Comment: Is Reform Possible Without a Paradigm Shift?

John B. Carlin

Murdoch Children's Research Institute & The University of Melbourne

John Carlin is Director, Data Science Core and Clinical Epidemiology & Biostatistics Unit (CEBU), Murdoch Children's Research Institute & University of Melbourne Department of Paediatrics, and is Professorial Fellow, Centre for Epidemiology & Biostatistics, Melbourne School of Population & Global Health, University of Melbourne, Melbourne, Australia.

Looking back over the ASA Statement on Statistical Significance and P-values, which I think is an important and valuable contribution to a real and pressing problem, a striking feature is that so much effort appears to be needed to *counteract* the misuse and misinterpretation of what must have seemed to Fisher and other pioneer statisticians to be a simple set of ideas. Of course the originators of the concepts of the p-value and its far more invidious offshoot, the practice of reducing empirical comparisons to declarations of "statistical significance" (or otherwise), could not have anticipated how these ideas would become so embedded in the practice of non-statistical researchers. As has been pointed out and will continue presumably to be analysed by the philosophers and historians of science, there seems to be an irresistible urge to encode scientific conclusion-making into a rule-based activity (Gigerenzer & Marewski, 2015). This always seems curious to me because it appears obvious that conclusions about the empirical world can only be made tentatively (beware the black swan!). Thus inductive inference must always be couched in a language of uncertainty, in contrast with which the familiar phraseology of statistically based research ("an association was found; $P < 0.05$…", "no effect was observed") just doesn't make sense.

Can the general scientific usage of statistical inference methods be reformed? The outright ban on traditional tools such as the p-value and confidence interval by the journal *Basic and Applied Social Psychology* (Trafimow & Marks, 2015) has achieved some positive outcomes, by way of much broader recognition and discussion of the underlying problems (Ashworth, 2015). It is less clear whether the quality of scientific inference within the pages of the journal has improved (Lakens, 2016).

I believe that fundamental improvement will only be possible if and when we can agree on some broad principles about the inference task. In particular, we need to cultivate a viable language of uncertainty that is primarily focused on expressing uncertain knowledge conditional on observed data (Morey et al, 2016). To my mind the only general language that seems to have any reasonable

chance of fulfilling this goal is the Bayesian use of probability. Genuine post-data conclusions seem to be possible only within a Bayesian or perhaps closely related paradigm.

However, I am well aware of the difficulties of this path. The overwhelming challenge is that as soon as we enter this paradigm we appear to require god-like knowledge of the "true" or at least "appropriate" model that should be specified, including prior distributions that will be needed to kick-start the uncertainty calculus. The difficulty of creating broadly accepted conventions for how models should be specified and checked before any conclusions based on their application to the data at hand may be trusted often seems insuperable, despite some suggested strategies (Gelman & Shalizi, 2013). To many, the danger of the rules becoming even more malleable – and so even more likely to allow researchers to put arbitrary stamps of statistical authority on the conclusions they would like to draw – under this paradigm than under the traditional muddled modes of p-value-based inference outweigh its compelling inherent logic.

Although I understand this point of view I just don't see any real choice, as no-one seems to be coming close to proposing a way of salvaging the traditional muddle. Indeed in a companion paper to this discussion (Greenland et al, 2016) we see a long list of misconceptions and misinterpretations, which it can be hoped scientists may start to avoid. Yet the length and complexity of the list itself suggests that the fundamental ideas that it seeks to clarify are so convoluted – and inherently unsuited to the task of uncertain inductive post-data inference – that a solution might only be possible with a more fundamental paradigm shift.

REFERENCES

Ashworth A (2015). Veto on the use of null hypothesis testing and p intervals: right or wrong? *Taylor & Francis Editor Resources* [online], http://editorresources.taylorandfrancisgroup.com/veto-on-the-use-of-null-hypothesis-testing-and-p-intervals-right-or-wrong/, accessed Feb. 29, 2016.

Gelman A, Shalizi C (2013). Philosophy and the practice of Bayesian statistics. (With discussion). *British Journal of Mathematical and Statistical Psychology* (2013), 66: 8-80.

Gigerenzer G, Marewski JN (2015). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management*, 41: 421-440.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. (2016) Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations. *The American Statistician.* In press.

Lakens D (2016). So you banned p-values, how's that working out for you? *The 20% Statistician: A blog on statistics, methods and open science* [online], http://daniellakens.blogspot.com.au/2016/02/so-you-banned-p-values-hows-that.html, accessed Feb. 29, 2016.

Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23:103-123.

Trafimow D, Marks M (2015). Editorial. *Basic and Applied Social Psychology*, 37:1-2.