Received: 8 September 2019 | Accepted: 8 September 2019

DOI: 10.1111/eci.13170

EDITORIAL

WILEY

P-value thresholds: Forfeit at your peril

A key recognition among those who write on the statistical crisis in science is that the pressure to publish attention-getting articles can incentivize researchers to produce eye-catching but inadequately scrutinized claims. We may see much the same sensationalism in broadcasting metastatistical research, especially if it takes the form of scapegoating or banning statistical significance. A lot of excitement was generated recently when Ron Wasserstein, Executive Director of the American Statistical Association (ASA), and co-editors A. Schirm and N. Lazar, updated the 2016 ASA Statement on *P*-values and statistical significance (ASA I).[1] In their 2019 interpretation, ASA I "stopped just short of recommending that declarations of 'statistical significance' be abandoned," and in their new statement (ASA II) announced: "We take that step here….'statistically significant'—don't say it and don't use it".[2] To herald the ASA II, and the special issue "Moving to a world beyond '$P < 0.05$'," the journal *Nature* requisitioned a commentary from Amrhein, Greenland and McShane "Retire statistical significance" (AGM).[3] With over 800 signatories, the commentary received the imposing title "Scientists rise up against significance tests"!

(Note: By "ASA II" I allude only to the authors' general recommendations, not their summaries of the 43 papers in the issue.)

Hardwicke and Ioannidis[4] worry that recruiting signatories on such a paper politicizes the process of evaluating a stance on scientific method, and fallaciously appeals to popularity (argumentum ad populum) "because it conflates *justification* of a belief with the *acceptance* of a belief by a given group of people." Opposing viewpoints are not given a similar forum. Fortunately, John Ioannidis[5] can come out with a note in *JAMA* challenging ASA II and AGM, but the vast majority of stakeholders in the debate go unheard. Appealing to popularity gives a *prudential* reason to go along; it is risky to stand in opposition to the hundreds who signed, not to mention, the thought 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 and fanfare?

Ioannidis points out what may not be obvious at first: it is not just a word ban but a gatekeeper ban[5]:

> Many fields of investigation … have major gaps in the ways they conduct, analyse, and report studies and lack protection from bias. Instead of trying to fix what is lacking and set better and clearer rules, one reaction is to overturn the tables and abolish any gatekeeping rules (such as removing the term *statistical significance*). However, potential for falsification is a prerequisite for science. Fields that obstinately resist refutation can hide behind the abolition of statistical significance but risk becoming self-ostracized from the remit of science.

Among the top-cited signatories who respond to their questionnaire, Hardwicke and Ioannidis[4] find a heavy representation of fields with prevalent concerns about low reproducibility. Yet "abandoning the concept of statistical significance would make claims of 'irreproducibility' difficult if not impossible to make. In our opinion this approach may give bias a free pass."

I agree and will show why.

1 | STATISTICAL SIGNIFICANCE TESTS AND THRESHOLDS

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."[6] These may be called *error probabilities*. The one piece addressed by statistical significance tests concerns mistaking a pattern or difference that is due to ordinary or random variability as a genuine or systematic effect. The test controls at small values the probability of mistakenly inferring evidence of a real effect and mistakenly failing to find such evidence. 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*. The umbrella includes simple Fisherian tests, but I allude to a Neyman-Pearson (N-P) formulation because that is where the criticisms here are mostly directed. For instance, the notion of a test's power does not exist without a threshold for "rejecting" a test hypothesis (what Fisher called the null hypothesis).

It might be assumed I would agree to "retire significance" since I often claim "the crude dichotomy of 'pass/fail' or 'significant or not' will scarcely do" and because I reformulate tests so as to "determine the magnitudes (and directions) of any statistical discrepancies warranted, and the limits to any substantive claims you may be entitled to infer from the statistical ones."[7] (Genuine effects, as Fisher insisted,[8] require not isolated small *P*-values, but a reliable method to successfully generate them.) We should not confuse prespecifying minimal thresholds in each test, which I would uphold, with fixing a value to habitually use (which I would not). N-P tests call for the practitioner to balance error probabilities according to context, not rigidly fix a value like .05. Nor does having a minimal *P*-value threshold mean we do not report the attained *P*-value: we should, and N-P agreed!

2 | THE "NO THRESHOLD" VIEW IS NOT MERELY TO NEVER USE THE S WORD AND REPORT CONTINUOUS *P*-VALUES

These two rules alone would not lead Hardwicke and Ioannidis to charge, correctly, in my judgment that "this approach may give bias a free pass." ASA II and AGM decry using any prespecified *P*-value threshold as the basis for categorizing data in some way, such as inferring that results are, or are not, evidence of a genuine effect.

- "Decisions to interpret or to publish results will not be based on statistical thresholds" (AGM).[3]
- "Whether a p-value passes any arbitrary threshold should not be considered at all" in interpreting data (ASA II).[2]

Consider how far reaching the "no threshold" view is for interpreting data. For example, according to ASA II, in order for the US Food and Drug Administration (FDA) to comply with its "no threshold" position, it does not suffice that they report continuous *P*-values and confidence intervals. 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".[2]

The New England Journal of Medicine (*NEJM*) responds[9] to the ASA call to revise their guidelines, but insists 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."[10] In the article accompanying the revised guidelines:

> 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.[10]

Specifying "thresholds that have a strong theoretical and empirical justification"[9] escapes the ASA II ruling: "Don't conclude anything about scientific …importance based on statistical significance."[2]

Although less well advertised, the "no thresholds" view also torpedoes common uses of confidence intervals and Bayes factor standards.

> [T]he problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups… . Similarly, we need to stop using confidence intervals [CIs] as another means of dichotomizing.
>
> (ASA II)[2]

AGM's "compatibility intervals" are redolent of the "consonance intervals" of Kempthorne and Folks,[11] except that the latter use many thresholds, one for each of several consonance levels. Even these would seem to violate the rule that results should not be "categorized into any number of groups." An objection to taking a difference that reaches *P*-value .025 as evidence of a discrepancy from the null, would also be an objection to taking it as evidence the parameter exceeds the lower .025 CI limit (or is "incompatible," at that level, with the parameter values below it). They are identical, insofar as CIs retain their duality with tests (likewise for the upper limit).

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. A claim has not passed a genuine test if no results are allowed to count against it. 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" (p. 57).[12]

3 | WHAT ARGUMENTS ARE GIVEN TO ACCEPT THE NO THRESHOLD VIEW?

Getting past the appeals to popularity and fear, the reasons ASA II and AGM give are that thresholds can lead to well-known fallacies, and even to some howlers more extreme than those long lampooned. Of course, it's true:

> a statistically nonsignificant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment …). Nor do statistically significant results 'prove' some other hypothesis.
>
> (AGM)[3]

It is easy to be swept up in their outrage, but the argument "significance thresholds can be used very badly, therefore remove significance thresholds" is a very bad argument. Moreover, it would remove the very standards we need to call out the fallacies. A rule that went from any nonsignificant result to inferring no effect was proved, or to take something less extreme, to inferring it is well warranted or the like, would have extremely high type II error probabilities. They deal with a point null hypothesis, which makes it even worse.

Granted, N-P theorists, in their search for optimality, formulate tests as a binary classification: "reject *H*" and "do not reject *H*," even though they initially had a region of undecidable results. But as Neyman made clear, the meaning of "'do not reject *H*' is 'no evidence against *H* is found'".[13] He developed power, and power analysis, to block the very fallacy of nonsignificance that AGM consider. There, the power for detecting parameter values in the interval is not high enough to say the nonsignificant results are evidence of the absence of discrepancies that large. (I prefer a more data-sensitive way to block the fallacy, as developed elsewhere.[7,14,15]) Finally, N-P would tell you to arrange your test hypothesis so that the type I error is the more serious (considering costs), and that alone can scotch the problem in the examples described. The ASA II warns of "the seductive certainty falsely promised by statistical significance." This warning is puzzling, however, given that all error statistical inferences are qualified with error probabilities, unlike many other approaches.

4 | GIVING DATA DREDGERS A FREE PASS

The danger of removing thresholds on grounds they could be badly used is that they are not there when you need them. Ioannidis[5] zeroes in on the problem:

> The proposal to entirely remove the barrier does not mean that scientists will not often still wish to interpret their results as showing important signals and fit preconceived notions and biases. With the gatekeeper of statistical significance, eager investigators whose analyses yield, for example, $P = .09$ have to either manipulate their statistics to get to $P < .05$ or add spin to their interpretation to suggest that results point to an important signal through an observed 'trend'. When that gatekeeper is removed, any result may be directly claimed to reflect an important signal or fit to a preexisting narrative.

As against Ioannidis' anything goes charge, it might be said that even in a world without thresholds a largish *P*-value could not be taken as evidence of a genuine effect. For to do so would be to say something nonsensical. It would be to say: Even though larger differences would frequently be expected by chance variability alone (ie, even though the *P*-value is largish), I maintain the data provide evidence they are not due to chance variability.

But such a response turns on appealing to a threshold to block it, minimally requiring the *P*-value be rather small, for example $<.1$? (It also shows why *P*-values are apt measures for the job of distinguishing random error.) Thus, our eager investigators, facing a nonsmall *P*-value, are still incentivized to manipulate their statistics. Say they ransack the data until finding a non-prespecified subgroup that provides a nominally small enough *P*-value. In a world without thresholds, we would be hamstrung from highlighting, critically, *P*-values that breach (as opposed to uphold) preset thresholds.

> [W]hether a p-value passes any arbitrary threshold should not be considered *at all* when deciding which results to present or highlight.
>
> (my emphasis, ASA II)[2]

More important than keeping a specific word is keeping a filter for error control. The 2016 ASA I warned[1] in Principle 4: "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." That is because their interpretation in terms of error control would be altered by these biasing selection effects. Dropping the requirement to meet prespecified thresholds is at odds with this ASA I principle. An unanswered question is how Principle 4 is to operate in a world with ASA II.

The *NEJM*'s revised guidelines,[9] far from agreeing to use *P*-values without error probability thresholds, will now be stricter in their use. When no method to adjust for multiplicity of inferences or controlling the type I error probability is prespecified, the report of secondary endpoints:

> should be limited to point estimates of treatment effects with 95% confidence intervals. In such cases, the Methods section should note that the widths of the intervals have not been adjusted

> for multiplicity and that the inferences drawn may not be reproducible. No P values should be reported for these analyses.

Confidence intervals severed from their dualities with tests, from which they were initially developed, lose their error probability guarantees.

5 | CONCLUSION

The ASA *P*-value project is lately careering into recommendations on which there has been little balanced discussion and much disagreement. Hardwicke and Ioannidis[4] find that more than half of the respondents deny significance should be excluded from all science, and the 43 papers in the special issue "Moving to a world beyond '$p < 0.05$'" offer a cacophony of competing reforms.

It is hard to resist the missionary zeal of masterful calls: Do you want bad science to thrive? or Do you want to ban significance? (a false dilemma). A question to raise before jumping on the bandwagon: Are they asking the most unbiased questions about the consequences of removing thresholds currently ensconced into hundreds of legal statutes and best practice manuals? This needs to be carefully considered, if the reforms intended to improve credibility of statistics are not to backfire, as they may already be doing.

ASA II is part of a large undertaking; it contains plenty of sagacious advice. Notably, the M in ATOM: Modesty.[2]

> **[B]e 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.

ASA II regards its positions "open to debate."[2] An open debate is very much needed.

ACKNOWLEDGEMENT

I would like to thank D. Hand, N. Schachtman and A. Spanos for comments and corrections on earlier drafts.

Deborah G. Mayo

Department of Philosophy, Virginia Tech, Blacksburg, VA, USA
Email: mayod@vt.edu

REFERENCES

1. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat*. 2016;70:129-133. [ASA I].
2. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". *Am Stat*. 2019;73:1-19. [ASA II].
3. Amrhein V, Greenland S, McShane B. Retire statistical significance [Scientists rise up against statistical significance]. *Nature*. 2019;567:305-307. [AGM].
4. Hardwicke T, Ioannidis J. Petitions in scientific argumentation: dissecting the request to retire statistical significance. *Eur J Clin Invest*. 2019.
5. Ioannidis J. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA*. 2019;321:2067-2068.
6. Birnbaum A. Statistical methods in scientific inference (letter to the editor). *Nature*. 1970;225(5237):1033.
7. Mayo DG. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge, UK: Cambridge University Press; 2018.
8. Fisher RA. *The Design of Experiments*. Edinburgh, UK: Oliver and Boyd; 1947.
9. *NEJM* author guidelines. https://www.nejm.org/author-center/new-manuscripts. Accessed July 19, 2019.
10. Harrington D, D'Agostino R, Gatsonis C, et al. New guidelines for statistical reporting in the Journal. *N Engl J Med*. 2019;381:285-286.
11. Kempthorne O, Folks L. *Probability, Statistics, Data Analysis*. Ames, IA: Iowa State University Press; 1971.
12. Box G. An apology for ecumenism in statistics. In: Box G, Leonard T, Wu D, eds. *Scientific Inference, Data Analysis, and Robustness*. London, UK: Academic Press; 1983:51-84.
13. Neyman J. Tests of statistical hypotheses and their use in studies of natural phenomena. *Commun Stat Theory Methods*. 1976;5(8):737-751.
14. Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *Br J Philos Sci*. 2006;57(2):323-357.
15. Mayo DG, Cox DR. Frequentist statistics as a theory of inductive inference. In: Rojo J, ed. *Optimality: The Second Erich L. Lehmann Symposium*. Lecture Notes-Monograph series. Beachwood, Ohio: Institute of Mathematical Statistics (IMS); 2006:247-275.