The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1081–1083
https://doi.org/10.1214/21-AOAS1500

EDITORIAL: STATISTICAL SIGNIFICANCE, *P*-VALUES, AND REPLICABILITY

BY KAREN KAFADAR

Editor-in-Chief, 2019–2021

kk3ab@virginia.edu

The debate about the value of hypothesis testing, and the over-reliance on p-values as a cornerstone of statistical methodology, has persisted for well over a century. Many researchers, including statisticians, have commented on the frequent use and abuse of p-values. The American Statistical Association (ASA) published an issue of *The American Statistician* in March 2019 devoted entirely to this topic. The message in many of these articles is sensible: the "0.05 threshold" for p-values is often arbitrary, and the notion of "$p < 0.05$" as "statistically significant" may not be appropriate for many situations. Some have interpreted the articles in that issue, and the many that followed, as statisticians abandoning hypothesis tests entirely (*Nature* Editorial, 20 March 2019). Others have incorrectly assumed that the articles represented official ASA policy (*Scientific American*: Denworth ((2019), p. 64); *Nature*: Amrhein, Greenland and McShane (2019); *Significance*: Tarran ((2019), p. 14)).

As ASA President in 2019, I convened a Task Force to prepare a statement to clarify the role of hypothesis tests, p-values, and their relation to replicability. The Statement from that Task Force appears as the next article following this Editorial. The Task Force was intended to span a wide range of expertise, experience, and philosophy, and remarkable unanimity was achieved. All Task Force members are listed as authors of the Statement, as all participated in writing it and approved it for publication. The Task Force Statement is important: as with almost all methods, in statistics and elsewhere, concepts of hypothesis tests, p-values, and replicability can be misunderstood and misused, but they remain central to scientific inference.

Results of hypothesis tests are routinely reported in scientific studies. For example, Beigel et al. (2020) reported in their abstract the results of their "double-blind, randomized, placebo-controlled trial of intravenous remdesivir" in 1,062 adults hospitalized with Covid-19 and evidence of lower respiratory tract infection: "Those who received remdesivir had a median recovery time of 10 days (95% confidence interval [CI], 9 to 11), as compared with 15 days (95% CI, 13 to 18) among those who received placebo (rate ratio for recovery, 1.29; 95% CI, 1.12 to 1.49; $P < 0.001$, by a log-rank test)." P-values are also commonly calculated in large-scale genome-wide association studies (e.g., Storey and Tibshirani (2003)).

Courts of law also rely heavily on statistical methods in assessing the admissibility of scientific evidence (Kaye and Freedman (2011)). Rule 702, *Testimony by Expert Witnesses*, of the *Federal Rules of Evidence* (Legal Information Institute) was amended in 2000 to take into consideration several factors when assessing the reliability of scientific expert testimony, including "whether the technique or theory has been subject to peer review and publication" and "the known or potential rate of error of the technique or theory when applied." Statistical tests are often critical components in peer-reviewed articles, and judges look for them in making decisions about the admissibility of scientific expert testimony. The *Reference Manual for Scientific Evidence* (Federal Judicial Center (2011)) devotes four of its thirteen chapters to

Received June 2021.

statistical concepts (*Statistics, Multiple Regression, Survey Research, Epidemiology*), which are frequently cited in judicial decisions. In *Matrixx Initiatives v Siracusano* (2011), Justice Sotomayor wrote (footnote 6):

> "A study that is statistically significant has results that are unlikely to be the result of random error. . . ." Federal Judicial Center, *Reference Manual on Scientific Evidence* 354 (2d ed. 2000). To test for significance, a researcher develops a "null hypothesis" e.g., the assertion that there is no relationship between Zicam use and anosmia. See id., at 122. The researcher then calculates the probability of obtaining the observed data (or more extreme data) if the null hypothesis is true (called the p-value). Ibid. Small p-values are evidence that the null hypothesis is incorrect. See ibid. Finally, the researcher compares the p-value to a preselected value called the significance level. Id., at 123. If the p-value is below the preselected value, the difference is deemed "significant." Id., at 124.

The quotation in Justice Sotomayor's opinion comes from the chapter on statistical methods in the *Reference Manual on Scientific Evidence* (Kaye and Freedman). That chapter cites several court cases that refer to hypothesis tests, p-values, and "statistical significance" (as well as confidence intervals, random versus biased samples, etc.; cf. Section 4B, pp. 249–253). In one such case [Giles *v* Wyeth, Inc., 500 F. Supp. 2d 1048, 1056 (S.D. Ill. 2007)], the Court refers to several statistical concepts, including confidence intervals, replicability, and "cherry-picking":

> "[Plaintiff] also relies on cherry-picked data from the FDA's 2006 study on antidepressant-induced suicide. The FDA based its analysis on data collected from 372 RCTs involving nearly 100,000 individuals. . . . Despite the study's overall conclusion, [Plaintiff] grasps onto a subset of data that facially suggests that antidepressants cause suicidality in adults in the 45–54 age group. . . . The odds ratio for antidepressants versus a placebo for [a specific] age group was 2.29, with a 95% confidence interval between 0.73–7.14, and a p-value 0.15. While [Plaintiff] admits that a p-value of 0.15 is three times higher than what scientists generally consider statistically significant—that is, a p-value of 0.05 or lower—she maintains that this 'represents 85% certainty, which meets any conceivable concept of preponderance of the evidence.' (Doc. 103 at 16)."

(The Court proceeds in its decision to appropriately criticize Plaintiff's claim that "a p-value of 0.15. . . 'represents 85% certainty'.") Numerous court cases discuss results of studies in terms of p-values from hypothesis tests, "significance," confidence intervals, and replicability (Kaye and Freedman (2011)); the Task Force addresses these concepts in its Statement. While not immune from misuse or confusion, these notions (like many in science), when applied and interpreted properly, remain useful and valid, for guiding both scientists and consumers of science (such as courts of law) towards insightful inferences from data.

Not all statistical methods are appropriate for all instances, yet they are critical for answering the question: "How firm should evidence be, to take a result seriously?" Sir David Cox wrote in 1986 (p. 121), "Something like a significance test is needed for the essential task of checking and criticizing models and formulating improved ones, a key aspect of successful applied work." More recently, Cox (2020) wrote, "The mathematical clarity of Neyman's work is, of course, appealing, but it may be argued that its overformalization continues to lead to misunderstanding, unproductive discussion and rigidity concerning, in particular, the role of significance tests." (See also Reid and Cox ((2014), Section 4).) The "unproductive discussion" has been unfortunate, leading some to view tests as wrong rather than as valid and often informative. Even Sir Ronald Fisher himself, the purported source of the "0.05" threshold, may not have been so rigid: "Sir Ronald's firm knowledge was *not one extremely significant result, but rather the ability to repeatedly get results significant at 5%*" (Tukey (1969), p. 85). Indeed, Fisher (1935) himself wrote in his classic *The Design of Experiments*:

> "In order to assert that a natural phenomenon is experimentally demonstrable we 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."

Research in statistics advances when new problems and new data types inspire the development of new methods. A misuse of a tool ought not to lead to complete elimination of the tool from the toolbox, but rather to the development of better ways of communicating when it should, and should not, be used. While better approaches to explaining the concepts continue to be explored, new methods related to inference and replicability also should continue to be proposed and considered by the community. With the increasing availability of data and the temptation to explore thousands (or more) of possible connections and associations, some structure in the problem formulation and scientific communication is needed. The Statement of the Task Force reinforces the critical role of statistical methods to ensure a degree of scientific integrity. I hope that the principles outlined in the Statement will aid researchers in all areas of science, that they will be followed and cited often, and that the Statement will inspire more research into other approaches to conducting sound statistical inference.

Acknowledgements. My thanks to the Task Force members for their time and efforts, to Linda J. Young and Xuming He for chairing the Task Force, and to them and Nancy Reid and Barry Graubard for their comments on an earlier version of this Editorial.

REFERENCES

AMRHEIN, V., GREENLAND, S. and MCSHANE, B. (2019). Scientists rise up against statistical significance. *Nature* **20** 305–307.

BEIGEL, J. H., TOMASHEK, K. M., DODD, L. E. et al. (2020). Remdesivir for the treatment of Covid-19—final report. *N. Engl. J. Med.* **383** 1813–1826. https://doi.org/10.1056/NEJMoa200776

COX, D. R. (1986). Some general aspects of the theory of statistics. *International Statistical Review* **54** 117–126.

COX, D. R. (2020). Discussion of paper by Brad Efron. *Journal of the American Statistical Association* **115** 659–659.

DENWORTH, L. (2019). A Significant Problem. *Sci. Am.* **10** 63–67. 2019.

FEDERAL JUDICIAL CENTER (2011). *Reference Manual on Scientific Evidence*, 3rd ed., National Academies Press, Washington, DC.

FISHER, R. (1935). *The Design of Experiments*. Oliver and Boyd, London.

KAYE, D. H. and FREEDMAN, D. A. (2011). Reference guide on statistics. In *Reference Manual on Scientific Evidence* Third Edition 211–302, National Academies Press.

MATRIXX INITIATIVES, INC. v. SIRACUSANO, 563 U.S. 27 (2011). 131 S.Ct. 1309 Supreme Court of the United States, No. 09-1156 (179 L.Ed.2d 398, 79 USLW 4187, Fed. Sec. L. Rep. P 96,249.)

REID, N. and COX, D. R. (2014). On some principles of statistical inference. *Int. Stat. Rev.* **83** 293–308. https://doi.org/10.1111/insr.12067

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856 https://doi.org/10.1073/pnas.1530509100

TARRAN, B. (2019). The *S* word… and what to do about it. *Signficance*, August 2019: 14. [Correction to initial version published 24 July 2019: "Correction added on 26 February 2021, after publication: This article has been updated to clarify that the recommendation to abandon the term 'statistical significance' came from the editors of a special issue of The American Statistician. not the American Statistical Association (which publishes the journal). We apologise for any confusion."]

TUKEY, J. W. (1969). Analyzing data: Sanctification or detective work. *Amer. Psychol.* **24** 83–91.