DOI: 10.1037//0003-066X.57.1.70

Bayes Rules

Joachim Krueger
Brown University

When I reviewed the debate on null hypothesis significance testing (NHST; Krueger, January 2001), I did not expect to break the impasse between critics and defenders of this method. Nevertheless, I felt it was time to examine the arguments (i.e., the critical theses and the defensive antitheses) and to look for dialectical solutions (i.e., syntheses). If, I thought, the critics or the defenders of NHST could have won the debate, they already would have. The stalemate suggested that something was missing, and I concluded that this something might be old-fashioned Bayesianism made explicit. Aside from my 2001 article, I see a resurgence of Bayesianism as an alternative to NHST. After being prodded by Haig (2000), even the Task Force on Statistical Inference (2000) endorsed Bayesianism.

Three of the commentators on my 2001 article doubt that Bayesianism will succeed where NHST has failed. First, Guenther (2002, this issue) would limit probabilities to relative frequencies. If probabilities were not applied to single instances (e.g., a specific hypothesis), however, the asymmetry between hypotheses and data would be insurmountable. Whereas hypotheses would forecast the frequency distributions of certain observations, observations could not forecast the validity of hypotheses (excepting hypotheses of the all-or-none type). Not only Bayesianism but also NHST and other methods of induction would be moot. Researchers would have to conclude that nothing of theoretical interest can be concluded from the little *p*, and it would follow that no experiments need to be conducted. For these reasons, I cannot share the refusal to express levels of confidence and doubt probabilistically. Like Hofmann (2002), I think that Bayesian probabilities are useful because they point to the verisimilitude of hypotheses and theories.

Second, Markus (2002) preferred institutionalized norms of justification to the explicit subjectivism of Bayesianism. Wilkinson and the Task Force on Statistical Inference (1999) sought to steer researchers away from ritualistic applications of NHST and toward a more flexible use of statistical and graphic display tools. Although I do not intend to downplay the advantages of flexible and sophisticated data analysis, I see two risks. First, findings may become more difficult to compare when investigators choose without constraints from an increasingly well-stocked toolbox. Institutionalized norms may enhance the comparability of studies, but they do so by empowering journal editors and policy committees at the expense of individual investigators. Second, conventional analyses keep inverse inferences from data to hypotheses implicit and potentially chaotic. In other words, "what the [original task force] report amounts to is a vote of confidence for business as usual" (Sohn, 2000, p. 964). Bayesianism, in contrast, offers a principled way to integrate theories and data. It does not dictate what prior opinions researchers should have; it only encourages them to offer compelling justifications for those opinions. The maturity of a field can then be gauged by the degree to which initial differences in opinion have shrunk.

Third, Schmidt and Hunter (2002) repeated their condemnation of NHST, suggesting that effect sizes, confidence intervals, and meta-analyses are fully informative while avoiding the pitfalls of NHST. I did not claim, as they suggested, that NHST is the best procedure for induction. The value of NHST, as I see it, lies in its provision of one of the probabilities needed for the computation of likelihood ratios and posterior probabilities for specific hypotheses. Schmidt and Hunter did not address this pragmatic value of NHST but did register their opposition to its conventional, ritualistic application (a position I share). So why not dump both NHST and Bayesianism in favor of effect sizes cum confidence intervals cum meta-analyses? I hesitate to root for this approach because it yields research without hypotheses, and without hypotheses, there are no theories. If investigators relied entirely on point estimates and measurement error, they could not conclude anything because they did not entertain any hypotheses.

Schmidt and Hunter (2002) actually repudiated atheoretical empiricism when they stated that "no single study is ever sufficient to support a conclusion about the validity of a hypothesis" (p. 66) and that "such conclusions should be based on multiple studies as processed through meta-analysis methods" (p. 66). Some conclusions seem to be all right then, such as conclusions regarding the meta-analytic null hypothesis that "the variance of [the true] effect sizes is zero" (Hunter & Schmidt, 1990, p. 485). This null hypothesis may even be tested for significance (see Hunter & Schmidt, pp. 437–438), but how can investigators draw conclusions from meta-analyses while refusing to conclude anything from single studies and still be coherent?

One theme pervading my article (Krueger, 2001) was the idea that many researchers are closet Bayesians anyway. Whereas Schmidt and Hunter (2002), for example, share my belief that "the null hypothesis is typically false" (p. 65), they have also stated that the meta-analytic hypothesis of effect size homogeneity is often true (Hunter & Schmidt, 1990). Like others, then, they distinguish between risky and safe null hypotheses. Brand (2002) noted that a significant result increases one's confidence in the effectiveness of the experimental manipulation. I would add that without stating explicit priors, researchers never know what this increase might be. Guenther (2002) noted that a significant effect discounts all hypotheses predicting effects of the opposite direction. If, for example, researchers find a correlation of .3 between variables *X* and *Y* with an associated *p* of .01, they not only reject the null hypothesis of a zero correlation but also reject the hypothesis that the true correlation is –.3, only more so. Bayesian belief revision gives expression to this intuition. After the null hypothesis has been rejected once, its alternative appears to be a safer bet than it used to. I maintain that investigators hold beliefs about the replicability of empirical findings, and it is worth repeating that power analyses alone cannot sustain these beliefs. The probability of replication is not the probability of getting significance assuming that the first result is the true population effect; it is simply the probability of getting significance, given that the first result was significant.

A second theme of my 2001 article was that pragmatism is necessary because inductive inferences do not withstand logical scrutiny (as Hofmann, 2002, elaborated). I suggested pragmatically that researchers can attach nonzero Bayesian probabilities to point-specific hypotheses, although these hypotheses must logically be false. Similarly, Brand (2002) and Guenther (2002) argued for the pragmatic truth of many null hypotheses. Again, many scientists are intuitive Bayesians when they reject far-fetched claims by pointing to microscopic effect sizes and large *p* values. Should researchers really continue the Fisherian mantra of saying that they know nothing about the healing powers of homeopathy or about communication with the dead because there have been no significant results? A Bayesian frame of mind allows them to say that after so many failed tests (and the lack of plausible mechanisms), the respective null hypotheses are pretty darn probable. As Brand (2002) put it, "most realize in their heart of hearts that theory development (knowledge) integrated across many single investigators . . . represents the true value of science to society" (p. 67).

REFERENCES

Brand, J. L. (2002). Why chance is a good theory. *American Psychologist, 57,* 66–67.

Guenther, R. K. (2002). How probable is the null hypothesis? *American Psychologist, 57,* 67–68.

Haig, B. D. (2000). Explaining the use of statistical methods. *American Psychologist, 55,* 962–963.

Hofmann, S. G. (2002). Fisher's fallacy and NHST's flawed logic. *American Psychologist, 57,* 69–70.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16–26.

Markus, K. A. (2002). Beyond objectivity and subjectivity. *American Psychologist, 57,* 68–69.

Schmidt, F., & Hunter, J. (2002). Are there benefits from NHST? *American Psychologist, 57,* 65–66.

Sohn, D. (2000). Significance testing and the science. *American Psychologist, 55,* 964–965.

Task Force on Statistical Inference. (2000). Narrow and shallow. *American Psychologist, 55,* 965–966.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Correspondence concerning this comment should be addressed to Joachim Krueger, Department of Psychology, Brown University, Box 1853, Providence, RI 02912. E-mail: joachim_krueger@brown.edu