# Adaptive revised standards for statistical evidence

Johnson (1) argues for decreasing the bar of statistical significance from 0.05 and 0.01 to 0.005 and 0.001, respectively. There is growing evidence that the canonical fixed standards of significance are inappropriate. However, the author simply proposes other fixed standards. The essence of the problem of classical testing of significance lies in its goal of minimizing type II error (false negative) for a fixed type I error (false positive). A real departure instead would be to minimize a weighted sum of the two errors, as proposed by Jeffreys (2). Significance levels that are constant with respect to sample size do not balance errors. Size levels of 0.005 and 0.001 certainly will lower false positives (type I error) to the expense of increasing type II error, unless the study is carefully designed, which is not always the case or not even possible. If the sample size is small, the type II error can become unacceptably large. Conversely, for large sample sizes, 0.005 and 0.001 levels may be too high. Consider the psychokinetic data (3): the null hypothesis is that individuals cannot change by mental concentration the proportion of 1s in a sequence of $n = 104{,}490{,}000$ 0s and 1s, generated originally with a proportion of $1/2$. The proportion of 1s recorded was 0.5001768. The observed $P$ value is $P = 0.0003$; therefore, according to the present revision of standards, the null hypothesis is still rejected and a psychokinetic effect is claimed. This is contrary to intuition and to virtually any Bayes factor. Conversely, to make the standards adaptable to the amount of information [see also Raftery (4)], Pérez and Pericchi (5) approximate the behavior of Bayes factors by

$$\alpha_{\text{ref}}(n) = \alpha \times \frac{\sqrt{n_0 \times \left[\log(n_0) + \chi^2_\alpha(1)\right]}}{\sqrt{n \times \left[\log(n) + \chi^2_\alpha(1)\right]}}. \quad [1]$$

This formula establishes a bridge between carefully designed tests and the adaptive behavior of Bayesian tests. The value $n_0$ comes from a theoretical design for which a value of both errors has been specified, and $n$ is the actual (larger) sample size. In the psychokinetic data $n_0 = 44{,}529$ for a type I error of 0.01, a type II error of 0.05 is needed to detect a difference of 0.01. The $\alpha_{\text{ref}}(104{,}490{,}000) = 0.00017$ and the null of no psychokinetic effect is accepted.

A simple constant recipe is not the solution to the problem. The standard how to judge the evidence should be a function of the amount of information. Johnson's main message is to toughen the standards and design the experiments accordingly. This is welcomed whenever possible. However, it does not balance type I and type II errors: it would be misleading to pass the message—now use significance levels divided by 10, regardless of either type II errors or sample sizes. This would change the problem without solving it.

***Luis Pericchi*[a,1], *Carlos A. B. Pereira*[b], *and María-Eglée Pérez*[a]**

[a]*Department of Mathematics, University of Puerto Rico, Rio Piedras Campus, San Juan, Puerto Rico, PR 00931; and* [b]*Instituto de Matematica e Estatistica, University of Sao Paulo, CEP 05508-090, Sao Paulo, Brazil*

**1** Johnson VE (2013) Revised standards for statistical evidence. *Proc Natl Acad Sci USA* 110(48):19313–19317.
**2** Jeffreys H (1939) *Theory of Probability* (Oxford Univ Press, Oxford, UK).
**3** Good IJ (1992) The Bayes/non-Bayes compromise. A brief review. *J Am Stat Assoc* 87(419):597–606.
**4** Raftery AE (1995) Bayesian model selection in social research. *Sociol Methodol* 25(1):111–196.
**5** Pérez ME, Pericchi LR (2014) Changing statistical significance with the amount of information: The adaptive *alpha* significance level. *Stat Probab Lett* 85(1):20–24.

Author contributions: L.P., C.A.B.P., and M.-E.P. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: luis.pericchi@upr.edu.