The Annals of Statistics
1987, Vol. 15, No. 4, 1439–1465

BELIEF FUNCTION REPRESENTATIONS OF STATISTICAL EVIDENCE

BY PETER WALLEY

Cornell University

In Glenn Shafer's theory of parametric statistical inference, observational evidence and prior evidence are separately represented by belief or commonality functions Q and R, which are then combined by Dempster's rule. We characterise, for finite parameter spaces, the functionals Q and R for which statistically independent observations may be combined by Dempster's rule, and those for which Dempster's rule is consistent with Bayes' rule. The functionals are determined up to an arbitrary partition of the parameter space and an arbitrary scale parameter, which might be chosen to reflect aspects of the evidence on which the statistical model is based. Our results suggest that Dempster's rule is not generally suitable for combining evidence from independent observations nor for combining prior beliefs with observational evidence.

1. Shafer's model for parametric statistical inference.

1.1. *Introduction.* In recent years, starting with *A Mathematical Theory of Evidence*, Glenn Shafer has developed a mathematical framework in which assessments of evidence can be represented by *belief functions*, set functions which generalise additive probability measures. The basic strategy of the theory is that the available evidence should be broken down into simpler "entirely distinct bodies of evidence," each of these bodies assessed separately and the resulting belief functions combined by *Dempster's rule of combination* to give an overall assessment of evidence.

The problem of parametric statistical inference is perhaps the most immediate and important testing ground for Shafer's theory. This paper aims to answer questions raised by previous discussions of the parametric statistical model, notably those of Shafer [(1976a), Chapter 11, (1976b) and (1982)] and the earlier work of Dempster [e.g., (1966) and (1968)].

Suppose we observe the outcome of a statistical experiment known to be governed by one of a finite set of probability models parametrised by Θ. In Shafer's approach we separate our assessment of the evidence concerning Θ provided by the statistical observation, to be represented by some commonality function Q defined on subsets of Θ, from our assessment of prior evidence, represented by a commonality function R. (Commonality functions are in one-to-one correspondence with the more usual belief functions.) Then Q and R are combined by Dempster's rule to give a posterior assessment of the total evidence.

Received November 1984; revised March 1987.

AMS 1980 *subject classifications.* Primary 62A99; secondary 60A05.

Key words and phrases. Belief functions, Dempster's rule, Bayesian inference, likelihood, likelihood principle, prior probabilities, Bayes' rule, upper and lower probabilities.

This strategy follows the familiar Bayesian line, but the generalisation of additive probabilities to belief functions promises to provide some important advantages over the Bayesian approach.

(a) Prior ignorance may be represented in a natural way by the *vacuous* belief function. Unlike Bayesian "noninformative" priors, the vacuous function has all the invariance properties and other properties that one would expect of a representation of ignorance [see Walley (1987), Chapter 5]. At the other extreme, a large amount of prior evidence can give rise to a Bayesian (additive) prior and the theory reduces essentially to the Bayesian theory. In most practical problems, the prior evidence will be translated into a belief function intermediate between these extremes.

(b) Both prior and observational evidence are represented by commonality functions (R and Q). Bayesians, on the other hand, represent prior and observational evidence by two functions (prior probability distribution and observed likelihood function) that are conceptually different; the likelihood function can generate probabilities only when it is combined, by Bayes' rule, with some prior distribution. In Shafer's theory, observational evidence alone may give rise to nontrivial beliefs, represented by Q, without any reference to prior beliefs.

(c) By admitting nonadditive probability models, the theory can model imprecision in beliefs and avoid the arbitrary choices of precise prior probabilities that are needed in applying the standard Bayesian theory. [There are, of course, other approaches which allow imprecise probability models, including Bayesian sensitivity analysis and other theories of upper and lower probability. See Walley (1987) for a critical survey.]

(d) Shafer's theory aims to provide constructive methods for translating evidence into numerical probability assessments. The basic strategy is to decompose the evidence into smaller, more manageable pieces, to assess probabilities based on each piece, and to combine the assessments through Dempster's rule. For example, a prior commonality function R may be constructed in this way through some decomposition of prior evidence. This basic strategy is to be supplemented by specific methods for assessing particular types of evidence, e.g., evidence in the form of likelihoods. A crucial problem for the subjective Bayesian theory, and for any other theory of epistemic probability, is to show us how to construct probabilities from a given body of evidence. (In our view, this problem remains largely unresolved.) Because of the insights it promises concerning the relationship between probabilities and evidence, the theory of belief functions deserves careful attention.

There are three basic problems for a belief-function model of parametric statistical inference.

(i) How do the statistical model $(\mathcal{X}, \Theta, P)$ and the observation x give rise to a commonality function Q?

(ii) How can prior evidence be represented by a commonality function R? In particular, how should an assessment already in the form of a Bayesian prior distribution be translated into the form of a commonality function?

(iii) How should Q and R be combined to give an overall (posterior) assessment of evidence about Θ?

A fundamental principle of Shafer's theory is that Dempster's rule of combination is the appropriate method for combining beliefs based on "unrelated bodies of evidence." In this paper we investigate the implications of Dempster's rule for problems (i) and (ii). Statisticians are used to regarding statistically independent observations as "unrelated" or "noninteracting" (given the parametric model). This suggests that Dempster's rule be used to combine Q-functions based on observations that are statistically independent for all θ in Θ. We might require that the Q-function resulting from combination agree with the Q-function based directly on the joint observation. [See Shafer (1976a), page 247, and (1982), page 338 for further discussion.] In Section 2, we show that this requirement (plus weak regularity conditions) leads either to violation of the sufficiency principle or to Bayesian Q-functions which are (essentially) Bayesian posteriors with respect to a uniform prior. We conclude that Dempster's rule is not suitable for combining evidence from statistically independent observations.

Since prior and observational evidence are intuitively "distinct" or "unrelated," a second role for Dempster's rule is in combining Q and R, thus dealing with problem (iii). In case R is induced by a Bayesian prior distribution, we might require that the combination of Q and R be induced in the same way by the Bayesian posterior, i.e., that Dempster's rule be "consistent" (in a weak sense) with Bayes' rule. In Section 3 we show that this requirement implies that R be Bayesian and, in fact, a power transformation of the Bayesian prior which induces it. Then Q is determined on singletons as the same power transformation of the likelihood function. In Section 4 we investigate extensions of Q to all subsets of Θ. A class of extensions is derived which includes formalised versions of fiducial and likelihood inference as extreme cases, as well as reducing to Bayesian inference in the case of a Bayesian prior R.

The class of extensions suggests a new rule, different from Dempster's rule, which can be used both for combining evidence from statistically independent observations and for combining prior with observational evidence, and we are finally led to question the central role of Dempster's rule in Shafer's theory. We argue in Section 5 that both Dempster's rule and the new rule are unsuitable for combining prior beliefs with statistical evidence; both rules can lead to incoherence when belief functions are interpreted as betting rates. Our general conclusion is that there are serious objections to any theory of statistical inference which is based on Dempster's rule of combination.

1.2. *Belief and commonality functions.* The mathematical theory of belief functions on finite spaces is presented by Shafer (1976a). The basic definitions are summarised next; for details, see Chapters 2 and 3 of Shafer's book. We assume throughout that the parameter space Θ is finite, $\Theta = \{\theta_1, \ldots, \theta_N\}$, where $N \geq 2$.

There are four mathematically equivalent descriptions of a belief function, denoted by BEL, PL, m and Q. All four are set functions defined on all subsets

of Θ. In this paper we concentrate on the commonality function Q, since Dempster's rule has a simple multiplicative form [see (6)] in terms of Q.

To motivate and interpret these definitions, suppose that a subset B of Θ is chosen randomly according to some probability mass function m defined on the class of all subsets of Θ, so that $m(B) \geq 0$, $\sum_{B \subset \Theta} m(B) = 1$ and $m(\phi) = 0$. We can regard the chosen B as a "set observation" or "randomly coded message" that carries the meaning that the true value of θ belongs to B, but supplies no further information about θ [Shafer (1982)]. The function m is called the *probability assignment*. Naturally, we will say that the observation B is *consistent with* a particular $\theta \in \Theta$ when $\theta \in B$.

The *commonality function* Q is defined for all $A \subset \Theta$ by

$$Q(A) = \sum_{B \supset A} m(B). \tag{1}$$

Thus $Q(A)$ is just the probability of obtaining a set observation that is consistent with *every* element of A. Clearly $Q(\phi) = 1$ and $Q(A) \geq Q(B)$ whenever $A \subset B$. Let $|B|$ denote the cardinality of the set B. By the Möbius inversion theorem, m can be recovered from Q by

$$m(B) = \sum_{A \supset B} (-1)^{|A-B|} Q(A) \quad \text{for all } B \subset \Theta. \tag{2}$$

Setting $B = \phi$, we obtain the normalization condition

$$\sum_{A \neq \phi} (-1)^{|A|+1} Q(A) = 1. \tag{3}$$

The *belief function* BEL is defined for all $C \subset \Theta$ by

$$\mathrm{BEL}(C) = \sum_{B \subset C} m(B) = \sum_{A \subset C^c} (-1)^{|A|} Q(A). \tag{4}$$

$\mathrm{BEL}(C)$ is interpreted as the probability of obtaining a set observation that implies the occurrence of C.

The *plausibility function* PL is defined for all $D \subset \Theta$ by

$$\mathrm{PL}(D) = 1 - \mathrm{BEL}(D^c) = \sum_{B \cap D \neq \phi} m(B) = \sum_{\substack{A \subset D \\ A \neq \phi}} (-1)^{|A|+1} Q(A). \tag{5}$$

$\mathrm{PL}(D)$ is interpreted as the probability of obtaining a set observation that is consistent with *some* element of D. Note that the plausibility and commonality functions agree on singleton sets.

PL and BEL are in fact the upper and lower envelopes of a class of probability measures on Θ, so that $\mathrm{PL}(A) \geq \mathrm{BEL}(A)$ for all $A \subset \Theta$, $\mathrm{PL}(A) \geq \mathrm{PL}(B)$ and $\mathrm{BEL}(A) \geq \mathrm{BEL}(B)$ whenever $A \supset B$. This suggests that PL and BEL might be interpreted as upper and lower betting rates, but that turns out to be inconsistent with the use of Dempster's rule of conditioning as a way of updating betting rates; an example is given in Section 5. Shafer's theory has been criticised for its failure to supply a *behavioural* interpretation for the set functions m, Q, BEL and PL; see Williams (1978) and the response of Shafer (1981).

Two extreme types of commonality function can now be defined. Call the commonality function Q *vacuous* when $Q(A) = 1$ for all $A \subset \Theta$ or, equivalently,

$$m(B) = \mathrm{BEL}(B) = \begin{cases} 1 & \text{if } B = \Theta, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathrm{PL}(B) = \begin{cases} 0 & \text{if } B = \phi, \\ 1 & \text{otherwise.} \end{cases}$$

This corresponds to always receiving the uninformative set observation $B = \Theta$ and is a natural model for complete ignorance about the true value of θ.

At the other extreme, call Q *Bayesian* if $Q(A) = 0$ whenever $|A| > 1$. Equivalently, $m(B) = 0$ whenever $|B| > 1$, and BEL and PL are identical additive probability measures on Θ that agree on singletons with m and Q. In that case, the possible set observations are all point observations that leave no doubt about the true value of θ. The usual Bayesian account of probability may be identified with this special case, by identifying a Bayesian probability measure with BEL or PL.

The combination of commonality functions Q_1 and Q_2 by *Dempster's rule* is denoted by $Q_1 \oplus Q_2$, and is defined whenever $Q_1(\{\theta\})Q_2(\{\theta\}) > 0$ for some $\theta \in \Theta$ by

$$(Q_1 \oplus Q_2)(A) = kQ_1(A)Q_2(A) \quad \text{for nonempty } A \subset \Theta, \tag{6}$$

where the normalizing constant k is determined by (3),

$$k^{-1} = \Sigma_{A \neq \phi}(-1)^{|A|+1}Q_1(A)Q_2(A)$$

[Shafer (1976a), Theorem 3.3].

Clearly, $Q_1 \oplus Q_2 = Q_2 \oplus Q_1$. If Q_1 is vacuous, $Q_1 \oplus Q_2 = Q_2$. If Q_1 is Bayesian, so is $Q_1 \oplus Q_2$.

1.3. *The basic axioms.* We now introduce the axioms whose implications are studied in the rest of the paper. Consider a standard parametric model $\{P_\theta: \theta \in \Theta\}$, where the P_θ are probability mass functions on a sample space $\mathcal{X}$. The discussion will refer only to a single finite parameter space Θ, but we will consider different models P_θ which describe different statistical experiments governed by the same parameter θ.

The basic idea is that an observation $x \in \mathcal{X}$ gives rise to beliefs about Θ that are represented by some commonality function Q_x defined on subsets of Θ. It is assumed that the function Q_x depends on the parametric model and the observation only through the values $P_{\theta_1}(x), \ldots, P_{\theta_N}(x)$. That assumption, which is a weak form of the likelihood principle, means that we can write $Q_x(A) = Q(A, \tau)$ for all $A \subset \Theta$, where τ is the N-vector $(P_{\theta_1}(x), \ldots, P_{\theta_N}(x))$ and Q is a functional that describes the general translation of observational evidence into beliefs about Θ. We will, therefore, drop reference to the observation x and write our axioms and theorems in terms of the functional Q. Of course, the axioms are plausible only if one bears in mind the interpretation of τ as a vector of likelihoods.

Define $\mathscr{S} = \{(\tau_1, \ldots, \tau_N)\colon 0 \leq \tau_j \leq 1 \text{ for all } j,\ \tau_j > 0 \text{ for some } j\}$ to be the set of likelihood vectors that are not identically zero. (Throughout the discussion, all vectors considered are N-vectors.) Our first axiom is

(A1) $Q(\cdot, \tau)$ is a commonality function on Θ whenever $\tau \in \mathscr{S}$.

This axiom formalises the basic requirement of Shafer's approach, that statistical evidence can be represented by a commonality function $Q(\cdot, \tau)$ or (equivalently) by a belief function. It is assumed that $Q(\cdot, \tau)$ can be defined for all conceivable likelihood vectors τ. (Different likelihoods τ might be generated by different observables in various conceivable experiments governed by the same parameter θ.)

(A1) relies on the weak version of the likelihood principle formulated earlier. The two statistical methods discussed by Shafer [(1982), Sections 3.1 and 3.2] which apply to finite spaces Θ do satisfy the weak version of the likelihood principle and (A1), as well as (A2). (The third method discussed by Shafer [(1982), Section 3.3], following Dempster (1966), is outside the scope of this paper because it applies only to an infinite space Θ which parametrises all multinomial distributions.)

Write $\tau\sigma$ for the product vector $(\tau\sigma)_j = \tau_j\sigma_j$. We require that the commonality function $Q_{(x, y)}$ based on two statistically independent observations x and y agrees with the combination of Q_x and Q_y by Dempster's rule. In view of our assumption that observations affect Q only through their likelihood functions, this becomes, since independent likelihoods multiply,

(A2) $Q(\cdot, \tau) \oplus Q(\cdot, \sigma) = Q(\cdot, \tau\sigma)$ whenever $\tau \in \mathscr{S}$, $\sigma \in \mathscr{S}$ and $\tau\sigma \in \mathscr{S}$.

The functionals Q satisfying (A1) and (A2) (plus weak regularity conditions) are characterised in Section 2. In Sections 3 and 4 we drop (A2) but add two more axioms concerning the commonality function $R(\cdot, \rho)$ induced by a Bayesian prior distribution ρ, where ρ_j is the prior probability of $\{\theta_j\}$. Let $\mathscr{P} = \{\rho\colon (\forall\, j)\ \rho_j \geq 0, \sum_{j=1}^{N}\rho_j = 1\}$ denote the class of all probability mass functions on Θ. The next axiom requires that any Bayesian prior ρ be translated into a commonality function $R(\cdot, \rho)$ on Θ that represents the prior evidence about Θ.

(A3) $R(\cdot, \rho)$ is a commonality function on Θ whenever $\rho \in \mathscr{P}$.

The final axiom demands a weak sort of consistency between Dempster's rule and Bayes' rule. If ρ denotes the Bayesian prior, $\tau_j = P_{\theta_j}(x)$ denotes the likelihood function and $(\rho \circ \tau)_j = \rho_j\tau_j/\sum\rho_i\tau_i$ denotes the posterior by Bayes' rule, we require that the combination of $R(\cdot, \rho)$ and $Q(\cdot, \tau)$ agree with $R(\cdot, \rho \circ \tau)$.

(A4) If $\rho \in \mathscr{P}$, $\tau \in \mathscr{S}$ and $\rho_j\tau_j > 0$ for some j, then

$$R(\cdot, \rho) \oplus Q(\cdot, \tau) = R(\cdot, \rho \circ \tau).$$

In other words, if we translate the Bayesian prior and the observational evidence separately into commonality functions, their combination by Dempster's rule should agree with our translation of the Bayesian posterior.

It might be reasonable to identify $R(\cdot, \rho)$ directly with the Bayesian commonality function that agrees with ρ on singletons, especially if we wish to carry

over the behavioural interpretation of Bayesian probabilities to the corresponding belief functions. The behavioural interpretation of belief functions is problematical, however, as it can lead to incoherence. The identification of $R(\cdot, \rho)$ with ρ is therefore not entirely compelling and will not be assumed here. We see in Section 3 that the weaker axioms (A3) and (A4) are enough to force $R(\cdot, \rho)$ to be Bayesian and, in fact, to be a power transformation of ρ on singletons.

Four weak regularity conditions will be assumed throughout the paper. These seem uncontroversial and may be skipped by readers who wish to proceed to our main results. The first says that the plausibility of a singleton $\{\theta\}$ based on observation x is bounded away from 0 provided θ assigns probability 1 to x.

(R1) For each $1 \leq i \leq N$, there is $c_i > 0$ such that $Q(\{\theta_i\}, \tau) \geq c_i$ whenever $\tau_i = 1$.

The next condition says that the plausibility of a singleton $\{\theta\}$ goes to 0 as $P_\theta(x)$ goes to 0, provided all other models assign probability 1 to x.

(R2) For all $\tau \in \mathscr{S}$ such that $\tau_j = 1$ whenever $j \neq i$, $Q(\{\theta_i\}, \tau) \to 0$ as $\tau_i \to 0$.

Define I_A to be the vector whose jth coordinate is 1 if $\theta_j \in A$ and 0 otherwise. Write $\mathbf{1} = I_\Theta$ for the unit likelihood. (R3) and (R4) are nontriviality conditions concerning the sets A which receive nonzero commonality values under the unit likelihood.

(R3) If $A \subset \Theta$ and $Q(A, \mathbf{1}) > 0$, then $Q(A, I_A) > 0$.

(R4) If $A \subset \Theta$, $Q(A, \mathbf{1}) > 0$ and $1 \leq i \leq N$, then there is some $\tau \in \mathscr{S}$ with $\tau_i < 1$ and $Q(A, \tau) > 0$.

The effect of dropping any of (R2)–(R4) is to admit further "trivial" functions in Theorem 1 and is clear from the proof of Lemma 2.

2. Dempster's rule for combining independent observations. We first investigate the implications of (A2), that the commonality functions derived from two statistically independent observations should combine by Dempster's rule to give the commonality function derived from the joint observation. Theorem 1 characterises the mathematical form of the functionals Q that satisfy (A2) [plus (A1) and regularity conditions] in terms of a partition of Θ and N positive real constants.

THEOREM 1. *A commonality functional Q satisfies* (A1), (A2) *and regularity conditions* (R1)–(R4) *if and only if there is a partition $\{A_1, \ldots, A_s\}$ of Θ and positive real numbers $\lambda_1, \ldots, \lambda_N$ such that, for all $\tau \in \mathscr{S}$,*

$$Q(A, \tau) = k(\tau) \prod_{\theta_j \in A} \tau_j^{\lambda_j} \quad \text{when } A \in \mathscr{A}^+, \tag{7}$$

$$Q(\phi, \tau) = 1 \quad \text{and} \quad Q(A, \tau) = 0 \quad \text{otherwise},$$

where

$$\mathscr{A}^+ = \{A: A \subset A_j \text{ for some } 1 \leq j \leq s,\ A \neq \phi\}$$

and

$$k(\tau) = \left(\sum_{i=1}^{s} \left[1 - \prod_{\theta_j \in A_i} \left(1 - \tau_j^{\lambda_j}\right) \right] \right)^{-1} \geq s^{-1}.$$

Proofs of Theorem 1 and the other theorems are given in Section 6. In the rest of this section we investigate the functionals (7) as possible representations of statistical evidence. In particular, we consider further specialisation of the functional form, sources of the partition and constants λ_j, consistency with the likelihood principle and a numerical example.

In the representation (7), the sets A_i in the partition are characterised by the fact that each A_i is assigned probability $m(A_i, \mathbf{1}) = 1/s$ under the unit likelihood. Generally, the functions in (7) are *additive across the partition* $\{A_1, \ldots, A_s\}$ in the sense that (for all $\tau \in \mathscr{S}$, $A \subset \Theta$, $1 \leq i \leq s$),

$$\mathrm{PL}(A_i, \tau) = \mathrm{BEL}(A_i, \tau),$$

$$\mathrm{PL}(A, \tau) = \sum_{i=1}^{s} \mathrm{PL}(A \cap A_i, \tau)$$

and

$$\mathrm{BEL}(A, \tau) = \sum_{i=1}^{s} \mathrm{BEL}(A \cap A_i, \tau).$$

PL and BEL are thence given by their values for $A \subset A_i$:

$$\mathrm{PL}(A, \tau) = k(\tau) \left[1 - \prod_{\theta_j \in A} \left(1 - \tau_j^{\lambda_j}\right) \right], \tag{8}$$

$$\mathrm{BEL}(A, \tau) = k(\tau) \prod_{\theta_j \in A_i - A} \left(1 - \tau_j^{\lambda_j}\right) \left[1 - \prod_{\theta_j \in A} \left(1 - \tau_j^{\lambda_j}\right) \right]. \tag{9}$$

It is easily seen that, for all $\tau \in \mathscr{S}$ and $A \subset \Theta$, $Q(A, \tau)$, $\mathrm{PL}(A, \tau)$ and $\mathrm{BEL}(A, \tau)$ are each continuous in τ, nondecreasing in τ_j when $\theta_j \in A$ and nonincreasing in τ_j when $\theta_j \notin A$.

Can additional axioms be justified to further restrict the functional representation (7), i.e., to restrict the admissible partitions and λ_j? Corollaries 1 and 2 describe the effect of two further axioms, (A5) and (A6). To force λ_j to be constant over j, it is enough to require the following weak invariance axiom,

(A5) $Q(\{\theta_i\}, \tau) = Q(\{\theta_j\}, \tau)$ whenever $\tau_i = \tau_j$.

COROLLARY 1. *The commonality functional Q satisfies the assumptions of Theorem* 1, *together with* (A5), *if and only if there is a partition* $\{A_1, \ldots, A_s\}$ *of* Θ *and a positive* λ *such that Q is given by* (7) *with each* λ_j *equal to* λ.

If we strengthen (A5) as follows to require that Q be invariant under permutations of the indices of Θ, we can show also that there must be either 1 or N sets in the partition.

(A6) $Q(\{\theta_1, \ldots, \theta_k\}, \tau) = Q(\{\theta_{j(1)}, \ldots, \theta_{j(k)}\}, \pi)$ whenever $\tau \in \mathscr{S}$, $1 \leq k \leq N$, $j(\cdot)$ is a permutation of $\{1, \ldots, N\}$ and π is defined by $\pi_{j(i)} = \tau_i$.

COROLLARY 2. *Q satisfies the assumptions of Theorem* 1, *together with* (A6), *if and only if there is some positive* λ *such that* either

$$Q(A,\tau) = k(\tau)\prod_{\theta_j\in A}\tau_j^{\lambda} \quad \text{for all } \tau\in\mathcal{S} \text{ and } A\neq\phi, \tag{10}$$

$$\text{where } k(\tau)^{-1} = 1 - \prod_{\theta_j\in\Theta}\left(1-\tau_j^{\lambda}\right),$$

or

$$Q\left(\{\theta_j\},\tau\right) = \tau_j^{\lambda}\Big/\sum_{i=1}^{N}\tau_i^{\lambda} \quad \text{for all } \tau\in\mathcal{S} \text{ and } 1\le j\le N, \tag{11}$$

$$\text{with } Q(A,\tau) = 0 \text{ when } |A|>1.$$

The invariance axiom (A6) is perhaps too strong in that it ignores additional structure of Θ that may arise from the evidence on which Θ is based, as indicated in the following section. The two extreme cases ($s = 1$ and N) are, however, especially interesting. The representation (10) corresponds to the trivial partition ($s = 1$), with $Q(\cdot,1)$ vacuous, and (11) to the finest possible partition ($s = N$), with $Q(\cdot,1)$ Bayesian, corresponding to the uniform probability distribution on Θ. The representation (11) appears to be related to the fiducial argument, since it generates additive probabilities from likelihood evidence alone, without reference to prior beliefs; see Shafer [(1982), Section 3.2] and Dempster (1966, 1968). (Of course Fisher applied the fiducial argument only to continuous spaces, whereas Θ is finite here.) On most interpretations of "plausibility," we would expect any θ which assigns probability 1 to the actual observation to be fully plausible, so that $\mathrm{PL}(\{\theta\}) = Q(\{\theta\}) = 1$, and this condition leads to $s = 1$ in (7).

2.1. *Sources of the partition.* If we take $\lambda = 1$ in (10) and (11), we obtain two of the models suggested by Shafer [(1982), Sections 3.1 and 3.2] as representations of statistical evidence. Shafer regards the two models as appropriate in different circumstances, depending on the sort of evidence on which our parametric model $\{P_\theta\colon \theta\in\Theta\}$ is based. He suggests representation (10) for the case in which the different P_θ are based on independent empirical data and (11) when the P_θ are based on a single frequency distribution (e.g., the distribution of errors e when $x = \theta + e$). Of course, intermediate cases are possible and some of these might give rise to the intermediate representations ($1 < s < N$) in Theorem 1. As a simple example, consider the model $x = \theta + e$, where x is a measurement of a real variable on a communications channel, $\theta = 0$ or 1 represents the absence or presence of a signal and e is random noise whose distribution P_ϕ depends on the unknown type of transmission medium ($\phi = 0$ or 1). Here $\Theta = \{(\theta,\phi)\colon \theta = 0,1;\ \phi = 0,1\}$. Suppose that the two distributions P_0 and P_1 are established empirically through two separate sets of observations of the noise e, one for each medium. Under Shafer's interpretation, the structure of this evidence for our

model could be reflected by partitioning Θ into $A_1 = \{(0,0), (0,1)\}$ and $A_2 = \{(1,0), (1,1)\}$. (So A_2 is the event that a signal is present; A_1 that it is absent.) Then conditional on θ, the separate evidence for P_0 and P_1 gives representation (10) for ϕ; while conditional on ϕ, the single frequency distribution P_ϕ gives representation (11) for θ. Thus the marginal beliefs about the signal θ will be Bayesian, i.e., $\mathrm{BEL}_x([\theta = 0]) + \mathrm{BEL}_x([\theta = 1]) = 1$, for any observation x.

Typically, the class Θ of possible models is constructed by comparing the experiment at hand with "analogous" experiments whose statistical behaviour is known. An alternative method of constructing the partition $A_1, \ldots, A_s$ is to take each A_j to correspond to a single *analogy*. Several different analogies may be considered and each may suggest several different parametric models. In the coin-tossing example of Section 2.4, with θ the bias of the coin, our experience with ordinary coins suggests that $\theta \sim 0.5$, giving $A_1 = [\theta = 0.5]$. A second analogy is that the coin is not ordinary, giving $A_2 = [\theta \neq 0.5]$. The two analogies are quite different because of both our different past experience of biased versus ordinary coins and the different mechanisms by which A_1 and A_2 might have arisen (e.g., deliberate deception seems likely if the coin is biased). We might further split A_2 into $[\theta > 0.5]$ and $[\theta < 0.5]$ if we believed the two directions of bias would arise through different mechanisms.

This method of constructing the partition is similar to Shafer's suggestion in that it reflects the source of the parametric model Θ, but is quite different in effect. In the communications example, it seems most natural to take the different transmission media as different analogies, since our past experience treats the different media separately but does not distinguish cases in which a signal was present from those in which it was absent. Thus we take $A_1 = [\phi = 0]$ and $A_2 = [\phi = 1]$, giving Bayesian marginal for ϕ but not for θ, whereas Shafer's method gives Bayesian marginal for θ but not for ϕ. When the observation x provides little information about each parameter, however, it seems unreasonable that either marginal should be Bayesian. For this reason, both methods of constructing the partition seem less satisfactory than adopting the trivial partition $s = 1$, which gives representation (10) and non-Bayesian marginals for each parameter.

2.2. *Interpretation of* λ. The significance of the parameters λ_j in Theorem 1 can be clarified by looking at the case $s = 1$, for which Q can be expressed in terms of a "weight of evidence" function w [Shafer (1976a), Chapter 5] as

$$Q(A, \tau) = k(\tau)\exp\Big(-\sum_{B \not\supset A} w(B, \tau)\Big), \tag{12}$$

where $w(\{\theta_j\}^c, \tau) = -\lambda_j \log \tau_j$ for $1 \leq j \leq N$ and $w(B, \tau) = 0$ for other sets B. Here $w(B, \tau)$ is interpreted as the total weight of evidence focused on B by τ. Thus τ provides evidence of weight $-\lambda_j \log \tau_j$ *against* $\{\theta_j\}$. [Note that these weights of evidence are additive across independent observations. In fact, from (6) and (12), Dempster's rule of combination is additive in weights of evidence.] It is clear that λ_j determines the *scale* for measuring the weight of evidence

against $\{\theta_j\}$. Small values of λ_j will "discount" the failure of θ_j to predict what is observed, while values $\lambda_j > 1$ will "emphasise" this failure.

Theorem 1 allows the λ_j to be different, but it appears that we would rarely want to discount different parameter values at different rates. So we concentrate on the case $\lambda_j \equiv \lambda$, which is implied by any of axioms (A5)–(A7) or by the assumptions of Theorem 2. If we require that the combination by Dempster's rule of $Q(\cdot, \tau)$ with a Bayesian prior π agrees with the Bayesian posterior of π and τ, we obtain $\lambda = 1$. The same conclusion follows from the embedding arguments in Shafer (1982).

2.3. *Difficulties with the representation.* We have assumed a weak version of the likelihood principle, that the commonality function produced by observation x depends only on the probabilities assigned to x by each parameter value. In the present context this assumption is significantly weaker than the usual likelihood principle. In fact, the following version of the strong likelihood principle, requiring invariance under multiplication of likelihoods by a positive constant, forces the "fiducial" representation (11).

(A7) $Q(\cdot, c\tau) = Q(\cdot, \tau)$ whenever $\tau \in \mathscr{S}$ and $0 < c < 1$.

COROLLARY 3. *Q satisfies the assumptions of Theorem* 1 *plus* (A7) *if and only if it has the form* (11) *for some positive* λ.

Indeed, consider an experiment with possible outcomes x and y such that $P_\theta(x) = cP_\theta(y)$ for all $\theta \in \Theta$, where $0 < c < 1$. If x and y generate Q_x and Q_y by (7), it is easy to see that $Q_x \neq Q_y$ unless $s = N$ and λ_j is constant. Thus even the sufficiency principle (or weak likelihood principle), which asserts that possible observations from the same experiment which generate proportional likelihood functions should give rise to the same inference about θ, is violated unless $s = N$.

To see the effect on Q of renormalizing likelihoods, consider $Q(\{\theta\}, c1) = c^\lambda / \sum_{i=1}^{s} [1 - (1 - c^\lambda)^{|A_i|}]$. This tends to $1/s$ as $c \to 1$ and tends to $1/N$ as $c \to 0$. More generally, for any $\tau \in \mathscr{S}$ and $1 \leq s \leq N$, $Q(\cdot, c\tau)$ approaches the Bayesian function given in (11) as $c \to 0$. Thus an observation with uniformly low probability generates near-Bayesian beliefs. Now one way of obtaining such an observation is to include the outcome of an independent ancillary experiment, say the results of 100 tosses of a fair coin, in the observational report. Doing so will evidently change the commonality function Q generated by the observation. If this change in Q has any empirical significance, the change is disturbing and calls into question our understanding of an "observation." (If such "uninformative" ancillaries may alter our inferences, we need to know which to report and which to ignore.) If the change in Q has no operational significance, it appears, by letting $c \to 0$, that any Q is operationally equivalent to the Bayesian Q' obtained by normalizing $Q(\{\theta\})$ (11). Since the combination of Q' with *any* prior belief function is Bayesian, one reaches the Bayesian conclusion that any posterior beliefs can be represented in the form of a probability measure,

frustrating our original aim to generalise the Bayesian approach by admitting imprecise probabilities.

Further problems with representations (10) and (11) arise when we try to embed them in belief functions defined on subsets of $\Theta \times \mathscr{X}$, as in Shafer (1982). All their extensions to belief functions on $\Theta \times \mathscr{X}$ have properties which contradict basic intuitions about "plausibility" and "belief" [see the discussion of Shafer (1982), page 344].

In view of these difficulties, especially the violation of the sufficiency principle by any non-Bayesian Q-function in Theorem 1, we do not find these functions attractive as representations of statistical evidence. We are then forced to reject at least one of axioms (A1) and (A2). If (A1) is rejected, then (A2) must also be rejected. So we are led to give up the requirement (A2) that the combination by Dempster's rule of Q-functions based on independent observations agrees with the Q-function based on the joint likelihood.

This can be done in several ways. First, we might retain Dempster's rule to combine evidence from independent observations, but not require that the combination be obtainable from the joint likelihood, as in Shafer [(1976a), Chapter 11]. This might be reasonable when we are uncertain that the parametric model applies to each observation, so that the individual likelihood functions contain information additional to that in the joint likelihood (which might, for example, be used to identify outliers). But if the statistical model is fully accepted, the joint likelihood of the observations does seem to contain all the relevant information about θ. Reduction of the data using sufficiency is a common practice in statistics and a theory of evidence that always requires "individual" likelihood functions rather than just the joint likelihood faces serious problems over what constitute "individual" observations. Obviously, a given likelihood function can be factored into independent components in many ways and thus can arise as a joint likelihood from many dissimilar sets of observations.

A second, more appealing, line is to give up Dempster's rule for combining evidence from independent observations. We can then try to retain (A2), but with some other rule of combination replacing Dempster's rule. This line is pursued in Section 4.

2.4. *Coin-tossing example.* In an example given by Shafer [(1976a), page 243], we obtain 5 heads in 10 tosses of a possibly biased coin, in which the chance of heads is known to be some multiple $\theta/10$ of $1/10$. Assuming a Bernoulli model, $\Theta = \{1, 2, \ldots, 9\}$. In Table 1 we give $\mathrm{PL}(A)$ and $\mathrm{BEL}(A)$ for selected subsets A of Θ under various belief-function representations. The given PL and BEL values are generated by just the statistical observation. In most practical contexts they would be combined with a prior strongly concentrated near $\theta = 5$ to take account of our past experience with coins. Columns A–D are based on Corollary 1 with $\lambda = 1$. Column A gives the Bayesian function (11), with $\mathrm{PL}(A) = \mathrm{BEL}(A)$ for all sets A. Model B is based on (10) using the observation consisting of the 10 ordered outcomes and is already very close to Bayesian. Column C is also based on (10), but using the observed number of heads (a

TABLE 1

Values of $PL(A)$ (upper) and $BEL(A)$ (lower) for selected subsets $A \subset \Theta$ in the coin-tossing example. Columns A–D *are based on* (7) *with* $\lambda_j \equiv 1$ (B *on the* 10 *ordered outcomes,* C *and* D *on the number of heads). Columns* E–J (*and* A) *are based on* (17) *with* $\lambda = 1$ *or* $\frac{1}{2}$. *The partitions used were* $A_1 = \{5\}$, $A_2 = \{5\}^c$ *for* $s = 2$ *and* $A_1 = \{5\}$, $A_2 = \{1,2,3,4\}$, $A_3 = \{6,7,8,9\}$ *for* $s = 3$.

	A	B	C	D	E	F	G	H	I	J
s	9	1	1	2	1	2	3	9	2	1
λ	1	1	1	1	1	1	1	0.5	0.5	0.5
{1}	0.0016	0.00164 0.00163	0.0023 0.0009	0.0020 0.0009	0.0060 0	0.0033 0	0.0023 0	0.016	0.041 0	0.078 0
{2}	0.029	0.0291 0.0290	0.042 0.016	0.035 0.017	0.107 0	0.059 0	0.041 0	0.067	0.172 0	0.328 0
{3}	0.113	0.1134 0.1130	0.162 0.066	0.135 0.073	0.418 0	0.230 0	0.159 0	0.132	0.340 0	0.647 0
{4}	0.221	0.2210 0.2204	0.317 0.145	0.264 0.160	0.815 0	0.449 0	0.310 0.151	0.184	0.474 0	0.903 0
{5}	0.271	0.2711 0.2704	0.388 0.189	0.324 0.324	1 0.185	0.551 0.551	0.380 0.380	0.204	0.525 0.525	1 0.097
{4, 5}	0.491	0.4919 0.4910	0.627 0.381	0.588 0.484	1 0.185	1 0.551	0.690 0.531	0.388	1 0.525	1 0.097
{3, 4, 5}	0.605	0.6051 0.6042	0.725 0.491	0.696 0.576	1 0.185	1 0.551	0.690 0.649	0.519	1 0.525	1 0.097
{4, 5, 6}	0.712	0.7126 0.7118	0.818 0.622	0.799 0.685	1 0.582	1 0.770	1 0.682	0.571	1 0.660	1 0.353
{3, 4, 5, 6}	0.825	0.8256 0.8251	0.896 0.760	0.885 0.800	1 0.582	1 0.770	1 0.800	0.703	1 0.660	1 0.353
{3, 4, 5, 6, 7}	0.939	0.9386 0.9384	0.966 0.913	0.963 0.928	1 0.893	1 0.941	1 0.918	0.835	1 0.828	1 0.672
{2, 3, 4, 5, 6, 7, 8}	0.9967	0.99673 0.99672	0.9983 0.9953	0.9981 0.9961	1 0.9939	1 0.9967	1 0.9954	0.968	1 0.959	1 0.922
{1, 2}	0.031	0.0308 0.0307	0.044 0.017	0.037 0.018	0.107 0	0.059 0	0.041 0	0.083	0.172 0	0.328 0
{1, 2, 3}	0.144	0.1441 0.1437	0.202 0.085	0.168 0.094	0.418 0	0.230 0	0.159 0	0.214	0.340 0	0.647 0
{1, 2, 3, 4}	0.365	0.3651 0.3642	0.478 0.251	0.398 0.278	0.815 0	0.449 0	0.310 0.310	0.398	0.474 0	0.903 0

sufficient statistic for θ). The sufficiency principle is violated and model C is far from Bayesian. Column D is also based on the observed number of heads, but with $s = 2$, $A_1 = \{5\}$ in Corollary 1. It is somewhat closer to Bayesian than model C is.

3. Consistency of Dempster's rule with Bayes' rule. We now consider combination of the "statistical" commonality function $Q(\cdot, \tau)$, based on an observation x with $P_\theta(x) = \tau_j$, with a "prior" commonality function R based on prior evidence separate from that provided by x. Q and R are combined by

Dempster's rule to give a posterior $Q \oplus R$. Shafer's theory generalises the Bayesian theory by allowing prior evidence that is too weak to generate a Bayesian R. For example, with a complete lack of prior evidence, R is vacuous and the posterior is just Q. In those cases where the prior evidence is strong enough to generate a Bayesian prior ρ, where ρ_j is the prior probability for $\{\theta_j\}$, we require (A3) and (A4), as well as (A1), from Section 1.3: Thus $R(\cdot, \rho)$ is a commonality function, and if $\rho \circ \tau$ denotes the posterior under Bayes' rule, then

$$R(\cdot, \rho) \oplus Q(\cdot, \tau) = R(\cdot, \rho \circ \tau),$$

whenever the posterior is well defined. Theorem 2 asserts that these axioms plus weak regularity conditions imply that R must be Bayesian and establishes the form of Q on singleton sets. We require the following additional regularity condition, where π_0 denotes the uniform probability distribution on Θ.

(R5) $R(\{\theta_i\}, \pi_0)$ does not depend on i.

This requires the prior commonality function based on π_0 to be constant on the singletons of Θ, since the uniform distribution π_0 does not distinguish between the singletons.

THEOREM 2. *Suppose* (A1), (A3), (A4) *and* (R1)–(R5) *hold. Then, for some positive* λ:

(a) *For all* $\rho \in \mathscr{P}$, $R(\cdot, \rho)$ *is Bayesian with*

$$R(\{\theta_i\}, \rho) = \rho_i^\lambda \Big/ \sum_{j=1}^{N} \rho_j^\lambda \quad \text{for } 1 \le i \le N. \tag{13}$$

(b) *For all* $\tau \in \mathscr{S}$ *and* $1 \le i \le N$,

$$Q(\{\theta_i\}, \tau) = k(\tau)\tau_i^\lambda \quad \text{for some } k(\tau) \ge N^{-1}. \tag{14}$$

REMARKS. (i) Given λ, R is completely determined by (13).

(ii) Q is not fully determined by (14). (A1) is, therefore, not necessary for the representation in Theorem 2. It is necessary that (R3) and (R4) hold for singleton sets A, but not for general A. The other conditions (A3), (A4), (R1), (R2) and (R5) are necessary for (13) and (14).

(iii) A special case of Theorem 2 follows from the results of Krantz and Miyamoto (1983), who studied the case where Θ contains only two points. They imposed stronger conditions than the preceding, including the likelihood principle (A7), which lead to the same formulas (13) and (14), but fully determine Q through formula (16) in Section 4.

(iv) If we weaken (R5) to the condition that $r_i = R(\{\theta_i\}, \pi_0)$ is nonzero for each $1 \le i \le N$, formula (13) generalises to $R(\{\theta_i\}, \rho) \propto r_i \rho_i^\lambda$, with $\Sigma_i r_i = 1$.

If we require that the commonality function R generated by a Bayesian prior ρ agree with ρ on singletons, then we obtain $\lambda = 1$ in Theorem 2. The commonality function Q is then proportional to the likelihood function on singletons. This case is generalised in Theorem 2 only through the scale parameter λ. As λ tends to zero, $R(\cdot, \rho)$ tends to the uniform distribution on $\{\theta_i\colon \rho_i > 0\}$.

As λ tends to infinity, $R(\cdot, \rho)$ tends to the uniform distribution on $\{\theta_i\colon\ \rho_i = \max \rho_j\}$.

Values of λ between 0 and 1 enable us to "discount" the prior ρ, replacing it by a prior which is proportional to ρ^λ and closer to the uniform prior. But, for consistency with Bayes' rule, we must discount the likelihood function and posterior to just the same degree λ. This method of discounting might be useful in a wider class of problems if we allowed the parameter λ to vary between formulas (13) and (14) to reflect differing degrees of confidence in our prior and likelihood evidence. We then lose consistency with Bayes' rule (A4), but that is no longer appropriate when we have incomplete confidence in the prior or likelihood.

Models which allow different degrees of discounting for the prior and likelihood have already appeared in the psychological and statistical literature. The phenomenon of "conservatism," which has been widely observed in empirical studies of probability revision [Kahneman, Slovic and Tversky (1982)], can be modelled as discounting the likelihood function by taking $\lambda < 1$ in (14) but $\lambda = 1$ in (13). The models with λ different from 1 in (13) but $\lambda = 1$ in (14) may be seen as special cases of the "obstinacy–timidity impediment functions" proposed by Lad (1978) as descriptions of non-Bayesian learning patterns.

The Bayesian regression models of West (1985) employ scale parameters to discount discrepant components of the prior or observational evidence. Different observations can be discounted to different degrees in order to reduce the influence of "outliers" on posterior beliefs.

Wolfenson and Fine (1982) study models in which a Bayesian prior is "discounted" by replacing it by prior upper and lower probabilities, but the likelihood function is not. Typically, we will have more confidence in the Bayesian posterior than in the Bayesian prior and this is reflected in different discount parameters for the prior and posterior.

We next turn to the problem of extending $Q(\cdot, \tau)$ from the singletons to all subsets of Θ. Clearly, the representations in Theorems 1 and 2 are consistent with each other, so one way to extend Q is to add (A2) to the conditions of Theorem 2.

COROLLARY 4. *Q and R satisfy* (A1)–(A4) *and* (R1)–(R5) *if and only if there is some partition $\{A_1, \ldots, A_s\}$ of Θ and some positive λ such that R is given by* (13) *and Q by* (7) *with $\lambda_j \equiv \lambda$.*

Adding the likelihood principle (A7) forces both Q and R to be Bayesian, with essentially the same form, $Q(\{\theta_i\}, \tau) \propto \tau_i^\lambda$, $R(\{\theta_i\}, \rho) \propto \rho_i^\lambda$. We next look for non-Bayesian Q which extend (14), satisfy (R1)–(R4) and (A7) and therefore violate (A2).

4. Partial consonance.

4.1. *Extensions of Q.* We require extensions of Q to all subsets of Θ which satisfy (14). The next result gives two possible extensions.

PROPOSITION. (a) *For each* $A \subset \Theta$, $Q(A, \tau)$ *is minimised over all commonality functions* Q *satisfying* (14) *by the Bayesian commonality function* Q_b, *with*

$$Q_b(\{\theta_i\}, \tau) = \tau_i^\lambda \Big/ \sum_{j=1}^{N} \tau_j^\lambda. \tag{15}$$

(b) *For each* $A \subset \Theta$, $Q(A, \tau)$ *is maximised over all commonality functions* Q *satisfying* (14) *by the consonant commonality function*,

$$Q_c(A, \tau) = \min\{\tau_i^\lambda : \theta_i \in A\} / \max\{\tau_j^\lambda : 1 \le j \le N\}. \tag{16}$$

Thus the Bayesian Q_b uniformly minimises $Q(A, \tau)$ and the consonant Q_c uniformly maximises $Q(A, \tau)$ amongst all possible extensions. In that sense, the consonant Q_c is as "far from Bayesian" as possible, given its relative values for singletons. The system Q_c with $\lambda = 1$ formalises likelihood inference and was suggested by Shafer [(1976a), Chapter 11]. We now derive a class of commonality functions which has Q_b and Q_c as extreme cases and contains *partially consonant* functions whose behaviour is intermediate between Q_b and Q_c.

We require two further axioms. First, we require that $Q(\cdot, \mathbf{1})$ act as an identity under Dempster's rule, i.e., that an "uninformative" observation giving unit likelihoods has no effect on beliefs.

(A8) $Q(\cdot, \mathbf{1}) \oplus Q(\cdot, \tau) = Q(\cdot, \tau)$ whenever $\tau \in \mathscr{S}$.

The final axiom is a version of Dempster's rule of conditioning [Shafer (1976a), Chapter 3].

(A9) If $\tau \in \mathscr{S}$ and $\tau I_B \in \mathscr{S}$, then $Q(A, \tau I_B) \propto Q(A, \tau)$ when $A \subset B$, and $Q(A, \tau I_B) = 0$ otherwise.

(A9) requires agreement between two ways of adjusting the commonality function $Q(\cdot, \tau)$ on learning that θ belongs to B. One way is to adjust the likelihood τ to τI_B, thereby assigning zero likelihood to each θ outside B and leaving the other likelihoods unchanged, giving commonality function $Q(\cdot, \tau I_B)$. A second way is to define the prior commonality function $R(\cdot, I_B)$ by $R(A, I_B) = 1$ when $A \subset B$, and $R(A, I_B) = 0$ otherwise, to represent the prior information that θ belongs to B, and to combine $R(\cdot, I_B)$ with $Q(\cdot, \tau)$ using Dempster's rule (6). The two answers agree just when (A9) holds.

We also require one further regularity condition, which asserts that the plausibility of any hypothesis θ_i should not increase when the likelihood of a different hypothesis is increased.

(R6) $Q(\{\theta_i\}, \tau)$ is nonincreasing in τ_j whenever $i \neq j$.

THEOREM 3. *The assumptions of Theorem* 2, *together with* (A7)–(A9) *and* (R6), *hold if and only if there is some* $\lambda > 0$ *and some partition* $\{A_1, \ldots, A_s\}$ *of*

Θ *such that* R *is given by* (13) *and for all* $\tau \in \mathscr{S}$,

$$Q(A,\tau) = k(\tau)\min\{\tau_i^\lambda : \theta_i \in A\} \quad \textit{when } A \in \mathscr{A}^+,\ Q(\phi,\tau) = 1, \tag{17}$$

and $Q(A,\tau) = 0$ *otherwise,*

where $\mathscr{A}^+ = \{A : A \subset A_j \text{ for some } 1 \le j \le s,\ A \ne \phi\}$ *and*

$$k(\tau) = \left(\sum_{j=1}^{s} \max\{\tau_i^\lambda : \theta_i \in A_j\} \right)^{-1} \ge s^{-1}.$$

REMARKS. (i) The case $s = N$ gives the Bayesian function (15) and $s = 1$ gives the consonant function (16). For general s, $Q(\cdot,\tau)$ is additive across the partition $A_1,\ldots,A_s$, and is consonant conditional on each A_j. In this case we say that $Q(\cdot,\tau)$ is *consonant over the partition* or *partially consonant.* The requirement that θ should have plausibility 1 whenever it assigns probability 1 to the observation leads to $s = 1$ and formula (16). Then θ has plausibility 1 whenever it maximises the likelihood function. The effect of varying s can be seen by comparing columns E, F, G and A of Table 1. Increasing s tends to bring $\mathrm{PL}(A)$ and $\mathrm{BEL}(A)$ closer together. Comparison of C with E and D with F shows that, for the same partition and λ, (17) tends to give much wider interval widths $\mathrm{PL}(A) - \mathrm{BEL}(A)$ than does (7).

(ii) The functions BEL and PL corresponding to (17) are

$$\begin{aligned} \mathrm{PL}(A,\tau) &= k(\tau) \sum_{j=1}^{s} \max\{\tau_i^\lambda : \theta_i \in A \cap A_j\} \\ &= \sum_{j=1}^{s} \max\{\mathrm{PL}(\{\theta_i\},\tau) : \theta_i \in A \cap A_j\}, \end{aligned} \tag{18}$$

$$\mathrm{BEL}(A,\tau) = k(\tau) \sum_{j=1}^{s} \left(\max\{\tau_i^\lambda : \theta_i \in A_j\} - \max\{\tau_i^\lambda : \theta_i \in A^c \cap A_j\} \right).$$

(iii) Each of $Q(A,\tau)$, $\mathrm{PL}(A,\tau)$ and $\mathrm{BEL}(A,\tau)$ is continuous in τ, nondecreasing in τ_j for $\tau_j \in A$ and nonincreasing in τ_j for $\tau_j \in A^c$. $Q(\{\theta_i\},\tau) = \mathrm{PL}(\{\theta_i\},\tau)$ is strictly increasing in τ_i unless $Q(\{\theta_i\},\tau) = 1$.

(iv) Theorem 3 remains valid when (A7) is replaced by the weaker axiom:

(A10) $Q(\cdot, c\mathbf{1}) = Q(\cdot,\mathbf{1})$ whenever $0 < c < 1$,

and also when (A9) is replaced by the weaker axiom:

(A11) When $A \subset B$, $A \ne \varnothing$, $\tau \in \mathscr{S}$ and $\tau I_B \in \mathscr{S}$, $Q(A,\tau I_B) = Q(B,\tau I_B)$ if and only if $Q(A,\tau) = Q(B,\tau)$.

(Only these weaker axioms are used in the proof.)

4.2. *Combination of independent observations.* When $s < N$, the functions given by (17) are not of the form (7) in Theorem 1 and so cannot satisfy (A2). Thus, Dempster's rule of combination cannot be used to combine evidence from

independent observations. [A restricted version of (A2) does hold in the case where one observation has the form $[\theta \in B]$, i.e., $Q(\cdot, \tau) \oplus Q(\cdot, I_B) = Q(\cdot, \tau I_B)$ whenever $\tau \in \mathscr{S}$ and $\tau I_B \in \mathscr{S}$.]

Next we define a new rule for combining partially consonant commonality functions which, unlike Dempster's rule, can be used to combine evidence from independent observations. Our purpose in introducing the rule is not to advocate its use—indeed, we will argue later that it shares some of the defects of Dempster's rule [see especially (v) and (vi) of Section 5]—but rather to show that there are alternative rules for combining belief functions which have advantages over Dempster's rule.

The new rule $\boxplus$ is defined for partially consonant commonality functions Q_1 and Q_2 for all nonempty sets A by

$$(Q_1 \boxplus Q_2)(A) = \begin{cases} 0 & \text{if } Q_1(A)\, Q_2(A) = 0, \\ k \min\{Q_1(\{\theta\})Q_2(\{\theta\}) \colon \theta \in A\} & \text{otherwise,} \end{cases} \tag{19}$$

where $k > 0$ is uniquely determined so that $Q_1 \boxplus Q_2$ is a commonality function by (3). Then $Q_1 \boxplus Q_2$ is well defined provided $Q_1(\{\theta\})Q_2(\{\theta\}) > 0$ for some $\theta \in \Theta$. If Q_1 is partially consonant over $A_1, \ldots, A_s$ and Q_2 is partially consonant over $B_1, \ldots, B_r$, then $Q_1 \boxplus Q_2$ is partially consonant over the common refinement $C_1, \ldots, C_t$ with $C_i = A_j \cap B_k$. Thus the class of partially consonant commonality functions is closed under $\boxplus$. (But $\boxplus$ cannot be used to combine arbitrary commonality functions, since their combination under it need not be a commonality function.)

It is easily verified that for the Q-functions in Theorem 3, $Q(\cdot, \tau\sigma) = Q(\cdot, \tau) \boxplus Q(\cdot, \sigma)$ under this rule, i.e., combination of independent observations by the new rule agrees with direct use of joint likelihoods. Thus, (A2) is satisfied by (17) if Dempster's rule is replaced by the new rule (19). Note that the new rule, like both Dempster's rule and Bayes' rule, is multiplicative in the commonality numbers of singletons. Moreover, the new rule is symmetric in Q_1 and Q_2, $Q_1 \boxplus Q_2 = Q_2$ when Q_1 is vacuous, and $Q_1 \boxplus Q_2$ is Bayesian when Q_1 is Bayesian. It follows that the functions Q and R of Theorem 3 satisfy (A4) when Dempster's rule is replaced by the new rule. Thus, (19) may be used to combine prior with observational evidence as well as to combine evidence from independent observations.

Given $0 \leq \tau \leq 1$, write $\tau^{(j)}$ for the N-vector with $\tau_j^{(j)} = \tau$, $\tau_i^{(j)} = 1$ for $i \neq j$. Then the functions Q_j, defined by $Q_j(A, \tau) = Q(A, \tau^{(j)})$, have the same simple form under both representations (7) and (17),

$$Q_j(A, \tau) \propto \begin{cases} \tau^{\lambda} & \text{if } \theta_j \in A \in \mathscr{A}^+, \\ 1 & \text{if } \theta_j \notin A \in \mathscr{A}^+, \\ 0 & \text{if } A \notin \mathscr{A}^+ \text{ and } A \neq \phi. \end{cases} \tag{20}$$

Moreover, for both representations, $Q(A, \tau)$ can be obtained by combining N of

these functions:

$$Q(\cdot, \tau) = Q_1(\cdot, \tau_1) \oplus Q_2(\cdot, \tau_2) \oplus \cdots \oplus Q_N(\cdot, \tau_N). \tag{21}$$

The representation of Theorem 1 is obtained by taking $\oplus$ to be Dempster's rule (6) and that of Theorem 3 by replacing $\oplus$ by the new rule $\boxplus$ defined in (19).

4.3. *Discounted likelihoods.* As in Section 2.2, some insight into the role of the parameter λ can be gained by transforming the consonant function Q_c (16) to its corresponding weight of evidence function

$$w(B, \tau) = \lambda \max\{v(B, \tau), 0\},$$

where

$$v(B, \tau) = \log(\min\{\tau_i\colon\ \theta_i \in B\}/\max\{\tau_i\colon\ \theta_i \in B^c\})$$

and $\log(0/0)$ is taken as 0. Thus, only those subsets B for which $\min\{\tau_i\colon\ \theta_i \in B\} > \max\{\tau_i\colon\ \theta_i \in B^c\}$ are assigned positive weight of evidence. Shafer [(1976a), Section 11.2] shows that the sum of weights of evidence over all subsets containing θ_1 but not θ_2, which he interprets as the "weight of evidence favouring θ_1 over θ_2," is just $\lambda \log(\tau_1/\tau_2)$ when $\tau_1 > \tau_2$, and 0 otherwise. Indeed, this property characterises the functions Q_c [Shafer (1976a), Theorem 11.3]. Thus, weights of evidence are a constant multiple λ of positive log likelihood ratios.

For the consonant function (16), $Q(A, \tau)$, $\mathrm{PL}(A, \tau)$ and $\mathrm{PL}(A, \tau) - \mathrm{BEL}(A, \tau) = \min\{\mathrm{PL}(A, \tau), \mathrm{PL}(A^c, \tau)\}$ are all nonincreasing in λ. Their behaviour as λ varies is illustrated in Table 1. (Compare columns E and J. The partially consonant F and I behave similarly.)

Notice that while the partition $A_1, \ldots, A_s$ and the scale parameter λ in Theorem 3 may depend on the source of the models in Θ (since Θ has been fixed throughout the discussion), they may not depend on aspects of the Bayesian prior and likelihood such as their mutual consistency or the consistency of different observations. Thus, we must not only discount a Bayesian prior and likelihood function to the same degree λ, but also discount all possible likelihood functions to this same degree. At the cost of giving up (A4) and (A2) for rule (19), we can employ this form of discounting much more widely, as a way of reducing the "informativeness" of a commonality function to reflect our degree of confidence in the prior assessment or likelihood function on which it is based. That is, we might use the functional forms for R and Q given by (13) and (17), but allow λ to depend on the particular prior ρ and likelihood τ.

5. Conclusion. Equation (14) states that the commonality (or plausibility) function on singletons is proportional to some positive power of the likelihood function. This is compatible with likelihood inference, which takes the likelihood function to express the relative support provided by the data for single values of θ. But likelihood theorists have often insisted that the likelihood function cannot be used to define the support for larger subsets of Θ, whereas (17) does extend the commonality function in (14) to all subsets of Θ. At one extreme ($s = N$),

the plausibility function derived from statistical evidence alone is additive, as in fiducial inference. At the other extreme ($s = 1$), the plausibility of a set A is obtained from the maximum of the likelihood function over A. This corresponds to a version of likelihood inference in which the likelihood function L is extended to all subsets of Θ by $L(A) = \max\{L(\theta)\colon \theta \in A\}$, as in the definition of the generalised likelihood-ratio statistic, widely used in hypothesis testing. The other values of s in (17) give functions intermediate between these two extremes.

Note also that Theorem 3 leads to Bayesian inference (with the minor generalisation through the scale parameter λ) if we require that Q *always* be combined with a Bayesian prior distribution. Then all values of s in Theorem 3 give the same Bayesian posterior. (The case $s = N$ gives a Bayesian posterior for any prior commonality function, but non-Bayesian priors will typically produce non-Bayesian posteriors when $s < N$.)

The representations of statistical evidence defined in Theorem 3 seem more reasonable than those in Theorem 1, especially as they satisfy the likelihood principle (A7). The new rule of combination (19) can then be used to combine evidence from statistically independent observations, but Dempster's rule cannot and we are led to ask what role Dempster's rule does have in statistical problems. The following considerations suggest that, in problems of parametric statistical inference at least, Dempster's rule is not a satisfactory way of combining evidence.

(i) No convincing arguments have been advanced to support Dempster's rule, even as a method for combining specific types of evidence such as prior and likelihood evidence. [Compare with the coherence arguments in Walley (1987), which support a different rule, the generalised Bayes' rule discussed later.] In problems where we have a Bayesian prior, as in Theorem 2, Dempster's rule is consistent with Bayes' rule, but so are other rules such as (19) or the generalised Bayes' rule. In examples where Dempster's rule gives plausible results [Shafer (1976a)], so do other rules. Dempster's rule therefore seems somewhat arbitrary.

(ii) We saw in Section 2 that Dempster's rule cannot be used to combine evidence from statistically independent observations without violating the sufficiency principle in a serious way, unless the set functions involved are actually Bayesian (the case $s = N$).

(iii) By (ii), Dempster's rule cannot be used (without violating the sufficiency principle) to combine new observational evidence with prior evidence when the prior evidence itself includes earlier observations, even if these are statistically independent of the new observations. Thus, Dempster's rule cannot be used to sequentially update beliefs as new observations are obtained. More generally, it should not be used without careful attention to the source of prior beliefs.

(iv) On the other hand, the new rule (19) can be used consistently to combine new observations both with independent observations and with prior evidence, through the representations of Theorem 3. This seems a substantial advantage over Dempster's rule.

(v) If the prior R is non-Bayesian and we interpret the PL and BEL functions generated by both prior and posterior as upper and lower betting rates, then use of Dempster's rule or (19) to combine prior and likelihood can lead to a sure loss (or "Dutch book"). This can happen even when the possible likelihood functions take only values 0 and 1, so having the effect of conditionalizing the prior.

The following example of such a sure loss can be interpreted as a model for the "three prisoners" problem discussed by Walley [(1987), Chapter 6.3]. Let $N = 4$ and $\mathcal{X} = \{x, y\}$. Take the statistical models P_{θ_i} to be degenerate distributions at x when $i = 1$ or $i = 3$ and to be degenerate at y when $i = 2$ or $i = 4$. Consider the prior probability assignment $m(\{\theta_1\}) = m(\{\theta_2\}) = m(\{\theta_3, \theta_4\}) = \frac{1}{3}$, which corresponds to the prior commonality function $R(A) = \frac{1}{3}$ when $|A| = 1$ or $A = \{\theta_3, \theta_4\}$ and $R(A) = 0$ otherwise. Observing x has the effect of conditionalizing the prior on the event $B = \{\theta_1, \theta_3\}$. Dempster's rule of conditioning [Shafer (1976a), Chapter 3] defines the posterior commonality function $R(A|x)$ to be proportional to the prior $R(A)$ when $A \subset B$ and zero otherwise. Because $R(B) = 0$ and $R(\{\theta_1\}) = R(\{\theta_3\})$, this gives $R(\{\theta_1\}|x) = R(\{\theta_3\}|x) = \frac{1}{2}$, $R(A|x) = 0$ otherwise. The same posterior is generated when Dempster's rule of combination (6) or the alternative rule (19) is used to combine the prior R with some commonality function Q induced by the degenerate likelihood function, provided Q satisfies the minimal conditions $Q(\{\theta_1\}) = Q(\{\theta_3\})$ and $Q(A) = 0$ unless $A \subset B$ [which are satisfied, for example, by all the functions of form (7) or (17)]. Similarly, observing y has the effect of conditionalizing on $\{\theta_2, \theta_4\}$, giving posterior $R(\{\theta_2\}|y) = R(\{\theta_4\}|y) = \frac{1}{2}$, $R(A|y) = 0$ otherwise.

Now consider the event $D = \{\theta_1, \theta_2\}$. The prior satisfies $\mathrm{BEL}(D) = \mathrm{PL}(D) = \frac{2}{3}$, whereas the posterior has $\mathrm{BEL}(D|x) = \mathrm{PL}(D|x) = \frac{1}{2} = \mathrm{BEL}(D|y) = \mathrm{PL}(D|y)$. The prior and posterior are inconsistent; a bet *on* event D prior to the observation at lower betting rate $\mathrm{BEL}(D) = \frac{2}{3}$, together with a posterior bet *against* D at upper betting rate $\mathrm{PL}(D|x) = \mathrm{PL}(D|y) = \frac{1}{2}$, will produce a sure loss. Thus, PL and BEL cannot (coherently) be interpreted as upper and lower betting rates when Dempster's rule of conditioning is used or when rule (6) or (19) is used to combine prior and likelihood.

(vi) Only those lower probability functions with the mathematical properties of belief functions can be combined by Dempster's rule. Dempster [(1968), page 225] and Shafer [(1981), page 15] have used this as an argument for restricting attention to belief functions. But the class of belief functions is too small to model all reasonable belief states. Consider, for example, beliefs about the outcomes of two physically unrelated coin-tosses, one made with an ordinary coin (believed to be fair) and the other with a coin that is so deformed that we are completely ignorant about its bias. Beliefs about the two outcomes can be properly modelled by a lower probability function, the lower envelope of all additive probability measures under which the two tosses are independent and the possible outcomes of the first toss each have probability $\frac{1}{2}$. This lower probability function is not a belief function, however, and there seems to be no reasonable alternative model which is a belief function. Analogous examples can

be constructed in which prior beliefs concerning statistical parameters cannot be represented by a belief function; e.g., reinterpret the outcomes "heads" or "tails" as presence or absence of a signal on two communications channels, where each signal is generated by a random mechanism and itself generates a noisy observation. It seems, therefore, that in some statistical problems we must use lower probability models that are not belief functions and so cannot be combined by Dempster's rule. [The rule (19) is even less satisfactory in this respect as it applies only to those belief functions which are partially consonant.]

We have argued that Dempster's rule should not be used to combine evidence from statistically independent observations. Arguments (iii), (v) and (vi) suggest that neither Dempster's rule nor the new rule (19) is suitable for combining prior and likelihood evidence. There is, however, an attractive alternative rule for combining prior and likelihood, the generalised Bayes' rule studied by Walley [(1987), Chapter 8]. This generalises Bayes' rule by admitting a wide class of "coherent" upper and lower probability models for prior beliefs; this class is somewhat wider than the class of belief functions admitted by Dempster's rule [it includes the model defined in (vi), for example]. The generalised Bayes' rule can be justified as the uniquely coherent rule for combining prior and likelihood when upper and lower probabilities are interpreted as upper and lower betting rates. It meets our objections (i), (iii), (v) and (vi) to Dempster's rule of combination. In our view, the generalised Bayes' rule has substantial advantages over Dempster's rule in statistical problems and the theory of coherence is more promising than the theory of belief functions as a foundation for statistical inference.

6. Proofs of main results.

LEMMA 1. *Suppose* Q *satisfies* (A1), (R1) *and*

$$(22)\quad (\forall A \in \mathscr{A};\ \tau, \sigma, \tau\sigma \in \mathscr{S}) \qquad Q(A, \tau\sigma) = k(\tau, \sigma)Q(A, \tau)Q(A, \sigma),$$

where $\mathscr{A}$ *is a class of nonempty sets containing all singletons. Define* $\mathscr{A}^+ = \{A \in \mathscr{A}:\ A \neq \phi,\ Q(A, \mathbf{1}) > 0\}$. *Then*

$$(\forall A \in \mathscr{A} - \mathscr{A}^+, \tau \in \mathscr{S}) \qquad Q(A, \tau) = 0,$$

$$(23)\quad (\forall A \in \mathscr{A}^+)\,(\exists\, \lambda_1(A), \ldots, \lambda_N(A) \geq 0)\,(\forall\, \tau \in \mathscr{S})$$

$$Q(A, \tau) = k(\tau) \prod_{j=1}^{N} \tau_j^{\lambda_j(A)}$$

for some $k(\tau) \geq N^{-1}$, *where* $\lambda_j(A)$ *may take the values* 0^+ *and* ∞, *with the conventions* $(\forall\, x > 0)\, x^{0^+} = 1$ *and* $0^{0^+} = 0$; $(\forall\, x < 1)\, x^{\infty} = 0$ *and* $1^{\infty} = 1$. *Also,* $\lambda_j(A) \geq \lambda_j(B)$ *whenever* $A, B \in \mathscr{A}^+$, $A \supset B$ *and* $1 \leq j \leq N$.

PROOF. For $A \in \mathscr{A}$, $0 \leq x \leq 1$, define $Q_j(A, x) = Q(A, \sigma)$, where $\sigma_j = x$, $\sigma_i = 1$ for $i \neq j$. Since any $\tau \in \mathscr{S}$ can be expressed as a product of N such σ's,

repeated use of (22) gives

$$(24) \qquad (\forall A \in \mathscr{A}, \tau \in \mathscr{S}) \qquad Q(A,\tau) = k_0(\tau) \prod_{j=1}^{N} Q_j(A,\tau_j).$$

But $\Sigma_j Q(\{\theta_j\},\tau) \geq 1$ by (A1), hence $k_0(\tau) > 0$. Now fix j, $i \neq j$ and $A \in \mathscr{A}$. For all $0 \leq x \leq 1$, define $f(x) = Q_j(A,x)/Q_j(\{\theta_i\},x)$. By (R1) and (A1), this is well defined and $0 \leq f(x) \leq c_i^{-1}$ for $0 \leq x \leq 1$. Use (22) to show

$$(25) \qquad f(xy) = f(x)f(y) \quad \text{for all } x \text{ and } y \text{ in } [0,1].$$

Clearly $f(0)$ and $f(1)$ take the values 0 or 1. If $f(0) = 1$, then $(\forall\ 0 \leq x \leq 1)$ $f(x) = 1$. If $f(1) = 0$, then $(\forall\ 0 \leq x \leq 1)$ $f(x) = 0$. If $f(x) = 0$ for some $x > 0$, then $(\forall\ 0 \leq y \leq x)$ $f(y) = 0$ and $f(x^{1/n}) = 0$ by (25), hence $(\forall\ 0 \leq y < 1)$ $f(y) = 0$. Thus there are five types of solution to (25):

(i) $(\forall\ 0 \leq x \leq 1)$ $f(x) = 0$.
(ii) $(\forall\ 0 \leq x \leq 1)$ $f(x) = 1$.
(iii) $f(0) = 0$, $(\forall\ 0 < x \leq 1)$ $f(x) = 1$.
(iv) $(\forall\ 0 \leq x < 1)$ $f(x) = 0$, $f(1) = 1$.
(v) $f(0) = 0$, $f(1) = 1$, $(\forall\ 0 < x < 1)$ $f(x) > 0$, $(\exists\ 0 < x < 1)$ $f(x) \neq 1$.

In case (v), define $g(y) = \log f(e^y)$ for $y \leq 0$ and $g(y) = -g(-y)$ for $y > 0$. Then (25) gives Cauchy's equation $g(y+z) = g(y) + g(z)$ for all real y, z, with $g(y) \leq -\log c_i$ for $y \leq 0$, and g not identically zero. The only solutions are $g(y) = \lambda y$ for some real $\lambda > 0$. Hence, $f(x) = x^\lambda$ for $0 \leq x \leq 1$, and this also covers cases (ii)–(iv) by taking $\lambda = 0, 0+, \infty$, respectively. Taking $\lambda_j(A) = \lambda$, either $(\forall\ 0 \leq x \leq 1)$ $Q_j(A,x) = 0$ [case (i)] or $(\forall\ 0 \leq x \leq 1)$ $Q_j(A,x) = k_j(x)x^{\lambda_j(A)}$, where $k_j(x) = Q_j(\{\theta_i\},x) \geq c_i$. By (24), either $(\forall\ \tau \in \mathscr{S})$ $Q(A,\tau) = 0$ or $(\forall\ \tau \in \mathscr{S})$ $Q(A,\tau) = k(\tau)\prod_{j=1}^{N}\tau_j^{\lambda_j(A)}$, where $k(\tau) = k_0(\tau)\prod_{j=1}^{N}k_j(\tau_j) > 0$. By (A1), $1 \leq \Sigma_j Q(\{\theta_j\},\tau) \leq Nk(\tau)$. If $A, B \in \mathscr{A}^+$ with $A \supset B$, then $(\forall\ 0 \leq x \leq 1)$ $Q_j(A,x)/Q_j(B,x) = x^{\lambda_j(A)-\lambda_j(B)} \leq 1$ since $Q_j(\cdot,x)$ is a commonality function, so that $\lambda_j(A) \geq \lambda_j(B)$. □

LEMMA 2. *Suppose Q satisfies* (A1), (R1)–(R4) *and* (22). *Then Q has the form* (23), *with* (*for* $A \in \mathscr{A}^+$) $\lambda_j(A) = 0$ *unless* $\theta_j \in A$, *and* $\infty > \lambda_j(A) \geq \lambda_j(\{\theta_j\}) > 0$ *when* $\theta_j \in A$.

PROOF. By Lemma 1, Q satisfies (23). Suppose $A \in \mathscr{A}^+$. Using (24), (R3) gives $Q_j(A,0) > 0$, hence $\lambda_j(A) = 0$, when $\theta_j \notin A$. By (R4), $\lambda_j(A) < \infty$ for all j. All singletons are in $\mathscr{A}^+$ by (R1). By (R2), $Q_j(\{\theta_j\},x) \geq c_i x^{\lambda_j(\{\theta_j\})} \to 0$ as $x \to 0$, so that $\lambda_j(\{\theta_j\}) > 0+$. Cases (iii) and (iv) in Lemma 1 are thus ruled out by (R2) and (R4). Thus, whenever $\theta_j \in A \in \mathscr{A}^+$, $0 < \lambda_j(\{\theta_j\}) \leq \lambda_j(A) < \infty$. □

LEMMA 3. *A commonality function Q is idempotent, i.e., $Q \oplus Q = Q$, if and only if for some $1 \leq s \leq N$ there are disjoint $A_1, \ldots, A_s \subset \Theta$ such that $Q(A) = s^{-1}$ when A is nonempty and contained in one of the A_j, $Q(\phi) = 1$ and $Q(A) = 0$ otherwise.*

PROOF. If Q has the given form, $(Q \oplus Q)(A) \propto Q(A)^2 \propto Q(A)$ for $A \neq \phi$, so Q is idempotent. Conversely, suppose Q is idempotent. Let $A_1, \ldots, A_s$ be the distinct subsets of Θ with $m(A_j) > 0$. By (6), $(\forall A \neq \phi)\ Q(A) = kQ(A)Q(A)$ for some $k > 0$, so $(\forall A \neq \phi)\ Q(A) = 0$ or k^{-1}. Hence, $Q(A_j) = k^{-1}$. For any $i \neq j$, $A_i \neq A_j$ so without loss of generality $A_j \not\subset A_i$. Then

$$Q(A_i \cap A_j) = \sum_{B \supset A_i \cap A_j} m(B)$$

$$\geq \sum_{B \supset A_j} m(B) + m(A_i) = Q(A_j) + m(A_i) > k^{-1},$$

so that $A_i \cap A_j = \phi$. Thus $A_1, \ldots, A_s$ are disjoint. It easily follows that Q has the given form, with $m(A_j) = s^{-1} = k^{-1}$. □

COROLLARY 5. *A commonality functional Q satisfies* (A8), *i.e., $Q(\cdot, 1)$ acts as an identity under Dempster's rule, if and only if $Q(\cdot, 1)$ has the form given in Lemma* 3 *and $Q(A, \tau) = 0$ for all $\tau \in \mathscr{S}$ unless A is contained in one of the sets $A_1, \ldots, A_s$.*

PROOF. If Q satisfies (A8), $Q(\cdot, 1)$ must be idempotent and so have the form given in Lemma 3. If $(\forall\ j)\ A \not\subset A_j$, then $Q(A, 1) = 0$, so $Q(A, \tau) \propto Q(A, \tau)Q(A, 1) = 0$. The converse is easy to check. □

PROOF OF THEOREM 1. Suppose Q satisfies (A1), (A2) and (R1)–(R4). Let $\mathscr{A}$ be the class of all nonempty subsets of Θ. (A2) implies (A8), so by Corollary 5 there are disjoint $A_1, \ldots, A_s$ with $Q(A, \tau) = 0$ unless $A \in \mathscr{A}^+ = \{A\colon (\exists\ j)\ A \subset A_j\}$. By (R1), $A_1, \ldots, A_s$ must be a partition of Θ. (A2) implies (22), so the conditions of Lemma 2 hold. Let A be any proper subset of A_l that contains θ_j. To establish (7), we need to show that $\lambda_j(A) = \lambda_j(A_l)$. Let Q_j be as in (24) and let m_j be the probability assignment corresponding to Q_j. Then $A - \{\theta_j\}$ and $A_l - \{\theta_j\}$ are in $\mathscr{A}^+$, hence $Q_j(A - \{\theta_j\}, x) = k_j(x) = Q_j(A_l - \{\theta_j\}, x)$ from Lemma 1. Because A contains $A - \{\theta_j\}$ but not $A_l - \{\theta_j\}$, by (1) we must have $m_j(A, x) = 0$. Hence,

$$(\forall\ 0 \leq x \leq 1)\ Q_j(A, x) = k_j(x)x^{\lambda_j(A)} = \sum_{A_l \supset B \supset A} m_j(B, x) = m_j(A_l, x)$$

$$= Q_j(A_l, x) = k_j(x)x^{\lambda_j(A_l)}.$$

Thus $\lambda_j(A) = \lambda_j(A_l)$ is constant over $A \subset A_l$ for which $\theta_j \in A$. Writing $\lambda_j = \lambda_j(A_l)$, (7) holds. To obtain the expression for $k(\tau)$, use (3).

For the converse, suppose Q has the form (7). Then $m(A, \tau) = \Sigma_{B \supset A}(-1)^{|B-A|}Q(B, \tau)$ is clearly zero except when $A \in \mathscr{A}^+$, and is easily computed to be zero when $A = \phi$. $m(A, \tau) = k(\tau)\prod_{\theta_j \in A}\tau_j^{\lambda_j}\prod_{\theta_j \in A_l - A}(1 - \tau_j^{\lambda_j}) \geq 0$ when $A \subset A_l$, $A \neq \phi$. Thus $Q(\cdot, \tau)$ is a commonality function.

Since $k(\tau) \geq s^{-1}$, (R1) holds with $c_i = s^{-1}$. When some $\tau_i = 1$, $k(\tau) \leq 1$, so $Q(\{\theta_j\}, \tau) \leq \tau_j^{\lambda_j} \to 0$ as $\tau_j \to 0$. Thus (R2) holds. (R3) holds because $Q(A, I_A) = 1$ if $A \in \mathscr{A}^+$, and (R4) holds because $Q(A, \tau) > 0$ whenever $A \in \mathscr{A}^+$ and

$(\forall\ \theta_j \in A)\ \tau_j > 0$. Finally, (A2) follows from (7) and the multiplicative form of Dempster's rule (6). □

PROOF OF THEOREM 2. Assume the axioms and regularity conditions hold. Define $r(A) = R(A, \pi_0)$ and $\mathscr{A} = \{A\colon A \neq \phi,\ r(A) > 0\}$. We show that $\mathscr{A}$ contains just the singleton sets.

By (A4), whenever $\tau \in \mathscr{S}$, $\rho \in \mathscr{P}$ and $\rho \circ \tau$ is well defined,

$$(26)\quad (\forall A \neq \phi) \qquad R(A, \rho \circ \tau) = k_1(\rho, \tau)R(A, \rho)Q(A, \tau) \quad \text{for some } k_1 > 0.$$

For $\tau \in \mathscr{S}$, let $\tau' \in \mathscr{P}$ denote the normalized version of τ, $\tau_j' = \tau_j / \Sigma\tau_i$. Taking $\rho = \pi_0$ in (26) gives

$$(27)\quad (\forall \tau \in \mathscr{S},\ A \neq \phi) \qquad R(A, \tau') = k_2(\tau)r(A)Q(A, \tau) \quad \text{for some } k_2 > 0.$$

Now for any $\tau \in \mathscr{S}$, $\sigma \in \mathscr{S}$ with $\tau\sigma \in \mathscr{S}$ and $A \neq \phi$,

$$\begin{aligned} R(A, (\tau\sigma)') &= k_2(\tau\sigma)r(A)Q(A, \tau\sigma) \quad [\text{by (27)}] \\ &= k_1(\tau', \sigma)R(A, \tau')Q(A, \sigma) \quad [\text{by (26) since } (\tau\sigma)' = \tau' \circ \sigma] \\ &= k_1(\tau', \sigma)k_2(\tau)r(A)Q(A, \tau)Q(A, \sigma) \quad [\text{by (27)}]. \end{aligned}$$

Hence, $(\forall\ A \in \mathscr{A})\ Q(A, \tau\sigma) = k(\tau, \sigma)Q(A, \tau)Q(A, \sigma)$, where $k > 0$. Thus Q satisfies (22). Now by (A3) and (R5), $\Sigma_{i=1}^N r(\{\theta_i\}) = Nr(\{\theta_i\}) \geq 1$, so that $\mathscr{A}$ contains all singletons. Thus Q satisfies the conditions of Lemma 2. Taking $\tau = \mathbf{1}$ in (27) shows $Q(A, \mathbf{1}) > 0$ whenever $r(A) > 0$, so that $\mathscr{A}^+ = \mathscr{A}$. By Lemma 2,

$$(28)\qquad (\forall A \in \mathscr{A},\ \tau \in \mathscr{S})\ Q(A, \tau) = k(\tau) \prod_{\theta_j \in A} \tau_j^{\lambda_j(A)},$$

where $\lambda_j(A) \geq \lambda_j(\{\theta_j\}) > 0$ for $\theta_j \in A$. Now consider the constant likelihood $\tau = c\mathbf{1}$, where $0 < c < 1$. By (28), $(\forall\ A \in \mathscr{A})\ Q(A, c\mathbf{1}) = k(c\mathbf{1})c^{\mu(A)}$, where $\mu(A) = \Sigma_{\theta_j \in A}\lambda_j(A)$. By (27), $Q(A, c\mathbf{1}) = k_2(c\mathbf{1})^{-1}$ is constant over $A \in \mathscr{A}$. Hence, $\mu(A)$ is constant over $A \in \mathscr{A}$. But $\mathscr{A}$ contains all singletons, so $\mu(\{\theta_j\}) = \lambda$ is constant over j. Thus for $A \in \mathscr{A}$,

$$\lambda = \mu(A) \geq \sum_{\theta_j \in A} \lambda_j\big(\{\theta_j\}\big) = \sum_{\theta_j \in A} \mu\big(\{\theta_j\}\big) = \lambda|A|,$$

so $|A| = 1$. This shows that $\mathscr{A}$ consists of exactly the singletons, so $r(A) = 0$ when $|A| > 1$. By (27), $(\forall\ \rho \in \mathscr{P})\ R(A, \rho) = 0$ when $|A| > 1$, i.e., $R(\cdot, \rho)$ is Bayesian. (14) follows from (28). (R5), (27) and (28) imply $R(\{\theta_i\}, \rho) \propto Q(\{\theta_i\}, \rho) \propto \rho_i^\lambda$, giving (13). □

PROOF OF PROPOSITION. (a) $\Sigma_{j=1}^N Q(\{\theta_j\}, \tau) = k(\tau)\Sigma_{j=1}^N \tau_j^\lambda \geq 1$ for any commonality function Q. Hence,

$$Q(\{\theta_i\}, \tau) = k(\tau)\tau_i^\lambda \geq \tau_i^\lambda \Big/ \sum_{j=1}^N \tau_j^\lambda = Q_b(\{\theta_i\}, \tau).$$

Also, for $|A| > 1$, $Q(A, \tau) \geq 0 = Q_b(A, \tau)$.

(b) For any commonality function Q,

$$1 \geq \max\big\{Q\big(\{\theta_j\}, \tau\big)\colon 1 \leq j \leq N\big\} = k(\tau)\max\big\{\tau_j^\lambda\colon 1 \leq j \leq N\big\}.$$

Hence, whenever $A \neq \phi$,

$$Q(A,\tau) \leq \min\{Q(\{\theta\},\tau)\colon \theta \in A\} = k(\tau)\min\{\tau_i^\lambda\colon \theta_i \in A\}$$
$$\leq \min\{\tau_i^\lambda\colon \theta_i \in A\}/\max\{\tau_j^\lambda\colon 1 \leq j \leq N\} = Q_c(A,\tau). \qquad \square$$

PROOF OF THEOREM 3. Suppose that the assumptions hold. Then (13) and (14) hold by Theorem 2. From Corollary 5, using (A8), there are sets $A_1,\dots,A_s$ such that $Q(A,\mathbf{1}) = s^{-1}$ when $A \in \mathscr{A}^+$ and $Q(A,\tau) = 0$ otherwise, where $\mathscr{A}^+$ contains all nonempty subsets of each A_j. Using (14), $A_1,\dots,A_s$ form a partition.

Let A be any member of $\mathscr{A}^+$ that contains at least two points and let τ be in $\mathscr{S}$. By relabelling the points θ, we can assume that $A = \{\theta_1,\theta_2,\dots,\theta_j\}$, where $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_j$. Then $Q(C,\tau_j\mathbf{1}) = Q(C,\mathbf{1}) = s^{-1}$ for any nonempty $C \subset A$, using (A7). Hence, using (A9), $Q(C,\tau_j I_A)$ is constant over all nonempty $C \subset A$ and zero otherwise. Using (1), $m(A,\tau_j I_A) = Q(A,\tau_j I_A) = 1$. Hence,

$$1 = Q(A,\tau_j I_A) \leq Q(\{\theta_j\},\tau_j I_A) \leq Q(\{\theta_j\},\tau I_A),$$

using (R6) for the last inequality. Thus $Q(\{\theta_j\},\tau I_A) = 1$, giving $m(C,\tau I_A) = 0$ unless $\theta_j \in C$. Writing $B = \{\theta_1,\theta_2,\dots,\theta_{j-1}\}$, it follows that $Q(B,\tau I_A) = Q(A,\tau I_A)$. Applying (A9) again, $Q(B,\tau) = Q(A,\tau)$. Repeating this argument $j-1$ times and using (14), we see that

$$Q(A,\tau) = Q(\{\theta_1\},\tau) = k(\tau)\tau_1^\lambda = k(\tau)\min\{\tau_i^\lambda\colon \theta_i \in A\}.$$

This also holds when A is a singleton, by (14).

To obtain the formula for $k(\tau)$, apply the identities (3) and

$$\sum_{\phi \neq A \subset A_j} (-1)^{|A|+1}\min\{q_i\colon \theta_i \in A\} = \max\{q_i\colon \theta_i \in A_j\},$$

with $q_i = Q(\{\theta_i\},\tau)$, giving

$$1 = \sum_{j=1}^{s} \max\{q_i\colon \theta_i \in A_j\} = k(\tau)\sum_{j=1}^{s} \max\{\tau_i^\lambda\colon \theta_i \in A_j\}.$$

This establishes (17).

Conversely, suppose R and Q are defined by (13) and (17). Use (2) to show that the probability assignment corresponding to Q is

$$m(B,\tau) = \min\{Q(\{\theta_i\},\tau)\colon \theta_i \in B\} - \max\{Q(\{\theta_i\},\tau)\colon \theta_i \in A_j - B\}$$

when this is positive and $B \subset A_j$, $m(B,\tau) = 0$ otherwise. Since $m(\cdot,\tau)$ is nonnegative, $Q(\cdot,\tau)$ is a commonality function and (A1) holds. (A7) holds because $k(c\tau) = c^{-\lambda}k(\tau)$. (A8) follows from Corollary 5. (R6) holds since $k(\tau)$ is nonincreasing in each τ_j. The other axioms and conditions are easily verified. $\square$

Acknowledgments. This work developed from seminars organized by Terry Fine at Cornell University in spring 1977. I am grateful to Terry Fine, Mike West and Marco Wolfenson for comments on earlier drafts and to several referees and an Associate Editor.

REFERENCES

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

DEMPSTER, A. P. (1968). A generalization of Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247.

KAHNEMAN, D., SLOVIC, P. and TVERSKY, A., eds. (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge Univ. Press, London.

KRANTZ, D. H. and MIYAMOTO, J. (1983). Priors and likelihood ratios as evidence. *J. Amer. Statist. Assoc.* **78** 418–423.

LAD, F. (1978). Embedding Bayes' theorem in general learning rules: Connections between idealized behaviour and empirical research on learning. *British J. Math. Statist. Psych.* **31** 113–125.

SHAFER, G. (1976a). *A Mathematical Theory of Evidence.* Princeton Univ. Press, Princeton, N.J.

SHAFER, G. (1976b). A theory of statistical evidence (with discussion). In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (W. L. Harper and C. A. Hooker, eds.) **2** 365–436. Reidel, Dordrecht.

SHAFER, G. (1981). Constructive probability. *Synthese* **48** 1–60.

SHAFER, G. (1982). Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B* **44** 322–352.

WALLEY, P. (1987). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

WEST, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 531–558. North-Holland, Amsterdam.

WILLIAMS, P. M. (1978). On a new theory of epistemic probability. *British J. Philos. Sci.* **29** 375–387.

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

SCHOOL OF ELECTRICAL ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-5401