Lindley's Paradox

GLENN SHAFER*

A sharp null hypothesis may be strongly rejected by a sampling-theory test of significance and yet be awarded high odds by a Bayesian analysis based on a small prior probability for the null hypothesis and a diffuse distribution of one's remaining probability over the alternative hypothesis. The Bayesian analysis seems to interpret the diffuse prior as a representation of strong prior evidence, and this may be questionable. The theory of belief functions allows us to represent the strength of prior evidence more realistically. These ideas are illustrated by the problem of identifying glass by its refractive index.

KEY WORDS: Bayesian inference; Belief functions; Conflicting evidence; Dempster's rule; Discounting; Significance testing.

A sharp null hypothesis may be strongly rejected by a standard sampling-theory test of significance and yet be awarded high odds by a Bayesian analysis based on a small prior probability for the null hypothesis and a diffuse distribution of one's remaining probability over the alternative hypothesis. This disagreement between sampling-theory and Bayesian methods was first studied by Harold Jeffreys in his *Theory of Probability* (1939, Sec. 5.0 and Appendix I). And it has become widely known to the statistical community as a result of expositions by I. J. Good (1956), Dennis Lindley (1957), and Edwards, Lindman, and Savage (1963). Lindley was the first to call the disagreement a paradox, and it has come to be called "Lindley's paradox." (See, e.g., Leamer 1978, p. 105; Berger 1980, p. 107.) Lindley himself has recently (1980, p. 11) referred to the paradox as "Jeffreys' paradox," but I know of no other reference to it under this name.

In this article I contrast the Bayesian treatment of Lindley's paradox with a treatment using the theory of belief functions. In order to make the discussion as concrete as possible I concentrate on a simplified form of a practical problem in forensic science, the problem of identifying a glass fragment by its refractive index. This problem is particularly interesting because it has been claimed that the Bayesian prior under the alternative hypothesis can be based on an empirical frequency distribution. A Bayesian treatment of the problem has been given by Lindley (1977) and a sampling-theory treatment by Evett (1977).

From the point of view of the theory of belief functions, Lindley's paradox involves conflicting evidence. The statistical evidence, which is the only evidence taken into account by the sampling-theory test of significance, points rather strongly to a relatively small interval of parameter values that does not include the null hypothesis. But the diffuse distribution over the alternative hypothesis gives only a small probability to this interval, reflecting, presumably, prior evidence against it. As in every case of conflicting evidence, our final opinion will depend crucially on our judgments of the strength and reliability of the opposing items of evidence. The theme of this article is that the theory of belief functions provides a straightforward and meaningful way to express and combine these judgments, whereas the "likelihood ratios" of the Bayesian theory serve only to obscure them.

We study the Bayesian analysis of Lindley's paradox in Section 1, and its application to the refractive index problem in Section 2. We study the theory of belief functions in Section 3 and apply it to the refractive index problem in Section 4. In Section 5 we use the refractive index problem to illustrate the limitations of actual empirical frequency distributions. In Section 6 we compare the Bayesian theory and the theory of belief functions in their ability to represent evidence that falls short of providing a fully relevant empirical frequency distribution.

1. LINDLEY'S PARADOX IN ITS ABSTRACT FORM

Consider a random quantity Y that has a Gaussian distribution with unknown mean θ and known variance σ^2; we suppose that Y is either a measurement of θ or an estimate based on several measurements. Our null hypothesis is that $\theta = \theta_0$, and we wish to give a Bayesian assessment of the evidence for this null hypothesis. Following Jeffreys's advice, we do this by assigning a nonzero prior probability π_0 to the null hypothesis and distributing the rest of our prior probability over the real line according to a fairly flat probability density $\pi_1(\theta)$. Let us suppose, for simplicity, that $\pi_1(\theta)$ has a variance τ^2. Suppose $\tau^2 \gg \sigma^2$. And suppose we observe $Y = y$, where y is several σ from θ_0. A value so far from θ_0 is very unlikely on the null hypothesis, and so we will reject the null hypothesis if we use a standard sampling-theory

* Glenn Shafer is Associate Professor, Department of Mathematics, University of Kansas, Lawrence, KS 66045. Research was supported in part by National Science Foundation Grant MCS-8002213. The author has benefited from conversations with Robert Friauf, Dennis Lindley, and Amos Tversky and from written comments by the editors and referees. Figures 2, 3, and 4 are reprinted with permission of the American Society for Testing and Materials, Philadelphia, PA 19103, Copyright.

June 1982, Volume 77, Number 378
Theory and Methods Section

test based on the test statistic $(y - \theta_0)/\sigma$. But since $\tau^2 \gg \sigma^2$, the set of values of θ with high likelihood, those within several σ of y, will be given a very small prior probability by the prior density $\pi_1(\theta)$, and therefore the overall likelihood of the alternative hypothesis $\theta \neq \theta_0$,

$$L_1 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y - \theta)^2/2\sigma^2)\, \pi_1(\theta) d\theta,$$

may be even smaller—much smaller—than the likelihood of the null hypothesis $\theta = \theta_0$,

$$L_0 = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y - \theta_0)^2/2\sigma^2).$$

By choosing an example where the ratio of τ^2 and σ^2 is sufficiently large, we may make this effect as great as we wish, thus making the posterior odds

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

very large, even if the prior probability π_0 is very small. Thus we can construct examples where the Bayesian analysis shows the measurement to provide very large odds in favor of the null hypothesis θ_0 even though the test statistic $(y - \theta_0)/\sigma$ indicates that the measurement provides very strong evidence against it. This is Lindley's paradox.

Lindley's paradox is evidently of great generality; the effect it exhibits can arise whenever the prior density under an alternative hypothesis is very diffuse relative to the power of discrimination of the observations. The effect can be thought of as an example of conflicting evidence: the statistical evidence points strongly to a certain relatively small set of parameter values, but the diffuse prior density proclaims great skepticism (presumably based on prior evidence) towards this set of parameter values. If the prior density is sufficiently diffuse, then this skepticism will overwhelm the contrary evidence of the observations.

The paradoxical aspect of the matter is that the diffuse density $\pi_1(\theta)$ seems to be skeptical about all small sets of parameter values. Because of this, we are somewhat uneasy when its skepticism about values near the "observed interval" overwhelms the more straightforward statistical evidence in favor of those values. We are especially uneasy if the diffuseness of $\pi_1(\theta)$ represents weak evidence, approximating total ignorance; the more ignorant we are the more diffuse $\pi_1(\theta)$ is, yet this increasing diffuseness is being interpreted as increasingly strong evidence against the "observed interval" (compare Dempster 1971, pp. 60–61).

The paradox can arise whenever an estimate Y is exceedingly precise relative to the prior density $\pi_1(\theta)$. The source of the precision does not matter; Y may be a single measurement made with a very precise measuring instrument, or it may be precise because it is an average of many measurements. Most authors have emphasized the case where precision arises from a large sample size, but this may not be the most important case in practice. In real problems of measurement, an average does not become indefinitely more precise as the number of measurements is increased; there is always a limit to how large a sample size is sensible, and this limit is usually fairly low. (See Wilson 1952, pp. 252–254, or Baird 1962, pp. 38–40.) In the example from forensic science discussed in this article, the relative precision of Y is not due to a large sample size.

Non-Bayesian statisticians will agree, in general, with my criticism of the Bayesian analysis; they will agree that if the large dispersion of $\pi_1(\theta)$ merely expresses subjective ignorance about θ, then this dispersion should not be interpreted as evidence against values of θ near the observation y. There is one situation, however, where most statisticians will be sympathetic with the Bayesian analysis—the situation where $\pi_1(\theta)$ is based on an empirical frequency distribution. For if it seems reasonable, under the alternative hypothesis $\theta \neq \theta_0$, to regard θ as having been chosen at random from a population whose frequency distribution $\pi_1(\theta)$ is empirically known, then the Bayesian calculation of the odds L_0/L_1 will be unexceptionable even from a frequentist point of view.

In this article I argue that the Bayesian analysis is wrong-headed even in this apparently most favorable case where the prior density $\pi_1(\theta)$ is an empirical frequency distribution. I base the argument in part on the practical problems in finding an empirical frequency distribution with the reliability and detail required by the Bayesian argument. (And here that argument does turn on the fine detail of $\pi_1(\theta)$.) But I also argue that we should refuse in principle to give any frequency distribution such complete credence as the Bayesian argument requires.

2. A PROBLEM IN FORENSIC SCIENCE

In a 1977 paper in *Biometrika*, Lindley studies the following problem. We measure the refractive index of the glass of a window broken during a burglary and the refractive index of a fragment of glass found on a suspect's clothing. Relative to the known frequency distribution of refractive indices for window glass, the two measurements are remarkably close. But relative to measurement error, they are rather far apart. Are the two indices the same? How do we combine the conflicting evidence?

Lindley gives a Bayesian solution to this problem under the quite reasonable assumption that both refractive indices are measured with error, the error possibly being greater in the case of the fragment. We will find it useful, however, to think about the problem under the simplifying assumption that the refractive index of the window has been measured without error, for then the Bayesian solution takes exactly the form that we have just studied.

Suppose, indeed, that the refractive index of the broken window at the scene of the burglary has been measured without error and found to have the value θ_0, uniformly throughout the window. Denote by θ the unknown value of the refractive index of the fragment found on the sus-

pect. A Bayesian will assign a certain prior probability π_0 to the null hypothesis $\theta = \theta_0$, equal, presumably, to the prior probability that the fragment on the suspect's clothing came from the broken window. Conditionally on the alternative hypothesis $\theta \neq \theta_0$, Bayesians will suppose that θ is a random refractive index from the known distribution $\pi_1(\theta)$ of refractive indices of window glass, and they will distribute the prior probability $1 - \pi_0$ assigned to the alternative hypothesis according to $\pi_1(\theta)$. We may assume that the measurement or measurements of θ result in an estimate Y that has a Gaussian distribution with mean θ and known variance σ^2, and this variance σ^2 will in fact usually be much smaller than the variance τ^2 of $\pi_1(\theta)$. So Lindley's paradox can arise: because of the large dispersion of $\pi_1(\theta)$ relative to σ^2, an observation $Y = y$ may produce impressive odds L_0/L_1 in favor of $\theta = \theta_0$ even though y is several σ from θ_0. What are we to think in such a case?

Lindley argues, essentially, that the Bayesian result should be taken at face value; we should sometimes conclude that the odds strongly favor $\theta = \theta_0$ even though this hypothesis would be rejected by the usual significance test. (He cautions, though, that in such cases, where "the data are unusual on both hypotheses," it would be sensible "to check that all the assumptions made were reasonably satisfied, or whether some hitherto unexpected hypothesis obtained" (see Lindley 1977, p. 209.) Is this reasonable? Can we really ever credit such a Bayesian analysis? Will our confidence in the relevance of the purported frequency distribution $\pi_1(\theta)$ ever be sufficient to support such reasoning?

Imagine a forensic expert who bases his courtroom testimony on such a Bayesian analysis. Measurements of the refractive index of the fragment, he testifies, result in very great odds for the hypothesis that it came from the broken window at the scene of the crime. Pressed by defense counsel, he goes into more detail. His measurements of the refractive index of the fragment gave a random value that was distinctly different from the refractive index of the broken window, he concedes, so distinctly different that the two indices can be the same only if there were extraordinary errors in his measurements. But it would be an even more extraordinary coincidence, he continues, for the fragment's refractive index to come as close as it does to the window's refractive index were it from some random source other than the window. On balance, then, the odds favor its being from the window.

Why is this so unconvincing? It is unconvincing because it fails to address the questions raised by the conflict in the evidence. The experts are weighing two conflicting items of evidence—the evidence from the measurements and the evidence from the empirical frequency distribution. But they are doing so only in terms of the relative magnitudes of the coincidences suggested by the two. They ought also to weigh the reliability of the measurements. And they ought to weigh the relevance, as a guide to belief, of a frequency distribution that asks us to regard every possibility as an improbable coincidence. (It is interesting, in this connection, to note that the testimony we have imagined would probably not be admissible in a court in the United States, precisely because no witness, not even an expert witness, is allowed to weigh conflicting evidence. This is the prerogative of the jury. See Louisell, Kaplan, and Waltz 1976, pp. 13, 1023.)

If we reject the Bayesian analysis, then how do we combine the two items of evidence?

Lindley notes (and criticizes; see p. 209 of this paper) a sampling-theory approach developed by a British forensic scientist, I. W. Evett (1977). In this approach we begin by performing a standard sampling-theory test of significance of the hypothesis $\theta = \theta_0$ (or, in the realistic case where both refractive indices are measured with error, the hypothesis that the two are the same). If this hypothesis is rejected, then the evidence is disregarded; if it is not rejected, then attention is drawn to the small intrinsic probability of a random index coming so close to θ_0. Thus if the evidence is internally conflicting it will not be brought to court. But if θ is within measurement error of θ_0, then it will be suggested that the null hypothesis is very probable.

Evett's approach has a certain appeal, but it is awkwardly abrupt in its switch from totally discounting the frequency distribution $\pi_1(\theta)$ (when $(y - \theta_0)/\sigma$ is significant) to discounting it not at all (when $(y - \theta_0)/\sigma$ is not significant). Let us see if the theory of belief functions can do better.

3. INTRODUCTION TO BELIEF FUNCTIONS

The theory of belief functions is described in detail in my monograph *A Mathematical Theory of Evidence* (1976). It differs from the Bayesian theory in that it permits nonadditive degrees of belief and emphasizes the combination of evidence instead of conditioning.

Both the Bayesian theory and the theory of belief functions can be thought of as constructive theories. In both theories, we construct probability judgments by locating our evidence on a scale of canonical examples. The Bayesian theory uses canonical examples where the truth is generated according to known chances; when we select a Bayesian probability distribution P we are saying that our evidence is comparable in strength to knowledge that the truth is generated according to chances given by P. The theory of belief functions uses canonical examples where the evidence is a message whose meaning depends, in a certain sense, on known chances.

Suppose someone chooses a code at random from a list of codes, uses the chosen code to encode a message, and then sends us the result. We know the list of codes and the chance of each code being chosen—say the list is $c_1, \ldots, c_n$, and the chance of c_i being chosen is p_i. We decode the encoded message using each of the codes and find that this always produces a message of the form "the truth is in A" for some nonempty subset A of a finite set of possibilities Θ. Let A_i denote the subset we get

when we decode using c_i, and set

$$m(A) = \sum\{p_i \mid 1 \le i \le n; A_i = A\}$$

for each $A \subset \Theta$. The number $m(A)$ is the sum of the chances for those codes that indicate A was the true message; it is, in a sense, the total chance that that true message was A. Notice that $m(\phi) = 0$ and that the $m(A)$ sum to one. The quantity

$$\text{Bel}(A) = \sum_{B \subset A} m(B) \tag{3.1}$$

is, in a sense, the total chance that the true message implies A. If the true message is infallible and the coded message is our only evidence, then it is natural to call Bel(A) our degree of belief that the truth lies in A. (See Shafer 1981 for a fuller discussion.)

Any function Bel of the form (3.1) is called a *belief function*. Since we can tell the story of the coded message with whatever values of the $m(A)$ we please (provided $m(\phi) = 0$ and the $m(A)$ sum to one), this story provides a canonical example for every belief function on a finite set. Of course we will seldom or never encounter in practice a situation in which our evidence really does consist of a coded message and all the assumptions of the canonical example are satisfied. But is also rare that our evidence amounts to knowledge of a chance distribution according to which the truth has been or will be generated. In both cases the canonical examples are meant not as realistic examples but as standards for comparison.

The subsets A of Θ for which $m(A) > 0$ are called the *focal elements* of the belief function Bel. The relations among these focal elements say something about how we are thinking about our evidence. Focal elements A and B that are disjoint suggest internal conflict in our evidence: we are comparing that evidence to a message that might mean A or might mean B, where B contradicts A. Focal elements A and B that are nested, say $A \supset B$, suggest more consonant evidence: there is evidence for the truth being in A and further evidence that does not disagree but is more specific in that it points to the truth being in B. When all Bel's focal elements are nested, say $A_1 \subset A_2 \subset \cdots \subset A_n$, we call Bel *consonant*; we see from (3.1) that the degree of belief given to a subset A by such a belief function is the same as the degree of belief given to the largest A_i contained in A.

We might, if the evidence is both very conflicting and very specific, choose the numbers $m(A)$ so that only singletons are focal elements. If we do this, then we can write (3.1) as

$$\text{Bel}(A) = \sum_{\theta \in A} p(\theta),$$

where $p(\theta) = m(\{\theta\})$. And this means that Bel is a Bayesian probability distribution. In general, however, a belief function will not be a Bayesian probability distribution. It will satisfy Bel(ϕ) = 0 and Bel(Θ) = 1, but it will not always be additive—it will not always satisfy Bel($A \cup B$) = Bel(A) + Bel(B) when $A \cap B = \phi$. In particular, there may be A for which Bel(A) and Bel($\bar{A}$) add to less than one.

Our task, when we assess evidence using belief functions, is to choose values of $m(A)$ that make the canonical example most like that evidence. But how do we do this? In complicated problems it is absurd, surely, to suppose that we can simply look at our evidence holistically and write down the best values for the $m(A)$. So we need a theory of belief functions—a set of tools for constructing complicated belief functions from simpler, more elementary judgments.

Dempster's rule of combination (Dempster 1967, pp. 335–338) is the most important single tool of the theory of belief functions. This rule tells us how to combine a belief function Bel_1 (with m values $m_1(A)$, say) representing one body of evidence with a belief function Bel_2 (with m values $m_2(A)$) representing an unrelated body of evidence so as to obtain a belief function Bel (with m values $m(A)$) representing the pooled evidence. The idea underlying the rule is that the unrelatedness of the two bodies of evidence makes pooling them like combining two stochastically independent randomly coded messages. We should, that is to say, combine the canonical examples corresponding to the two bodies of evidence by supposing that the two random choices of codes are stochastically independent. It is easy to see how this leads to a rule for obtaining the $m(C)$ from the $m_1(A)$ and the $m_2(B)$. Denote by $c_1, \ldots, c_n$ and by $p_1, \ldots, p_n$ the codes and their chances in the case of the first message, and by $c'_1, \ldots, c'_m$ and $p'_1, \ldots, p'_m$ the codes and their chances in the case of the second. Then independence means that there is a chance $p_i p'_j$ that the pair (c_i, c'_j) of codes will be chosen. But decoding may now tell us something. If the message A_i we get by decoding the first message with c_i contradicts the message B_j we get by decoding the second message with c'_j (i.e., if $A_i \cap B_j = \phi$), then we know that (c_i, c'_j) could not be the pair of codes actually used. So we must condition the chance distribution, eliminating such pairs and multiplying the chances for the others by κ, where

$$\kappa^{-1} = \sum\{p_i p'_j \mid 1 \le i \le n; 1 \le j \le m; A_i \cap B_j \ne \phi\}$$
$$= \sum\{m_1(A)m_2(B) \mid A \subset \Theta; B \subset \Theta; A \cap B \ne \phi\}.$$

Notice also that if the first message is A and the second message is B, then the overall message is $A \cap B$. Thus the total chance of the overall message being C is

$$m(C) = \kappa \sum\{p_i p'_j \mid 1 \le i \le n; 1 \le j \le m; A_i \cap B_j = C\}$$
$$= \kappa \sum\{m_1(A)m_2(B) \mid A \subset \Theta; B \subset \Theta; A \cap B = C\}. \tag{3.2}$$

Formula (3.2) is Dempster's rule.

Suppose we construct a belief function Bel but then have second thoughts about the soundness of the judgments that went into the construction and wish to discount the degrees of belief Bel gives. This is easily done within the theory. If we want to discount by a factor α,

$0 < \alpha < 1$, we simply reduce each of Bel's m values $m(A)$ to $(1 - \alpha)m(A)$ and then increase the m value for Θ by α. The result is a belief function Bel^α related to Bel by

$$\text{Bel}^\alpha(A) = (1 - \alpha)\text{Bel}(A)$$

for all proper subsets A of Θ. ($\text{Bel}^\alpha(\Theta) = \text{Bel}(\Theta) = 1$, of course.) Discounting is an essential element in the belief-function treatment of conflicting evidence. For discount factors are the theory's way of expressing judgments about the relative reliability of conflicting items of evidence.

In the next section we use the theory of belief function in a continuous rather than a discrete setting. In the continuous setting we do not, of course, add m values. Instead, we integrate. But the basic ideas remain the same.

4. ANALYSIS BY BELIEF FUNCTIONS

In order to apply the theory of belief functions to our problem in forensic science, we must represent each of our two items of evidence—the evidence from the measurement and from the empirical frequency distribution—by a belief function, and then combine the two belief functions by Dempster's rule.

In this section we begin by representing each item of evidence in the simplest possible way: the measurement is represented by a belief function corresponding to the usual nested confidence intervals, and the frequency distribution is taken at face value. The results of this approach are in qualitative agreement with those of the Bayesian approach. We next explore what happens when our reservations about the relevance of the frequency distribution are expressed by discounting the belief function corresponding to it. This, it turns out, gives results that are very intuitive and that resemble in some respects the results of Evett's sampling-theory procedure.

As evidence, the measurement y gives us reason to believe that θ is close to y, and no particular reason to believe that it is far from y. So it seems reasonable to represent this evidence by a consonant belief function, say by

$$\text{Bel}_1(A) = \int_a^a \eta(x)dx,$$

where $\eta(x)$ is the density for the Gaussian distribution with mean zero and variance σ^2, and

$$a = \sup\{\delta \mid [y - \delta, y + \delta] \subset A\}.$$

This belief function can be described by saying that our belief is divided into increments of the form $\eta(x)dx$ and that the increment $\eta(x)dx$ is committed to θ being within $|x| + dx$ of the measurement y. The interpretation in terms of the metaphor of a randomly coded message is that the measurement is like a message that says θ is in a symmetric interval about y. Our uncertainty about the message is uncertainty about the width of the interval referred to; for every $x > 0$, there is a chance $2\eta(x)dx$ that the reference is to an interval of width between $2x$ and $2(x + dx)$.

We are supposing that the refractive index of the fragment is either θ_0 or else is drawn at random from the distribution $\pi_1(\theta)$ of refractive indices. The beliefs about θ corresponding to this supposition can be represented by the belief function

$$\text{Bel}_2(A) = \begin{cases} 0 & \text{if } \theta_0 \notin A \\ \int_A \pi_1(z)dz & \text{if } \theta_0 \in A. \end{cases}$$

This belief function corresponds to distributing one's belief according to the density $\pi_1(z)$ except that all the belief is allowed to move to the point θ_0; in other words, the increment $\pi_1(z)dz$ is committed to θ being in the set $[z, z + dz] \cup \{\theta_0\}$. In terms of randomly coded messages: the message is that the index is either θ_0 or some other value z; the uncertainty about the meaning of the message is uncertainty about the identity of z; and this uncertainty is described by the density $\pi_1(z)$.

In order to simplify later formulas, let us assume that $y < \theta_0$, and let us shift the scale of measurement so that $y = 0$.

Let us first consider the case where Bel_1 and Bel_2 are combined without discounting Bel_2. In this case Dempster's rule amounts, except for a renormalization, to multiplying each increment $\eta(x)dx$ by each increment $\pi_1(z)dz$ and committing each resulting increment $\eta(x)dx\ \pi_1(z)dz$ to θ being in the set (see Figure 1)

$$[-|x| - dx, |x| + dx] \cap ([z, z + dz] \cup \{\theta_0\}).$$

This results in total belief of measure

$$R = 2\int_{x>\theta_0} dx \int_{|z|<x} dz\ \eta(x)\ \pi_1(z)$$

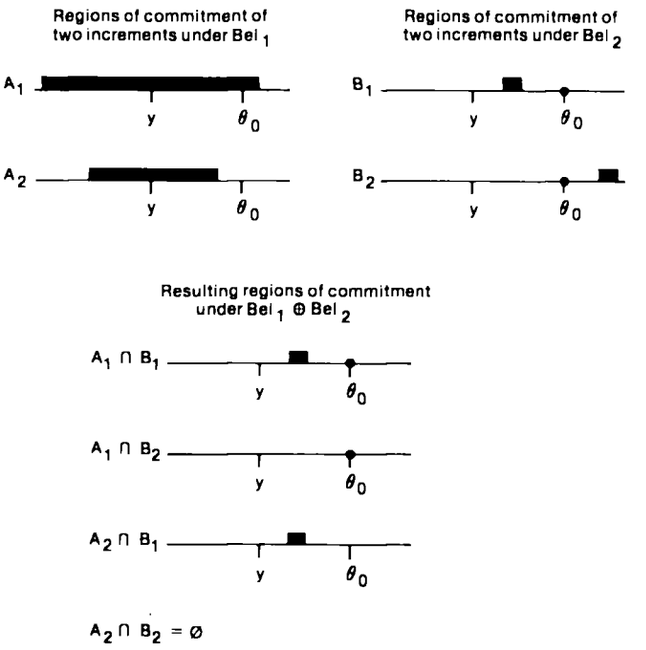

Figure 1. Some Aspects of the Combination of Bel_1 *and* Bel_2

committed to $\theta = \theta_0$, but also in total belief

$$S = 2\int_{0<x<\theta_0} dx \int_{|z|>x} dz\, \eta(x)\, \pi_1(z)$$

committed to θ being in the empty set. Dempster's rule requires that we discard the belief committed to the empty set and renormalize; so our final degree of belief in the hypothesis $\theta = \theta_0$ is $R/(1 - S)$.

Notice that as $\pi_1(z)$ is made more and more diffuse,

$$R \to 2\int_{x>\theta_0} \eta(x)dx = \hat{p}$$

and

$$S \to 2\int_{0<x<\theta_0} \eta(x)dx = 1 - \hat{p},$$

where $\hat{p}$ is the observed significance level achieved by the observation $Y = 0$ with respect to a two-sided test of $\theta = \theta_0$. And hence our degree of belief in $\theta = \theta_0$, $R/(1 - S)$, tends to 1 regardless of the distance of θ_0 from 0, the value of the measurement. This is in qualitative agreement with the Bayesian result previously discussed.

Now suppose we discount Bel_2 by the discount rate α before applying Dempster's rule. Then α of our belief is assigned in accordance with Bel_1 and the remaining $1 - \alpha$ is assigned as described in the preceding paragraph; this results in only $(1 - \alpha)S$ of our belief being assigned to the empty set, so that the renormalizing constant is $1/(1 - (1 - \alpha)S)$.

If $\pi_1(\theta)$ is very diffuse and the observation $Y = 0$ is many σ from θ_0, then S will be nearly one; this means that nearly all of the $(1 - \alpha)$ will be assigned to the empty set and therefore discarded. Only the α assigned in accordance with Bel_1 will survive, and since the renormalizing constant $1/(1 - (1 - \alpha)S)$ will be approximately $1/\alpha$, the final result will approximate Bel_1. In particular, we will very strongly doubt the hypothesis $\theta = \theta_0$.

If, on the other hand, $\pi_1(\theta)$ is very diffuse but θ_0 is relatively close to 0, then we may obtain a fairly strong degree of belief in $\theta = \theta_0$. That degree of belief is, in general,

$$\frac{(1 - \alpha)R}{1 - (1 - \alpha)S},$$

and as $\pi_1(\theta)$ becomes more diffuse this tends to

$$\frac{(1 - \alpha)\hat{p}}{1 - (1 - \alpha)(1 - \hat{p})} = \frac{(1 - \alpha)\hat{p}}{\alpha + (1 - \alpha)\hat{p}}.$$

This degree of belief cannot exceed $1 - \alpha$, but if α is not too great and $\hat{p}$ is fairly large relative to α, then it may be impressively large.

Evett's procedure, as we mentioned above, consists of drawing attention to the very improbable coincidence suggested by the diffuse frequency distribution $\pi_1(\theta)$ only if $\hat{p}$ is large enough to avoid conventional significance, say $\hat{p} > .1$. Since the improbability of this coincidence is then understood as a measure of the force of the evidence for $\theta = \theta_0$, Evett's procedure seems fully appropriate only if the condition $\hat{p} > .1$ is sufficient to assure that the degree of belief in $\theta = \theta_0$ is impressively large, say greater than .999. This would require, however, that

$$\frac{(1 - \alpha)(.1)}{\alpha + (1 - \alpha)(.1)} \geq .999,$$

or, approximately, $\alpha \leq .0001$, a very low discount rate. If the frequency distribution seems to deserve more severe discounting than this, then we may feel that Evett's procedure, though far more conservative than the Bayesian solution, is not conservative enough.

There may be occasions when Bel_1, the belief function based on the measurements, also seems to deserve discounting. This will, of course, make a substantial degree of belief in $\theta = \theta_0$ yet more difficult to attain. But so long as the discount rate for Bel_1 is an order of magnitude smaller than the discount rate for Bel_2, our qualitative conclusions will remain unchanged.

5. THE ELUSIVE EMPIRICAL FREQUENCY DISTRIBUTION

We have been taking for granted that a frequency distribution $\pi_1(\theta)$ has been empirically observed, that its detail is sufficiently fine for the purposes of the Bayesian analysis, and that its relevance to our alternative hypothesis is clear enough for it to deserve our trust except for a modest discounting. But when we look more closely at the available evidence, we find that it falls far short of this ideal.

Heideman suggested (1975, p. 107) that a large data bank should be organized to collect results of refractive index measurements from forensic scientists. Apparently no such data bank now exists, but several histograms based on modest surveys have been published by British government scientists. One of these histograms, shown in Figure 2, is based on the refractive indices of 939 specimens of broken window glass collected by fire brigades in England and Wales in 1968 and 1969. Another, shown in Figure 3, is based on the refractive indices of 551 glass fragments found in 100 men's suits examined at random during a three month period around 1970 in a large drycleaning establishment in southern England. Only 37 of the 100 suits were apparently free of glass fragments, but two of the suits contained 46 percent of the fragments, and one of these contained 166. Figure 4 is a histogram of the 166 fragments found in the single suit.

In his 1977 paper on this problem (p. 212), Lindley uses the histogram in Figure 2 in applying his Bayesian analysis to an example from Evett (1977). In this example, the refractive indices of both the broken window and the fragment are measured with error. In order to make the role of the histogram in the Bayesian analysis as easily understood as possible, I will adapt the example, using similar numerical values but assuming that the value θ_0 of the refractive index of the broken window is known without error.

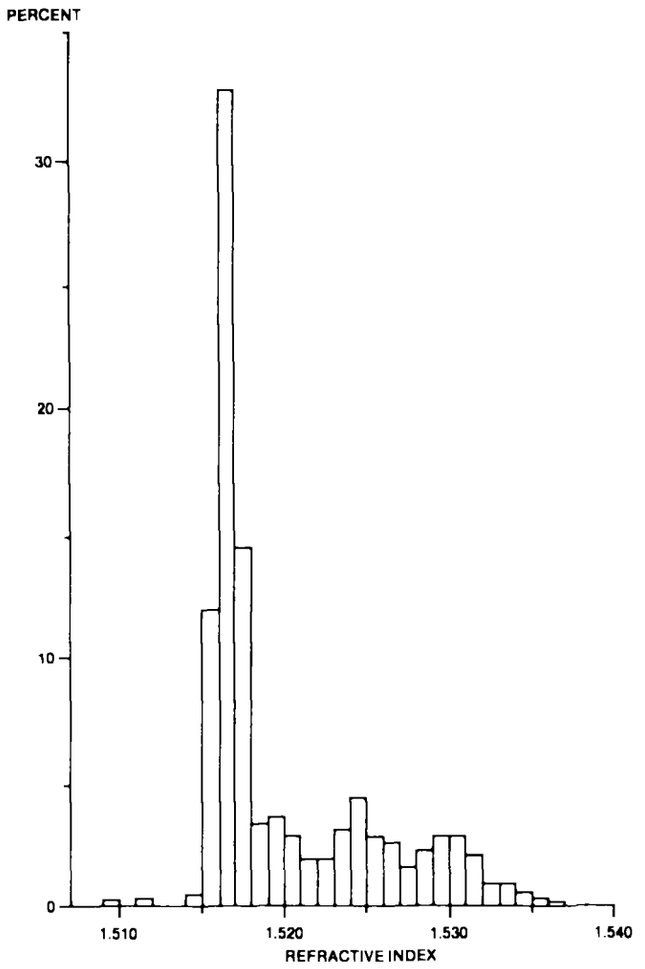

Figure 2. Refractive Indices for 939 Specimens From British Fire Survey (Dabbs and Pearson 1972)

Suppose, then, that $\theta_0 = 1.518458$, while measurement of the fragment yields an estimated refractive index $y = 1.518472$, with standard error $\sigma = .0000219$. Then, following the Bayesian analysis described in Section 1, we calculate

$$L_0 = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\theta_0)^2}{2\sigma^2}\right) \approx 14{,}850$$

and

$$L_1 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right) \pi_1(\theta)d\theta \approx \pi_1(1.518472).$$

The histogram in Figure 2 shows about 3 percent of the refractive indices to fall in the interval from 1.518 to 1.519. Assuming that $\pi_1(\theta)$ is indeed flat in this region, the value of $\pi_1(1.518472)$ will be about $.03/.001 \approx 30$. So on the Bayesian analysis this evidence yields odds $L_0/L_1 \approx 14{,}850/30 \approx 500$ for the hypothesis $\theta = \theta_0$. (This conclusion in favor of the null hypothesis does not conflict with the classical significance test in this case, for $y - \theta_0/\sigma \approx .64$.)

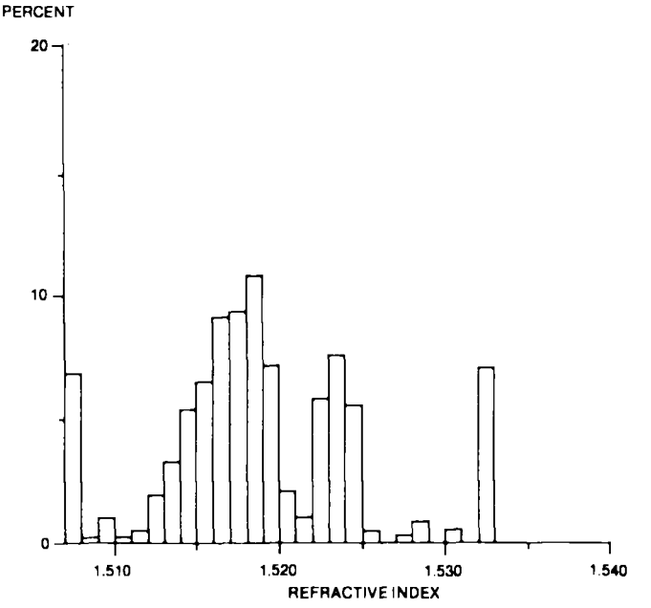

Figure 3. Refractive Indices for 551 Fragments From 100 Men's Suits (Pearson et al. 1971)

The Bayesian analysis turns on the fact that the observation y is very unlikely under the diffuse distribution $\pi_1(\theta)$. This becomes clearer if we recognize that the observation is actually discrete rather than continuous. Our estimate $y = 1.518472$ is recorded to the nearest .000001. Perhaps it would be reasonable to say we have observed y to be in the interval from 1.5184715 to 1.5184725, which has width $.000001 \approx .05\sigma$. But this may be an exagger-

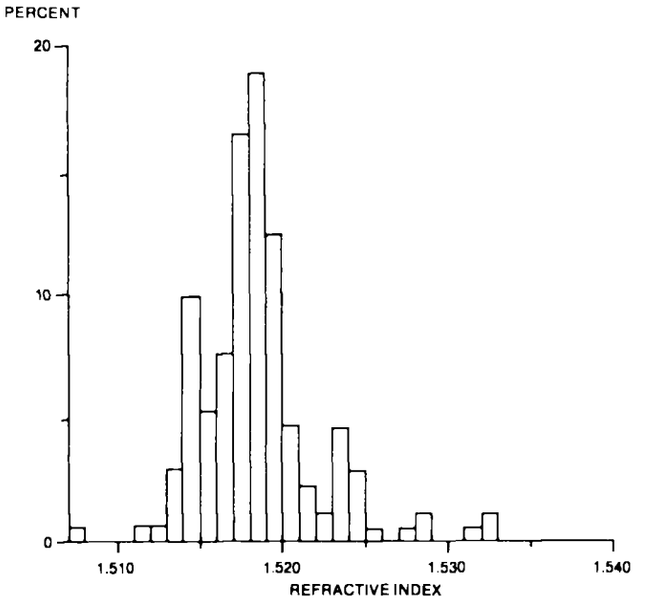

Figure 4. Refractive Indices for 166 Fragments From One of the 100 Suits of Figure 3 (Pearson et al. 1971)

ation. To be on the conservative side in the discussion that follows, let us say that we have observed y to be in the interval 1.51847 to 1.51848, which has width .00001 $\approx .5\sigma$. The crux of the Bayesian analysis is that the probability of this "observed interval" is .00001 $\times L_0 = .1485$ under the null hypothesis and .00001 $\times L_1 = .0003$ under the alternative hypothesis.

Let us review the argument leading to the probability .0003. First we must conclude that the observation $y = 1.518472$ is meaningful to five decimal places. This is reasonable, for if y is not meaningful to at least this many decimals then its standard deviation $\sigma = .0000219$ is fraudulent. Then we must conclude that the chance of the measured index being 1.51847 to five decimals is about the same as the chance of the true index being 1.51847 to five decimals. This is reasonable if the distribution of true indices is flat at this scale. Then we must conclude from the histogram that about 3 percent of all true refractive indices fall between 1.518 and 1.519. This is reasonable provided we grant the general relevance of the histogram. (There are some minor difficulties: the figure 3 percent is obtained visually from the histogram, the histogram is based on a random sample rather than on a population, and the histogram actually represents measured values rather than true values. But the uncertainties arising from these difficulties are probably not great enough to make the figure of 3 percent any less precise than its one significant figure suggests.) Finally, from the datum that 3 percent of indices fall between 1.518 and 1.519, we must infer that a hundredth as many fall between 1.51847 and 1.51848. We must, that is to say, assume that the distribution of true indices is flat rather than lumpy from the third to the fifth decimal places.

Two objections to the Bayesian analysis emerge from our review: (a) We do not know that the frequency distribution estimated by Figure 2 is flat on a fine scale. (b) This distribution has only limited relevance to our problem.

5.1 Is the True Frequency Distribution Lumpy?

The histogram in Figure 2 hardly provides any evidence for the distribution of refractive indices being flat in the fifth and sixth decimal places. Nor could we hope for such evidence from a closer examination of the fire brigade survey; that survey produced only about .03 $\times$ 939 $\approx$ 28 measurements in the interval from 1.518 to 1.519, and these were measurements, not true refractive indices. Clearly any belief that the distribution is flat on this scale must be based on an understanding of glass manufacture rather than on this survey or other random observations.

When we turn to what forensic scientists do know about current practice in glass manufacture, we find reason to think that the distribution may actually be lumpy. The refractive index of glass is sensitive to the composition of the glass and thus does tend to fluctuate continuously in the process of manufacture. But modern manufacturers use the refractive index as a monitor in quality control and have tended to control it more and more precisely. Dabbs and Pearson (1972, p. 75) accounted for the concentration of refractive indices from 1.515 to 1.518 in Figure 2 by the fact that the window glass manufactured in England and Wales since 1945 has been held in this range. And recent papers in the forensic science literature (Dabbs et al. 1973; Reeve, Mathiesen, and Fong 1976) have demonstrated that manufacturers now produce large quantities of glass with variation in refractive index too small to be detected. Because of these large quantities of glass with precise refractive indices, the overall distribution of indices may now be lumpy rather than flat.

5.2 Is Figure 2 Relevant?

We have been following Evett and Lindley in supposing that the distribution of refractive indices of window glass estimated by Figure 2 is an appropriate model for our distribution of belief concerning the refractive index θ under the alternative hypothesis $\theta \neq \theta_0$. But a moment's reflection shows this is not so. Figure 2 is based on a survey of window glass, and the fragment found on the suspect's clothing need not, under the alternative hypothesis, be window glass.

A comparison of Figures 2 and 3 shows that the distribution of refractive indices of window glass may in fact be quite different from the distribution of refractive indices of glass found in clothing. And this difference may render a Bayesian analysis based on Figure 2 very misleading. In our numerical example, Figure 2 led us to think that there was only a 3 percent chance of a random refractive index falling in the interval from 1.518 to 1.519, and this relative rarity of refractive indices in this interval contributed sharply to the 500 to 1 odds in favor of $\theta = \theta_0$. But Figure 3 would lead us to believe that the interval 1.518 to 1.519 is the most common interval, containing over 10 percent of the refractive indices of fragments found on clothing. The Bayesian odds based on this figure would be less than 150 to 1.

Should we then use Figure 3 as our frequency distribution $\pi_1(\theta)$? Most of us would be reluctant to take even this histogram too seriously. Comparison of it with Figure 4 will be enough to make us uneasy about having any very definite expectations even as to what the refractive indices from the next 100 suits at this same dry cleaning establishment will be like. And what relevance does this establishment's experience have for fragments found in clothing that is not drycleaned—or in clothing in the U.S.—and so forth and so on? All in all, our empirical frequency distribution has proven a chimera.

5.3 Other Complications

In the preceding paragraphs we have not attempted a comprehensive or fully informed account of the problem of identifying glass fragments by their refractive index. We have merely used the problem to illustrate how difficult it is for an empirical frequency distribution to carry

the weight assigned to the prior density $\pi_1(\theta)$ by the Bayesian analysis of Lindley's paradox. But in the course of discussing the refractive index problem we have passed over some complications that call out for acknowledgment.

First, it is not always true that the variation of refractive indices within a single pane of window glass is undetectable. See Dabbs and Pearson (1970) and Evett (1977).

Second, as Figures 3 and 4 suggest, it is common to find several glass fragments, with various refractive indices, on a suspect's clothing. And the chance of one of the fragments accidentally matching a broken window is obviously a function of the number of fragments.

Both Bayesian and non-Bayesian analyses of concrete examples would no doubt encounter these or other complications. But these complications do not affect our basic point: no empirical frequency distribution is available that can sensibly be used in the role of $\pi_1(\theta)$.

6. PROBABILITY JUDGMENTS FROM THE FREQUENCY EVIDENCE

It is, of course, no embarrassment to the Bayesian theory that convincing empirical frequency distributions are not always available to serve as priors. Bayesians require only degrees of belief. In this problem they need only to supply a density $\pi_1(\theta)$ that gives their prior degrees of belief about the refractive index under the alternative hypothesis, and this density need not have any frequency interpretation.

But with a subjective density $\pi_1(\theta)$ we still face Lindley's paradox. This subjective density $\pi_1(\theta)$ will have to be spread over the same range, from 1.51 to 1.54, as the empirical frequency distribution. If we make $\pi_1(\theta)$ flat rather than lumpy down to the fifth and sixth decimal places, then we will be indicating strong evidence against the index being in the observed interval. The lesson of Section 5 is that this indication is not justified by the actual evidence. But how can Bayesians avoid it? They will be hard put to justify a lump in $\pi_1(\theta)$ at the precise position of the observed interval.

(The Bayesian can, of course, incorporate into the analysis the idea that there is a true frequency distribution and that this true distribution might have a lump at the precise position of the observed interval. But if this idea of a true frequency distribution is brought in, then the subjective distribution $\pi_1(\theta)$ will presumably be an average over all possibilities for the true frequency distribution. And if we have no specific evidence as to where the lumps are, then averaging over the different possible positions for the lumps will lead us back to a flat density for $\pi_1(\theta)$.)

In Section 3 I suggested that the Bayesian theory should be thought of as a constructive theory that compares all evidence to knowledge of chances. No matter what our evidence is, a Bayesian analysis will end up saying it is equivalent in weight to knowledge that the truth is generated according to certain chances. In our problem, this means that any evidence about what refractive indices might be expected under the alternative hypothesis will end up being equated with a definite distribution for refractive indices under the alternative hypothesis. And once we are committed to such a distribution, we seem compelled to make it relatively flat and diffuse.

Table 1. Example of Degrees of Beliefs for Intervals of Various Lengths in the Range from 1.51 to 1.53

ϵ	.001	.0005	.0001	.00005	.00001
$P^*(A_\epsilon)$	.1180	.0956	.0786	.0774	.0774
$\text{Bel}(\bar{A}_\epsilon)$	.8820	.9044	.9214	.9226	.9226

The theory of belief functions, with its more flexible scale of canonical examples, allows us to escape from this trap. It allows us to compare our evidence to an uncertain message that warrants only limited degrees of belief as to what refractive index should be expected under the alternative hypothesis.

Here is a crude illustration of how evidence to the effect that refractive indices tend to be distributed extensively within the limits 1.51 to 1.53 might be represented by a randomly coded message. Divide the interval from 1.51 to 1.53 into 4 intervals of width .005 each. Divide it similarly into 20 intervals of width .001, 40 intervals of width .0005, 200 intervals of width .0001, and 400 intervals of width .00005. This gives a total of 4 + 20 + 40 + 200 + 400 = 664 intervals. Let us say that the randomly coded message has a chance .0014 of meaning that the refractive index of a "random" fragment from someone's clothing will be in any one of these 664 intervals and a chance $1 - (664)(.0014) = .9296$ of meaning merely that it will be somewhere in the whole interval 1.51 to 153. Table 1 shows some of the values of the resulting belief function. In this table, A_ϵ denotes an interval of length ϵ, and $\text{Bel}(\bar{A}_\epsilon)$ denotes the degree of belief that the refractive index is not in the interval. We call $1 - \text{Bel}(\bar{A})$ the *plausibility* of the interval and denote it $P^*(A)$. The table suggests that we give an interval of length .001 a plausibility $P^*(A_{.001}) \approx .12$, a rather large value reflecting our uncertainty about the gross features of the most relevant distribution of refractive indices. The table shows the plausibility of the index's being in a small interval declining at less than a proportionate rate as the interval is shrunk. This reflects the plausibility of lumpiness in the distribution of refractive indices at an intermediate scale, as well as the eventual meaninglessness of the idea of the index's randomness as it is specified more and more exactly.

[*Received March 1979. Revised August 1981.*]

REFERENCES

BAIRD, D.C. (1962), *Experimentation: An Introduction to Measurement Theory and Experiment Design*, Englewood Cliffs, N.J.: Prentice-Hall.

BERGER, JAMES O. (1980), *Statistical Decision Theory*, New York: Springer-Verlag.

DABBS, M.D.G., and PEARSON, E.F. (1970), "The Variation of Refractive Index and Density Across Two Sheets of Window Glass," *Journal of the Forensic Science Society*, 10, 139–150.
——— (1972), "Some Physical Properties of a Large Number of Window Glass Specimens," *Journal of Forensic Sciences*, 17, 70–78.
DABBS, M.D.G., GERMAN, B., PEARSON, E.F., and SCAPLEHORN, A.W. (1973), "The Use of Spark Source Mass Spectrometry for the Analysis of Glass Fragments Encountered in Forensic Applications," *Journal of the Forensic Science Society*, 13, 281–286.
DEMPSTER, A.P. (1967), "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals of Mathematical Statistics*, 38, 325–339.
——— (1971), "Model Searching and Estimation in the Logic of Inference," in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, 56–81.
EDWARDS, WARD, LINDMAN, HAROLD, and SAVAGE, LEONARD J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.
EVETT, I.W. (1977), "The Interpretation of Refractive Index Measurements," *Forensic Sciences*, 9, 209–217.
GOOD, I.J. (1956), "Discussion of Paper by G. Spencer Brown," in *Information Theory: Third London Symposium 1955*, ed. Colin Cherry, London: Butterworths, 13–14.
HEIDEMAN, D.H. (1975), "Glass Comparisons Using a Computerized Refractive Index Base," *Journal of Forensic Sciences*, 20, 103–108.
JEFFREYS, HAROLD (1939, 1948, 1961), *Theory of Probability*, Oxford University Press.
LEAMER, EDWARD E. (1978), *Specification Searches*, New York: John Wiley.
LINDLEY, D.V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192. (Comments by M.S. Bartlett and M.G. Kendall appear in Vol. 45, 533–534.)
——— (1977), "A Problem in Forensic Science," *Biometrika*, 64, 207–213.
——— (1980), "L.J. Savage—His Work in Probability and Statistics," *The Annals of Statistics*, 8, 1–24.
LOUISELL, DAVID W., KAPLAN, JOHN, and WALTZ, JON R. (1976), *Cases and Materials on Evidence* (3rd ed.), Mineola, Minn.: Foundation Press.
PEARSON, E.F., MAY, R.W., and DABBS, M.D.G. (1971), "Glass and Paint Fragments Found in Men's Outer Clothing—Report of a Survey," *Journal of Forensic Sciences*, 16, 283–300.
REEVE, VICTOR, MATHIESEN, JIM, and FONG, WILKAAN (1976), "Elemental Analysis by Energy Dispersive X-Ray: A Significant Factor in the Forensic Analysis of Glass," *Journal of Forensic Sciences*, 21, 291–306.
SHAFER, GLENN (1976), *A Mathematical Theory of Evidence*, Princeton University Press.
——— (1981), "Constructive Probability," *Synthese*, 48, 1–60.
WILSON, E. BRIGHT, JR. (1952), *An Introduction to Scientific Research*, New York: McGraw-Hill.

Comment

D. V. LINDLEY*

The supporter of a theory should welcome good criticism—and I know of no better critic of the Bayesian viewpoint than Shafer. If the theory survives the criticism, then it is enhanced the more the better the critique. In my view, Bayesian ideas come out of Shafer's analysis rather well.

1. RELIABILITY OF EVIDENCE

It is not always recognized that only the relevant probability matters: whether that probability is based on strong or weak evidence is immaterial. Shafer is wrong when he says "he ought also to weigh the reliability of the evidence." Consider the following example. An urn contains a large number of balls each of which is colored either red or black; one of them is to be drawn at random and a prize awarded if the ball is red. Contrast two situations. In the first, the proportion of red balls is known to be $\frac{1}{2}$. In the second, the proportion p is unknown but is described by a probability density $f(p)$ with mean $\frac{1}{2}$. As far as the prize is concerned, the relevant probability is that of a red ball being drawn, which is $\frac{1}{2}$ in both situations. The fact that the knowledge of p is less reliable in the second case is irrelevant. Tversky (1974) reports that in a choice between the two situations subjects incoherently prefer the first. Shafer appears to share their view when he discounts the histogram evidence, for only the probability of guilt is relevant.

The reason for the confusion is that the irrelevant aspects can become relevant if the problem is changed and a different probability required. To see this, modify the examples so that *two* balls are to be drawn and the prize awarded if they are of the same color. The relevant probability for a given p is $p^2 + (1 - p)^2$. This is $\frac{1}{2}$ in the first case but $\int[p^2 + (1 - p)^2]f(p)dp$ in the second. This is easily evaluated to give $\frac{1}{2} + 2\sigma^2$, where σ^2 is the variance of p. Now the situations are distinguishable. Similarly, there are aspects of the histogram evidence that would be relevant for some questions, but for the question of guilt the strength of that evidence does not matter any more than did that about p in the example.

2. BEHAVIORAL ASSESSMENT

In discounting the histogram evidence, Shafer uses a rate α. What does this number mean? He argues that a

* D.V. Lindley was formerly Professor and Head of the Department of Statistics at University College London. He is now retired and lives at 2 Periton Lane, Minehead, TA24 8AQ, England. The research was sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

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