incentives underlying irreplication. [8]

The 2016 ASA Statement declared itself concerned that irreplication would lead to "doubt about the validity of science". To say now that the method supplied for statistical falsification is unsound would increase those doubts. David Hand puts it this way:

> *Proposals to abandon the use of significance testing and play down the role of p-values risk implying that the statistical community accepts that those tools are unsuitable, rather than that misuse of those tools is the problem*
>
> *…the most dramatic example of a scientific discipline shooting itself in the foot.*
>
> *With consequent damage to science, public policy, industry, medicine, and everywhere that statistical tools are used – which is just about everywhere.* (David Hand 2020**)**

[8] The 6 Principles from the 2016 ASA Statement on *P*-values:

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 the 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.

Acknowledgments

I am extremely grateful to my co-panelists, Karen Kafadar, Yaacov Ritov, and Stanley Young for their inspiration and encouragement in relation to this work. I thank Jean Miller for numerous recommendations and corrections on earlier drafts.

References

Adaptive COVID-19 Treatment Trial (ACTT), Clinical Trials.gov registry, available at https://clinicaltrials.gov/ct2/show/NCT04280705?titles=ACTT&draw=2

Amrhein, V., Greenland, S., and McShane, B. (2019), "Comment: Retire Statistical Significance," *Nature*, 567, 305-308.

Benjamini, Y. (2016), "It's Not the *P*-values' Fault" comment on *Wasserstein, R. and Lazar, N. (2016), "The ASA's Statement on P-values: Context, Process and Purpose" (and supplemental materials), The American Statistician, 70(2), 129-133.*

Birnbaum, A. (1970), "Statistical Methods in Scientific Inference" (Letter to the Editor), *Nature* 225(5237), 1033.

Box, G. E. P. (1983), "An Apology for Ecumenism in Statistics," in *Scientific Inference, Data Analysis, and Robustness*, Box G, Leonard, T, Wu D, eds. London, UK: Academic Press; 1983:51-84.

Cook, J., Fergusson, D., Ford, I, Gonen, M, Kimmelman, J, Korn, E., and Begg, C. (2019), "There Is Still a Place for Significance Testing in Clinical Trials," *Clin Trials*, 2019 June 16(3), 223-224. doi: 10.1177/1740774519846504. Epub 2019 May 9. PMID: 31068002; PMCID: PMC6533134.

Cox, D. R. (1977) "The Role of Significance Tests" (with Discussion), *Scandinavian Journal of Statistics* 4, 49–70.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge: Cambridge University Press.

Cox, D. R. (2019 online/2020), "Statistical Significance" *Annual Review of Statistics and Its Application* 7(1), 1-10.

Cox, D. R. and Hinkley, D. (1974), *Theoretical Statistics*, London: Chapman and Hall LTD.

Fisher, R. A. (1935/reprinted 1990), *The Design of Experiments*, Oxford: Oxford University Press.

Fisher, R. A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference*, (ed.), Bennett, J. H. Oxford: Oxford University Press.

Gelman, A. and Hennig, C. (2017), "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society*, Series A

180(4), 967–1033.

Hand, D. (2020), "Trustworthiness of Statistical Data," recorded slide presentation for D. Mayo's 2020, LSE PH500 research seminar in *Current Controversies in Philosophy of Statistics,* available at https://wp.me/abBgTB-n4

Harkonen v. United States, No. 18– (Supreme Court of the United States, filed October 1, 2018), *Petition for a Writ of Certiorari*, available at https://errorstatistics.files.wordpress.com/2019/06/harkonenv-us-scotus-2018-petn-cert.pdf

Harrington, D., D'Agostino, R., Gatsonis, C., et al. (2019), "New Guidelines for Statistical Reporting in the *Journal*," *New England Journal of Medicine* 381(3), 285-286, (July 18, 2019).

Hurlbert, S., Levine, R. and Utts, J. (2019), "Coup de Gr ce for a Tough Old Bull: 'Statistically Significant' Expires," *The American Statistician*, 73:sup1, 352-357, DOI: 10.1080/00031305.2018.1543616

Ioannidis J. (2019), "The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance," *JAMA*, 321(21), 2067-2068.

Kafadar, K. (2019), "The Year in Review…And More to Come," President's Corner. *AMSTATNEWS*, December 2019, Issue 510, 3-4. https://magazine.amstat.org/blog/2019/12/01/kk_dec2019/.

Kafadar, K. (2020), "Task Force on Statistical Significance and Replicability Created," President's Corner. *AMSTATNEWS*, February 2020, Issue 512, 7.

Kolata, G., Baker, P. and Weiland, N. (2020), "Remdesivir Shows Modest Benefits in Coronavirus Trial," *New York Times*, (April 29, 2020). https://www.nytimes.com/2020/04/29/health/gilead-remdesivir-coronavirus.html.

Lehmann, E. (1993), "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?," *Journal of the American Statistical Association* 88(424), 1242–1249.

Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Mayo, D. G. (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge: Cambridge University Press.

Mayo, D. G. (2019), **"**P-value Thresholds: Forfeit at Your Peril," *European Journal of Clinical Investigation* 2019 49(10): e13170. (https://doi.org/10.1111/eci.13170)

Mayo, D. G. (2020), "P-Values on Trial: Selective Reporting of (Best Practice Guides Against) Selective Reporting," (2020) *Harvard Data Science Review*, 2.1.

(https://doi.org/10.1162/99608f92.e2473f6a)

Mayo, D. G., and Cox, D. R. (2006), "Frequentist Statistics as a Theory of Inductive Inference," in *The Second Erich L. Lehmann Symposium: Optimality*, ed. J. Rojo, 77-97. *Lecture Notes-Monograph Series*, Volume 49, Institute of Mathematical Statistics.

Mayo, D. G., and Spanos. A. (2006), "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction," *British Journal for the Philosophy of Science*, 57, 323–357. https://doi.org/10.1093/bjps/axl003.

Neyman, J. (1976), "Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena," *Communications in Statistics: Theory and* Methods, 5(8), 737–51.

Neyman, J. and Pearson, E. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I," *Biometrika* 20A(1/2), 175–240. Reprinted in *Joint Statistical Papers*, 1–66.

Neyman, J. and Pearson, E. (1967), *Joint Statistical Papers of J. Neyman and E. S. Pearson*, Berkeley, CA: University of California Press.

Pearson, E. (1966). *The Selected Papers of E. S. Pearson*. Berkeley, CA: University of California Press.

Pearson, E. and Chandra Sekar, C. (1936). "The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations," *Biometrika* 28 (3/4), 308–20. Reprinted 1966 in *The Selected Papers of E. S. Pearson*, pp. 118–30.

Wasserstein, R. and Lazar, N. (2016), "The ASA's Statement on p-Values: Context, Process and Purpose" (and supplemental materials), *The American Statistician*, 70(2), 129-133.

Wasserstein, R., Schirm, A. and Lazar, N. (2019), "Moving to a World Beyond '$p < 0.05$'" (Editorial), *The American Statistician* 73(S1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wehrwein, P. (2020), "Remdesivir: Fauci Thumbs Up, Lancet Study Thumbs Down," *Managed Healthcare Executive* website, available at https://www.managedhealthcareexecutive.com/view/remdesivir-fauci-thumbs-lancet-study-thumbs-down)

Xie, M. and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review* 81(1), 3–39.