Are *p*-values under attack? Contribution to the discussion of 'A critical evaluation of the current "*p*-value controversy"'

Walter W. Piegorsch*,[1,2]

[1] Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ 85721, USA
[2] BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA

Received 28 February 2017; revised 28 February 2017; accepted 4 March 2017

Keywords: Bayesian model averaging; Empirical Bayes; Hypothesis testing; *p*-values.

It is a pleasure to congratulate Prof. Wellek on his intriguing and thought-provoking paper. He reminds us of some instructive themes on this topic, and I must admit that I found little with which to quibble. My take-away message from the exposition is that while some high-visibility sources have called into question use of *p*-values in modern, data-rich, scientific discourse, their complaints may be overblown: the *p*-value is as indispensable (Prof. Wellek's term) as ever in contemporary medical applications and in associated areas such as regulatory affairs.

The issue appears to be that (i) recalcitrant objectors have argued against *p*-values—and null hypothesis based testing in general—as they are applied in far too-automated a fashion, and that (ii) in the end *p*-values provide little information about the actual experimental hypotheses under study. In reply, the American Statistical Association issued a Statement in support of "statistical significance and *p*-values" (Wasserstein and Lazar, 2016), the useful guiding principles from which are aptly summarized, and in a few cases criticized, in Prof. Wellek's Discussion Section. In fact, from my reading I found the agreement rather substantial between the ASA's published principles and Prof. Wellek's critique. I think all players would agree with Prof. Wellek's convincing conclusion (end of his Sec. 3.3) that

> ". . . criticizing the methodology of statistical testing for an undue lack of flexibility with respect to its coverage of different types of questions raised in scientific research, is unwarranted."

(Also see Kuffner and Walker, 2017.) Indeed, my own experience in a large variety of environmental settings and with some selected biomedical applications leads to essentially similar conclusions. So, why all the controversy?

Perhaps these various contributors—or, at least those raising disparaging complaints—are focusing on one component, viz. *p*-values and hypothesis tests, of a larger problem: ambitious but poorly informed use of statistical tools when analyzing complex, modern data. Few would disagree that *p*-values and hypothesis tests are, unfortunately, misused in some settings, and that this misuse harms scientific progress. I am unconvinced, however, that the methodology is to blame and for that matter, that the discussion needs to be elevated to a "controversy." As in any field of endeavor, applying best practices leads to, well, best practice (usually), and failure to do so leads to poor practice. Simply put, we should not do lazy statistics, and we should not condone it when done by others.

It is worth highlighting one of Prof. Wellek's calls for elevating the conversation here: greater reliance on and, I would add, ongoing development of more flexible quantitative techniques for extracting information from data. His opening call (cf. his Sec. 4) for greater use of confidence intervals is hard to argue against. His other suggestions are also for the most part unassailable, and I am sure any reader

*Corresponding author: e-mail: piegorsch@math.arizona.edu, Phone: +1-520-621-2357

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

could extend the list further and more profitably if given the chance. One particular item that caught my attention is the call for greater appeal to Bayesian inferences (see his Sec. 4.4, and also other comments throughout the exposition). Prof. Wellek makes repeated, compelling reference to the confusion many underinformed users feel when interpreting (frequentist) *p*-values and interval estimators, and how Bayesian alternatives provide more direct answers to the questions those users often ask. I have seen the effect first-hand. Full disclosure here: I do not profess to be a "dyed-in-the-wool" Bayesian, but I nonetheless will argue for Bayesian hierarchical modeling whenever a data-analytic scenario warrants its use. Such warrant can be tricky, of course: as Prof. Wellek intones, proper construction of the hierarchy and in particular of pertinent prior distribution(s) remains as much an art as a science. As different stakeholders may have different concerns regarding choice of the prior, apprehensions can arise regarding prior sensitivity. This can be particularly onerous in regulatory or other governmental settings: it is not beyond belief to imagine an unscrupulous entity who perniciously manipulates a prior to produce posterior inferences suited to parochial, *a priori* interests. Informed onlookers would quickly recognize this, of course, but they are not always at the ready, nor are they the intended audience of such chicanery.

Admittedly, counterarguments to this "subjectivity" of a Bayesian hierarchical model are strong. Certainly as the sample size increases the data should override any incongruous prior specification, unscrupulous, or not. And, echoing Prof. Wellek's comments, I have witnessed more than a few instances where a Jeffreys' noninformative prior operated with admirable resiliency. Indeed, if use of noninformative priors were to become even more widespread, strong deviations from their use could raise the sorts of red flags needed to produce more-skeptical, or at least deeper, inquiry and examination into validity of the hierarchy's components.

It is worth mention that at present some intriguing strategies lie "under the radar" that can help assuage concerns with Bayesian prior subjectivity. Among my own favorites are the increasing applications of Bayesian model averaging (BMA), where a class of different priors is formally employed and their corresponding posteriors are averaged across the class, weighted by their posterior model probabilities (Hoeting et al., 1999). In effect, prior uncertainty is quantified and then incorporated into the hierarchical calculations. Indeed, one could arguably include even the most senseless of priors, where an ostensibly low posterior weight essentially dismisses its posterior impact. (Or, if upon inspection a strong posterior weighting appeared for such a prior, the senselessness of the specification could be called into question. Remember, do not do lazy statistics!) The BMA method holds room for improvement, admittedly, but further study into its uses and features is warranted to help develop its wider potential.

Another under-examined approach to address uncertain prior specification is the use of (parametric) empirical Bayes methods. Here, the data are employed to estimate unknown/uncertain components of the prior using concomitant features of the hierarchical construction (van Houwelingen, 2014). Since it does not completely specify the prior before seeing the data, the method is often frowned upon as not being strictly Bayesian (e.g., Bernardo, 2008). This concern may be more philosophical than practical, however. I have seen empirical Bayes strategies slip operably into large hierarchies and complex prior formulations, not always under the empirical Bayes moniker but nonetheless with the same style and function. And as Efron (2010) notes, the empirical Bayes approach holds great potential for bridging the larger frequentist/Bayesian gap, particularly for the sorts of large-scale data problems we more regularly encounter. (Of course, he also notes that it has held such promise for well over 50 years, so coalescence into a coherent theory is still wanting.) I look forward to seeing continued developments in both BMA and empirical Bayes for hierarchical modeling and analysis.

As a practicing statistician and as an interdisciplinary data-science educator, I am grateful to Prof. Wellek for asking the questions he asks, and for encouraging us to find answers to them. Personally, I think the *p*-value "controversy" is overblown, and that in the end cooler heads will prevail. Prof. Wellek's paper is a step in the right direction. To keep moving forward, Mayo's (2016) titular aphorism, also mentioned by Prof. Wellek, is salient. Paraphrasing it slightly: we *can* save the statistical tools baby, even enhance it, while we throw out the bad statistics bathwater. It may seem obvious, but as

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

statistical educators we must do a (much) better job of informing and training our students in the proper interpretation and use of p-values, and of statistical testing in general; as statistical practitioners we must do a better job of "practicing what we preach" and not haphazardly employ p-values where they are inappropriate, misleading, and/or engender further confusion; as statistical consultants we must do a better job of working with—and where necessary, (re)educating—our domain-specific clients to keep them from making the sorts of missteps that have led to this squabble; and as (theoretical and methodological) statistical developers we must keep imagining, inventing, and improving effective tools—frequentist, Bayesian, et al.!—that expand our ability to extract real knowledge from data, some examples of which Prof. Wellek gives in his Sec. 4. p-values are not dead, at least for the present perhaps, and they need not be viewed that way; we simply need to avoid doing lazy statistics and lazy data science.

Acknowledgements The author is indebted to Drs. Edward J. Bedrick, Melanie Bell, D. Dean Billheimer, and Gabriel Huerta for helpful deliberations on this topic. This discussion was supported in part by Grant #R03ES027394 from the U.S. National Institute of Environmental Health Sciences. All commentary herein is solely the responsibility of the author and does not necessarily reflect the official views of this agency.

Conflict of interest
The author has declared no conflict of interest.

References

Bernardo, J. M. (2008). Comment on article by Gelman. *Bayesian Analysis* **3**, 451–454.

Efron, B. (2010). The future of indirect evidence (with discussion). *Statistical Science* **25**, 145–171.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–401 (corr. vol. 15, no. 3, pp. 193–195).

Kuffner, T. A. and Walker, S. G. (2017). Why are p-values controversial? *American Statistician* (in press). https://doi.org/10.1080/00031305.2016.1277161.

Mayo, D. G. (2016). Don't throw out the error control baby with the bad statistics bathwater. Online discussion: ASA statement on p-values and statistical significance. Available at http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm4621.pdf.

van Houwelingen, H. C. (2014). The role of empirical Bayes methodology as a leading principle in modern medical statistics. *Biometrical Journal* **56**, 919–932.

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

Wellek, S. (2017). A critical evaluation of the current "p-value controversy". *Biometrical Journal* **59**, 854–872.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim