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.