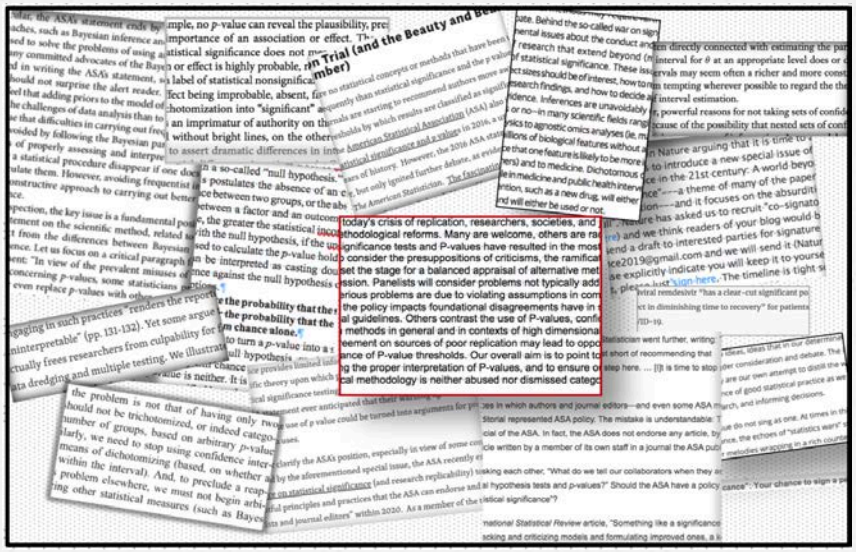
P-VALUES AND "STATISTICAL SIGNIFICANCE": DECONSTRUCTING THE ARGUMENTS



Deborah G Mayo

JSM 2020

August 6, 2020

- Mounting failures of replication give a new urgency to critically appraising proposed statistical reforms.
- Many are welcome (preregistration, replication)
- Others are radical and might actually obstruct the practices known to improve on replication.

- The problem calls for a mix of statistical, and philosophical considerations
- Although I'm not a statistician but a philosopher of science and statistics, I hope to add some useful reflections

American Statistical Association (ASA): 2016 Statement on P-values

"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." (ASA 2016 Statement on p-values, Wasserstein & Lazar)

- "Nothing in the ASA statement is new”—it declares
- It is merely a “statement clarifying several widely agreed upon principles [for interpreting] the *p*-value”.

- From this outsider's view, it was a surprise to hear the authors of the 2016 Statement together with a third guest editor (Schirm) of a special issue in March 2019 declare that:

- The 2016 Statement "stopped just short of recommending that declarations of 'statistical significance' be abandoned"

(2019) Editorial: Don't say 'significance', don't use P-value thresholds

- The new statement declares: "We take that step here….[I]t is time to stop using the term 'statistically significant' entirely."
- When we refer to the March 2019 Editorial, we are referring to this special issue of TAS ("a world beyond $p < 0.05$") where it introduced over 40 papers

- We agree we should move away from unthinking uses of thresholds (in significance or confidence levels)
- Agreed as well is that the actual P-value should be reported (as all the founders of tests recommended)
- But the 2019 Editorial goes much further

- In its view: Prespecified P-value thresholds should not be used at all in interpreting results
- The 2019 Editorial is not just a word ban but a gatekeeper ban
- For example, the “no threshold” view precludes the FDA's long-established drug review procedures, as the authors recognize

- Even if the 2019 Editorial only reflects the views of its authors, that it includes ASA officials gives it a great deal of impact

- That's one of the reasons we put together this forum

- But for insiders as well, the 2019 Editorial was sufficiently perplexing to call for a New ASA Task Force on Significance Tests and Replication

- to “prepare a …piece reflecting “good statistical practice,” without leaving the impression that *p*-values and hypothesis tests…have no role in “good statistical practice.” (K. Kafadar 2019)

- 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.

- Significance testers have an argument to block fishing and data dredging–they wreck error probability guarantees of tests

- It’s 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

- To be fair, it might be argued that by removing P-value thresholds, researchers lose an incentive to data dredge, and otherwise exploit researcher flexibility

- Even without the word ‘significance’, eager researchers can’t take the large (non-significant) P-value to indicate a genuine effect—and they will still want to show this

- To do so would be to say something non-sensical:

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

- In short, eager researchers would still need to claim a reasonably small P-value
- The eager investigators will need to "spin" their results, ransack, data dredge, outcome-switch

- In a world without predesignated thresholds, it would be hard to hold the data dredgers accountable for reporting a *nominally* small P-value:

- “whether a *p*-value passes any arbitrary threshold should not be considered at all" in interpreting data (2019 Editorial)

Principle 4 (from ASA 2016 Statement)

- The 2016 ASA statement warned (Principle 4) that data dredging “renders the reported *p*-values essentially uninterpretable”.
- However, the same p-hacked hypothesis can occur in Bayes factors, likelihood ratios, and a number of alternative methods
- The 2019 Editorial doesn't say if Principle 4 holds for these other methods

- One paper in the special issue takes this up: You might wonder how we can control Type I error rates (with multiple testing) if we abandon fixed thresholds? ”The short and happy answer is: 'you can't. And shouldn't try!’” (Hurlbert et al., 2019)

- We lose the intrinsic property enjoyed by statistical significance tests

No tests, no falsification

- The “no thresholds” view also blocks common uses of confidence intervals and Bayes factor standards as tests
- 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
- What’s the point of insisting on replications if at no point can you say, the effect has failed to replicate?

Fallacy of the Beard

- A common fallacy is to suppose that because we have a continuum, that we cannot distinguish points at the extremes
- We *can* distinguish results readily produced by random variability from cases where there is evidence of incompatibility with the chance variability hypothesis alone
- We daily see Covid treatments discriminated in this way

- Yet the 2019 Editorial rejects any number of categories

- [T]he problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups…. (2019 Editorial)

Philosophy of Statistics

- Underlying the statistical significance test wars is a long-standing philosophical controversy about the 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?

How believable vs how-well tested

- 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 today' s allegations that P-values must be misinterpreted to be relevant

How believable vs how-well tested

- Some of us think there are contexts for both, 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 don't want to lost that distinction
- Regardless of your philosophy of statistics, it will not do to declare by fiat that science should reject the falsification or testing view

- 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
- To argue we shouldn't use them because they may be used badly is itself a bad argument
- It's also wrong to suppose that banning P-value thresholds would diminish P-hacking—just the opposite

The 2016 ASA Statement declared itself concerned about irreplication leading to “doubt about the validity of science”

To say now that the method supplied for statistical falsification is unsound (“No *p*-value can reveal the …presence of an association or effect” 2019 Editorial) would increase those doubts

Statistician David Hand:

“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”

Extra: The ASA 2016 Statement's Six Principles

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.