Revised evidence for statistical standards

Andrew Gelman [*] **and Christian P. Robert** [†] [‡]

[*]Department of Statistics, Columbia University,[†]Université Paris-Dauphine, CEREMADE, Paris, France, and [‡]Department of Statistics, University of Warwick, UK

Submitted to Proceedings of the National Academy of Sciences of the United States of America

In (1), Johnson proposes replacing the usual $p = 0.05$ standard for significance with the more stringent $p = 0.005$. This might be good advice in practice but we remain troubled by Johnson's logic because it seems to dodge the essential nature of any such rule, that it expresses a tradeoff between the risks of publishing misleading results and of important results being left unpublished. Ultimately such decisions should depend on costs, benefits, and probabilities of all outcomes.

Johnson's minimax prior is not intended to correspond to any distribution of effect sizes; rather it represents a worst-case scenario under some mathematical assumptions. Minimax and tradeoffs do not play well together (3), and it is hard for us to see how any worst-case procedure can supply much guidance on how to balance between two different losses.

Johnson's evidence threshold is chosen relative to a conventional value, namely Jeffreys' target Bayes factor of 1/25 or 1/50, for which we do not see any particular justification except with reference to the tail-area probability of 0.025, traditionally associated with statistical significance.

To understand the difficulty of this approach, consider the hypothetical scenario in which R. A. Fisher had chosen $p = 0.005$ rather than $p = 0.05$ as a significance threshold. In this alternative history, the discrepancy between p-values and Bayes factors remains and Johnson could have written a paper noting that the accepted 0.005 standard fails to correspond to 200-to-1 evidence against the null. Indeed, a 200:1 evidence in a minimax sense gets processed by his fixed-point equation $\gamma = \exp[z\sqrt{2\log(\gamma)} - \log(\gamma)]$ at the value $\gamma = 0.005$, into $z = \sqrt{-2\log(0.005)} = 3.86$, which corresponds to a (one-sided) tail probability of $\Phi(-3.86)$, approximately 0.0005. Moreover, the proposition approximately divides any small initial p-level by a factor of $\sqrt{-4\pi\log(p)}$, roughly equal to 10 for the p's of interest. Thus, Johnson's recommended threshold p = 0.005 stems from taking 1/20 as a starting point; p = 0.005 has no justification on its own (any more than does the $p = 0.0005$ threshold derived from the alternative default standard of 1/200).

One might then ask, was Fisher foolish to settle for the $p = 0.05$ rule that has caused so many problems in later decades? We would argue that the appropriate significance level depends on the scenario, and that what worked well for agricultural experiments in the 1920s might not be so appropriate for many applications in modern biosciences. Thus, Johnson's recommendation to rethink significance thresholds seems like a good idea that needs to include assessments of actual costs, benefits, and probabilities, rather than being based on an abstract calculation.

References

1. Johnson V (2013) Revised standards for statistical evidence. *Proc Natl Acad Sci USA*.
2. Kane TJ (2013) Presumed averageness: The mis-application of classical hypothesis testing in education. The Brown Center Chalkboard, Brookings Institution. (`http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/12/04-classical-hypotesis-testing-in-education-kane`).
3. Berger J (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, New York), Second edition.