J. R. Statist. Soc. B (1982),
44, *No.* 3, *pp.* 322–352

Belief Functions and Parametric Models

By GLENN SHAFER

University of Kansas, Kansas, USA

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, 5th May, 1982, Professor R. N. CURNOW in the Chair]

SUMMARY

The theory of belief functions assesses evidence by fitting it to a scale of canonical examples in which the meaning of a message depends on chance. In order to analyse parametric statistical problems within the framework of this theory, we must specify the evidence on which the parametric model is based. This article gives several examples to show how the nature of this evidence affects the analysis. These examples also illustrate how the theory of belief functions can deal with problems where the evidence is too weak to support a parametric model.

Keywords: BAYESIAN STATISTICS; BELIEF FUNCTIONS; CONSTRUCTIVE PROBABILITY; DEMPSTER'S RULE; EXCHANGEABILITY; FIDUCIAL INFERENCE; LOWER PROBABILITIES; LIKELIHOOD; PARAMETRIC MODELS

1. CONSTRUCTIVE PROBABILITY

IN *A Mathematical Theory of Evidence* (1976), I discussed the possibility that the mathematical structure for upper and lower probabilities that Dempster developed in his attempt to deal with parametric models might be used more widely as a structure for probability judgements. I suggested that we call set functions that have the structure of Dempster's lower probabilities *belief functions*, and I developed the implications of Dempster's rule for combining belief functions based on different bodies of evidence.

The central role of Dempster's rule of combination in the theory of belief functions is merely one aspect of the theory's emphasis on the decomposition and description of evidence. In general, the theory allows probability judgements to depend not only on the overall strength of the evidence on which they are based but also on the structure of that evidence.

In this paper I turn this general emphasis on evidence back onto the problem of parametric models. I argue that belief-function analyses of these models should depend not just on the models themselves but also on the nature of the evidence for them. I give several examples of this dependence.

Before taking up the problem of parametric models, I briefly review the theory of belief functions and its relation to other constructive theories of probability judgement.

The exposition that follows is based on the idea, first developed in unpublished work by Amos Tversky and myself, that all theories of probability judgement, including both the theory of belief functions and the Bayesian theory, should be thought of in terms of canonical examples to which the theories compare evidence. For a further development of this theme, see Shafer (1981a, b).

1.1. *Three Constructive Theories*

Probability judgement, like all judgement, involves comparison. In order to judge whether given evidence makes something practically certain, very probable, fairly probable, or not at all probable, say, we must compare this evidence to examples where it is agreed that these adjectives fit. We must, in other words, fit our evidence to a scale of canonical examples.

0035-9246/82/44322$2.00

Numerical probability judgement similarly involves fitting our evidence to a scale of canonical examples. Different choices of this scale produce different constructive theories of probability.

Here are three such theories.

The Bayesian theory. Suppose our scale consists exclusively of examples where the truth is generated according to known chances. Then when we make a probability judgement $P(A) = p$ we are saying that our evidence provides support for A comparable to what would be provided by knowledge that the truth is generated by a chance set-up that produces a result in A exactly p of the time. And these probability judgements will obey the usual Bayesian rules.

If we are working with a set of possibilities Ω, then our scale of canonical examples will include, for each chance distribution P over Ω, an example where the truth is generated according to the chances given by P. Usually we will not, of course, be able to fit our evidence to this scale by means of a single holistic judgement. Instead we will break the overall comparison down into many simpler comparisons and then construct P from these simpler judgements.

Lower probabilities. Suppose we know that a certain process is governed by chance, but instead of knowing precisely the chance distribution P governing it, we know only that P is in a class $\mathscr{P}$ of chance distributions. Denote by Ω the set of possible outcomes for the process. Then we might set our probability or degree of belief that the outcome of a particular trial will be in a subset A of Ω equal to

$$P_*(A) = \inf\{P(A) \mid P \in \mathscr{P}\}.$$

This seems natural because we know the chance of A is at least $P_*(A)$. Notice that the probabilities or degrees of belief obtained in this way will, in general, be non-additive: $P_*(A)$ and $P_*(\bar{A})$ may add to less than one.

By varying the class $\mathscr{P}$ in this story we obtain a scale of canonical examples. Let us call the constructive theory that uses this scale the theory of lower probabilities.

Belief functions. Suppose someone chooses a code at random from a list of codes, uses the 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 $c_1, ..., c_n$, 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 non-empty subset A of the set of possibilities Ω. Let A_i denote the subset we get when we decode using c_i, and set

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

for each $A \subset \Omega$. 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 the true message was A. Notice that $m(\phi) = 0$ and that the $m(A)$ sum to one. The quantity

$$\text{BEL}(A) = \sum_{B \subset A} m(B) \tag{2}$$

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 $\text{BEL}(A)$ our probability or degree of belief that the truth lies in A.

A function BEL is called a *belief function* if it is of the form (2), with the $m(A)$ non-negative and summing to one and with $m(\phi) = 0$. The subsets A of Ω for which $m(A) > 0$ are called the *focal elements* of the belief function.

It is easily seen from (2) that $\text{BEL}(A) + \text{BEL}(\bar{A}) \leqslant 1$, or $\text{BEL}(A) \leqslant 1 - \text{BEL}(\bar{A})$. The quantity $1 - \text{BEL}(\bar{A})$ is called the *plausibility* of A and is denoted by $\text{PL}(A)$; it can be large even if the evidence for A is slight, provided that the evidence against A is also slight.

The equation $\text{BEL}(A) + \text{BEL}(\bar{A}) = 1$, which is equivalent to $\text{BEL}(A) = \text{PL}(A)$, holds for all subsets A of Ω if and only if BEL's focal elements are all singletons. In this case, BEL is an additive probability distribution.

We can tell the story of the coded message with any values for the $m(A)$ we please. So this story provides a canonical example corresponding to each possible belief function. It is sometimes helpful to vary the story slightly; what is essential is that some chance experiment with outcomes $c_1,\ldots,c_n$ has been carried out, that we know these outcomes had chances $p_1,\ldots,p_n$, and that we receive a message that means A_i if c_i was the outcome.

1.2. *Elements of the Theory of Belief Functions*

Belief functions, we have suggested, are obtained by fitting evidence to a certain scale of canonical examples. In order to turn this idea into a practical tool, we need rules for breaking the fitting task down into simpler judgements, and techniques for making these simpler judgements feasible. Here we will review some of these rules and techniques for the case where Ω is finite. For an introduction to the case where Ω is infinite, see Shafer (1979).

The vacuous belief function. Consider the belief function BEL obtained by setting $m(\Omega) = 1$ and $m(A) = 0$ for every proper subset A of Ω. We see by (2) that BEL also satisfies $\text{BEL}(A) = 0$ for every proper subset A; BEL indicates no positive beliefs at all as to where in Ω the truth lies. This belief function is appropriate when the evidence being considered does not, by itself, tell us anything about which element of Ω is the truth.

Simple support functions. Consider the following variation on the story of the randomly coded message. A certain mechanism that produces messages has two modes of operation: reliable and unreliable. It is in its reliable mode with chance p_1, and then it produces only true messages. It is in its unreliable mode with chance $p_2 = 1-p_1$, and then it is completely unpredictable; we have no idea whether or how often the messages it produces will be true or false. Suppose this mechanism produces the message that the truth is in the subset of E of Ω. Then we will say that the message has a chance p_1 of meaning E and a chance p_2 of meaning nothing—i.e. meaning only that the truth is in Ω. And so we will adopt a belief function with focal elements E and Ω, with $m(E) = p_1$ and $m(\Omega) = p_2$. This belief function, given by

$$\text{BEL}(A) = \begin{cases} 0 & \text{if } E \not\subset A, \\ p_1 & \text{if } E \subset A \neq \Omega, \\ 1 & \text{if } A = \Omega, \end{cases}$$

is called a *simple support function.*

It is often natural to compare evidence to a mechanism that is only sometimes reliable and thus to represent it by a simple support function. The reliability of a witness can obviously be taken into account in this way. The strength of an argument can often be assessed in the same way; this means we compare the argument to one that has a definite and known chance of being reliable.

Dempster's rule of combination. One of the basic strategies of the theory is to decompose our evidence into two or more unrelated bodies of bodies of evidence, make probability judgements separately on the basis of each of these bodies of evidence, and then combine these judgements by Dempster's rule. This rule tells us how to combine a belief function BEL_1 representing one body of evidence with a belief function BEL_2 representing an unrelated body of evidence so as to obtain a belief function $\text{BEL}_1 \oplus \text{BEL}_2$ representing the pooled evidence. The rule is most easily stated in terms of m-values: If the m-values for BEL_1 and BEL_2 are denoted by $m_1(A)$ and $m_2(B)$, respectively, then $\text{BEL}_1 \oplus \text{BEL}_2$ is the belief function with m-values $m(C)$, where $m(\phi) = 0$ and

$$m(C) = \frac{\sum\{m_1(A)\,m_2(B) \mid A\subset\Omega;\ B\subset\Omega;\ A\cap B = C\}}{\sum\{m_1(A)\,m_2(B) \mid A\subset\Omega;\ B\subset\Omega;\ A\cap B \neq \phi\}} \tag{3}$$

for all non-empty subsets C of Ω. (Notice that the focal elements of $\text{BEL}_1 \oplus \text{BEL}_2$ consist of all the non-empty intersections of focal elements of BEL_1 with focal elements of BEL_2.)

The idea underlying Dempster's rule is that the unrelatedness of two bodies of evidence makes pooling them like combining two stochastically independent randomly coded messages. Suppose BEL$_1$ and BEL$_2$ do correspond to two such messages. Denote by $c_1,...,c_n$ and $p_1,...,p_n$ the codes and their chances in the case of the first message, and by $c'_1,...,c'_m$ and $p'_1,...,p'_m$ the codes and their chances in the case of the second. 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 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) cannot be the pair of codes actually used. So we condition the chance distribution, eliminating such pairs and multiplying the chances for the others by K, where

$$\begin{aligned} K^{-1} &= \sum \{p_i p'_j \mid 1 \leqslant i \leqslant n;\ 1 \leqslant j \leqslant m;\ A_i \cap B_j \neq \phi\} \\ &= \sum \{m_1(A) m_2(B) \mid A \subset \Omega;\ B \subset \Omega;\ A \cap B \neq \phi\}. \end{aligned}$$

If the first message is A and the second message is B, then the overall message is $A \cap B$. So the total chance of the overall message being C is

$$\begin{aligned} m(C) &= K \sum \{p_i p'_j \mid 1 \leqslant i \leqslant n;\ 1 \leqslant j \leqslant m;\ A_i \cap B_j = C\} \\ &= K \sum \{m_1(A) m_2(B) \mid A \subset \Omega;\ B \subset \Omega;\ A \cap B = C\}, \end{aligned}$$

which is indeed equal to (3).

We may call BEL$_1 \oplus$ BEL$_2$ the "orthogonal sum" of BEL$_1$ and BEL$_2$. Here are some elementary properties of the operation $\oplus$: (i) BEL$_1 \oplus$ BEL$_2$ exists unless there is a subset A of Ω such that BEL$_1(A) = 1$ and BEL$_2(\bar{A}) = 1$. (ii) Commutativity: BEL$_1 \oplus$ BEL$_2$ = BEL$_2 \oplus$ BEL$_1$, (iii) Associativity: (BEL$_1 \oplus$ BEL$_2$) $\oplus$ BEL$_3$ = BEL$_1 \oplus$ (BEL$_2 \oplus$ BEL$_3$). (iv) In general: BEL $\oplus$ BEL $\neq$ BEL; BEL $\oplus$ BEL will favour the same subsets as BEL but with, as it were, twice the weight of evidence. (v) If BEL$_1$ is Bayesian, then so is BEL$_1 \oplus$ BEL$_2$. (vi) If BEL$_1$ is vacuous, then BEL$_1 \oplus$ BEL$_2$ = BEL$_2$.

Dempster's rule can be seen as a generalization of rules formulated in the eighteenth century by James Bernoulli and Johann Heinrich Lambert. (See Shafer, 1978.)

Conditioning. Consider evidence which establishes conclusively that the truth is in a subset E of Ω but which does not tell us anything more specific. Such evidence can be compared to a randomly coded message which has chance one of meaning E, and so we can represent it by a belief function BEL$_E$ whose m-value for E is one. The values of BEL$_E$ are

$$\text{BEL}_E(A) = \begin{cases} 0 & \text{if } A \not\supset E, \\ 1 & \text{if } A \supset E. \end{cases}$$

An important property of BEL$_E$ is its idempotence: BEL$_E \oplus$ BEL$_E$ = BEL$_E$.

If BEL is a belief function satisfying BEL$(\bar{E}) < 1$, then BEL $\oplus$ BEL$_E$ exists. It is natural to call BEL $\oplus$ BEL$_E$ the result of *conditioning* BEL on E and to denote (BEL $\oplus$ BEL$_E$)(A) by BEL$(A \mid E)$. Notice that conditioning an orthogonal sum is equivalent to conditioning each term in the sum before combining: since BEL$_E$ is idempotent,

$$(\text{BEL}_1 \oplus \text{BEL}_2) \oplus \text{BEL}_E = (\text{BEL}_1 \oplus \text{BEL}_E) \oplus (\text{BEL}_2 \oplus \text{BEL}_E).$$

The process of conditioning can be described directly in terms of focal elements: to condition BEL on E, reduce the focal elements of BEL to their intersections with E and then renormalize the m-values to take into account the elimination of those focal elements that have been reduced to ϕ. If BEL is an additive probability distribution, then this reduces to the usual Bayesian rule of conditioning.

Minimal extension. Suppose the set of possibilities Ω has n elements: $\Omega = \{\omega_1, ..., \omega_n\}$. And suppose Λ is a finer set of possibilities. This means that the elements $\omega_1, ..., \omega_n$ of Ω correspond to a partition $E_1, ..., E_n$ of Λ: "ω_i is the truth" means the same as "the truth is in E_i", and, more generally, a subset $\{\omega_{i_1}, ..., \omega_{i_k}\}$ of Ω has the same meaning as the subset $E_{i_1} \cup ... \cup E_{i_k}$ of Λ.

Given a belief function BEL over Λ we can speak of its *marginal* over Ω: the belief function BEL$|\Omega$ given by

$$(\text{BEL}\,|\,\Omega)(\{\omega_{i_1},\ldots,\omega_{i_k}\}) = \text{BEL}\,(E_{i_1} \cup \ldots \cup E_{i_k}).$$

Marginalization can be described in terms of focal elements by saying that a focal element A of BEL is reduced to the subset $\{\omega_i \,|\, E_i \cap A \neq \phi\}$ of Ω. In general, there will be many belief functions over Λ having a given marginal over Ω. Or, to put the matter another way, a belief function over Ω will extend in many ways to a belief function over Λ.

Suppose we use a given body of evidence to construct a belief function BEL$_0$ over Ω. And suppose we judge that this evidence bears on the questions discerned by Λ only insofar as it bears on those already discerned by Ω. In terms of the randomly coded message to which we are comparing our evidence, this says that if $\{\omega_{i_1},\ldots,\omega_{i_k}\}$ is the meaning of the message relative to Ω, then $E_{i_1} \cup \ldots \cup E_{i_k}$ is its meaning relative to Λ. This suggests that BEL$_0$ should be extended to the belief function BEL over Λ whose m-values are given by $m(E_{i_1} \cup \ldots \cup E_{i_k}) = m_0(\{\omega_{i_1},\ldots,\omega_{i_k}\})$ and $m(A) = 0$ for all $A \subset \Lambda$ which are not unions of elements of the partition $E_1,\ldots,E_n$. This belief function BEL does have BEL$_0$ as its marginal. And for each $A \subset \Lambda$, BEL(A) is less than or equal to the degree of belief given to A by any other extension of BEL$_0$ to Λ. So we call BEL the *minimal extension* of BEL$_0$.

Conditional embedding. Sometimes we rule out some of the possibilities in a set of possibilities Λ, thus reducing it to a smaller set of possibilities $\Omega \subset \Lambda$. If we have constructed a belief function BEL over Λ and we then reduce Λ to Ω because of new evidence that establishes that the truth is in Ω without saying anything more specific, then we will, of course, replace BEL by its *conditional given* Ω—i.e. by the belief function over Ω that assigns to each $A \subset \Omega$ the degree of belief BEL$(A\,|\,\Omega)$. In general, there will be many belief functions over Λ having a given conditional given Ω.

Suppose we begin by taking it for granted that the truth is in Ω and construct a belief function BEL$_0$ over Ω, but we later decide that all the elements of Λ must be admitted as possibilities. And suppose we judge that the evidence on which BEL$_0$ is based does not impugn any of the possibilities in $\Lambda - \Omega$. In terms of the randomly coded message to which we are comparing the evidence, this means that if $A \subset \Omega$ is the meaning of the message relative to Ω, then $A \cup (\Lambda - \Omega)$ is its meaning relative to Λ. This suggests that BEL$_0$ should be replaced by the belief function BEL over Λ whose m-values are given by $m(A \cup (\Lambda - \Omega)) = m_0(A)$ for all $A \subset \Omega$ and $m(A) = 0$ for all subsets A of Λ that do not contain $\Lambda - \Omega$. This belief function BEL has BEL$_0$ as its conditional given Ω. And for each $A \subset \Lambda$, BEL(A) is less than or equal to the degree of belief given to A by any other belief function over Ω that has BEL$_0$ as its conditional given Ω. We call BEL the conditional embedding of BEL$_0$ in Λ.

The idea of conditioning embedding was first developed by Smets (1978).

Discounting. Suppose that after observing a randomly coded message and calculating the belief function BEL by (1) and (2) we discover that our understanding of the process producing the message is not fully reliable: say there is a chance $1 - \alpha$ that our understanding is correct, so that the message is indeed the result of choosing among the codes $c_1,\ldots,c_n$ with chances $p_1,\ldots,p_n$, but a chance α that the message was produced in some other way about which we know nothing and must therefore be counted as meaning nothing. Then we must change the chance associated with the code c_i from p_i to $(1-\alpha)p_i$, and we must, in effect, introduce a new "code" that is used with chance α and which decodes any message to the non-informative statement that the truth is in Ω. This means reducing each m-value $m(A)$ to $(1-\alpha)m(A)$ and then increasing the m-value for Ω by α. The result is a belief function BEL$^\alpha$ related to BEL by BEL$^\alpha(A) = (1-\alpha)\,BEL(A)$ for all proper subsets A of Ω. (BEL$^\alpha(\Omega)$ = BEL(Ω) = 1, of course.) We say that BEL$^\alpha$ is the result of *discounting* BEL. Discounting is the natural way to take account of doubts or second thoughts about belief functions constructed by ourselves or others.

1.3. *The Constructive View of Probability*

By saying that probability judgements are made by fitting given evidence to a scale of canonical examples, we are able to bring together two ideas that have sometimes been set up in opposition to one another: the idea that probabilities are subjective judgements, and the idea that probabilities can be based on a limited body of evidence.

The idea that probability judgements can be based on limited evidence is essential, of course, to a proper understanding of the theory of belief functions. Ultimately, we are always interested in judgements based on our total evidence. But the motivation for using Dempster's rule of combination is the idea that we might gain in clarity of thought by weighing different items of evidence separately before thinking about how they reinforce or contradict each other.

I do not wish to suggest that the idea of basing subjective probability judgements on limited evidence is utterly new. But consider the typology of views on the interpretation of probability that Savage presents in *The Foundation of Statistics* (1954, p. 3). Savage distinguishes three main classes of views: objectivistic, personalistic and necessary. Objectivistic views hold that probability is an objective property of certain repetitive events; personalistic views hold that probability measures the confidence that a particular individual has in the truth of a particular proposition; necessary views hold that probability measures the extent to which one set of propositions, out of logical necessity and apart from human opinion, confirms the truth of another. This typology obviously does not accommodate the idea of probability judgement based on limited evidence. Personalistic views focus on the attitudes a person actually has towards a proposition, and these attitudes are presumably based on his total evidence. Necessary views allow us to delimit the evidence, but they insist that this evidence be cast in the form of propositions, and they exclude any role for judgement in assessing it.

I would like to suggest that our *constructive* view of probability—the view that probability judgement amounts to fitting given evidence to a scale of canonical examples—should be recognized as a fourth view of probability, distinct from and on a par with the objectivistic, personalistic and necessary views.

2. GENERALIZATIONS OF BAYESIAN PARAMETRIC INFERENCE

Let us adopt the now standard general notation for parametric statistical models: Θ denotes the set of possible values for the parameter, θ, $\mathscr{X}$ denotes the set of possibilities for the data x and $\{P_\theta \colon \theta \in \Theta\}$ denotes the model. How do we make probability judgements about θ after observing x?

The Bayesian answers this question by representing prior evidence about θ by an additive probability distribution P_0 over Θ and by using this distribution, together with the P_θ, to construct a distribution, say P, over $\Theta \times \mathscr{X}$; P is the unique probability distribution over $\Theta \times \mathscr{X}$ that has P_0 as its marginal for θ and the P_θ as its conditionals given θ. Once the Bayesian has observed x, he will condition P on x to obtain posterior probabilities for θ.

How should our constructive generalizations of the Bayesian theory generalize this Bayesian treatment of parametric statistical inference?

Lower probabilities. The natural lower-probability generalization is to replace the prior distribution P_0 by a class $\mathscr{P}_0$ of additive probability distributions. This leads in turn to a class $\mathscr{P}$ of additive probability distributions over $\Theta \times \mathscr{X}$, and conditioning this class on the observed data x gives posterior lower probabilities for θ. The weakness of this approach is that if $\mathscr{P}_0$ is a reasonably broad class, then the posterior lower probabilities are not very informative. If, for example, we judge that we have no cogent prior evidence about θ and so allow $\mathscr{P}_0$ to be the class of all additive probability distributions on Θ, then our posterior degrees of belief will not indicate any evidence for any proper subset of the set of P_θ which are possible in light of the observed data. For a review of the literature on this lower-probability approach to parametric inference, see DeRobertis (1978).

Belief functions. Suppose we represent our prior evidence about θ by a belief function BEL_0 over Θ. Then it seems natural to generalize the Bayesian approach by asking for a belief function over $\Theta \times \mathscr{X}$ that has BEL_0 as its marginal for θ and P_θ as its conditional given θ. Such a belief function could then be conditioned on the observed data x to yield a posterior belief function over Θ.

This line of thought brings us immediately to a fundamental difference between additive probability measures and more general belief functions: a belief function is not, in general, uniquely determined by its marginal for a given partition and its conditionals given elements of that partition. There may be many belief functions over $\Theta \times \mathscr{X}$ having a given marginal BEL_0 and given conditionals P_θ. And there may be no reason to prefer one to the others. In his original work on "generalized Bayesian inference", Dempster (1968) proposed a particular method of constructing a belief function with a given marginal BEL_0 and given conditionals P_θ, but both he and his critics were uncomfortable with the seemingly arbitrary character of the method. (There are general principles from which Dempster's method can be derived (see Shafer, 1976b) but I now believe the method is appropriate only in the case where the evidence about a random experiment is limited to evidence for its randomness; see Section 4 below.)

But it is no embarrassment to the general theory of belief functions that a belief function is not fully determined by a given marginal and corresponding conditionals. Belief functions are not meant, in general, to be constructed from such elements. They are meant to be constructed from analyses of evidence. And so long as we are working within the theory of belief functions we expect to represent individual items of evidence by belief functions, not by objects like conditional belief functions or parametric models.

So the general spirit of the theory of belief functions leads us to look beyond the parametric model $\{P_\theta\colon \theta\in\Theta\}$ to the evidence on which the model is based. Our goal should be to represent this evidence directly by a belief function over $\Theta \times \mathscr{X}$, and it will be this belief function, say BEL, that we should regard as a full account of the effect of this evidence on $\Theta \times \mathscr{X}$. The model $\{P_\theta\colon \theta\in\Theta\}$ will be only a partial account: P_θ will be BEL's conditional given θ.

Once we have constructed such a belief function, BEL, we can take the prior evidence about θ into account by combining BEL with BEL_0's minimal extension to $\Theta \times \mathscr{X}$, which we may denote by $\overline{\text{BEL}_0}$. If the evidence for the parametric model does not by itself give any indication as to the value of θ (so that BEL's marginal for θ is vacuous), then the resulting belief function $\text{BEL} \oplus \overline{\text{BEL}_0}$ will satisfy the conditions formulated above: BEL_0 will be its marginal for θ, and P_θ will be its conditional given θ.

3. Some Examples of Evidence for Parametric Models

Here we shall consider three possible ways a parametric model $\{P_\theta\colon \theta\in\Theta\}$ might arise:

(1) Perhaps the values of the parameter θ have a substantive significance, and our knowledge of each P_θ derives from actual observations, the observations affording our knowledge of one P being distinct and independent of those affording our knowledge of another. In a problem of medical diagnosis, for example, each θ might correspond to the hypothesis that the patient has a particular disease, with P_θ giving the frequencies with which that disease has been observed to give rise to various symptoms.

(2) Perhaps the model arises from a single empirical frequency distribution—an "error distribution". This possibility is often mentioned in textbooks.

(3) Perhaps we are convinced that a phenomenon is random without having any evidence as to the frequency distribution of its outcomes, so that the model includes all additive probability distributions on $\mathscr{X}$.

These three ways suggest, as we shall see, quite different belief functions on $\Theta \times \mathscr{X}$, though in each case the belief function has the P_θ as its conditionals and has a vacuous marginal for θ. For another example of the use of belief functions in statistical problems see Shafer (1982).

3.1. *Models Composed of Independent Frequency Distributions*

Suppose our model consists of finitely many P_θ and each is based on independent empirical data—i.e. each P_θ is an empirical frequency distribution which we would be willing to translate into degrees of belief about x if we knew θ to be true, and the P_θ for different θ are based on independent observations. Then how should we combine them to obtain a belief function BEL on $\Theta \times \mathscr{X}$?

Smets (1978, pp. 145–190) has pointed out that the method of conditional embedding can be used to answer this question. We represent each P_θ by its conditional embedding, say BEL_θ, in $\Theta \times \mathscr{X}$, and then we set BEL equal to the orthogonal sum of all the BEL_θ.

Let us show that BEL is vacuous for θ and has P_θ for its conditional given θ. We begin with the fact that BEL_θ's focal elements are in one-to-one correspondence with the elements of $\mathscr{X}$; corresponding to $x \in \mathscr{X}$ is the focal element

$$\{(\theta, x)\} \cup ((\Theta - \{\theta\}) \times \mathscr{X}), \tag{4}$$

with m-value equal to $P_\theta(x)$. (i) A focal element for BEL is obtained by intersecting focal elements from the different BEL_θ's: in other words, it is of the form

$$\bigcap_{\theta \in \Theta} [\{(\theta, x_\theta)\} \cup ((\Theta - \{\theta\}) \times \mathscr{X})] = \bigcup_{\theta \in \Theta} \{(\theta, x_\theta)\} \tag{5}$$

for some choice of x_θ's. But any subset of $\Theta \times \mathscr{X}$ of the form (5) has a non-empty intersection with every cylinder set $\{\theta\} \times \mathscr{X}$. So BEL has a vacuous marginal for θ. (ii) Intersecting the focal element (4) with $\{\theta\} \times \mathscr{X}$ yields $\{(\theta, x)\}$, while intersecting it with $\{\theta'\} \times \mathscr{X}$, where $\theta' \neq \theta$, yields $\{\theta'\} \times \mathscr{X}$. So BEL_θ yields P_θ when conditioned on θ and yields the vacuous belief function on $\mathscr{X}$ when conditioned on $\theta' \neq \theta$. Since the conditioning of an orthogonal sum can be achieved by conditioning each component before combining, it follows that BEL yields P_θ when conditioned on θ.

It should be stressed that Smet's method depends on the assumption that Θ is finite. Moreover, it gives sensible results only when the number of elements in Θ is fairly small, for enlarging Θ has the effect of weakening the posterior degrees of belief. It is only when Θ is small, of course, that we could hope to satisfy the assumption that each P_θ be based on independent empirical data.

Example 1. Consider, for simplicity, the case where $\mathscr{X}$ and Θ have only two elements; say $\mathscr{X} = \{0, 1\}$, $\Theta = \{\theta_1, \theta_2\}$, $P_{\theta_1}(1) = p_1$ and $P_{\theta_2}(1) = p_2$. Then the belief function BEL on $\Theta \times \mathscr{X}$ has the m-values given in Table 1. Conditioning BEL on the observation $x = 1$ yields the degrees of belief

$$\mathrm{BEL}(\theta_1 \mid x = 1) = \frac{p_1(1-p_2)}{1-(1-p_1)(1-p_2)} \quad \text{and} \quad \mathrm{BEL}(\theta_2 \mid x = 1) = \frac{(1-p_1)p_2}{1-(1-p_1)(1-p_2)}. \tag{6}$$

Some insight into these formulae may be gained by fixing p_2 at some value equal neither to 0 nor to 1 and considering extreme values of p_1. If $p_1 = 0$, then the observation $x = 1$ tells us that $\theta = \theta_2$; we have $\mathrm{BEL}(\theta_1 \mid x = 1) = 0$ and $\mathrm{BEL}(\theta_2 \mid x = 1) = 1$. If $p_1 = 1$, then the observation $x = 1$ is evidence in favour of $\theta = \theta_1$; we have $\mathrm{BEL}(\theta_1 \mid x = 1) = 1 - p_2$ and $\mathrm{BEL}(\theta_2 \mid x = 1) = 0$.

TABLE 1

Focal element	*m-value*
$\{(\theta_1, 1), (\theta_2, 1)\}$	$p_1 p_2$
$\{(\theta_1, 1), (\theta_2, 0)\}$	$p_1(1-p_2)$
$\{(\theta_1, 0), (\theta_2, 1)\}$	$(1-p_1)p_2$
$\{(\theta_1, 0), (\theta_2, 0)\}$	$(1-p_1)(1-p_2)$

The combination of observations. Smet's method can be applied, of course, to the case of multiple observations. If we expect to make n independent observations from P_θ, then we simply construct the product distributions P_θ^n on $\mathscr{X}^n$, conditionally embed these to obtain belief functions BEL_θ^n on $\Theta \times \mathscr{X}^n$, and then combine by Dempster's rule to obtain a belief function BEL^n on $\Theta \times \mathscr{X}^n$ that can be conditioned on the observations $x_1, \dots, x_n$ to yield a posterior belief function on Θ.

An alternative approach to assessing independent observations $x_1, \dots, x_n$ is to use each x_i to construct a posterior belief function $\mathrm{BEL}(\cdot \mid x_i)$ on Θ and then to combine these posterior belief functions by Dempster's rule. This, it turns out, gives the same result (Smets, private communication).

Proof. For each $(x_1, \dots, x_n) \in \mathscr{X}^n$, BEL_θ^n assigns the m-value $P_\theta(x_1) \dots P_\theta(x_n)$ to the focal element

$$\{(\theta, x_1, \dots, x_n)\} \cup ((\Theta - \{\theta\}) \times \mathscr{X}^n). \tag{7}$$

Let BEL_θ denote, as before, the conditional embedding of P_θ in $\Theta \times \mathscr{X}$. And let $\mathrm{BEL}_{i\theta}$ denote the result of conditionally embedding BEL_θ in $\Theta \times \mathscr{X}^n$, with the $\mathscr{X}$ in $\Theta \times \mathscr{X}$ identified with the ith copy of $\mathscr{X}$ in $\Theta \times \mathscr{X}^n$. Then $\mathrm{BEL}_{i\theta}$ assigns, for each $x_i \in \mathscr{X}$, the m-value $P_\theta(x_i)$ to the focal element

$$(\{\theta\} \times \mathscr{X}_1 \times \dots \times \mathscr{X}_{i-1} \times \{x_i\} \times \mathscr{X}_{i+1} \times \dots \times \mathscr{X}_n) \cup ((\Theta - \{\theta\}) \times \mathscr{X}^n). \tag{8}$$

We see, by comparing (7) and (8), that $\mathrm{BEL}_\theta^n = \mathrm{BEL}_{1\theta} \oplus \dots \oplus \mathrm{BEL}_{n\theta}$. So

$$\mathrm{BEL}_n = \bigoplus_\theta \mathrm{BEL}_\theta^n = \bigoplus_\theta (\mathrm{BEL}_{1\theta} \oplus \dots \oplus \mathrm{BEL}_{n\theta}) = (\bigoplus_\theta \mathrm{BEL}_{1\theta}) \oplus \dots \oplus (\bigoplus_\theta \mathrm{BEL}_{n\theta}).$$

But $\bigoplus_\theta \mathrm{BEL}_{i\theta}$ is the conditional embedding in $\Theta \times \mathscr{X}^n$ of $\mathrm{BEL} = \bigoplus_\theta \mathrm{BEL}_\theta$. So conditioning $\bigoplus_\theta \mathrm{BEL}_{i\theta}$ on $(x_1, \dots, x_n)$ yields the same belief function on Θ as conditioning BEL on x_i. So

$$\mathrm{BEL}_n(\cdot \mid x_1, \dots, x_n) = \mathrm{BEL}(\cdot \mid x_1) \oplus \dots \oplus \mathrm{BEL}(\cdot \mid x_n)$$

for all $(x_1, \dots, x_n) \in \mathscr{X}^n$.

Example 1 *continued.* Suppose k of our observations $x_1, \dots, x_n$ are equal to 1 and $n-k$ are equal to 0. Then $\mathrm{BEL}_n(\cdot \mid x_1, \dots, x_n)$ is obtained by using Dempster's rule to combine k copies of $\mathrm{BEL}(\cdot \mid x = 1)$ and $n-k$ copies of $\mathrm{BEL}(\cdot \mid x = 0)$. The result is

$$\mathrm{BEL}_n(\theta_1 \mid x_1, \dots, x_n) = \frac{p_1^k(1-p_1)^{n\ k} - (p_1 p_2)^k((1-p_1)(1-p_2))^{n-k}}{p_1^k(1-p_1)^{n-k} + p_2^k(1-p_2)^{n-k} - (p_1 p_2)^k((1-p_1)(1-p_2))^{n-k}}$$

and

$$\mathrm{BEL}_n(\theta_2 \mid x_1, \dots, x_n) = \frac{p_2^k(1-p_2)^{n-k} - (p_1 p_2)^k((1-p_1)(1-p_2))^{n-k}}{p_1^k(1-p_1)^{n-k} + p_2^k(1-p_2)^{n-k} - (p_1 p_2)^k((1-p_1)(1-p_2))^{n-k}}.$$

Notice that for large values of n and $n-k$,

$$\mathrm{BEL}_n(\theta_1 \mid x_1, \dots, x_n) + \mathrm{BEL}_n(\theta_2 \mid x_1, \dots, x_n) \approx 1,$$

and

$$\frac{\mathrm{BEL}_n(\theta_1 \mid x_1, \dots, x_n)}{\mathrm{BEL}_n(\theta_2 \mid x_1, \dots, x_n)} \approx \left(\frac{p_1}{p_2}\right)^k \left(\frac{1-p_1}{1-p_2}\right)^{n-k}. \tag{9}$$

This agrees with the posterior Bayesian odds that would result from equal prior probabilities for θ_1 and θ_2.

Medical diagnosis. Smets' work was inspired by the problem of medical diagnosis. Here Θ is a list of possible diseases from which a patient might be suffering, $\mathscr{X}$ is a list of symptoms he might exhibit, and we assume that study of each disease θ has resulted in a distribution P_θ that gives the frequency with which that disease produces the various symptoms. Conditional embedding seems reasonable because P_θ bears on the set of possibilities $\Theta \times \mathscr{X}$ regarding our patient only conditionally on his having disease θ, and the use of Dempster's rule seems reasonable because the different frequency distributions can be regarded as independent items of evidence.

The assumption that one's evidence in a problem of medical diagnosis consists of complete and clearly relevant frequency distributions of symptoms is, of course, very unrealistic. But, as Smets points out (p. 160), the method of conditional embedding can still be used when the evidence about each disease justifies only a relatively weak belief function instead of a full frequency distribution. The following example illustrates some of the possibilities.

Example 2. Imagine a disorder called "ploxoma", which comprises two distinct "diseases": θ_1 = "virulent ploxoma", which is invariably fatal, and θ_2 = "ordinary ploxoma", which varies in severity and can be treated. Virulent ploxoma can be identified unequivocally at the time of a victim's death, but the only way to distinguish between the two diseases in their early stages seems to be a blood test with three possible outcomes, labelled x_1, x_2 and x_3. The following evidence is available: (i) Blood tests of a large number of patients dying of virulent ploxoma showed the outcomes x_1, x_2 and x_3 occurring 20, 20 and 60 per cent of the time, respectively. (ii) A study of patients whose ploxoma had continued so long as to be almost certainly ordinary ploxoma showed outcome x_1 to occur 85 per cent of the time and outcomes x_2 and x_3 to occur 15 per cent of the time. (The study was made before methods for distinguishing between x_2 and x_3 were perfected.) There is some question whether the patients in the study represent a fair sample of the population of ordinary ploxoma victims, but experts feel fairly confident (say 75 per cent) that the criteria by which patients were selected for the study should not affect the distribution of test outcomes. (iii) It seems that most people who seek medical help for ploxoma are suffering from ordinary ploxoma. There have been no careful statistical studies, but physicians are convinced that only 5–15 per cent of ploxoma patients suffer from virulent ploxoma.

We can represent each of these three items of evidence by a belief function on $\Theta \times \mathscr{X} = \{\theta_1, \theta_2\} \times \{x_1, x_2, x_3\}$. (i) The first item of evidence can be represented by the conditional embedding in $\Theta \times \mathscr{X}$ of the frequency distribution P_{θ_1}, where $P_{\theta_1}(x_1) = 0\cdot2$, $P_{\theta_1}(x_2) = 0\cdot2$ and $P_{\theta_1}(x_3) = 0\cdot6$. (ii) For the second item of evidence, we begin with a belief function BEL_{θ_2} on $\mathscr{X}$ that has focal elements $\{x_1\}$ and $\{x_2, x_3\}$ with m-values 0·85 and 0·15, respectively. We discount this belief function at rate $\alpha = 0\cdot25$, and then conditionally embed it in $\Theta \times \mathscr{X}$. (iii) For the third item of evidence we begin with a belief function BEL_0 on Θ that has m-values $m_0(\{\theta_1\}) = 0\cdot05$, $m_1(\{\theta_2\}) = 0\cdot85$ and $m_0(\Theta) = 0\cdot10$, and we minimally extend BEL_0 to $\Theta \times \mathscr{X}$.

Combining these three belief functions by Dempster's rule results in the belief function on $\Theta \times \mathscr{X}$ with the m-values given in Table 2. Table 3 shows the posterior degrees of belief that result when this belief function is conditioned on the result of the patient's blood test. As these numbers indicate, the blood test is not as informative as one might hope. The physician's initial 85 per cent degree of belief that a given ploxoma is ordinary is raised only to 96·5 per cent by a test that comes out x_1 and lowered only to 78·2 per cent by a test that comes out x_3.

3.2. *Models derived from a Single Frequency Distribution*

Let us turn from the case where there is a different frequency distribution underlying each P_θ to an opposite extreme: the case where all the P_θ are derived from a single frequency distribution. And let us think about the tritest example: the parametric model generated by an error distribution.

Consider a measuring instrument whose propensities to err are thoroughly known to us; we have used it to measure many known quantities and recorded its errors in these cases so as to obtain a frequency distribution $P(e)$ which we are willing to translate into degrees of belief about what our error $e = x - \theta$ will be when we shortly use the instrument to obtain a measurement x of an unknown quantity θ. Consider $\Theta \times \mathscr{X}$, where Θ is the set of possible values of θ and $\mathscr{X}$ is the set of possible values of x; we assume that $\Theta = \mathscr{X}$. Each possible error e

TABLE 2

Focal element	*m-value*	*Focal element*	*m-value*
$\{(\theta_2, x_1)\}$	0·541875	$\{(\theta_1, x_1)\}$	0·01
$\{(\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0·2125	$\{(\theta_1, x_2)\}$	0·01
$\{(\theta_2, x_2), (\theta_2, x_3)\}$	0·095625	$\{(\theta_1, x_3), (\theta_2, x_2), (\theta_2, x_3)\}$	0·00675
$\{(\theta_1, x_3), (\theta_2, x_1)\}$	0·03825	$\{(\theta_1, x_1), (\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0·005
$\{(\theta_1, x_3)\}$	0·03	$\{(\theta_1, x_2), (\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0·005
$\{(\theta_1, x_3), (\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0·015	$\{(\theta_1, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0·00225
$\{(\theta_1, x_1), (\theta_2, x_1)\}$	0·01275	$\{(\theta_1, x_2), (\theta_2, x_2), (\theta_2, x_3)\}$	0·00225
$\{(\theta_1, x_2), (\theta_2, x_1)\}$	0·01275		

TABLE 3

	BEL$(\theta_1 \mid x)$	BEL$(\theta_2 \mid x)$
x_1	0·014	0·965
x_2	0·062	0·918
x_3	0·165	0·782

will correspond to a subset $\Theta \times \mathscr{X}$; namely, $\{(\theta, x) \mid x - \theta = e\}$. So we can accomplish the translation of the error distribution $P(e)$ into degrees of belief about $x - \theta$ by minimally extending P to $\Theta \times \mathscr{X}$. This means adopting the belief function BEL on $\Theta \times \mathscr{X}$ that assigns the m-value $P(e)$ to the focal element $\{(\theta, x) \mid x - \theta = e\}$. It is evident that BEL is vacuous for θ. And its conditional on $\mathscr{X}$ given θ is given by BEL$(x \mid \theta) = P(x - \theta)$. The belief function BEL$(\cdot \mid \theta)$ is an additive probability distribution, and so it may denote it by P_θ, thus obtaining a parametric model $\{P_\theta : \theta \in \Theta\}$ on $\mathscr{X}$.

The preceding paragraph merely translates into the language of belief functions a traditional account of how a parametric model arises from an error distribution. Moreover, the result of conditioning the belief function BEL on the actual measurement x is the additive probability distribution BEL$(\cdot \mid x)$ on Θ given by BEL$(\theta \mid x) = P(\theta - x)$, and this is the familiar fiducial solution to the problem of inference for this model. Notice, however, that the belief-function argument depends on the model having really arisen from the error distribution; the argument gives no sanction to fiducial methods in cases where one begins with an abstract model $\{P_\theta : \theta \in \Theta\}$ and then notices a pivotal quantity $x - \theta$. (This belief-function treatment of the fiducial method was given by Dempster (1966). The only novelty in the present exposition is my insistence that the criterion for the method's validity should be sought in the origin of the parametric model.) This lack of sanction for the use of arbitrary pivotal quantities appears to rule out marginalization paradoxes of the type discussed by Dawid *et al.* (1973).

The belief function BEL on $\Theta \times \mathscr{X}$ is non-additive, even though its conditionals BEL$(\cdot \mid \theta)$ and BEL$(\cdot \mid x)$ are all additive. Notice also that BEL can, in some circumstances, lead to posterior probabilities for θ that are non-additive. If instead of observing the measurement x we observe only that x is in some subset A of $\mathscr{X}$, then we will condition BEL on $\Theta \times A$, and the resulting conditional belief function will have a non-additive marginal for θ.

The combination of observations. Here, as in the case of Smets' method, there are two approaches to combining independent observations. We can construct the product distribution P^n, conditionally embed it in $\Theta \times \mathscr{X}^n$, and then condition on the observations $(x_1, \ldots, x_n)$. Or we can construct a posterior belief function BEL$(\cdot \mid x_i)$ for each observation and then combine these by Dempster's rule. It can be shown, here as in the case of Smets' method, that both approaches give the same final belief function BEL$_n(\cdot \mid x_1, \ldots, x_n)$ on Θ. In this case, BEL$_n(\cdot \mid x_1, \ldots, x_n)$ is an additive probability distribution.

Example 3. Suppose $\mathscr{X}$ and Θ are both equal to the set of all integers. and P is given by $P(e) = c0{\cdot}8^{e^2}$, where $c \approx 0{\cdot}26651$. Table 4 gives the values of $P(e)$ that exceed 10^{-5}. If we observe $(x_1,\ldots,x_n)$, then $\text{BEL}(\theta \mid x_i) = c0{\cdot}8^{(\theta - x_i)^2}$, and $\text{BEL}_n(\cdot \mid x_1,\ldots,x_n) = \text{BEL}(\cdot \mid x_1) \oplus \ldots \oplus \text{BEL}(\cdot \mid x_n)$ is the additive probability distribution specified by

$$\text{BEL}_n(\theta \mid x_1,\ldots,x_n) \propto \prod_{i=1}^{n} \text{BEL}(\theta \mid x_i) \propto (0{\cdot}8)^{\Sigma(\theta - x_i)^2}.$$

If, for example, $n = 4$ and $(x_1,\ldots,x_4) = (-2, 1, 0, 9)$, then we obtain

$$\text{BEL}_4(\theta \mid -2, 1, 0, 9) \propto 0{\cdot}8^{4(\theta-2)^2}.$$

Table 5 gives the values of $\text{BEL}_4(\theta \mid -2, 1, 0, 9)$ that exceed 10^{-5}.

TABLE 4

e	$P(e)$	e	$P(e)$
−6	0·00009	1	0·21321
−5	0·00101	2	0·10916
−4	0·00750	3	0·03577
−3	0·03577	4	0·00750
−2	0·10916	5	0·00101
−1	0·21321	6	0·00009
0	0·26651		

TABLE 5

θ	$\text{BEL}_4(\theta \mid -2, 1, 0, 9)$
−1	0·00017
0	0·01500
1	0·21832
2	0·53300
3	0·21832
4	0·01500
5	0·00017

Example 4. Let us suppose, in order to construct an example that is comparable to Example 1 above, that $\Theta = \{0, 1\}$, that 0 and 1 are also the possible errors, with frequencies $P(0) = p$ and $P(1) = 1-p$, and that the addition to obtain $x = \theta + e$ is modulo 2. This means that $\mathscr{X} = \{0, 1\}$, and that P_θ assigns 0 and 1 the frequencies p and $1-p$, respectively, when $\theta = 0$ and the frequencies $1-p$ and p, respectively, when $\theta = 1$.

The belief function BEL on $\Theta \times \mathscr{X}$ has focal elements $\{(0,0),(1,1)\}$ and $\{(0,1),(1,0)\}$ with m-values p and $1-p$, respectively. So conditioning on $x = 1$ yields $\text{BEL}(\theta = 0 \mid x = 1) = 1-p$ and $\text{BEL}(\theta = 1 \mid x = 1) = p$. Notice that these posterior degrees of belief do not agree with the posterior degrees of belief that we obtained using Smets' method in Example 1. In order to make the comparison, we set $\theta_1 = 1$, $\theta_2 = 0$, $p_1 = p$ and $p_2 = 1-p$ in (6), thus obtaining

$$\text{BEL}(\theta = 1 \mid x = 1) = p^2\{1-p(1-p)\}^{-1} \quad \text{and} \quad \text{BEL}(\theta = 0 \mid x = 1) = (1-p)^2\{1-p(1-p)\}^{-1}.$$

There is asymptotic agreement, however. If we have measurements $x_1,\dots,x_n$, k of which equal 1 and $n-k$ of which equal 0, then we obtain

$$\text{BEL}_n(\theta = 1\,|\,x_1,\dots,x_n) = p^k(1-p)^{n-k}(p^k(1-p)^{n-k}+(1-p)^k p^{n-k})^{-1},$$

$$\text{BEL}_n(\theta = 0\,|\,x_1,\dots,x_n) = (1-p)^k p^{n-k}(p^k(1-p)^{n-k}+(1-p)^k p^{n-k})^{-1}$$

and

$$\frac{\text{BEL}_n(\theta = 1\,|\,x_1,\dots,x_n)}{\text{BEL}_n(\theta = 0\,|\,x_1,\dots,x_n)} = \frac{p^k(1-p)^{n-k}}{(1-p)^k p^{n-k}},$$

which agrees with (9).

Practical complications. The premises for our justification of the fiducial method through belief functions will rarely be fully satisfied. Usually our experience with a measuring instrument will be inadequate for us to credit fully a frequency distribution and the possibility of systematic errors will always limit the extent to which we are willing to treat successive errors as independent. However, these complications, though they do push us away from the fiducial method, need not push us away from the use of belief functions.

Example 3 *continued.* Suppose we take seriously the possibility of outliers and therefore discount the frequency distribution $P(e)$, using the discount rate $\alpha = 0{\cdot}01$. This results in the $\text{BEL}(\cdot\,|\,x_i)$ also being discounted at this rate. When we combine these four discounted belief functions by Dempster's rule, we obtain a belief function $\text{BEL}_4^{0\cdot01}(\theta\,|-2,1,0,9)$ that is very nearly an additive probability distribution; the whole set Θ is a focal element, but its m-value is only 0·00005, and all the other focal elements are singletons. Values of $\text{BEL}_4^{0\cdot01}(\theta\,|-2,1,0,9)$ that exceed 10^{-5} are shown in Table 6. Notice the sharp disagreement with the values of $\text{BEL}_4(\theta\,|-2,1,0,9)$ given in Table 5. When we do not discount, we obtain a probability of 0·53300 for $\theta = 2$, but when we do discount, we obtain a probability of only 0·02517 for $\theta = 2$ and a probability of 0·87302 for $-1 \leqslant \theta \leqslant 1$. This disagreement can be explained by saying that discounting leads us to treat the measurement $x_4 = 9$ as a probable outlier. (See pp. 251–255 of Shafer, 1976.)

TABLE 6

	$\text{BEL}_4^{0\cdot01}(\theta\,\vert-2,1,0,9)$		$\text{BEL}_4^{0\cdot01}(\theta\,\vert-2,1,0,9)$
−7	0·00001	4	0·00042
−6	0·00004	5	0·00013
−5	0·00022	6	0·00022
−4	0·00119	7	0·00059
−3	0·00961	8	0·00116
−2	0·08154	9	0·00144
−1	0·32160	10	0·00116
0	0·39843	11	0·00059
1	0·15299	12	0·00019
2	0·02517	13	0·00004
3	0·00320	14	0·00001

Now suppose we admit the possibility that there may be a systematic error f affecting all our measurements. And suppose we make the following probability judgements about f, based on our knowledge of the measuring instrument and process: we consider it certain that $|f| \leqslant 2$, and we feel there is a chance 0·8 that $|f| \leqslant 1$ and a chance 0·6 that $f = 0$. In other words, we adopt a belief function BEL_f that has focal elements $\{0\}$, $\{-1,0,1\}$ and $\{-2,-1,0,1,2\}$, with m-values 0·6, 0·2 and 0·2, respectively.

We are now assuming that $x_i = \theta + f + e_i$, or $\theta + f = x_i - e_i$. So the belief function $\text{BEL}_4^{0\cdot01}(\cdot\,|-2,1,0,9)$ must now be interpreted as giving degrees of belief about $\theta + f$ rather than

TABLE 7

Focal element	*m-value*	*Focal element*	*m-value*	*Focal element*	*m-value*
$\{-7\}$	0·00000	$\{-8, -7, -6\}$	0·00000	$\{-9, -8, -7, -6, -5\}$	0·00000
$\{-6\}$	0·00003	$\{-7, -6, -5\}$	0·00001	$\{-8, -7, -6, -5, -4\}$	0·00001
$\{-5\}$	0·00013	$\{-6, -5, -4\}$	0·00004	$\{-7, -6, -5, -4, -3\}$	0·00004
$\{-4\}$	0·00071	$\{-5, -4, -3\}$	0·00024	$\{-6, -5, -4, -3, -2\}$	0·00024
$\{-3\}$	0·00577	$\{-4, -3, -2\}$	0·00192	$\{-5, -4, -3, -2, -1\}$	0·00192
$\{-2\}$	0·04892	$\{-3, -2, -1\}$	0·01631	$\{-4, -3, -2, -1, 0\}$	0·01631
$\{-1\}$	0·19297	$\{-2, -1, 0\}$	0·06432	$\{-3, -2, -1, 0, 1\}$	0·06432
$\{0\}$	0·23906	$\{-1, 0, 1\}$	0·07969	$\{-2, -1, 0, 1, 2\}$	0·07969
$\{1\}$	0·09179	$\{0, 1, 2\}$	0·03060	$\{-1, 0, 1, 2, 3\}$	0·03060
$\{2\}$	0·01510	$\{1, 2, 3\}$	0·00503	$\{0, 1, 2, 3, 4\}$	0·00503
$\{3\}$	0·00192	$\{2, 3, 4\}$	0·00064	$\{1, 2, 3, 4, 5\}$	0·00064
$\{4\}$	0·00025	$\{3, 4, 5\}$	0·00008	$\{2, 3, 4, 5, 6\}$	0·00008
$\{5\}$	0·00008	$\{4, 5, 6\}$	0·00003	$\{3, 4, 5, 6, 7\}$	0·00003
$\{6\}$	0·00013	$\{5, 6, 7\}$	0·00004	$\{4, 5, 6, 7, 8\}$	0·00004
$\{7\}$	0·00036	$\{6, 7, 8\}$	0·00012	$\{5, 6, 7, 8, 9\}$	0·00012
$\{8\}$	0·00069	$\{7, 8, 9\}$	0·00023	$\{6, 7, 8, 9, 10\}$	0·00023
$\{9\}$	0·00087	$\{8, 9, 10\}$	0·00029	$\{7, 8, 9, 10, 11\}$	0·00029
$\{10\}$	0·00069	$\{9, 10, 11\}$	0·00023	$\{8, 9, 10, 11, 12\}$	0·00023
$\{11\}$	0·00035	$\{10, 11, 12\}$	0·00012	$\{9, 10, 11, 12, 13\}$	0·00012
$\{12\}$	0·00012	$\{11, 12, 13\}$	0·00004	$\{10, 11, 12, 13, 14\}$	0·00004
$\{13\}$	0·00002	$\{12, 13, 14\}$	0·00001	$\{11, 12, 13, 14, 15\}$	0·00001
$\{14\}$	0·00000	$\{13, 14, 15\}$	0·00000	$\{12, 13, 14, 15, 16\}$	0·00000
Θ	0·00005				

TABLE 8

A	BEL*(A)
$\{0\}$	0·23906
$\{-1\}$	0·19297
$\{1\}$	0·09179
$\{-2\}$	0·04892
$\{2\}$	0·01510
$\{-1, 0, 1\}$	0·60351
$\{-2, -1, 0\}$	0·54527
$\{0, 1, 2\}$	0·37655
$\{-2, -1, 0, 1, 2\}$	0·84214
$\{-3, -2, -1, 0, 1, 2, 3\}$	0·96609

about θ. When we combine these degrees of belief about $\theta + f$ with the degrees of belief about f given by BEL$_f$, we obtain a belief function BEL* with the m-values given (to the nearest 0·00001) in Table 7. A few values of BEL* are given in Table 8.

3.3. *Pure Randomness*

Suppose we know an unknown quantity X must take one of a finite set, say $\mathscr{X} = \{1, \ldots, k\}$, of possible values, and we feel it does so randomly. We can express this by saying that X is governed by some frequency distribution. But there are only so many frequency distributions on $\mathscr{X}$—so many as there are vectors $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ of non-negative numbers that add to one. Setting Θ equal to the set of all these vectors and letting P_θ denote the frequency distribution corresponding to θ (i.e. $P_\theta(x) = \theta_x$ for all $x \in \mathscr{X}$), we obtain a parametric model $\{P_\theta\colon \theta \in \Theta\}$. This model, it seems fair to say, arises solely from the idea that X is random.

As a result of work by de Finetti (1964), Hewitt and Savage (1955) and others, many Bayesians subscribe to a purely subjective interpretation of the idea that X is random and is governed by one of the frequency distributions P. This interpretation involves thinking of X as one of a sequence $\mathbf{X} = (X_1, X_2, \ldots)$ of unkown quantities, each of which takes values in $\mathscr{X}$, and considering a countably additive† probability distribution P that represents a Bayesian's beliefs about $\mathbf{X}$ and that is symmetric—i.e. invariant under permutations of finitely many of the X_i's. As it turns out, the countable additivity and symmetry of P imply that for each $x \in \mathscr{X}$, $P(\lim_{n \to \infty} f(x,n) \text{ exists}) = 1$, where $f(x,n)$ is the proportion of the quantities $X_1, \ldots, X_n$ that equal x. The vector $\lim_{n \to \infty}(f(1,n), \ldots, f(k,n))$ can be identified, of course, with the unknown parameter θ; conditioning P on this vector being equal to θ reduces P to the product distribution P_θ^∞. The Bayesian's prior distribution for θ is implicitly contained in P; it is P's marginal for the vector $\lim_{n \to \infty} f(1,n), \ldots, f(k,n))$. The distribution P is fully determined, moreover, by this prior distribution; there is only one symmetric and countably additive distribution for $\mathbf{X}$ having a given marginal for $\lim_{n \to \infty}(f(1,n), \ldots, f(k,n))$.

How might we give a treatment of randomness via belief functions which is analogous to this Bayesian treatment? The obvious goal is to capture the aspects of our idea of randomness (belief in the existence of limiting frequencies and recovery of $\{P_\theta: \theta \in \Theta\}$ by conditioning on the limiting frequencies) captured by the Bayesian treatment while avoiding opinions about the value of the limiting frequency. This means we should try to construct a symmetric belief function BEL for $\mathbf{X} = (X_1, X_2, \ldots)$ that satisfies

$$\text{BEL}(\lim_{n \to \infty} f(x,n) \text{ exists}) = 1 \tag{10}$$

for all $x \in \mathscr{X}$,

$$\text{BEL}(X_1 = x_1, \ldots, X_n = x_n \mid \lim_{n \to \infty} (f(1,n), \ldots, f(k,n)) = \theta) = P_\theta(x_1), \ldots, P_\theta(x_n) \tag{11}$$

for all $x_1, \ldots, x_n \in \mathscr{X}$, and

$$\text{BEL}(\lim_{n \to \infty} (f(1,n), \ldots, f(k,n)) \in A) = 0 \tag{12}$$

for every proper subset A of Θ. As it turns out, this goal can be achieved; there are belief functions satisfying these conditions.

The dichotomous case. The construction of a belief function BEL satisfying (10), (11) and (12) is most easily carried out in the case where $\mathscr{X}$ has only two elements. In this case it is convenient to use $\{0,1\}$ rather than $\{1,2\}$ to label the elements of $\mathscr{X}$ and to use $[0,1]$ as the parameter space Θ, with $P_\theta(1) = \theta$ and $P_\theta(0) = 1-\theta$. Let us also write $S_n = \Sigma_{i=1}^n X_i$. Then (10), (11) and (12) become

$$\text{BEL}(\lim_{n \to \infty} (S_n/n) \text{ exists}) = 1, \tag{13}$$

$$\text{BEL}(X_1 = x_1, \ldots, X_n = x_n \mid \lim_{n \to \infty} (S_n/n) = \theta) = P_\theta(x_1) \ldots P_\theta(x_n) \tag{14}$$

and

$$\text{BEL}(\lim_{n \to \infty} (S_n/n) \in A) = 0 \tag{15}$$

for all $A \subset [0,1]$.

The construction of a belief function BEL satisfying (13), (14) and (15) begins with the construction of a belief function BEL_n for the finite sequence $(X_1, X_2, \ldots, X_n)$. We construct BEL_n, which is a belief function over $\{0,1\}^n$, by assigning m-values $1/n!$ to each of the $n!$ subsets of $\{0,1\}^n$ of the form

$$A_\sigma = \{(x_1, \ldots, x_n) \in \{0,1\}^n \mid x_{\sigma(1)} \geqslant x_{\sigma(2)} \geqslant \ldots \geqslant x_{\sigma(n)}\},$$

where σ is a permutation of $\{1, \ldots, n\}$. (Here is an example of a set A_σ. If $n = 3$ and $(\sigma(1), \sigma(2), \sigma(3)) = (1,3,2)$, then

$$A_\sigma = \{(0,0,0), (1,0,0), (1,0,1), (1,1,1)\}.)$$

† De Finetti prefers the weaker condition of finite additivity. But we can neglect this subtlety in the present brief exposition.

A permutation of $(X_1, \ldots, X_n)$ merely permutes the A_σ. So BEL_n is symmetric—i.e. it satisfies

$$\mathrm{BEL}_n((X_1, \ldots, X_n) \in A) = \mathrm{BEL}_n((X_{\sigma(1)}, \ldots, X_{\sigma(n)}) \in A) \tag{16}$$

for all permutations σ. It is also easy to see that each A_σ has exactly one representative for each possible frequency of ones—i.e. for each k, $0 \leqslant k \leqslant n$, there is exactly one element $(x_1, \ldots, x_n) \in A_\sigma$ such that $\Sigma_{i=1}^n x_i = k$. This means that BEL_n's marginal for S_n is vacuous—i.e.

$$\mathrm{BEL}_n(S_n \in A) = 0 \tag{17}$$

for every proper subset A of $\{0, 1, \ldots, n\}$. It also means that conditioning on $S_n = k$ reduces the A_σ to singletons and hence reduces BEL_n to an additive (i.e. Bayesian) belief function. Thus, by the symmetry of BEL_n,

$$\mathrm{BEL}_n(X_1 = x_1, \ldots, X_n = x_n \mid S_n = k) = 1 \Big/ \binom{n}{k}, \tag{18}$$

provided that $\Sigma_{i=1}^n x_i = k$.

The belief functions BEL_n "cohere", in the sense that if $m < n$ then BEL_m is BEL_n's marginal for $X_1, \ldots, X_m$. And it is fairly easy to show that this set of coherent belief functions is the only one satisfying (16), (17) and (18). All the BEL_n together can be regarded as a belief function for the infinite sequence $(X_1, X_2, \ldots)$. More precisely, they can be regarded as defining a belief function on the algebra of subsets of $\{0, 1\}^\infty$ consisting of all "finite cylinder sets". This belief function can then be minimally extended to a belief function on the algebra of all subsets of $\{0, 1\}^\infty$. It turns out that if we use a form of minimal extension that preserves "sequential continuity" (a condition equivalent to countable additivity in the presence of finite additivity), then the resulting belief function BEL on $\{0, 1\}^\infty$ does indeed satisfy (13). Since BEL_n is BEL's marginal, (16) says that BEL is symmetric. And, as it turns out, (17) implies (15) and (18) implies (14). (The proofs of the assertions in this paragraph have not been published. But the concepts of continuity and minimal continuous extension are discussed in Shafer, 1979.)

Since the belief function BEL, like the Bayesian's additive probability distribution P, gives degree of belief one to the existence of the limit $\theta = \lim_{n \to \infty}(S_n/n)$, we can examine BEL's marginal for (θ, X_1), which is a belief function on $\Theta \times \mathcal{X}$. By (15), this belief function has a vacuous marginal for θ. And by (14), its conditional given θ is P_θ. Thus the construction of BEL yields a solution to our general problem of constructing a belief function on $\Theta \times \mathcal{X}$—a solution which seems appropriate when the specification is based purely on the idea of randomness. As it turns out, this solution is Dempster's original "generalized Bayesian" method. (See Dempster, 1968, or Shafer, 1976b).

Instead of considering the marginal just for (θ, X_1), we could also consider the marginal for $(\theta, X_1, \ldots, X_n)$, thus obtaining a belief function on $\Theta \times \mathcal{X}^n$ which is vacuous for θ and has P_θ^n as its conditional given θ. It is also true, here as in the case of Smets' method and the fiducial method, that the belief function $\mathrm{BEL}_{\mathbf{x}}$ on Θ obtained by conditioning on a vector $\mathbf{x} = (x_1, \ldots, x_n)$ of actual observations is the same as the belief function $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$, where BEL_{x_i} is the belief function on obtained by conditioning on a single observation x_i. See Section 4 of Dempster (1966) for some calculations of values of $\mathrm{BEL}_{\mathbf{x}}$.

The general case. The results in the dichotomous case generalize to the case where $\mathcal{X} = \{1, \ldots, k\}$ in that there does exist a symmetric belief function on $\mathcal{X}^\infty$ that satisfies (10), (11) and (12) and has a marginal for $\mathcal{X} \times \Theta$ corresponding to Dempster's generalized Bayesian method. It appears, however, that when $k > 2$ there are other symmetric belief functions on $\mathcal{X}^\infty$ that satisfy (10), (11) and (12) but have different marginals for $\mathcal{X} \times \Theta$. It would be interesting to obtain an understanding of these belief functions.

It should be noted, in any case, that the justification for Dempster's generalized Bayesian method offered here depends on the idea of pure randomness and hence only applies when the parametric model consists of all the distributions on $\mathscr{X}$. This rules out Aitchison's counter-example to the method. (See Aitchison, 1968, or Lindley, 1972, p. 9.)

4. PARAMETRIC MODELS NOT BASED ON EVIDENCE

In Chapter 11 of *A Mathematical Theory of Evidence* I suggested a general belief-function treatment of statistical evidence which, in contrast to the methods just discussed, does not depend on the nature of the evidence establishing the parametric model and does not condition on the observations. This method simply translates each observation x into the consonant belief functions on Θ guven by

$$\mathrm{BEL}_x(A) = \sup\{s \mid f_x(\theta) \geqslant 1-s \text{ implies } \theta \in A\}, \tag{19}$$

where $f_x(\theta)$ is the normalized likelihood function:

$$f_x(\theta) = P_\theta(x)/\sup_{\theta' \in \Theta} P_{\theta'}(x).$$

(BEL_{x_i} is determined by the conditions that it be consonant and that it award degree of belief s to each "likelihood interval" $\{\theta \mid f_x(\theta) \geqslant 1-s\}$.)

Many statisticians have discussed the idea of determining degrees of belief by (19). (See, for example, Hudson, 1971, and Edwards, 1972.) But the usefulness of the idea seems to be limited, for one can construct examples where the likelihood function cannot be normalized, or where the normalized likelihood function seems to be misleading. (See Lindley, 1972, pp. 12–13.) I emphasized likelihood intervals in *A Mathematical Theory of Evidence* because of their simple relation to the idea of weights of evidence. But I now think (19) should be rejected as a general method of statistical inference because it does not take into account the origin of the model.

If we do use (19), then how should we combine physically independent observations $x_1, \ldots, x_n$? For each of the three methods we considered above (Smets' method, the fiducial method and the model of pure randomness) there are two different ways of combining observations: (1) A belief function can be constructed on $\Theta \times \mathscr{X}^n$ that has P_θ^n as its marginal given θ, and this belief function can be conditioned on $\mathbf{x} = (x_1, \ldots, x_n)$ to yield a belief function $\mathrm{BEL}_{\mathbf{x}}$ on Θ. (2) A belief function can be constructed on $\Theta \times \mathscr{X}$ that has P_θ as its marginal given θ, for each x_i this belief function can be conditioned on x_i to yield a belief function BEL_{x_i} on Θ, and Dempster's rule can be used to obtain the orthogonal sum $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$. These two ways of combining $x_1, \ldots, x_n$ give the same final result for all three methods: we always find that $\mathrm{BEL}_{\mathbf{x}} = \mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$. In the case of (19) we are not conditioning belief functions constructed on $\Theta \times \mathscr{X}$ or $\Theta \times \mathscr{X}^n$, but we can still distinguish two ways of combining observations: (1) We can represent the physical independence of $x_1, \ldots, x_n$ by constructing the product model $\{P_\theta^n: \theta \in \Theta\}$ and apply (19) directly to this model to obtain a belief function $\mathrm{BEL}_{\mathbf{x}}$. (2) We can apply (19) for each x_i and then combine the resulting belief functions, obtaining the orthogonal sum $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$. And in this case $\mathrm{BEL}_{\mathbf{x}}$ and $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$ will, in general, be different.

Several reviewers of *A Mathematical Theory of Evidence* (see Diaconis, 1977, p. 678; Fine, 1978, p. 671; and Williams, 1978, pp. 384–385) have found the divergence between $\mathrm{BEL}_{\mathbf{x}}$ and $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$ in the case of (19) unacceptable. I am now inclined to agree with them. The choices that a theory of evidence asks us to make ought always to be judgements based on our evidence—i.e. choices for which we can look to our evidence for guidance. And it is not clear how we can use our evidence to choose between $\mathrm{BEL}_{\mathbf{x}}$ and $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$.

The use of likelihood intervals, though unacceptable where the evidence for a parametric model can be spelled out, may still be of interest in cases where there is no evidence for the model—in cases, that is to say, where one is merely trying out the model to see how it fits and what it suggests about θ. Here the arbitrariness of the choice between $\mathrm{BEL}_{\mathbf{x}}$ and $\mathrm{BEL}_{x_1} \oplus \ldots \oplus \mathrm{BEL}_{x_n}$ can be seen as a consequence of the arbitrariness of the model itself.

5. BEYOND THE PARAMETRIC MODEL

In the preceding pages we have seen several examples where evidence conventionally used to justify parametric models can further be used to justify belief-function analyses of those models. The purpose of presenting these examples was to illustrate how the choice of a belief-function analysis depends on the nature of the evidence for the model, not just on the model itself. But a second lesson also emerged from our discussion—the lesson that the evidence for a parametric model often does not justify the model very well and that a belief-function analysis that makes weaker claims on behalf of the evidence may often be appropriate.

It is here, I believe, that the theory of belief functions has the most to offer. There is no great need for new methods of statistical inference for traditional problems where we have well-supported parametric models involving few parameters. But there is a need for new methods for problems where such models are not available. Some Bayesians have sought to address this need by constructing models that have so many parameters that they could not possibly fail to fit the data and then pretending to have prior beliefs about these parameters. The theory of belief function offers an approach that better respects the realities and limitations of our knowledge and evidence.

ACKNOWLEDGEMENT

Research supported by National Science Foundation Grant No. MCS 8002213.

REFERENCES

AITCHISON, J. (1968). In discussion of "A generalization of Bayesian inference", by A. P. Dempster. *J. R. Statist. Soc.* B, 30, 234–237.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. R. Statist. Soc.* B, 35, 189–233.

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.*, 37, 355–374.

—— (1967a). Upper and lower probabilities induced by a multivariate mapping. *Ann. Math. Statist.*, 38, 325–339.

—— (1967b). Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 54, 515–528.

—— (1968a). Upper and lower probabilities generated by a random closed interval. *Ann. Math. Statist.*, 39, 957–966.

—— (1968b). A generalization of Bayesian inference (with discussion). *J. R. Statist. Soc.* B, 30, 205–247.

—— (1969). Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Ann. Math. Statist.*, 40, 953–969.

DEROBERTIS, L. (1978). The use of partial prior knowledge in Bayesian inference. Ph.D. Dissertation, Yale University.

DIACONIS, P. (1978). Review of *A Mathematical Theory of Evidence*. *J. Amer. Statist. Ass.*, 73, 677–678.

EDWARDS, A. W. F. (1972). *Likelihood*. Cambridge: University Press.

FINE, T. L. (1977). Review of *A Mathematical Theory of Evidence*. *Bull. Amer. Math. Soc.*, 83, 667–672.

DE FINETTI, B. (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds). New York: Wiley.

HEWITT, E. and SAVAGE, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, 80, 470–501.

HUDSON, D. J. (1971). Interval estimation from the likelihood function. *J. Roy. Statist. Soc.* B, 256–262.

LINDLEY, D. V. (1972). *Bayesian Statistics, A Review*. Philadelphia: SIAM.

SAVAGE, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

SHAFER, G. (1976a). *A Mathematical Theory of Evidence*. Princeton: University Press.

—— (1976b). A theory of statistical evidence. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (W. Harper and C. A. Hooker, eds), Vol. II, 365–436.

—— (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19, 309–370.

—— (1979). Allocations of probability. *Ann. Prob.*, 7, 827–839.

—— (1981a). Constructive probability. *Synthese*, 48, 1–60.

—— (1981b). Jeffrey's rule of conditioning. *Philos. Sci.*, 48, 337–362.

—— (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* 77, 325–351.

SMETS, P. (1978). Un modèle mathématico-statistique simulant le processus du diagnostic médical. Doctoral Dissertation at the Free University of Brussels, Presses Universitaires de Bruxelles.

WILLIAMS, P. M. (1978). On a new theory of epistemic probability. (Review of *A Mathematical Theory of Evidence*.) *The British Journal for the Philosophy of Science*, 29, 375–387.

DISCUSSION OF PROFESSOR SHAFER'S PAPER

Professor G. A. BARNARD (Retired): The first thing I like about this paper is the development of the author's idea that theories of probability judgement should be thought of in terms of canonical examples with which the theories compare evidence. Indications of such an idea are to be found in Ramsey's 1928 essay and in Koopman's 1930's papers on the foundations of probability; and Allan Birnbaum came close with his term "evidential meaning" used to denote an equivalence class of experimental data, since it is merely a technical step to denote such a class by a canonical member of it. But Shafer's formulation has the great merit of encouraging us to look for evidence of various types, corresponding to various canonical examples. This removes the temptation, so prevalent nowadays, to force all the Protean forms in which uncertainty can present itself into two or three simple frames. The worst offenders in this respect are the strict personal Bayesian, with their insistence that all uncertainties are to be uniformly expressed as personal probabilities. But adherents of the Neyman–Wald ideology are not much better, insisting on reducing all problems to decision making. And when we teach statistical inference with textbooks limiting their categories to "hypothesis testing", point and interval estimation, and perhaps a few frills such as discrimination, we may easily fall guilty of encouraging students to develop similar blinkered views.

Although, as Mahalanobis was fond of pointing out, Hindu philosophers perceived something of the complexity of uncertainty, and such perception of complexity continues to be expressed in the works of J. Bernoulli and J. H. Lambert, from the mid-eighteenth century until Fisher's 1912 paper, with only a suggestion of dissent in S. D. Poisson's treatise, the idea that there may be more than one kind of measurable uncertainty was dormant.

In his 1912 paper Fisher distinguished likelihood from probability though it was not, perhaps, until the 1930s that he fully recognized the concept as of value in itself rather than as a means of arriving at the maximum likelihood estimate and its properties. (Though already in the 1925 (first) edition of *Statistical Methods for Research Workers* he says that "inferences respecting populations . . ." are to be expressed in terms of likelihood, which "does not obey the laws of probability"). In his 1956 *Statistical Methods and Scientific Inference* he explicitly distinguished the "hypothetical" probabilities involved in tests of significance from the (in some sense) real probabilities encountered in games of chance. In his 1957 paper "The underworld of probability", published in *Sankhyā*, he went much further, pointing to uncertainties of Rank A, Rank B and Rank C, with practical examples, and calling for further exploration of other forms. He would, I think, have welcomed Professor Shafer's paper as a development of the kind he was calling for, unless he were unduly put off by Shafer's tolerance for Bayesian notions.

Another attractive feature of Shafer's theory is his way of expressing absence of knowledge by means of a belief function. When Fisher realized the importance for fiducial theory of the condition of ignorance he saw that it would be a major advance if it became possible to give such ignorance formal expression. He also recognized that such formal expression would be difficult though not, as the personal Bayesians suggest, impossible.

It may be worth pointing out in this connection that in suitable circumstances we can give expression to ignorance by a marginalization step. If, from the joint distribution of X and Y, we derive the distribution of X by marginalization this implies that nothing is known about Y other than what is implied by this joint distribution. If, for example, it were known, aside from the joint distribution, that Y lay within a narrow range, the set of associated conditional probabilities might well contradict the marginal distribution for X. In the derivation of inferences about location, using Student's t, the marginalization with respect to the scale parameter expresses ignorance of this parameter.

As another connection with other inferential theories one may mention the work of Graham Wilkinson where, in some problems, he leaves some of the fiducial probability unassigned. This may be compared with the fact that, in general, $\text{BEL}(A)+\text{BEL}(\bar{A})$ is less than 1. In both theories, it may be remarked, the set upon which the uncertainty is defined is taken to be a Boolean algebra, closed under negation. This differentiates the theory from, for example, those of likelihood and plausibility (Barndorff-Nielsen), defined, as these are, on a set not closed under negation, nor, necessarily, disjunction. It is too often forgotten that the negation of a proposition is, typically, a wholly different kind of thing from the proposition itself.

It took nearly 50 years to reach a reasonably clear understanding of the meaning and applications of likelihood; we must therefore be prepared to allow at least a similar period to reach clarity on belief functions. Certainly to me they currently present puzzling features. With the author's Example 1, we have, if $p_1 = 1$ and p_2 is between 0 and 1, the observation $x = 1$ produces belief $1-p_2$ in θ_1 and belief 0 in

θ_2. It could perhaps be argued that the corresponding likelihood ratio $1/p_2$ in this situation underestimates the "degree of confirmation" of θ_1, since it fails to allow for the much greater specificity of θ_1– that is, the fact that just one observation could contradict θ_1, though this is impossible with θ_2. But the comparison of $1-p_2$ for θ_1 with zero for θ_2 seems to err in the other direction. Of course we have to bear in mind that zero does not correspond with total disbelief, and this may be the source of my difficulty. But this difficulty of scaling is for me reinforced by the ploxoma example. Ignoring the qualifications made in the paper concerning the reliability of the data, if we split the 15 per cent probability for (x_2 or x_3), given ordinary ploxoma into 10 per cent for x_2 and 5 per cent for x_3, we obtain a likelihood ratio for virulent ploxoma of 12 : 1, given x_3. This outweighs the prior odds, taking these at their central values of 10 and 90 per cent, giving odds of 1 : 9. The belief function, on the other hand, suggests that the ordinary form is considerably more plausible than the virulent form even when x_3 has been observed. This suggests to me that a set of rough estimates of odds along lines well described in Spiegelhalter's recent paper (*The Statistician*, 31 (1982), pp. 19–36)—incidentally a good illustration of the application of likelihood ideas, in spite of the author's Bayesian prolegomena—would be more usefully interpretable than the belief functions.

This is, I think, the case for a further reason. The analysis using odds ratios shows clearly that the way in which the 15 per cent proportion is split between x_2 and x_3 greatly influences the interpretation of the x_3 result, and so suggests it would be very much worth while to check, if possible, on how this division should be made. It may be that the belief function analysis can indicate equally well where we should look for further information; but for me, for the present, the calculations are too complex. Perhaps one day we shall have hand-held computer programs to get over this difficulty.

Another problem I have with the theory is also perhaps due to unfamiliarity. The connections between the logical structure of the ploxoma example and the logical structure of the story about the uncertain codes are not at all clear to me. While the rules of procedure—Dempster's rule, and the marginal and conditional rules—have reasonably clear justifications when the coding model is known to apply, it is therefore not clear to me how far it is reasonable to carry the rules over to the ploxoma case. Professor Shafer's brilliant presentation of his ideas tonight, and his paper, make it clear that he is not suggesting he has cut and dried solutions for these problems, only that he is giving suggestions which may help us explore largely uncharted territory. If I, for my part, feel the best use of what time I have lies in smoothing existing roads and exploring more byways in the area of more standard statistical models, that is all the more reason for hoping that he and his friends will continue to work along the lines he has opened up and that they will send us back more and more messages in the form of more and more examples of the application of these ideas. If I might suggest one possible such example, it occurs to me that the recent pair of papers by Bickel and Doksum, and by Box and Cox in rebuttal, concerning the use of transformations in the analysis of linear models, could perhaps be looked at in the light of the notions here presented. While it is clear to me that the rebuttal is fully effective, the question arises whether a belief function analysis could provide a formal framework within which the issues involved here could be more explicitly dealt with. In the hope of more to come, I have much pleasure in proposing a hearty vote of thanks.

Dr P. M. Williams (University of Sussex): A belief function in Professor Shafer's sense is intended to measure the degrees of support a body of evidence provides for the various propositions in its domain. Since there need be no positive evidence in favour of either of two alternatives even when there is positive evidence for their disjunction, belief functions are generally non-additive. We may have positive evidence that either A or B is the culprit, if only they hold keys and the safe was not forced, though we may have no positive evidence against either individually. Every belief function is nonetheless the lower envelope of a family of additive distributions but only those lower envelopes for which there exist associated m-functions in the sense of Professor Shafer's identity (2) count as belief functions.

The central tool of the theory is Dempster's rule of combination. It is this that principally distinguishes it in both its aims and methods from either the Bayesian theory or the theory of lower probabilities. Conditioning is a special case of the rule. If we learn that the truth lies in a subset E of Ω then the old value of $m(A)$ is reassigned to $A \cap E$, if this is non-empty, with suitable renormalization. Expressed in terms of its associated plausibility function, the resulting belief function when defined is given by

$$\mathrm{PL}(A \mid E) = \mathrm{PL}(A \cap E)/\mathrm{PL}(E).$$

Example 1 provides a simple example of the strength of one of the techniques Professor Shafer recommends. The posterior marginal for θ depends non-trivially on the two conditionals in accordance with relations (6) even though the prior marginal for θ is vacuous. In such a case the method of lower probabilities would leave us with a vacuous posterior marginal for θ no matter what observations were made. Again in the treatment of "pure randomness" in Section 3.3 the symmetric belief functions BEL_n have vacuous marginals for the frequency but yield the usual additive schemes when conditioned on its value. Incidentally, I doubt whether a belief in the existence of a limiting frequency should be assumed to be already part of the intuitive idea of randomness. It is possibly more interesting to point out that this follows necessarily from the idea of pure randomness expressed by conditions (16)–(18) if "sequential continuity" is assumed.

The techniques which Professor Shafer has developed are undoubtedly powerful and the results which they yield in numerical form appear intuitively plausible. But is there any finer way of judging the theory? Where does its justification lie?

Professor Shafer has presented us with three theories which he calls *constructive*. By this he means that the mathematical formalism of each is supplemented by a class of models or examples which are intended to be "canonical". For the theory of belief functions these concern situations in which the meaning of a message depends on a randomly chosen code. Professor Shafer argues that the quantity $m(A)$ defined by the relation (1) is then, in a sense, the total chance that the true message was A. This is correct from a Bayesian point of view if all possible true messages were *a priori* equiprobable. Then the same is true of all possible coded messages if we think of a coded message as a function from the list of codes to the list of possible true messages. But in that case one who was obliged to specify rates at which he would then have to risk bets concerning the events in Ω, either on or against at the choice of an opponent, might pay more attention to the additive distribution assigning to each element ω of Ω the value

$$\sum_A \{|A|^{-1} m(A) : \omega \in A\},$$

where $|A|$ is the number of elements in A. Under the assumption of a uniform prior this will be the posterior distribution over Ω given certain plausible assumptions. Naturally this reduction of a belief function to an additive distribution discards the structure of the evidence. Both the vacuous belief function and the uniform additive belief function, for example, reduce to the same distribution. We lose the distinction between a complete lack of relevant evidence and the existence of positive evidence favouring each of the possibilities equally. I do not believe, however, that the theory of belief functions should stand or fall on the question of the material adequacy of this analogy with randomly coded messages.

There is an important sense in which Professor Shafer's is the only constructive theory amongst the three mentioned. It seeks to construct a belief function representing the effect of total evidence by decomposing it into separate unrelated items. Each item by itself is supposed to determine a belief function, in the judgement of some individual, without relation to any other item of evidence or prior belief. The overall belief function is obtained by combining these separate belief functions by Dempster's rule. (See the remarks on "limited evidence" in Section 1.3.) It is important to emphasize that this is very different from the Bayesian theory according to which nothing can normally be said about the effect of new evidence without reference to prior beliefs. This is because in the Bayesian theory, or extensions of it based on principles of maximum entropy or minimum information for example, the stimulus to which change of belief is the response takes the form of a direct constraint on the posterior distribution which expresses the effect of the total evidence. This is equally true of the theory of lower probabilities. In this respect Professor Shafer's theory is attempting to answer a question which the other theories choose not to ask, namely one concerning the effect of a limited portion of the evidence taken by itself prior to combination. There are certainly cases where it is intuitively natural and possibly unavoidable to raise such a question—legal proceedings for example—and Professor Shafer has now shown that problems of statistical reasoning can also be put in this form. He has shown that the analysis yields mathematically significant and intuitively plausible results. I believe nonetheless that a deeper justification of the method is still required. It may need to provide a further treatment of the idea of "unrelated bodies of evidence" or else to show how the discovery of relations between various items of evidence can itself function as an item of evidence. If we agree that a principle is a general rule that treats similar cases in a similar way, such a justification would best of all demonstrate that Dempster's rule of combination is the unique principle governing processes of this type. These are interesting and important problems. We owe

Professor Shafer a debt of gratitude for the energy and insight he has shown in drawing attention to them and to their possible solution, both in the present paper and in a series of recent publications, and I am sincerely pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

Professor P. SMETS (Brussels): I have been most interested in the comparative study of the various inferences presented by Professor Shafer. The idea of selecting an appropriate model for the representation of ignorance through canonical examples is most useful but I think, nevertheless, that the underlying axiomatization is also needed in order to understand the basic assumptions of each model.

I want to comment first on Shafer's claim that the evidential inference as developed in Smets (1978) gives practical results essentially when Θ is finite. The equation is

$$\textsc{bel}(\theta\,|\,x) = (\prod_{t\in\bar\theta} P_t(\bar x) - \prod_{t\in\Theta} P_t(\bar x))/(1-\prod_{t\in\Theta} P_t(\bar x)), \qquad (1)$$

where t are the singletons of Θ, and the prior belief function on X is vacuous. Its extension to the case where Θ is infinite is straightforward but $\textsc{bel}(\theta\,|\,x) = 0$ for all θ in Θ except when $\bar\theta$ is finite. It represents a highly uninformative belief function. When one starts with total ignorance on an infinite space, it is indeed predictable that the *a posteriori* belief function one could derive from some observation should not be much different from total ignorance: some kind of infinite information is needed to transform infinite ignorance into some knowledge.

One might feel that my evidential inference method suffers from a serious weakness inasmuch as the *a posteriori* belief function can only be computed and applied when Θ is finite. One could start with infinite Θ but relation (1) applies only when one conditions the *a posteriori* belief function on some finite subset T of Θ. The *a posteriori* belief function can only be described explicitly on such finite T. This fact is shared by another highly respectable measure of belief: the probability function (Fine, 1973). Let N be the set of integers and select an integer at random, any integer being equiprobable. A probability function cannot be described on N, but whenever one conditions on any finite subset of N, one obtains a well-defined conditional probability function. Of course, this example does not justify our model, but shows that the so-called weakness is also encountered with probability functions.

My second remark deals with the normalized likelihood function that Shafer used in Chapter 11 of *A Mathematical Theory of Evidence* and now rejects as unfit for a general method of statistical inference. Indeed this model leads to contradictory results when one considers two independent observations and combines the informations either before or after computing the *a posteriori* belief function. This contradiction was the reason why I developed my method. Nevertheless the likelihood function is an excellent method for statistical inference when placed in the theory of possibility as developed by Zadeh (1978). A possibility function POSS is a set function such that for all A and B in its domain

$$\textsc{poss}(A \cup B) = \max(\textsc{poss}(A), \textsc{poss}(B))$$

POSS is defined by its value on the singletons of its domain. It obeys the same rules as Shafer's consonant plausibility functions, but covers different situations. It deals with problems like "the possibility to put n tennis balls in a given wooden box", a situation where randomness and belief are irrelevant.

If one starts with a possibility function on Θ and, if for all singletons θ of Θ one has a probability function $P_\theta(x)$ on X, then the *a posteriori* possibility function for the singletons θ of Θ given x is

$$\textsc{poss}(\theta\,|\,x) = P_\theta(x)/\max_{t\in\Theta} P_t(x),$$

the normalized likelihood function (Smets, 1982). Of course Dempster's rule for the combination of evidence does not apply to the possibility function and is replaced by the following possibilistic rule of combination

$$\textsc{poss}_{1,2}(\theta) = \textsc{poss}_1(\theta)\,\textsc{poss}_2(\theta)/\max_{u\in\Theta}(\textsc{poss}_1(u)\,\textsc{poss}_2(u)).$$

With such a combination rule and statistical inference based on the normalized likelihood function, the results are identical when one considers two independent observations and decides to take inference directly from $\Theta \times X$ or to combine the *a posteriori* possibility function that can be inferred from each individual observations (Smets, 1982). Likelihood function can be perfectly integrated in a well-structured theory and one might wonder if this possibility theory could not be used to justify inference based on the likelihood.

Dr P. WALLEY (University of Warwick): This is an adventurous and stimulating paper. Reading Professor Shafer's book in 1977 led me to study belief function representations of statistical evidence. Reluctantly, I reached the conclusion that an approach along the lines of this paper, based on Dempster's rule of combination, is unlikely to succeed. I have space here only to outline the difficulties, and must omit discussion of many important points of agreement with Shafer—notably his emphasis on *constructing* probability judgements.

(i) Assume $\Theta = \{\theta_1, \ldots, \theta_N\}$ is finite. PL_x denotes the plausibility function on Θ induced by an observation x, and is regarded as a function of $P_{\theta_1}(x), \ldots, P_{\theta_N}(x)$. (The discussion is a little simpler in terms of PL_x rather than BEL_x.) We look for functions with the property that $\mathrm{PL}_{(x,y)}$ based on independent observations x and y agrees with the combination $\mathrm{PL}_x \oplus \mathrm{PL}_y$ by Dempster's rule. Formally, we require

(1) $\mathrm{PL}_x(A) = \mathrm{PL}(A, P_{\theta_1}(x), \ldots, P_{\theta_N}(x))$ is defined for all $A \subseteq \Theta$ and $0 \leqslant P_{\theta_j}(x) \leqslant 1$, provided $P_{\theta_j}(x) > 0$ for some j.

(2) PL_x is a plausibility function on Θ.

(3) When x and y are independent under each $\theta \in \Theta$, $\mathrm{PL}_{(x,y)} = \mathrm{PL}_x \oplus \mathrm{PL}_y$.

(4) $\mathrm{PL}_x(\{\theta_j\})$ is non-increasing in $P_{\theta_i}(x)$ when $i \neq j$.

(5) When PL_0 is additive on Θ, $\mathrm{PL}_0 \oplus \mathrm{PL}_x$ agrees with the combination of PL_0 and $\{P_\theta(x) \colon \theta \in \Theta\}$ by Bayes' rule.

There is a one-to-one correspondence between functions PL satisfying 1–5 and partitions of Θ. Given a partition $\{A_1, \ldots, A_r\}$, the corresponding PL_x have additive restrictions to the partition, with

$$\mathrm{PL}_x(B) = k^{-1}(1 - \prod_{\theta \in B} [1 - P_\theta(x)]) \quad \text{for } B \subseteq A_j,$$

where

$$k = \sum_{j=1}^{r} (1 - \prod_{\theta \in A_j} [1 - P_\theta(x)]).$$

This result brings together two of Shafer's methods. The trivial partition ($r = 1$) gives Smets' method (Section 3.1). For the finest partition ($r = N$), PL_x is additive with $\mathrm{PL}_x(\{\theta\}) \propto P_\theta(x)$, which agrees with the "fiducial" method (Section 3.2).

(ii) Consider the constant likelihood $P_\theta(x) \equiv \rho$. Apart from the fiducial case, PL_x depends on ρ, violating the Strong Likelihood Principle; e.g. each $\mathrm{PL}_x(\cdot \mid A_j)$ is vacuous for $\rho = 1$ but not for $\rho < 1$. This presents difficulties in specifying an experimental outcome. In Shafer's Example 2, suppose we observe the patient's sex (y), and this is unrelated to the result of the blood test (x) and type of ploxoma (θ). Except in the fiducial case, $\mathrm{PL}_{(x_1,y)} \neq \mathrm{PL}_{x_1}$. It is disturbing that our conclusions might depend on whether we report the "uninformative" ancillary y.

(iii) Shafer constructs BEL and PL on subsets of $\Theta \times \mathcal{X}$. Consider

(6) If $P_\theta(D) = \rho$ for all $\theta \in \Theta$ then $\mathrm{BEL}(\Theta \times D) = \mathrm{PL}(\Theta \times D) = \rho$.

This condition seems reasonable under a "constructive" interpretation of probabilities, and also under frequentist or betting-rate interpretations. The next examples show that Shafer's first two methods violate (6).

(a) Let $N = 2$, $P_{\theta_1}(x_1) = P_{\theta_2}(x_1) = 0{\cdot}5$. For Smets' method, $\mathrm{BEL}(\Theta \times \{x_1\}) = \prod_\theta P_\theta(x_1) = 0{\cdot}25$, $\mathrm{PL}(\Theta \times \{x_1\}) = 1 - \prod_\theta [1 - P_\theta(x_1)] = 0{\cdot}75$.

(b) Let $\mathcal{X} = \Theta = \{0, 1, 2, 3\}$, $x = \theta + e \pmod 4$, $P(0| = 0{\cdot}4$, $P(1) = 0{\cdot}3$, $P(2) = 0{\cdot}1$, $P(3) = 0{\cdot}2$, with BEL induced by $P(e)$ as in Section 3.2. If $D = \{0, 2\}$ then $P_\theta(D) = 0{\cdot}5$ for each θ, but the fiducial method gives $\mathrm{BEL}(\Theta \times D) = 0$, $\mathrm{PL}(\Theta \times D) = 1$.

(iv) Such counter-intuitive properties of the solutions in (i) suggest that we weaken condition (3) by allowing other rules of combination. For example, simple axioms imply $\mathrm{PL}_x(A) = \sup_{\theta \in A} f_x(\theta)$, where f_x is the normalized likelihood function. This is equivalent to Shafer's equation (19). Condition (3) then implies a new rule of combination, which disagrees with Dempster's rule in general, but agrees in the case of conditioning. We might then try to define PL on $\Theta \times \mathcal{X}$, with vacuous Θ-marginal, so that PL_x is obtained by conditioning (by Dempster's rule) on $\Theta \times \{x\}$, and P_θ by conditioning on $\{\theta\} \times \mathcal{X}$. That cannot be done if PL is required to be a plausibility function, but can be achieved by the upper probability function $\mathrm{PL}(A) = \sup_{\theta \in \Theta} P_\theta(A_\theta)$, where $A_\theta = \{x \colon (\theta, x) \in A\}$, which also satisfies (6). In order to reconcile basic intuitions about "plausibility", it seems necessary to abandon both Dempster's rule of combination and the restriction to plausibility or belief functions.

(v) Even Dempster's rule of conditioning must be abandoned if one wants to interpret PL and BEL as sensible upper and lower betting rates, as Smith (1961). But there is a promising alternative—the theory of upper *and* lower probability, developed by Smith, Williams (1976) and Walley (1981, 1982). Shafer mentions one interpretation of upper and lower probabilities: instead of knowing precisely a distribution P, we know only that P is in some class of distributions. That is the "Bayesian sensitivity analysis" interpretation. An alternative "direct" interpretation, in terms of betting behaviour or other actions, seems more promising because it does not invoke an "underlying" additive P. When the direct interpretation is followed through we find important divergences from Bayesian sensitivity analysis, e.g. concerning independence and exchangeability, with serious consequences for statistical inference.

Professor J. F. C. Kingman (Oxford University): It may be helpful to remark that the quantities $m(B)$ in (2) define the distribution of a random subset X of Ω, and that

$$\mathrm{BEL}(A) = \mathrm{Prob}(X \subset A),$$

$$\mathrm{PL}(A) = \mathrm{Prob}(X \text{ hits } A).$$

If the examples in the paper are interpreted in this way, they become easier to visualize, and the great variety of different belief functions becomes more evident. Moreover, it is easier to see the necessary adjustments to make the theory work smoothly when Ω is infinite.

The following contributions were received in writing, after the meeting.

Professor A. P. Dawid (University College London): The situation considered in Section 3.2 would appear to cover the general *functional model* (Dawid and Stone, 1982) in which we have, for each value of an "error" variable e, a one–one correspondence between the values of θ and of x, although x and θ together may not determine e. An example, with $e = (e_1, e_2)$, is given by $x = (\theta+e_1)/e_2$, $\theta = xe_2-e_1$. Assigning to e its fixed distribution (e.g. $e_1 \sim N(0,1)$, $e_2 \sim \surd(\chi_\nu^2/\nu)$ independently), the first equation induces the model distributions for x given θ, and the second the fiducial distributions for θ given x, here justified, it seems, by the theory of belief functions. But in fact this is a vacuous justification, since an unjustified fiducial step lies at the heart of that theory. We can consider θ as a message, e as a code and x as the coded message. Then probabilities are assigned to θ, after observing x, on the basis that the original distribution assigned to e is still relevant—a step willingly taken by Fisher, Fraser and Shafer, but by few others.

Can the theory of belief functions, suitably extended to conditioning on probability-zero events, extricate fiducial inference from the following anomaly (Dempster, 1963, Dawid and Stone, 1982)? We start with the functional model:

$$\left.\begin{aligned} \bar{x} &= \mu+\sigma e_1 \\ s &= \sigma e_2 \end{aligned}\right\}.$$

What is the appropriate distribution for (μ, σ) based on data $(\bar{x}, s)$ and the *extra information* that $\mu/\sigma = k$? If the data become available first, we might calculate the joint fiducial distribution of (μ, σ), and then condition this on the value k for μ/σ. Alternatively, inserting the information $\mu/\sigma = k$ into the model, we discover that $(k+e_1)/e_2 = \bar{x}/s$. We have therefore learned something from the information and the data about the code e used, and so can proceed with the fiducial analysis after first conditioning the distribution of e on the value $\bar{x}/s$ for $(k+e_1)/e_2$. However, these two routes yield different answers, although the logic underlying Dempster's rule appears to apply with equal force to both.

I foresee equally difficult problems of conditioning in other continuous problems: e.g. with $x = (\theta+e_1)/e_2$, $y = (\theta+f_1)/f_2$, how do we condition, after seeing (x, y), on the discovered fact that $xe_2-e_1 = yf_2-f_1$? To do so, we should embed this information in a suitable partition, but this is problematical (see Brenner and Fraser, 1979). Indeed conditioning even in finite problems is not immune from such considerations (Dawid and Dickey, 1977), so that Dempster's rule must be treated with caution.

Professor Terrence L. Fine (Cornell University): Professor Shafer is to be congratulated on his sustained and thoughtful development of the theory and application of belief functions. I welcome the present contribution in which he clarifies how the weight and structure of evidence come to bear on the selection of probability models. The constructive view of probability that is espoused is plausible, particularly in the Bayesian/subjectivist/personalist context wherein individuals are often encouraged to scale their degree of conviction through reference to familiar chance setups. However, I am not clear about the explicit bearing of the constructive viewpoint on either lower probabilities or belief functions.

A given lower probability P_* can be derived as the lower envelope of rather different sets of probability measures, and this absence of unicity suggests that the class $\mathscr{P}$ does not constitute a suitable canonical example. Furthermore, the coded message interpretation suggests for belief functions, while providing a canonical scale, is ignored when Professor Shafer actually constructs belief function models in Section 3, thereby calling into question the relevance of the canonical scale.

The Dempster rule of combination for belief functions, while staunchly defended by Professor Shafer, numbers me among its opponents. I continue to fail to see how a single such prescription can account for the interactions between such different sets of evidence as those leading to the parametric model and those informing us about the parameter value.

The defence or justification of new approaches to probabilistic reasoning is a difficult and poorly understood process. I presume that Professor Shafer means us to bring our intuitions to bear on his examples and thence find his conclusions satisfactory in some private epistemic sense. I would prefer to see more explicit arguments for the acceptability of the suggested processes of constructing belief functions, although I recall that in his earlier work Professor Shafer expressed little hope for such arguments. For example we might wish to inquire into the use to which belief functions will be put. After all, a great strength of Bayesian probability is the close articulation between an individual's assessments of probabilities and the use which the individual makes of these probabilities in forming rational decisions. Considerations of belief function-based decision-making lead to the use of upper and lower expectations. The upper and lower expectations in turn lead us to consider the probability measures dominating the belief function; given a random variable X its, say, lower expectation E_*X is in fact equal to the usual expectation $E_\Gamma X$ for some Γ dominating $((\forall A)\Gamma(A) \geqslant \text{BEL}(A))\text{BEL}$. While Professor Shafer may not feel that the set of measures dominating BEL need have any direct meaning, it might nonetheless be disturbing to find, say, in a model of independent observations that E_*X is achieved by a measure Γ that is not a product measure.

The preceding cavils, notwithstanding, I believe that the theory of belief functions is well worth study, and I agree with Professor Shafer's emphasis on the structure of evidence.

Professor D. V. LINDLEY (Somerset): I will discuss Example 2 and try to show that probability concepts are adequate. The first piece of evidence (i) establishes in the usual way that the chances for a person with virulent ploxoma to have blood-test results of types x_1, x_2 and x_3 are 0·2, 0·2 and 0·6. The second (ii) is subtler for two reasons: x_2 and x_3 are not distinguished in the data, and the patients in the study are not judged exchangeable with other patients so that the chances β in the study and γ for the new patients are not necessarily equal. The first presents no difficulty since the likelihood for the data is $\beta_1^r(\beta_2+\beta_3)^{n-r}$ where $r = 0{\cdot}85n$ and n is the number of patients in the study. The distribution of β given the data can therefore be found. Let $p(\gamma\,|\,\beta)$ be the conditional distribution of γ, given β. This concept replaces the single figure of 75 per cent quoted by Shafer and which yields a discount rate of $\alpha = 0{\cdot}25$. It would be possible to suppose $\gamma = \beta$ with probability 0·75 and is otherwise uniform in the unit interval in imitation of belief functions; but this may be an unrealistic description of the situation. The third piece of evidence (iii) says the distribution of the chance θ that a patient has virulent ploxoma, $p(\theta)$, is essentially confined to the range (0·05. 0·15). We are now ready to perform the requisite probability calculations.

Let G be the event that a new patient, George, has virulent ploxoma and let g_i be the result of his blood test. We require $p(G\,|\,g_i, E)$ where E is the evidence. From (iii) $p(G) = \int \theta p(\theta)\,d\theta$. From (i) $p(g_i\,|\,G, E) = 0{\cdot}2$ for $i = 1$, 2 and 0·6 for $i = 3$. From (ii)

$$p(g_i\,|\,\bar{G}, E) = \int\!\!\int \gamma_i\, p(\gamma\,|\,\beta)\, p(\beta\,|\,E)\, d\beta\, d\gamma$$

$$= \int E(\gamma_i\,|\,\beta)\, p(\beta\,|\,E)\, d\beta$$

and the calculations can be completed in the usual way using Bayes' theorem. If $E(\theta) = 0{\cdot}10$, $E(\gamma_i\,|\,\beta) = \beta_i$ and $E(\beta_2\,|\,\beta_1) = \tfrac{1}{2}(1-\beta_1)$ then the probabilities of G given g_i are respectively 0·025, 0·229 and 0·471.

It may be objected that this analysis virtually ignores the uncertainty about the study and about θ. It does so because they are irrelevant. The interested reader may like to consider the case of George and

Henry and their blood tests. Then the uncertainties will matter: for example, $E(\gamma_i^2 \mid \beta)$, involving the conditional variance of γ_i, will arise.

My view is that, as here, anything belief functions can do probability can do as well, but with the advantage that the results will be coherent and have an operational meaning. The only description of uncertainty is probability.

Professor C. A. B. Smith (University College London): Professor Shafer's paper represents a new, ingenious and attractive attempt to overcome problems in statistical inference. For example, one would very much like to turn observational evidence directly into a probability, or at least something resembling Shafer's belief functions. Yet, according to subjective probability theory, observational evidence gives only likelihoods, which are factors modifying degrees of belief according to Bayes' theorem. Can one do better? The world's energy problem would be solved if one could find an infinite source of energy. But the law of conservation of energy forbids that. However attractive it would be to obtain probabilities which are functions only of the observations, if such probabilities are to be used for decisions (by one person), e.g. medical treatment following diagnosis, either Bayes' theorem effectively holds, or decisions are "incoherent", i.e. contradictory (Savage, 1954; Smith, 1961). Even when no decisions are involved, this conclusion still seems to hold (Smith, 1977). These considerations appear to impose uncomfortable limits on any "reasonable" theory of inference, just as conservation of energy or the Arrow paradox impose uncomfortable but unavoidable restrictions. This point was of course made by Savage in 1954; it does not lose its validity with age.

Professor David H. Krantz (Bell Laboratories): Professor Shafer's theory of belief functions is a valuable generalization of the Bayesian theory, chiefly because it provides a natural account of prior ignorance or weak evidence. In the paper under discussion, this is illustrated by a variety of examples in which the marginal belief functions over parameter spaces are vacuous; thus, the proposed statistical analyses can be used on a background of complete prior ignorance or any degree of weak or strong prior evidence.

Shafer's methods would be even more useful, and more readily accepted, if a satisfactory interpretation could be devised for numerical degrees of belief. Section 1 of the paper suggests such an interpretation: evaluate a belief function numerically by comparing a body of evidence to a probabilistically coded message. I find this suggestion unsatisfactory for two reasons. First, such comparisons seem strained; second, the procedure only seems numerical, whereas in fact it is merely ordinal.

Consider an analogy: suppose one attempted to measure the "degree of aesthetic pleasure" from viewing a painting by comparing the experience to the sweetness of a graded series of sucrose solutions of known concentration. First, the two realms of experience do not match very well. And, second, the numerical value of sucrose concentration yields (at best) an ordinal scale of "aesthetic pleasure", because no empirical relations in the aesthetic domain are represented, except ordering. Similarly, measurement of mass by comparison of objects with a graded series of weights in a pan balance would only yield ordinal measurement, were it not for the fact that the mass numbers are required to represent an additional empirical relation: combination of parts into a whole is represented by addition of the corresponding mass values. This leads to a ratio scale of mass.

Returning to Shafer's suggested interpretation, the comparison of evidence to a probabilistically coded message seems strained; moreover, such comparisons run counter to the basic rationale for belief functions. For example, consider the reliability of a witness. One's judgement about such a question would typically be based on just the sort of evidence that Shafer analyses elsewhere using belief functions rather than additive probability. If one could justify comparing the judgement of reliability to a random drawing of "reliable" or "unreliable" balls from an urn, why not do the same with one's judgement of the fact about which the witness is testifying, eliminating altogether the need for non-additive belief functions?

In place of probabilistically coded messages, I would suggest two kinds of canonical examples that could provide comparisons to many, though not all, evidential situations. First, when the evidence consists of (a) the similarity of the current situation to many past situations and (b) the record of past experience in the aforesaid similar situations, then it seems natural to compare one's knowledge to an additive probability distribution. For example, "He's late about 60 per cent of the time" summarizes such a comparison. Second, when evidence consists of an observation that would be usual under hypothesis H but rarer or astonishing under H', then it is the likelihood ratio that provides a family of apt canonical examples. For example, "10 to 1 it wasn't an accident" ordinarily should be interpreted as a judgement of likelihood ratio, rather than of posterior odds.

The use of additive probabilities and likelihoods as canonical examples does not, however, require me, or even tempt me, to be Bayesian. The advantages of belief functions are great and need not be surrendered. The comparison of given evidence with the above examples is merely ordinal, as argued above. That is, a numerical measure of strength of evidence should be a monotonic function of the matching probability value, in the first type of example, and should be monotonic with likelihood in the second one. The measurement on a cardinal scale can be obtained from the requirement that strength-of-evidence values be combined by Dempster's rule when independent lines of evidence are combined, just as ratio-scale measurement of mass is obtained from the requirement that combination of parts be represented by additivity of mass values. Details of such a measurement scheme for belief functions have been worked out by John Miyamoto and me, but are not yet published. How well the scheme will work in practice is not known at present.

I do not agree with Shafer that the constructive view of probability should be regarded as an alternative to the objectivistic, personalistic and necessary views. Rather, it seems to be a correction applicable to each of them, since some matching process, involving canonical examples, may play an important role in a practical realization of each.

Turning to Sections 2 and 3 of Shafer's paper, I see a difficulty with the method of conditional embedding: equation (6) implies that a likelihood ratio of 1 has non-trivial effects on belief! Since conditional embedding seems attractive, this paradox may be hard to resolve. Perhaps it is a mistake to try ever to incorporate evidence into the domain of belief functions; statistical evidence, in particular, may be better handled via methods based on likelihood, e.g. equation (19): I find it unnecessary to reject such methods just because statistical and epistemic combination rules yield conflicting results (end of Section 4). This difficulty can be resolved another way; epistemic combination (Dempster's rule) is simply inappropriate when applied to different observations governed by the same chance model. This is not an *ad hoc* adjustment, for there is no reason to apply Dempster's rule except under the limited conditions of bodies of evidence that are entirely unrelated except via their bearing on common hypotheses. Observations that are statistically independent are by definition related by a common chance model in which this statistical independence is expressed, and they should be combined within the model, not by Dempster's rule.

In my view, the major puzzle about parametric models lies not so much in the direction of statistical inference about parameter values—since that can be dealt with through the likelihood function—as in the direction of statistical prediction or explanation. What kind of belief function over X is generated by a family of parametric models M, each providing a probability distribution over X, together with a belief function over M? The method of conditional embedding suggests an interesting possible answer, one which cries out for an axiomatic analysis.

The AUTHOR replied later, in writing, as follows.

I would like to thank Professor Barnard, Dr Williams and the other contributors to the discussion for their thoughtful and constructive comments. I would also like to express my gratitude to the Royal Statistical Society for the opportunity to present my paper and for the very pleasant reception they gave me.

Probability analyses as arguments. I am pleased by the positive reaction to the general idea of constructive probability. There are aspects of this idea which were not developed in my paper but which are relevant to the discussion. One of these is best expressed in the words of Amos Tversky: "The result of a probability analysis is only an argument." Suppose we compare various items of evidence to the canonical examples for a particular theory, combine the results by the rules of the theory, and end up with a degree of belief 0·99 that A is true. Then we have constructed an argument for A. We have said, "Look, our evidence, when put together in this way, strongly supports A." Other people may or may not find this argument convincing, and if they do not find it convincing they may try to improve on it or try to construct a different argument with a different conclusion. Thinking of probability analyses in this way can help us resolve some of the difficulties which arise if we look for probability judgements that are based on logic alone or are otherwise apodictic.

Consider, for example, Professor Dawid's comments on the fiducial step he sees at the heart of the theory of belief functions. Suppose I receive a coded message and I know there was a chance 0·99 that it was encoded using the code c. I decode using c and obtain the message A. This strikes me as a strong argument for A. Perhaps Dawid is right that few people would agree that A now has, in some apodictic sense, probability 0·99. But most people would agree that there is a strong case for A, and this is all a constructive theory can hope for.

Seeing probability analyses as arguments also opens the way to a pragmatism that allows our approach to probability judgement to be influenced by our purposes. My remark that Smets' method is sensible only when Θ is small might be better expressed by saying that the method is not appropriate for the purposes we usually have in mind when Θ is large. Similarly, we may find Dr Walley's condition (6) attractive if we are interested in prediction but irrelevant if we are interested in inference.

Belief functions—canonical examples and axiomatic foundations. Professors Barnard and Fine are troubled because they do not see a very clear fit between randomly coded messages and the items of evidence that I represent by belief functions in Section 3 of the paper. I share their misgivings to some extent. It is not clear, for example, that a study of symptoms of victims of virulent ploxoma has the same structure, as evidence, as a message that has a certain chance of meaning that a given victim will have a given symptom if his ploxoma is virulent. It does not, at any rate, seem obligatory that we should think of this evidence as having this structure. To a certain extent our problem is, as Barnard suggests, one of unfamiliarity. The abandon with which Bayesians assimilate all sorts of evidence to the familiar model where truth is generated by chance suggests that with practice we could fit any evidence to a randomly coded message. Still, one would like the fit to be clearer.

David Krantz, in private conversation, has pointed out to me that we want the fit to be clear because the canonical examples for belief functions carry so much weight—they justify the whole calculus of the theory. If this calculus could be given a more fundamental justification, then some of this weight would be removed. We would still need canonical examples in order to scale the strength of individual items of evidence, but we would not need to match the structure of our evidence to these canonical examples. I am intrigued, therefore, by Dr Williams' comments about how a deeper justification of Dempster's rule might be sought. I hope that both Williams and Krantz will further develop their ideas on these issues.

In the absence of axiomatic foundations of the sort to which Williams and Krantz point, I will continue to rely on the canonical examples of the theory of belief functions as the source of the theory's structure. This approach is often adequate for the construction of convincing probability analyses and since the canonical examples use traditional probability ideas, this approach has the virtue of drawing the theory closer to the statistical tradition.

Alternative canonical examples for belief functions. In his contribution to the discussion, Krantz emphasizes not the need for axiomatic justification of Dempster's rule, but rather the possibility, once we have accepted Dempster's rule, of giving a fully cardinal meaning to the values of belief functions. The argument he has in mind involves adopting an alternative set of canonical examples for belief functions: additive probability distributions and likelihood ratios. I find the idea of a cardinal scale intriguing, but I have not yet convinced myself of its importance. Does a probability analysis based on using randomly coded messages as the canonical examples for belief functions lack effectiveness as an argument because the degrees of belief it produces have only ordinal meaning? I am inclined to resist a reliance on likelihood ratios as canonical examples for belief functions because I believe in many problems the theory of belief functions can do better than the Bayesian theory precisely because it allows us to avoid forcing our evidence into the likelihood mould. One example is discussed in Shafer (1981d). Another is provided by the ploxoma example in the paper under discussion.

Aleatory vs epistemic combination. Several participants in the discussion touch on the question of whether Dempster's rule should be used to combine belief functions based on statistically independent observations or whether such observations should be combined within one's chance model before undertaking a belief-function analysis. The fundamental question is whether the statistical independence of observations can be taken as satisfying the intuitive criterion of independence required by Dempster's rule. Smets and Walley say yes, and so regard the discrepancy between the two methods of combination in the case of the likelihood method (formula (19)) as an unacceptable "contradiction". Krantz says no, on the grounds that belief functions based on statistically independent observations represent overlapping evidence; both rely on the evidence for the statistical model. I have changed my own opinion on this issue several times, and I continue to waver. I do believe that there are many problems where features of the evidence that statisticians are accustomed to representing as statistical independence can be alternatively thought of as justifying Dempster's rule. And so I find the concordance between the two methods of combination in the examples considered in Section 3 of my paper reassuring. On the other hand, I see the justice of Krantz's view once we have accepted the statistical model.

Canonical examples for other theories. Professor Fine notes that a given lower probability can be the lower envelope of many different sets of probability measures. I do not see that this fact renders an example involving partial knowledge of chances inappropriate as a canonical example for a constructive theory, though it does mean that this theory can be richer than the name "lower probability" might suggest. (See Shafer, 1981a, pp. 11–12.) Perhaps there are other constructive theories that can also be called theories of lower probabilities—i.e. other useful sets of canonical examples for lower probability functions. I would be interested in Fine's thoughts on this. Are there, for example, canonical examples that better fit the techniques developed in Wolfenson and Fine (1982)?

I am puzzled by Dr Walley's suggestion that a betting interpretation can replace the canonical examples I have suggested for lower probabilities and by his related claim that Dempster's rule must be abandoned if we are to interpret BEL(A) and PL(A) as lower and upper betting rates. It seems to me that the degrees of belief constructed within the theory of belief functions and those constructed within the theory of lower probabilities can equally well be used as betting rates. The idea of betting cannot distinguish between these two very different constructive theories. (See Shafer, 1981a, pp. 16–40.)

Professor Smets mentions Zadeh's theory of possibility. What is meant when one says, "The possibility of putting 10 tennis balls in this box is $\frac{1}{2}$"? Perhaps it would be helpful to have canonical examples for this theory.

Canonical functional models. I am fascinated by Professor Dawid's example of a continuous functional model that yields different fiducial distributions depending on how the problem of conditioning on an event of zero probability is handled. I do think the theory of belief functions can cast some light on this example. In contrast to Fisher's fiducial argument, the theory of belief functions allows us to take discrete models as basic and to think of continuous models as mathematically convenient approximations. This might mean that the correct result from conditioning a continuous belief function on an event of zero plausibility could only be found by referring to the exact discrete model. But presumably a continuous belief function would be counted as useful only if this correct result could be found as the limit of the results from conditioning on a sequence of events of smaller and smaller positive plausibility. This idea is familiar from standard probability theory, as is the concomitant idea that in order to condition on an event of zero plausibility we need to specify which decreasing sequence of events of positive plausibility we have in mind. In this context, Dawid's example runs as follows. We have a continuous belief function for $(\bar{X}, \bar{S}, \mu, \sigma)$ such that $\{\bar{X} = \bar{x},\ S = s,\ \mu/\sigma = k\}$ has zero plausibility, yet (i) $\{\bar{X} = \bar{x},\ S = s,\ |\mu/\sigma - k| \leqslant \varepsilon\}$ and (ii) $\{\bar{X} = \bar{x},\ |\bar{X}/S - \bar{x}/s| \leqslant \varepsilon.\ \mu/\sigma = k\}$ both have positive plausibility for $\varepsilon > 0$. And the limit obtained after conditioning on (i) differs from the limit obtained after conditioning on (ii). This example shows that life is more complicated with continuous belief functions than with continuous additive probability distributions; in a well-behaved additive probability space we would take it for granted that a conditional distribution is uniquely determined by values of the random variables $\bar{X}$, S and μ/σ. But there is no fundamental paradox; it is simply necessary to say whether (i) or (ii) is meant. Presumably a convincing story about the evidence that led us to construct the functional model would say which is meant.

I agree with Dawid that Bayesian conditioning is subject to pitfalls even in finite problems and that we need to watch for these same pitfalls when we use Dempster's rule of conditioning. One way of explaining the problem is to say that Bayesian conditioning is justified only when the event that we receive the information we want to condition on can be considered independent of the event that the information is true. (See Shafer, 1981c, 1982b.) Conundrums arise because this is not made explicit in the standard presentations of the Bayesian theory. I have tried to make the corresponding requirement of independence quite explicit in my presentations of Dempster's rule.

Ploxoma. Since I believe that theories of probability judgement must ultimately be compared in terms of examples, I am delighted to see that the discussion includes two alternative analyses of the ploxoma example—a likelihood analysis by Professor Barnard and a Bayesian analysis by Professor Lindley. Both these analyses disagree with the belief-function analysis given in Example 2 in that they show greater support for virulence in the case of a ploxoma patient whose test comes out x_3. Table D1 puts the results of Lindley's analysis and the results of the belief-function analysis in a form where they can be directly compared. When the test comes out x_1, the two analyses are in rough agreement, but when it comes out x_2 or x_3, the belief-function result is more conservative, in the sense that it stays closer to the prior 5–15 per cent for virulence. This is because the belief-function analysis discounts the results of the study of patients with ordinary ploxoma. Lindley insists that the uncertainties affecting this study are irrelevant and should be ignored. Is this reasonable? Suppose that instead of having only 75 per cent confidence in the study we have much less confidence. Is there not some point where even Lindley would chuck out the study and revert to the prior 5–15 per cent?

TABLE D1

		Lindley	*Example 2*	
		P (*virulence*)	BEL (*virulence*)	PL (*virulence*)
Prior		0·10	0·05	0·15
Posterior	x_1	0·025	0·014	0·035
	x_2	0·229	0·062	0·082
	x_3	0·471	0·165	0·218

Observational evidence. Professor Smith reminds us that "observational evidence gives only likelihoods". The constructive view suggests that we express this somewhat differently. We should say, "Observational evidence is evidence that we have decided to assess in terms of its likelihood." Evidence does not come to us labelled "observational" or "other"; even within the Bayesian framework, we must decide whether given evidence is to be assessed in terms of its likelihood or used in the construction of prior probabilities. We can treat any evidence as observational evidence if we insist on doing so, for we can always ask, "What is the likelihood of obtaining this evidence if such-and-such is true?" But an insistence on treating all evidence as observational would be incompatible with the completion of a successful Bayesian analysis; it would lead to an increasingly complicated model, where more and more detailed prior probabilities are required but less and less evidence is available on which to base them. So even within the Bayesian framework we do not assess evidence in terms of its likelihood every time it is possible to do so. The central theme of my paper is that evidence that can be assessed in terms of its likelihood can sometimes be better assessed in terms of belief functions.

Other comments. I would like to thank Professor Barnard for his review of the history of varieties of uncertainty, Dr Walley for his review of his work on belief functions, and Professor Kingman for his remarks on the connection with random sets. For more on the connection with random sets, see Nguyen (1978).

I would like to conclude by again thanking the Royal Statistical Society for the opportunity to bring the theory of belief functions before a broad statistical audience. The ideas in this theory will, I am certain, continue to arouse wide interest outside of statistics. These ideas are sufficiently natural to have been reinvented many time since they were first sketched by James Bernoulli; recent reinventions have appeared in the legal literature (Ekelöf, 1981) and in the literature on artificial intelligence (Friedman, 1981). Only within the statistical tradition is the theory likely to acquire the depth and clarity that will be needed for it to become widely useful.

REFERENCES IN THE DISCUSSION

BRENNER, D. and FRASER, D. A. S. (1979). On foundations for conditional probability with statistical models—when is a class of functions a function? *Statist. Hefte* (N.F.), 20, 148–159.

DAWID, A. P. and DICKEY, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Ass.*, 72, 845–850.

DAWID, A. P. and STONE, M. (1982). The functional-model basis of fiducial inference. *Ann. Statist.* (to appear).

DEMPSTER, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.* 34, 884–891.

EKELÖF, PER O. (1981). Beweiswert. In *Festschrift für Fritz Baur* (W. Grunsky *et al.*, eds). Tubingen: J. C. B. Mohr (Paul Siebeck).

FINE, T. (1973). *Theories of Probability*. New York: Academic Press.

FRIEDMAN, L. (1981). Extended plausible inference. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 487–495.

NGUYEN, H. T. (1978). On random sets and belief functions. *J. Math. Anal. & Applic's*, 65, 531–542.

SHAFER, G. (1981c). When are conditional probabilities legitimate? (Unpublished MS.)

—— (1981d). Two theories of probability. In *PSA 1978*, Vol. 2, P. D. Asquith and I. Hacking, eds). East Lansing, Mich.: Philosophy of Science Association.

——(1982b). Bayes's two arguments for the rule of conditioning. *Ann. Statist.* (in press).

SMETS, P. (1982). Possibilistic inference from statistical data (submitted for publication).

SMITH, C. A. B. (1961). Consistency in statistical inference and decision (with Discussion). *J. R. Statist. Soc.* B, 23, 1–37.

——(1977). The analogy between decision and inference. *Synthese*, 36, 71–85.

WALLEY, P. (1981). Coherent lower and upper probabilities. Statistics Research Report, University of Warwick.

——(1982). The elicitation and aggregation of beliefs. Statistics Research Report, University of Warwick.

WILLIAMS, P. M. (1976). Indeterminate probabilities. In *Formal Methods in the Methodology of Empirical Sciences* (M. Przelecki *et al.*, eds). Dordrecht: Reidel.

WOLFENSON, M. and FINE, T. L. (1982). Bayes-like decision making with upper and lower probabilities. *J. Amer. Statist. Ass.*, 77, 80–88.

ZADEH, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1, 3–28.