


concern in the paper was to emphasize other points.) All these methods use smoothing and the better ones estimate the smoothing hyperparameter. If there is additional evidence about the smoothing then this could be incorporated into the prior. Actually, it is clear that L_1 is not much affected by lumpiness. For example, L_1 will scarcely be altered if 30 values are all at y or 30 values are spread over $y \pm \sigma$, for L_1 is a smoothed version of $\pi_1(\theta)$, smoothed by the error in y.

There is a point at which the lumpiness does matter. In my paper the assumption was made that if the glass on the suspect's clothing did truly match θ_0 then he was guilty. But if there are lumps this may not be so, for there may be several windows with index θ_0 and all that the evidence could show is that the glass came from one of these, not necessarily from the window at the scene of the crime.

6. MISCELLANEOUS COMMENTS

Seheult (1978) and Grove (1980) have both commented on my paper and their criticism is worth studying, although neither makes reference to the fact that his proposals are incoherent.

It was assumed in that paper that the glass was window and not, for example, bottle glass. My understanding was that it was possible to distinguish between the various broad types of glass.

A problem that does need analysis is that suggested by Shafer in his second comment in Section 5.3, when more than one piece of window glass is found on the suspect. There are several possibilities: none of the glass came from the broken window, only one piece did, two pieces did, and so on. It becomes a little messy to compare all the possibilities.

Is Shafer correct when he refers to the precision of an average? Is he not confusing precision with accuracy? Precision may be measured by the inverse of the variance: accuracy by the inverse of the mean squared error. Because scientific measurements typically contain unknown and undetected biases, precision can increase without limit but accuracy cannot. Statisticians, with their emphasis on standard errors that ignore the bias, have confused the issue in some scientific experimentation because the error they quote is substantially less than the true error.

There is one unsatisfactory feature of the Bayesian analysis that Shafer does not mention. It is sensitive to the error distribution. For example, if $(Y - \theta)/\sigma$ has a t distribution on five degrees of freedom, then at $Y - \theta = 2\sigma$ the likelihood is .171 times its value at $Y = \theta$, compared with .135 for the normal: at 4σ the values are 1.35×10^{-2} and 3.35×10^{-4}, respectively. We need more information about the tails of the error distribution.

There is room for improvement in the details of the Bayesian analysis of forensic data but the basic principles seem untouched by the criticism offered by Shafer.

REFERENCES

COHEN, L. JONATHAN (1977), *The Probable and the Provable*, Oxford: Clarendon Press.

DE FINETTI, B. (1974), *Theory of Probability* (Vol. 1), New York: Wiley.

FINKELSTEIN, MICHAEL O. (1978), *Quantitative Methods in Law*, New York: The Free Press.

GROVE, D.M. (1980), "The Interpretation of Forensic Evidence Using a Likelihood Ratio," *Biometrika*, 67, 243–246.

LINDLEY, DENNIS V. (1982), "Scoring Rules and the Inevitability of Probability," *International Statistical Review*, to appear.

RAMSEY, FRANK PLUMPTON (1931), *The Foundations of Mathematics and Other Essays*, London: Kegan, Paul, Trench, Trubner.

SEHEULT, ALLAN (1978), "On a Problem in Forensic Science," *Biometrika*, 65, 646–648.

TVERSKY, AMOS (1974), "Assessing Uncertainty," *Journal of the Royal Statistical Society*, Ser. B, 36, 148–159.

Comment

MORRIS H. DEGROOT*

This stimulating and well-written article on Lindley's paradox by Glenn Shafer raises several interesting issues, some of which pertain to the Bayesian approach to statistical inference in general, some to the particular example of criminal evidence discussed in the article, and some to Shafer's theory of belief functions. Here I shall briefly discuss each of these types of issues in turn.

1. DIFFUSE PRIOR DISTRIBUTIONS AND BAYESIAN INFERENCE

Consider first what Shafer calls a "diffuse" prior density function for a real-valued parameter θ; that is, in his words, a "fairly flat" density function that assigns "very

* Morris H. DeGroot is Professor, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213.

© Journal of the American Statistical Association
June 1982, Volume 77, Number 378
Theory and Methods Section

small'' probability to every interval of some fixed small length. I do not agree with the notion expressed by Shafer, and by many others before him, that a diffuse prior distribution represents ignorance about θ, or with the statement that ''the more ignorant we are, the more diffuse'' the prior distribution should be. Indeed, a diffuse prior distribution, represented for example by a normal distribution with a very large variance, indicates not that I am totally ignorant about θ but that I am quite certain that $|\theta|$ is large. I doubt that the concept of total ignorance about θ has any precise meaning: If I am totally ignorant about θ, then I am also totally ignorant about $\eta = \exp(\theta)/[1 + \exp(\theta)]$, but a diffuse distribution for θ indicates that η is almost certainly near 0 or 1.

Nevertheless, diffuse prior distributions and even improper prior distributions have proven to be useful in estimation problems of the type that Savage has called problems of ''stable estimation'' or ''precise measurement'' (Savage 1961; Savage et al. 1962; Edwards, Lindman, and Savage 1963). In these problems, the posterior distribution of θ is relatively insensitive to the particular diffuse prior distribution that is used. It is also well known, however, as explicitly pointed out by Jeffreys (1961, p. 251) that diffuse prior distributions are not appropriate for tests of significance. In these problems, the posterior distribution of θ is extremely sensitive to the particular diffuse prior distribution that is used under the alternative hypothesis, and improper prior distributions cannot meaningfully be used at all.

To illustrate this last statement with a simple example, consider again the calculation of the posterior odds

$$\Delta = \frac{P(\theta = \theta_0 \mid Y = y)}{P(\theta \neq \theta_0 \mid Y = y)} = \frac{\pi_0}{1 - \pi_0} \cdot \frac{L_0}{L_1} \tag{1}$$

carried out by Shafer in Section 1 of his article. If the prior density $\pi_1(\theta)$ under the alternative hypothesis $\theta \neq \theta_0$ is taken to be the improper uniform density such that $\pi_1(\theta)d\theta = d\theta$ for $\theta \neq \theta_0$, then $L_1 = 1$. On the other hand, the *same* improper uniform distribution on the values $\theta \neq \theta_0$ can equally well be represented by the density $\pi_1(\theta)d\theta = kd\theta$, where k is any fixed positive constant. (To verify that the improper prior densities $d\theta$ and $kd\theta$ represent the same improper prior distribution under the alternative hypothesis, it is sufficient to note that both of these densities yield the same posterior distribution under the alternative hypothesis for any possible observations.) But under the density $kd\theta$, it is found that $L_1 = k$. Hence, the posterior odds Δ can be taken to have any desired value whatsoever, which means that the posterior distribution is not well defined for this improper prior distribution.

In summary, when diffuse prior distributions are used in Bayesian inference, they must be used with care. Although they can serve as convenient and useful approximations in some estimation problems, they are never appropriate for tests of significance. Under no circumstances should they be regarded as representing ignorance.

2. WHAT TO DO WHEN THE MODEL SEEMS TO BE WRONG

It should always be kept in mind that the assignment of a prior distribution to the parameter θ induces a predictive distribution for the observation Y. If an observed value $Y = y$ having very small relative likelihood under this predictive distribution is observed, then we are led to consider the following three possible explanations: (a) a rare event has occurred, (b) our prior distribution was poorly specified in the sense that it assigned relatively small likelihood to the true value of θ, or (c) our statistical model was incorrect in the sense that Y was generated in accordance with some probability distribution not included in the parametric family specified by the model.

All good Bayesian statisticians reserve a little pinch of probability for the possibility that their model is wrong. Usually this possibility does not have to be acknowledged formally in the statistical analyses that are carried out, but every now and then, when an unusual value of Y is observed, it is important to reconsider the model and to evaluate that little pinch of probability.

Suppose, for example, that in a test of significance the predictive distribution of Y is discrete under both the null hypothesis H_0 and the alternative hypothesis H_1, and let $L_0(y)$ and $L_1(y)$ be the probability mass functions representing these distributions. Suppose that if a value $Y = y$ is observed for which both $L_0(y)$ and $L_1(y)$ are very small, then we will explicitly acknowledge that instead of having assigned probabilities π_0 and $1 - \pi_0$ to H_0 and H_1, we had actually assigned probabilities $\pi_0(1 - \epsilon)$ and $(1 - \pi_0)(1 - \epsilon)$, and we had assigned the remaining probability ϵ to some third (otherwise unspecified) possibility under which the observed value $Y = y$ occurs with probability 1. In other words, if $Y = y$ is very unlikely under both H_0 and H_1, then we feel that there may be some other perfectly reasonable explanation of why we obtained this observed value (even though we don't know what that explanation might be). Then

$$P(H_0 \mid Y = y) = \frac{\pi_0(1 - \epsilon)L_0(y)}{\pi_0(1 - \epsilon)L_0(y) + (1 - \pi_0)(1 - \epsilon)L_1(y) + \epsilon}. \tag{2}$$

Suppose now, as in Section 1 of Shafer's article, that both $L_0(y)$ and $L_1(y)$ are very small, that $L_0(y)$ is of the order of magnitude of ϵ, and that $L_1(y)$ is of an even smaller order of magnitude. In symbols, suppose that $L_0(y) = k\epsilon + o(\epsilon)$ for some positive number k and that $L_1(y) = o(\epsilon)$. Then, from (2),

$$\lim_{\epsilon \to 0} \frac{P(H_0 \mid Y = y)}{P(H_1 \mid Y = y)} = \infty, \tag{3}$$

but

$$\lim_{\epsilon \to 0} P(H_0 \mid Y = y) = \frac{k\pi_0}{k\pi_0 + 1}. \tag{4}$$

Thus, if a third possibility other than H_0 or H_1 is allowed when the observed value y is unlikely under H_0 and even more unlikely under H_1, then the posterior probability of H_0 will be given by (4) rather than being close to 1. A subjective evaluation of k is essential here, where

$$k = \frac{P(Y = y \mid H_0)}{P(\text{neither } H_0 \text{ nor } H_1 \text{ is correct})}. \quad (5)$$

As previously mentioned, another explanation that we might consider when we obtain an unlikely or unexpected observation is that our prior distribution was poorly specified. If a statistician repeatedly obtains unlikely or unexpected observations in his experiments, then he should begin to suspect that there may be something wrong with his view of the world and with the process by which he is specifying his prior distributions. A good statistician cannot go through life repeatedly being surprised—after a while he will learn to expect the unexpected. In other, more precise, words, the statistician will learn about the world and about his own biases, so that his prior distributions will become more realistic and the observed outcomes will not be so surprising. I have considered problems and models of this type elsewhere (DeGroot 1981).

I agree strongly with Shafer that no prior distribution should be given "complete credence." Nevertheless, as I have tried to indicate here, many problems of conflicting evidence and unlikely observations can be resolved within the Bayesian framework.

3. GUILTY OR INNOCENT?

Consider now the particular problem in forensic science described by Shafer in Section 2 of his article. The question here seems to be how to specify an appropriate prior distribution for the refractive index θ of the glass fragment found on the suspect. It would appear to be inappropriate to use a general distribution $\pi_1(\theta)$ of the type given by Shafer in Figures 2 and 3 in this problem where it is possible to update the prior distribution $\pi_1(\theta)$ by gathering more information. Any good Bayesian detective would try to develop a more relevant individualized prior distribution for the suspect by considering the window glass in his residence and place of employment, as well as in the residences of his neighbors, friends, and relatives; the kind of work the suspect does; his hobbies; and so on. It is clearly important for the sake of justice that the state spend some time and effort in developing this prior distribution.

The preceding comments pertain to the investigation of the burglary and the suspect. If the suspect is on trial in a courtroom, the decision problem changes. In this setting, the null hypothesis is essentially equivalent to the suspect's guilt. Our law states that there must be a preponderance of the evidence or evidence beyond a reasonable doubt in order to decide that a defendant is guilty. This means that if the observed data have small likelihood under H_0 and we can think of a plausible alternative model under which the data would have greater likelihood, then we must reject H_0 and find the defendant not guilty. In other words, if there is some other reasonable source for the fragment of glass found on the suspect's clothing other than the broken window, and there typically would be such a source even if we don't know exactly what it is, then we must not find the suspect guilty.

4. BELIEF FUNCTIONS

In this article, Shafer has presented a lucid and elegant recapitulation of his novel theory of belief functions, and I enjoyed reading about it. I agree with many of the main points and disagree with some. I agree with Shafer that a Bayesian analysis based on diffuse prior distributions does not satisfactorily accommodate problems of conflicting evidence, but as I have tried to indicate in these comments, a careful and more realistic Bayesian approach can handle these problems.

I agree with Shafer that his theory of belief functions fills some gaps left by the sampling theory approach, but it seems that the dependence of his belief functions on "Dempster's rule of combination" is a shortcoming that limits their applicability in most statistical problems. As Seidenfeld (1979) and Diaconis (1978) have pointed out, the result obtained from Dempster's rule depends on the way in which sample data are partitioned.

Finally, I wonder whether the discount factor α introduced by Shafer in Section 3 might be related to (perhaps a distant cousin of) the factor ϵ introduced here in (2) of my comments.

In conclusion, I believe that the Bayesian approach is alive and well, and still appropriate in most problems. Like all powerful weapons, it must be used only with the utmost care and in accordance with the highest ethical standards. It is important for those who follow the Bayesian approach to keep in mind that even subjectivists have scientific responsibilities. Although the prior distributions that decision makers use in particular decision problems represent their own personal knowledge and information, they owe it to those who will be affected by their decisions to specify their prior distributions as carefully and precisely as possible: to utilize their knowledge fully and scientifically, to make use of any outside expertise that may be available, and to gather and incorporate as much additional data and information as possible.

REFERENCES

DEGROOT, M.H. (1981), "Improving Predictive Distributions," *Bayesian Statistics*, eds. Bernardo, DeGroot, Lindley, and Smith, Valencia, Spain: University Press, 383–395.

DIACONIS, P. (1978), Review of "A Mathematical Theory of Evidence" by Glenn Shafer, *Journal of the American Statistical Association*, 73, 677–678.

EDWARDS, W., LINDMAN, H., and SAVAGE, L.J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

JEFFREYS, H. (1961), *Theory of Probability* (3rd ed.), Oxford: Oxford University Press.
SAVAGE, L.J. (1961), *The Subjective Basis of Statistical Practice*, unpublished report, Dept. of Mathematics, University of Michigan, Ann Arbor.
SAVAGE, L.J., and others (1962), *The Foundations of Statistical Inference*, London: Methuen & Co.
SEIDENFELD, T. (1979), "Statistical Evidence and Belief Functions," Philosophy of Science Association, East Lansing, Michigan, *PSA 1978*, 2, 38–49.

Comment

A. P. DEMPSTER*

I am enthusiastic about Glenn Shafer's paper, "Lindley's Paradox." Shafer's specific message illuminates one of the most thought-provoking points of disagreement between frequentist and Bayesian statistics by showing that one can uphold the essential Bayesian principles of quoting posterior probabilities and using prior knowledge and still obtain answers qualitatively similar to frequentist answers. More important, I believe, is the general message that the theory of belief functions is a practical tool capable of providing simple and straightforward technological aids to professionals who must assess uncertainty in their daily work. The example comes from law, but the potential relevance to many fields such as choice of medical treatment, nuclear safety, strategic planning, and investment decision making is exciting.

HISTORY

In his historical writing Shafer (1979) traces the key elements of the theory of belief functions to James Bernoulli and J.H. Lambert in the 17th and 18th centuries. During the 1960's I reinvented the theory in the limited context of inference from a random sample, and when Shafer spent the year 1970–1971 at Harvard he started to work on his extensive development of the general theory. My original goal was to provide an alternative to R.A. Fisher's fiducial argument, remaining faithful to Fisher's reasoning principles, but weakening certain technical assumptions regarding pivotal quantities. It turned out by a happy accident that the generalized fiducial framework was big enough to include Bayesian inference as a special case (Dempster 1968). Thus, just as in Shafer's treatment of sharp null hypotheses in "Lindley's Paradox," the possibility of a unified treatment including sampling theory answers and Bayesian answers as limiting cases was exposed.

My active work in the theory of belief functions receded to a low level about the time Shafer took up the field with great vigor and skill. The main reason for my switch was a desire to devote myself to the development of parametric modeling tools, which I continue to see as the main thrust of useful statistical technology.

On the negative side, I was frustrated by the computational difficulties of carrying out inferences from my basic models (Dempster 1972). Also, it became apparent that the inferences from my models and sample data alone often yield P_* close to 0 and P^* close to 1, especially when multivariate complexity is assessed with limited data. Neither negative reason is compelling. Computing methods and power have improved to the point where effort put into the computational side of inference has a good prospect of payoff. Also, it has become increasingly clear to me that nonrobustness of statistical inferences against model failure can only sometimes be fixed by clever choice of statistical techniques. Increasingly often I encounter irreducible nonrobustness intrinsic to the data. The only solution is implicit or explicit introduction of prior knowledge. Belief functions offer a flexible and rich collection of models for representing necessary prior information. Thus, I continue to believe that belief function technology has important potential for statistical applications and for broader problems of assessing uncertainty.

MEANING OF PROBABILITY

To present a well-rounded picture of any theory of probability, I believe it is necessary to describe and analyze three basic dimensions of the theory. The first dimension is the *formal structure* of concepts, definitions, and relations together with their mathematical representations; the second is the *meaning* that the user of the theory ascribes to numerical probabilities; and the third is the *source* of any particular realization of the formal structure invoked in any application of the technology, especially the source of particular numbers representing assessed probabilities. Shafer (1976) did a masterful job along the first dimension. Recently, Shafer (1981) emphasized the second and third dimensions, but these need much more exposure and discussion of key ideas.

* Arthur P. Dempster is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138.

© Journal of the American Statistical Association
June 1982, Volume 77, Number 378
Theory and Methods Section