The scientific method and *p* values: response to Mayo 2021

Edward Ionides, Department of Statistics, University of Michigan, Ann Arbor, MI, U.S.A., email ionides@umich.edu

Ya'acov Ritov, Department of Statistics, University of Michigan

Article impact statements: Scientific judgment is required to assess statistical evidence to evaluate critically the extent to which results support study conclusions.

There is a need for articles, such as the recent *Conservation Biology* editorial by Mayo (2021), elaborating on and contextualizing the American Statistical Association President's Task Force statement on statistical significance (Benjamini et al, 2021). This statement speaks what seems to us like plain good sense. However, it avoids addressing why there is a debate in the first place, and the justifications and misconceptions that drive people's differing positions. Consequently, it may be ineffective at communicating with those swing voters who have sympathies with some of the insinuations in the Wasserstein and Lazar (2016) editorial. We use *insinuations* here because we consider that this editorial attacks *p* values forcefully, indirectly, and erroneously.

Wasserstein and Lazar (2016) start with a constructive discussion about the uses and abuses of *p* values before moving against them. This approach is good rhetoric, reminiscent of Shakespeare's Marc Anthony: I come to praise *p* values, not to bury them. Good rhetoric does not always promote good science, but Wasserstein and Lazar (2016) successfully frame and lead the debate, according to Google Scholar. We warned of the potential consequences of that article and its flaws (Ionides et al, 2017). Wasserstein et al. (2019) made their position

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/cobi.13984.

clearer and therefore easier to confront. We are grateful to Benjamini et al. (2021) and Mayo (2021) for rising to the debate. In support of their efforts, we rephrase a Churchill quotation. We contend that "many forms" of statistical methods "have been tried, and will be tried in this world of sin and woe. No one pretends that" the *p* value "is perfect or all wise. Indeed" (noting that its abuse has much responsibility for the replication crisis) "it has been said that" the *p* value "is the worst form of inference except all those other forms that have been tried from time to time."

Mayo (2021) started her editorial by asking for scientific editorial policy not to take sides in favor of a particular statistical point of view. The article then summarized her approach to the foundations of statistical reasoning (Mayo, 2018) based on the desire to make decisions supported by error control claims, which she considered desirable from the point of view of Popperian severe testing in scientific inquiry, as well as in nonscientific situations, including law and public policy. She showed that a frequentist Fisher-Neyman-Pearson framework can be consistent with this desire, whereas the calculus of Bayesian beliefs alone cannot in general guarantee this. It took Mayo (2018) a whole book to patiently disentangle a mass of arguments surrounding this topic. For example, she introduced the idea of an "audited error probability" to explain why one has to assess whether an asserted *p* value actually corresponds to the scientific process carried out to avoid sins such as *p* hacking. She also pointed out that *p* values provide a tool to detect and correct for abuses due to poor auditing.

Mayo (2018) did not dispute the potential utility of Bayesian calculations, she merely showed that they must be supplemented with non-Bayesian reasoning to have error control properties widely seen as favorable. The need to look beyond Bayesian posterior belief calculations in order to carry out model criticism, even when desiring a Bayesian inference, was championed

by Box (1983) and reinvigorated by Little (2011). An extension of the position of Mayo (2018, 2021) suggests that substantive conclusions of a Bayesian analysis should be similarly checked. Error probabilities can be supplied to support Bayesian belief statements. For example, one can calculate by simulation the chance of obtaining a belief at least as extreme as the presented inference under a specific hypothesis of interest. In particular, when a frequentist could construct a bootstrap sample, the bootstrapped Bayesian posterior belief provides a relevant test statistic. Huelsenbeck and Rannala (2004) demonstrated this in a complex biological model: if a conclusion based on Bayesian beliefs has undesirable frequentist properties, this is something that most careful scientists would want to know. Bayesian procedures can have favorable frequentist properties, though this cannot be taken for granted in complex models and may require the use of different priors for different questions even if the same statistical model is applied to the same data set (Ritov et al., 2014). Evidently, a conclusion based on Bayesian belief is stronger if the authors present numerical evidence concerning error probabilities for their proposed inference. Thus, the argument properly supported by Mayo (2021) is stronger than her opening request for the scientific community not to take sides in the ongoing debate over appropriate choices of statistical methodology: the presentation of error probabilities (including, but not necessarily limited to, *p* values) should be encouraged even when the author's main goal is to quantify beliefs.

Some scientific endeavors are concerned with exploratory or descriptive data analyses, aiming to generate hypotheses rather than resolve them. For such studies, formal statistical methods, such as *p* values, may not be required, though *p* values may be still useful. A ubiquitous example is the use of confidence intervals, or the corresponding *p* values, for providing a measure of uncertainty in descriptive data analyses. In this situation, the

hypotheses being examined are not predesignated. The logic of frequentist inference continues to apply without predesignation (Mayo 1996). Care is needed to distinguish between true *p* values (which make correct allowance for data-dependent decisions used to construct the proposed hypothesis) and naive *p* values (which suppose that a hypothesis was predesignated when this was not the case). For example, multiple testing corrections are appropriate when exploring a large number of hypotheses. Proper frequentist data analysis involves description of whether the hypotheses were predesignated, permitting the reader to audit the inferences presented. Scientific judgment is required to assess the statistical evidence, as for other aspects of the design and implementation of the study, to evaluate critically the extent to which the results support the study conclusions.

Acknowledgments

We acknowledge constructive feedback from the editor and 4 anonymous referees**.**

References

Benjamini, Y., De Veaux, R.D., Efron, B., Evans, S., Glickman, M., Graubard, B.I., He, X., Meng, X.L., Reid, N.M., Stigler, S.M. and Vardeman, S.B., 2021. ASA President's Task Force Statement on Statistical Significance and Replicability. *Annals of Applied Statistics*, 15(3), pp. 1084-1085.

Box, G. E. (1983). An apology for ecumenism in statistics. In *Scientific inference, data analysis, and robustness*, pages 51–84. Elsevier.

Huelsenbeck, J. P. and Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53(6):904–913.

Ionides, E.L., Giessing, A., Ritov, Y. and Page, S.E., 2017. Response to the ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 71(1), pp. 88-89.

Little, R., 2011. Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26(2):162–174.

Mayo, D.G., 1996. *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D.G., 2018. *Statistical inference as severe testing*. Cambridge University Press.

Mayo, D.G., 2021. The statistics wars and intellectual conflicts of interest. *Conservation Biology,* 36(1):e13861

Ritov, Y., Bickel, P. J., Gamst, A. C., and Kleijn, B. J. K., 2014. The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statistical Science*, 29(4):619–639.

Wasserstein, R.L. and Lazar, N.A., 2016. The ASA statement on p-values: context, process, and purpose. *The American Statistician,* 70(2), pp. 129-133.

Wasserstein, R.L., Schirm, A.L. and Lazar, N.A., 2019. Moving to a world beyond "p< 0.05". The American Statistician, 73(sup1), pp. 1-19.