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 $\pi_0 = ½$ 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.

subjective input and claims to the contrary are in my opinion, wrong. Every choice (including the model, prior, inferential procedure, etc.) needs to be justified via argumentation which will often be application specific. This is not to say that $\pi_0 = 1/2$ is necessarily an unreasonable choice, but I do decry the suggestion that there are automatic Bayesian methods that relieve the user and/or statisticians from justifying choices.

The use of automatic methods, be they Bayesian or frequentist, invite the user/statistician not to think hard about the application of statistical methods. Both user and statistician need to understand the strengths and weaknesses of the procedures they use and be aware of different inferential interpretations. It seems to me that anything less than this increases the incidence of inappropriate statistical analyses.

OBJECTIVITY

Each time I read a contemporary statistical piece in which the word "objective" (or some variant thereof) appears, I eagerly await the appearance of a few explanatory remarks that will anchor me to the author's understanding of what the word means. More often than not I wait in vain and am left to guess at the usage. This situation together with a specific concern about "objective Bayesianism" is what prompts most of my remarks in this section.

The recent article by Efron (1986) is a good example of the situation described above. His largely undefined notion of "objectivity" (along with "scientific objectivity," "strict objectivity," "complete objectivity," "objective Bayesianism," ...) leaves us to wonder just what is meant. None of the five commenters on Efron's paper was particularly attracted by the "objectivity argument," although only A. F. M. Smith raised the question of definition regarding the use of "objectivity." This is not the place to delve into a lengthy discussion of such a thorny subject, but to use "objectivity" without some hint of what is meant is incredibly naive in view of the vast literature on the subject. In particular the philosophy of science literature is filled with discussions of the "objectivity" issue (for example, see the anthology edited by Klemke, Hollinger and Kline (1980)). Such discussions have even made their way into the popular science literature (see Burke (1985), page 306).

Berger and Delampady use "objective" and "objective Bayesian" frequently in this paper, also without any discussion of what is meant. Let me focus on "objective Bayesianism." A common thread which runs through references to "objective Bayesianism" is the recommended use of "noninformative priors," although this term is typically ill-defined. In many situations these "noninformative" priors turn out to be improper. Now, it is well-known that some improper priors generate "good" stastistical procedures although others produce inferior answers, particularly in moderate or high dimensional problems. Thus, one cannot defend the use of improper priors without reference to something else—namely, the resulting statistical inference. Indeed it is completely proper that proposed statistical methods, be they frequentist, Bayesian or some ad hoc combination thereof, be evaluated by a comparison of the inferences they produce. If "objective Bayesianism" recommends the routine use of "noninformative" priors, then it is suggesting a demonstrably inferior strategy. If it refers to something else, I would surely like to know what, but ask that "objective" be defined before beginning the discussion.

The vagueness of the terms "objective" and "objective Bayesianism" does not seriously diminish the force of the Berger-Delampady argument for the following reason. The main thrust of the argument is that in the point null testing problem, the usual interpretations of P-values and $P(H_0 \mid x)$ can lead to radically different statements concerning evidence for H_0. Because $P(H_0 \mid x)$ has a direct probabilistic interpretation regarding evidence for H_0, they conclude that P-values should be eschewed in these point null problems. It is the direct comparison of these two inferential numbers that gives the argument much of its strength. This seems to me to be an appealing way to settle disagreements about different modes of inference.

ADDITIONAL REFERENCES

BURKE, J. (1985). *The Day the Universe Changed.* Little, Brown, Boston.

EFRON, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40** 1–11.

FREEDMAN, D., PISANI, R. and PURVES, R. (1978). *Statistics.* Norton, New York.

KLEMKE, E. D., HOLLINGER, R. and KLINE, D. A. (1980). *Introductory Readings in the Philosophy of Science.* Prometheus Books, Buffalo, N. Y.