Remove, rather than redefine, statistical significance

To the Editor — Benjamin et al.[1] propose to redefine statistical significance with a trichotomy: what was once 'highly significant' ($P < 0.005$) becomes 'significant', what was once significant ($P < 0.05$) becomes 'suggestive', and what was 'nonsignificant' ($P > 0.05$) remains nonsignificant. Trichotomization is better than dichotomization, and we agree that P values around 0.05 convey only limited evidence against the tested hypothesis (which is usually a 'null' hypothesis of no effect)[2].

We also agree that P hacking, selective reporting and publication bias "are arguably the bigger problems"[1] than false positives arising by chance. Nonetheless, imposing a more stringent significance threshold will aggravate those problems[3]. Benjamin et al. say that their proposal "should not be used to reject publications of novel findings with $0.005 < P < 0.05$". But rejections due to $P > 0.05$ will remain, and rejections due to $P > 0.005$ will now also occur, leading to more intense P hacking and selective reporting, with increased bias in reported effects (because estimates from studies that are selected for having $P < 0.005$ are usually more inflated than those selected for having $P < 0.05$)[3,4].

'Significance' and 'nonsignificance' are too often equated with 'falsity' and 'truth' of hypotheses, reflecting overconfidence about mathematical results and ignoring unmodelled uncertainties[2,3,5]. We believe that the proposed trichotomy will increase such overconfidence in nonsignificance and thus retard scientific progress[6]: depending on the context, the increase in false-negative conclusions from using more stringent thresholds may far outweigh in number or cost the false positives so avoided. Worse, lowering the significance threshold will probably aggravate the misinterpretation of $P > 0.05$ or even $P > 0.005$ as 'support' for the null hypothesis, rather than as mere failure to refute it[2,3,7,8].

To avoid perpetuating problems caused by discrete decision rules applied to single studies, we argue that presentation decisions should not be based on any P value threshold at all[2,3,8]. Reliable scientific conclusions require information to be combined from multiple studies and lines of evidence. To allow valid inference from literature syntheses, results must be published regardless of statistical significance, with the P value presented as a continuous summary[2,3,8] — for example, as an index of compatibility between the data and the model used to compute P, on a scale of 0 (completely incompatible) to 1 (completely compatible). The P value could even be replaced by a more intuitively scaled evidence measure, such as a likelihood ratio or a surprisal[2] $-\log(P)$, which are unbounded above and thus difficult to misinterpret as hypothesis probabilities. Interval estimates are also essential, along with an indication of their bias sensitivity[2].

In sum, lowering significance thresholds will aggravate several biases caused by significance testing[3]. Thus, while P values can be useful, we think statistics reform should involve completely discarding 'significance' and the oversimplified reasoning it encourages[2,3,8], instead of just shifting thresholds. Treating P values as continuous indices would emphasize that inferences do not "suddenly assume the mantle of reality"[9] once a threshold is crossed. Any study that reports methods and data honestly should be freely accessible regardless of the P value or other statistical results — keeping in mind that selective reporting based on study outcomes is a recipe for misleading conclusions and distorted literature[10]. ❐

Valentin Amrhein[1,2]* and Sander Greenland[3]*

[1]Zoological Institute, University of Basel, 4051 Basel, Switzerland. [2]Swiss Ornithological Institute, 6204 Sempach, Switzerland. [3]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA.
**e-mail: v.amrhein@unibas.ch; lesdomes@g.ucla.edu*

Published online: 25 September 2017
DOI: 10.1038/s41562-017-0224-0

References

1. Benjamin, D. J. et al. *Nature Hum. Behav.* **1**, 0189 (2017).
2. Greenland, S. The need for cognitive science in methodology. *Am. J. Epidemiol.* (in the press).
3. Amrhein, V., Korner-Nievergelt, F. & Roth, T. *PeerJ* **5**, e3544 (2017).
4. Button, K. S. et al. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
5. Gelman, A. & Loken, E. *Am. Sci.* **102**, 460–465 (2014).
6. Fiedler, K., Kutzner, F. & Krueger, J. I. *Persp. Psy. Sci.* **7**, 661–669 (2012).
7. Greenland, S. *Eur. J. Epidemiol.* **32**, 3–20 (2017).
8. Greenland, S. et al. *Eur. J. Epidemiol.* **31**, 337–350 (2016).
9. Oakes, M. *Statistical Inference: A Commentary for the Social and Behavioural Sciences* (Wiley, Chichester, 1986).
10. Wasserstein, R. L. & Lazar, N. A. *Am. Stat.* **70**, 129–133 (2016).

Competing interests

The authors declare no competing interests.