Comment

Morris L. Eaton

Let me begin by expressing my thanks for the opportunity to comment on this interesting and thought-provoking article. To put the remarks below into perspective I should say that my sympathies lie in the subjective Bayesian direction. Thus, for me inferential statements about the validity of an hypothesis are ideally expressed as conditional probabilities—that is, probabilities representing degree of belief, given everything known at the time. It is therefore no surprise that I strongly support the suggestion of reporting a posterior probability $P(H_0 \mid x)$ over a P-value. In the context of the paper, the authors have certainly shown that the common interpretation of a P-value of .05 as "strong evidence against H_0" is at best problematical, but I do think some alternative viewpoints on certain aspects of the paper are worthwhile.

WHAT'S THE QUESTION?

A wide range of believable situations in which P-values and $P(H_0 \mid x)$ differ dramatically are presented in this paper as well as elsewhere. The interpretation of $P(H_0 \mid x)$ as a subjective probability is certainly well-known; there is little debate about its meaning. In the same vein, the frequency interpretation of a P-value is well-known and very carefully explained in widely available sources. For example, Freedman, Pisani and Purves (1978) contains an excellent discussion, together with many relevant cautions, concerning P-values. But even in this reference, as elsewhere, the use of the set

$$E = \{\text{possible data } x\colon T(x) \geq T(x_0)\}$$

to interpret a P-value is not adequately justified and, as Berger and Delampady point out, this curious step certainly decreases the force of the "rare event" argument. Further, it is abundantly clear now that a P-value of .05 does not necessarily indicate a low subjective probability for H_0.

Because P-values are frequency-based measures of evidence, there is no compelling reason to think they should be directly comparable to subjective probability assessments. Thus, the direct comparison of the two seems to me somewhat inappropriate. However, describing a P-value of .05 as "strong evidence against H_0" while at the same time, a plausible assignment of prior probabilities leads to $P(H_0 \mid x)$ in the .2–.4 range, leads one to ponder—what question is being answered?

The number $P(H_0 \mid x)$ gives an easily interpretable numerical answer to the question:

$$\text{Q}\begin{cases}\text{What should one think about the truth of } H_0\\ \text{based on the model, the data } x \text{ and the prior}\\ \text{information available?}\end{cases}$$

Q is usually the relevant question, but P-values do not address Q, at least not directly; instead our attention is directed to the frequency interpretation of the set E whose relevance to Q is at best tangential.

The point is that the interpretation of a P-value and $P(H_0 \mid x)$ takes place in very different worlds and a direct numerical comparison may not be appropriate. However, concentrating on the question one wants to answer most often dictates the form and interpretation of the answer. Whereas $P(H_0 \mid x)$ gives a direct answer to Q, just what question the P-value addresses is not clear.

AUTOMATIC PROCEDURES

An oft advertised strength of many frequency-based statistical methods is the ease with which they can be applied. One simply plugs in the numbers and out comes an estimate coupled with a standard error, a P-value or some other frequency-based creation. The user is not required to supply any input except the model and the data; and in particular, knowledge based on previous work is not incorporated in the analysis (although it may be incorporated into the model). In this sense, such procedures might be called automatic procedures.

On the other hand, subjective Bayesian methods of analysis demand input of prior information. The inferential output is a posterior probability (or posterior distribution) that is supposed to represent an updated view of the world based on the model, current data and prior assessments. The Bayesian method is an attempt to quantify inductive inference and as such depends on both past and present evidence. The very act of selecting a prior distribution is a prior assessment, and thus to claim there are choices that are "objective" is misleading at best. In particular, choosing a conventional prior or choosing $\pi_0 = \frac{1}{2}$ as the prior probability for H_0 (as suggested by Berger and Delampady in Section 5 under Method 2) is a

Morris L. Eaton is Professor of Theoretical Statistics, University of Minnesota, 206 Church Street, S.E., Minneapolis, Minnesota 55455.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.*
www.jstor.org