

P-values: Assumptions, Replicability, “Significance”

Karen Kafadar

Department of Statistics

University of Virginia

`kkafadar@virginia.edu`

JSM 2020

August 6, 2020



Reporting p-values

1. Common practice vs *TAS Supplement* (March 2019)
2. Thresholds in science
3. Real issues: Assumptions, Replication, Interpretation
4. Going forward

Thresholds

- Your bone density T-score: www.nof.org
 - “Between -1.0 & -2.5 : low bone density or osteopenia”
 - “Below -2.5 : diagnosis of osteoporosis”
- Blood pressure and hypertension: www.acc.org
 - “Elevated BP: 120-129/ <80 ”
 - “hypertension stage 1 is 130–139 or 80–89 mm Hg”
 - “hypertension stage 2 is ≥ 140 or ≥ 90 mm Hg”
- Prostate Specific Antigen (PSA): www.mskcc.org
 - “Above 3ng/mL: talk with doctor about biopsy”
 - “1–3 ng/mL: see doctor for another test every 2-4 years”

Is anyone complaining about these thresholds?

“Significance”

NYTimes, 4/29/2020:

The trial sponsored by NIAID enrolled 1,063 patients who were given remdesivir or a placebo. The time to recovery averaged 11 days among those who got the drug, compared with 15 days for those who got the placebo.

There were fewer deaths in the remdesivir group, but the result did not reach statistical significance, Dr. Fauci said. Deaths were not a primary measure in the trial.

Was that sentence misleading? Or did it tell us something?

TAS Supplement, March 2019

Wasserstein, Schirm, Lazar: **Editorial**

“Moving to a World Beyond $p < 0.05$ ”

“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 [43] articles in this special issue and the broader literature, that it is time to stop using the term “statistically significant” entirely. Regardless of whether it was ever useful, a declaration of ‘statistical significance’ has today become meaningless.”

“Remdesivir for the Treatment of Covid-19—Preliminary Report”

Beigel et al. (2020), [nejm.org/doi/full/10.1056/NEJMoa2007764](https://www.nejm.org/doi/full/10.1056/NEJMoa2007764):

“1059 patients (538 remdesivir, 521 placebo) ... remdesivir [arm] had a median recovery time of 11 days (95% CI, 9–12), as compared with 15 days (95% CI, 13–19) in [placebo arm] (rate ratio for recovery, 1.32; 95% CI, 1.12–1.55; $P < 0.001$). The Kaplan-Meier estimates of mortality by 14 days were 7.1% with remdesivir and 11.9% with placebo (hazard ratio for death, 0.70; 95% CI, 0.47–1.04). Serious adverse events were reported for 114 of the 541 [remdesivir] patients ... who underwent randomization (21.1%) and 141 of the 522 [placebo] patients ...who underwent randomization (27.0%).” [95% CI, 0.7%-11.1%]

The “significant” sentence probably told us something.

We all know that:

- “ $p < 0.05$ ” has come to be synonymous with “significance”
- “ $p < 0.05$ ” does **not** mean “Result is true”
nor that “No further studies need to be conducted”
- p-values do not tell the whole story
- But they do tell us *something*

Fricker et al. (2019): What happened “one full year after the *BASP* editors banned the use of inferential statistics”?

BASP articles in 2016: “*We found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered.*”

We need replication (Cox, Tukey, Benjamini, ...)

The real problem: **The Editorial had no disclaimer.**

- Generally, no disclaimer is needed. (I do not speak for UVA.)
- First author: “*As Executive Director, Wasserstein also is an official ASA spokesperson*”

<https://www.amstat.org/ASA/About/Ronald-L-Wasserstein.aspx>

- Assumption: “The Editorial is an ASA recommendation.”
- An ASA member: “No confusion out there!”

Is there “confusion out there”?

- Invited talks: “*Time to Say Goodbye to Statistical Significance*” (SF Chapter, CDC, McGill, 2019 Fall Technical Conference*, Smith-Amherst*)
- Brian Tarran, Editor of *Significance*, August 2019: “*In fact, not long after the [ASA] issued its recommendation [1] ...*”
Reference [1] is the Editorial — **not** *ASA Recommendation*.

After 1.5 years, message does seem to be getting through:

Eric Gibson (2020), “The role of p-values in judging the strength of evidence and realistic replication expectations,” *Statistics in Biopharmaceutical Research*

(doi.org/10.1080/19466315.2020.1724560)

“The unintended consequence of these proposals has been confusion within the scientific community, especially in the absence of consensus or clear alternatives.”

“The adjoining editorial (Wasserstein, Schirm, and Lazar 2019) recommended that scientists stop using the term ‘statistically significant’ entirely, but noted the articles ‘do not sing as one’ and reflect ‘deep dissonance.’”

Gibson (2020), cont'd:

“In contrast to the ASA, the New England Journal of Medicine (Harrington et al. 2019) recommended, “despite the difficulties they pose, p-values continue to have an important role in medical research, and we do not believe that p-values and significance tests should be eliminated altogether.”

“An Editorial in the journal Clinical Trials (Cook et al. 2019) also cautioned that “there is still a place for significance testing in clinical trials.”

“Although I focus on p-values, it is important to note that other statistical tools such as confidence intervals and Bayes factors are also prone to misuse.”

“Statistics and Unintended Consequences”

(*Amstat News*, 1 June 2019, President’s Column):

- “Don’t trust statisticians. They told us for 100 years: calculate p-values. Now they’re telling us not to.”
- “Goody, I don’t have to do statistics! Never understood it anyway.”
- Robert Abelson (1997): “oboes, band saws, skis, and college educations” are misused. “*Will we want to ban effect sizes too, when their misuse escalates?*”

(also: President’s Columns, Apr 2019 & Dec 2019)

- Should we adopt the Editorial as ASA Policy?
- Should ASA be setting policy for the **practice** of statistics?
- What messages **do** we want to convey?
- In her last act as 2019 ASA President, she established a task force to address **p-values, significance, and replicability**

Xuming He*

Linda Young*

Yoav Benjamini

Richard De Veaux

Bradley Efron

Scott Evans

Mark Glickman

Barry Graubard

Xiao-Li Meng

Nancy Reid

Stephen Stigler

Stephen Vardeman

Chris Wikle

Tommy Wright

(KK, Ex-officio)

Task Force delivered to ASA Board (July 30, 2020):

- Proposed ASA Statement
- 5 recommendations:
 - Disclaimer on all publications, articles, editorials, ... authored by ASA Staff (e.g., as required for U.S. Govt employees)
 - Distinction between ‘Statement’ & ‘Response’ to current situations (e.g., SVEP, Transfer of Covid-19 data collection out of CDC’s surveillance program, ...)
 - Three others
- All are under discussion.

In the opinion of your current speaker:

- Should we adopt the Editorial as ASA Policy? *No.*
- Should ASA be setting policy for the **practice** of statistics?
No. Role of ASA is to foster & facilitate forums for discussions, but to take no stance itself.
- What messages **do** statisticians want to convey?
Good, sound statistical practice, multiple approaches, recognition of assumptions, continued research into better methods, need for replicability, dangers of failing to account for multiplicity

More to come

Thank you!