UPPER AND LOWER PROBABILITIES INDUCED BY A MULTIVALUED MAPPING[1]

By A. P. Dempster

Harvard University

0. Summary. A multivalued mapping from a space X to a space S carries a probability measure defined over subsets of X into a system of upper and lower probabilities over subsets of S. Some basic properties of such systems are explored in Sections 1 and 2. Other approaches to upper and lower probabilities are possible and some of these are related to the present approach in Section 3. A distinctive feature of the present approach is a rule for conditioning, or more generally, a rule for combining sources of information, as discussed in Sections 4 and 5. Finally, the context in statistical inference from which the present theory arose is sketched briefly in Section 6.

1. Introduction. Consider a pair of spaces X and S together with a multivalued mapping Γ which assigns a subset $\Gamma x \subset S$ to every $x \varepsilon X$. Suppose that μ is a probability measure which assigns probabilities to the members of a class $\mathfrak{F}$ of subsets of X. If μ is acceptable for probability judgments about an uncertain outcome $x \varepsilon X$, and if this uncertain outcome x is known to correspond to an uncertain outcome $s \varepsilon \Gamma x$, what probability judgments may be made about the uncertain outcome $s \varepsilon S$? The answer to this question would be a familiar one if Γ were single-valued, for under wide conditions a single-valued Γ would carry the measure μ over subsets of X into a unique probability measure over subsets of S. For multivalued Γ, however, one is led to consider upper and lower probabilities defined as follows over subsets of S.

For any $T \subset S$ define

$$T^* = \{x \varepsilon X, \Gamma x \cap T \neq \emptyset\} \tag{1.1}$$

and

$$T_* = \{x \varepsilon X, \Gamma x \neq \emptyset, \Gamma x \subset T\}. \tag{1.2}$$

In particular, $S^* = S_*$ is the domain of Γ. Define $\mathcal{E}$ to be the class of subsets T of S such that T^* and T_* belong to $\mathfrak{F}$. Suppose that $S \varepsilon \mathcal{E}$. Finally, define the *upper probability* of $T \varepsilon \mathcal{E}$ to be

$$P^*(T) = \mu(T^*)/\mu(S^*) \tag{1.3}$$

and the *lower probability* of $T \varepsilon \mathcal{E}$ to be

$$P_*(T) = \mu(T_*)/\mu(S^*). \tag{1.4}$$

$P^*(T)$ and $P_*(T)$ are defined only if $\mu(S^*) \neq 0$.

Received 28 July 1966.

[1] This research was supported by the Office of Naval Research. Reproduction in whole or in part is permitted by the U. S. Government. Distribution of the document is unlimited.

Since T^* consists of those $x \varepsilon X$ which can possibly correspond under Γ to an $s \varepsilon T$, one may naturally regard $\mu(T^*)$ to be the largest possible amount of probability from the measure μ which can be transferred to outcomes $s \varepsilon T$. Similarly T_* consists of those $x \varepsilon X$ which must lead to an $s \varepsilon T$, so that $\mu(T_*)$ represents the minimal amount of probability which can be transferred to outcomes $s \varepsilon T$. The denominator $\mu(S^*)$ in (1.3) and (1.4) is a renormalizing factor necessitated by the fact that the model permits, in general, outcomes in X which do not map into a meaningful subset of S. The offending subset $\{x \varepsilon X, \Gamma x = \emptyset\}$ must be removed from X and the measure of the remaining set S^* renormalized to unity. It would have been possible to restrict the formulation so that $\mu(S^*) = 1$, but it will be convenient in Sections 4 and 5 to have the general model.

The case of finite $S = \{s_1, s_2, \cdots, s_m\}$ will now be developed somewhat further. Suppose that $S_{\delta_1\delta_2\cdots\delta_m}$ denotes the subset of S which contains s_i if $\delta_i = 1$ and excludes s_i if $\delta_i = 0$, for $i = 1, 2, \cdots, m$. The 2^m subsets of S so defined are the possible Γx, and they determine a partition of X into

$$X = \bigcup_{\delta_1\delta_2\cdots\delta_m} X_{\delta_1\delta_2\cdots\delta_m} \tag{1.5}$$

where

$$X_{\delta_1\delta_2\cdots\delta_m} = \{x \varepsilon X, \Gamma x = S_{\delta_1\delta_2\cdots\delta_m}\}. \tag{1.6}$$

For any $T \subset S$, the subsets T^* and T_* are unions of subsets of the form $X_{\delta_1\delta_2\cdots\delta_m}$ and hence $P^*(T)$ and $P_*(T)$ are uniquely determined by the 2^m quantities

$$p_{\delta_1\delta_2\cdots\delta_m} = \mu(X_{\delta_1\delta_2\cdots\delta_m}). \tag{1.7}$$

It is assumed, of course, that each $X_{\delta_1\delta_2\cdots\delta_m}$ is in $\mathfrak{F}$. Note that *any* set of 2^m non-negative numbers $p_{\delta_1\delta_2\cdots\delta_m}$ with sum unity determines a possible set of upper and lower probabilities for all $T \subset S = \{s_1, s_2, \cdots, s_m\}$.

Table 1 displays formulas for all possible upper and lower probabilities when $m = 3$. For example, if $T = S_{110} = \{s_1, s_2\}$, then $T^* = X_{100} \cup X_{010} \cup X_{110} \cup X_{101}$

TABLE 1

Upper and lower probabilities when $S = \{s_1, s_2, s_3\}$.

T	$P^*(T)$	$P_*(T)$
$\emptyset$	0	0
$\{s_1\}$	$(p_{100} + p_{110} + p_{101} + p_{111})/(1 - p_{000})$	$p_{100}/(1 - p_{000})$
$\{s_2\}$	$(p_{010} + p_{110} + p_{011} + p_{111})/(1 - p_{000})$	$p_{010}/(1 - p_{000})$
$\{s_3\}$	$(p_{001} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$p_{001}/(1 - p_{000})$
$\{s_1, s_2\}$	$(p_{100} + p_{010} + p_{110} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$(p_{100} + p_{010} + p_{110})/(1 - p_{000})$
$\{s_1, s_3\}$	$(p_{100} + p_{001} + p_{110} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$(p_{100} + p_{001} + p_{101})/(1 - p_{000})$
$\{s_2, s_3\}$	$(p_{010} + p_{001} + p_{110} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$(p_{010} + p_{001} + p_{011})/(1 - p_{000})$
S	1	1

$\cup X_{011} \cup X_{111}$ and $T_* = X_{100} \cup X_{010} \cup X_{110}$, and therefore

$$\mu(T^*) = p_{100} + p_{010} + p_{110} + p_{101} + p_{011} + p_{111} \tag{1.8}$$

and

$$\mu(T_*) = p_{100} + p_{010} + p_{110}. \tag{1.9}$$

These need only be divided by

$$\mu(S^*) = 1 - p_{000} \tag{1.10}$$

to become upper and lower probabilities as defined in (1.3) and (1.4). Similar arguments yield the rest of Table 1.

This section closes with several more definitions. The term *variate* will be used for a real-valued function defined over S. Subject to measurability requirements, any variate V has an *upper distribution function* $F^*(v)$ and a *lower distribution function* $F_*(v)$ defined by

$$\begin{aligned} F^*(v) &= P^*(V \leqq v), \\ F_*(v) &= P_*(V \leqq v), \end{aligned} \tag{1.11}$$

for $-\infty < v < \infty$. The corresponding upper and lower expected values $E^*(V)$ and $E_*(V)$ are defined by

$$\begin{aligned} E^*(V) &= \textstyle\int_{-\infty}^{\infty} v\, dF_*(v); \\ E_*(V) &= \textstyle\int_{-\infty}^{\infty} v\, dF^*(v). \end{aligned} \tag{1.12}$$

(The interchange of upper and lower stars is necessary here in order to have both $F_*(v) \leqq F^*(v)$ and $E_*(V) \leqq E^*(V)$.)

The concepts of upper expected value and lower expected value generalize the concepts of upper probability and lower probability, respectively. For, if the variate Z is defined to be the indicator function of $T \subset S$, i.e., if

$$\begin{aligned} Z(s) &= 1 \quad \text{for } s \,\varepsilon\, T, \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{1.13}$$

then it follows from (1.12) that

$$\begin{aligned} E^*(Z) &= P^*(T); \\ E_*(Z) &= P_*(T). \end{aligned} \tag{1.14}$$

2. The class of compatible measures over S. Given a system of upper and lower probabilities for the subsets $\mathcal{E}$ of S determined as above from $(X, \mathfrak{F}, \mu)$ and Γ, it is natural to ask for the class $\mathcal{C}$ of probability measures P such that

$$P_*(T) \leqq P(T) \leqq P^*(T) \tag{2.1}$$

for all $T \,\varepsilon\, \mathcal{E}$. Clearly $\mathcal{C}$ is the same as the class of probability measures P such that

$$E_*(V) \leqq E(V) \leqq E^*(V) \tag{2.2}$$

for all variates V for which $E_*(V)$ and $E^*(V)$ are defined and finite, and where $E(\cdots)$ refers to expectation with respect to P. The class $\mathcal{C}$ will be called the class of measures *compatible* with the given system of upper and lower probabilities.

It is convenient to begin with a constructive definition of a class $\mathcal{C}_1$ of measures P and to prove ultimately that $\mathcal{C} = \mathcal{C}_1$. A general member of the class $\mathcal{C}_1$ is defined by specifying a probability measure $\gamma_{\Gamma x}$ over each possible $\Gamma x \subset S$ and taking $P(T) = \int \gamma_{\Gamma x}(T \cap \Gamma x)\, d\mu(x)$. To avoid topological complexities only the case of finite S will be considered in detail. Consider, therefore, the following method of constructing a probability measure P over the finite sample space $S = \{s_1, s_2, \cdots, s_m\}$ given the 2^m quantities $p_{\delta_1\delta_2\cdots\delta_m}$ defined in (1.7).

Suppose that each $p_{\delta_1\delta_2\cdots\delta_m}$ other than $p_{00\cdots0}$ is partitioned into a sum of m non-negative pieces

$$p_{\delta_1\delta_2\cdots\delta_m} = \sum_{i=1}^{m} p^{(i)}_{\delta_1\delta_2\cdots\delta_m} \tag{2.3}$$

where $p^{(i)}_{\delta_1\delta_2\cdots\delta_m} = 0$ unless $\delta_i = 1$. Define the measure P from

$$P\{s_i\} = \sum_{\delta_1\delta_2\cdots\delta_m} p^{(i)}_{\delta_1\delta_2\cdots\delta_m}/(1 - p_{00\cdots0}) \tag{2.4}$$

for $i = 1, 2, \cdots, m$. The motivation behind this definition of P is that in the logic of the situation $p_{\delta_1\delta_2\cdots\delta_m}$ is a piece of probability that may attach to any s_i for which $\delta_i = 1$. The partition (2.3) specifies the subpieces to be attached to each eligible s_i and (2.4) collects the appropriate subpieces from all $p_{\delta_1\delta_2\cdots\delta_m}$.

For example, when $m = 3$, one needs the decompositions

$$\begin{aligned} p_{110} &= p^{(1)}_{110} + p^{(2)}_{110}; \\ p_{101} &= p^{(1)}_{101} + p^{(3)}_{101}; \\ p_{011} &= p^{(2)}_{011} + p^{(3)}_{011}; \\ p_{111} &= p^{(1)}_{111} + p^{(2)}_{111} + p^{(3)}_{111}; \end{aligned} \tag{2.5}$$

and the corresponding measure P is defined from

$$\begin{aligned} P\{s_1\} &= (p_{100} + p^{(1)}_{110} + p^{(1)}_{101} + p^{(1)}_{111})/(1 - p_{000}); \\ P\{s_2\} &= (p_{010} + p^{(2)}_{110} + p^{(2)}_{011} + p^{(2)}_{111})/(1 - p_{000}); \\ P\{s_3\} &= (p_{001} + p^{(3)}_{101} + p^{(3)}_{011} + p^{(3)}_{111})/(1 - p_{000}). \end{aligned} \tag{2.6}$$

The class of all measures P determined by such partition schemes will be denoted by $\mathcal{C}_1$. These measures are compatible in the sense of (2.1); indeed,

$$\begin{aligned} P_*(T) &= \min_{P \in \mathcal{C}_1} P(T), \\ P^*(T) &= \max_{P \in \mathcal{C}_1} P(T) \end{aligned} \tag{2.7}$$

for each $T \subset S$. More generally,

$$\begin{aligned} E_*(V) &= \min_{P \in \mathcal{C}_1} E(V), \\ E^*(V) &= \max_{P \in \mathcal{C}_1} E(V) \end{aligned} \tag{2.8}$$

for any variate V.

Before proving (2.8), it is convenient to introduce a finite subclass of $\mathcal{C}_1$ with several important properties, including the property that the extremes in (2.7) and (2.8) are all attained within this finite subclass. Suppose that $\pi(1)$, $\pi(2), \cdots, \pi(m)$ is a permutation of $1, 2, \cdots, m$. The partition (2.3) may be determined in such a way that $p_{\delta_1\delta_2\cdots\delta_m} = p^{(i)}_{\delta_1\delta_2\cdots\delta_m}$ for that i which appears first in the permutation $\pi(1), \pi(2), \cdots, \pi(m)$ subject, of course, to the restriction $\delta_i = 1$. Determining this partition for each $\delta_1, \delta_2, \cdots, \delta_m$ determines a specific member of $\mathcal{C}_1$ associated with the permutation $\pi(1), \pi(2), \cdots, \pi(m)$. The $m!$ members of $\mathcal{C}_1$ which are determined in this way are not necessarily distinct. They will be called the *extremal* members of $\mathcal{C}_1$ for reasons to become evident.

Given any variate V there is at least one permutation $\pi(1), \pi(2), \cdots, \pi(m)$ such that

$$V(s_{\pi(1)}) \leqq V(s_{\pi(2)}) \leqq \cdots \leqq V(s_{\pi(m)}). \tag{2.9}$$

It will now be shown that $\min_{P\varepsilon\mathcal{C}_1} E_*(V)$ is achieved when P is the extremal measure associated with any such $\pi(1), \pi(2), \cdots, \pi(m)$. Note first that for any measure P and any permutation satisfying (2.9)

$$E(V) = V(s_{\pi(1)}) + \textstyle\sum_{j=2}^{m} [V(s_{\pi(j)}) - V(s_{\pi(j-1)})] \cdot P\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}. \tag{2.10}$$

Second, it is claimed that the $(m - 1)$ terms in the sum on the right side of (2.10) are simultaneously minimized by choosing P to be the extremal measure associated with any permutation satisfying (2.9). Indeed, $P\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}$ is minimized by requiring that the partition (2.3) concentrate as much as possible on $p^{(i)}_{\delta_1\delta_2\cdots\delta_m}$ with $i = \pi(1), \pi(2), \cdots, \pi(j-1)$. The partition defining the extremal measure corresponding to $\pi(1), \pi(2), \cdots, \pi(m)$ is clearly one means of assuring such a concentration. Furthermore, the definition of lower probability implies that this minimum of $P\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}$ is $P_*\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}$. The first half of (2.8) is thus proved; the other half follows similarly using the reverse permutation $\pi(m), \pi(m-1), \cdots, \pi(1)$.

Defining $\mathcal{C}_2$ to be the class of measures P formed by taking mixtures of the extremal measures, it is clear from their definitions that each of the classes $\mathcal{C}$, $\mathcal{C}_1$, and $\mathcal{C}_2$ are closed under the operation of mixing. It is also clear from the relations proved above that $\mathcal{C}_2 \subset \mathcal{C}_1 \subset \mathcal{C}$. This section concludes by showing that $\mathcal{C} = \mathcal{C}_1 = \mathcal{C}_2$, i.e., that these three possible definitions of compatibility are equivalent.

Any measure P determines a point

$$\mathbf{P} = (p^{(1)}, p^{(2)}, \cdots, p^{(m)}) \tag{2.11}$$

in the $(m - 1)$-dimensional simplex with the m vertices $(1, 0, \cdots, 0)$, $(0, 1, \cdots, 0), \cdots, (0, 0, \cdots, 1)$ where

$$p^{(i)} = P\{s_i\} \tag{2.12}$$

for $i = 1, 2, \cdots, m$. Any class of measures P defines a subset of the simplex and a class closed under mixing defines a convex subset. Thus $\mathcal{C}$, $\mathcal{C}_1$, and $\mathcal{C}_2$ may be identified with convex subsets of the simplex. Any convex set in $(m - 1)$-dimensional space is uniquely determined by the pairs of planes of support determined by all families of parallel planes of dimension $m - 2$. To show that $\mathcal{C} = \mathcal{C}_2$ one need only check that they have all the same planes of support.

In the present formulation, the intersection of the simplex with any plane of dimension $m - 2$ consists of all those measures P for which a variate V has the same expectation. For example, the plane of points $\mathbf{P}$ such that

$$a_1p^{(1)} + a_2p^{(2)} + \cdots + a_mp^{(m)} = c \tag{2.13}$$

contains all measures P such that $E(V) = c$ where V is defined by

$$V(s_i) = a_i \tag{2.14}$$

for $i = 1, 2, \cdots, m$. Of course, V is unique only up to a linear transformation of the form $a + bV$ and the family of planes parallel to (2.13) shares the family of variates $a + bV$. It follows that the planes of support of a closed convex subset of the simplex in the family of planes parallel to (2.13) are those which maximize and minimize $E(V)$ over choices of P in the closed convex subset. From (2.2) and (2.8), and because the extrema in (2.8) occur in $\mathcal{C}_2$, it follows that the closed convex subsets $\mathcal{C}$ and $\mathcal{C}_2$ have the same pairs of planes of support, as was required to prove.

From all this, it is seen that the class of compatible measures is a closed convex polygon in the simplex, having at most $m!$ vertices, namely, the extremal measures P. There may be as few as m distinct vertices; for example, the class of compatible measures may be the whole simplex in the "informationless" model where $p_{11\cdots1} = 1$ and all other $p_{\delta_1\delta_2\cdots\delta_m} = 0$.

3. Other approaches. The approach to upper and lower probabilities introduced above may be placed in a clearer perspective by considering a hierarchy of approaches, suggested to the author by L. J. Savage. Again consider for simplicity the case of finite S.

Any class $\mathcal{C}$ of probability measures P over the subsets $T \subset S$ defines upper and lower probabilities

$$P^*(T) = \sup_{P\varepsilon\mathcal{C}} P(T);$$
$$P_*(T) = \inf_{P\varepsilon\mathcal{C}} P(T). \tag{3.1}$$

Since the same upper and lower probabilities are yielded by the convex closure of $\mathcal{C}$ as by $\mathcal{C}$ itself, one might as well restrict $\mathcal{C}$ to be a closed convex set of measures.

Define Ω to be the class of all closed convex subsets of the simplex, i.e., all sets of probability measures over the subsets of S which are closed under mixing. Define $\Omega_1 \subset \Omega$ to consist of those closed convex sets of measures defined solely by inequalities on probabilities of events. Finally, define Ω_2 to consist of sets of

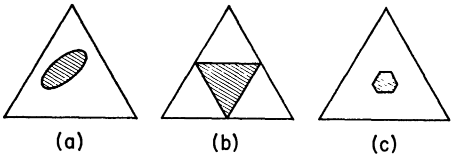

FIG. 1. Three types of convex subsets of the triangle: case (a) a general convex subset, case (b) a subset in Ω_1 but not in Ω_2 as described in the text, and case (c) a subset in Ω_2 with $p_{100} = p_{010} = p_{001} = \frac{1}{4}$ and $p_{110} = p_{101} = p_{011} = p_{111} = \frac{1}{16}$.

compatible measures as defined in Section 2, where the definition (2.1) assures that $\Omega_2 \subset \Omega_1$. It is clear, see for example Figure 1, that Ω_1 is properly contained in Ω. It will next be shown that Ω_2 is properly contained in Ω_1.

For any member of Ω_2, define

$$p'_{\delta_1\delta_2\cdots\delta_m} = p_{\delta_1\delta_2\cdots\delta_m}/(1 - p_{00\cdots0}) \tag{3.2}$$

if at least one $\delta_i = 1$, and define $p'_{00\cdots0} = 0$. The set of $p'_{\delta_1\delta_2\cdots\delta_m}$ determine the same $\mathcal{C}$ as do the original $p_{\delta_1\delta_2\cdots\delta_m}$ with the simplification that the normalizing factor $1 - p_{00\cdots0}$ may be ignored. Thus, for example when $m = 3$, all lower probabilities (and hence upper probabilities from (3.6)) may be formed from

$$\begin{aligned} P_*\{s_1\} &= p'_{100}\,; \\ P_*\{s_2\} &= p'_{010}\,; \\ P_*\{s_3\} &= p'_{001}\,; \\ P_*\{s_1, s_2\} &= p'_{100} + p'_{010} + p'_{110}\,; \\ P_*\{s_1, s_3\} &= p'_{100} + p'_{001} + p'_{101}\,; \\ P_*\{s_2, s_3\} &= p'_{010} + p'_{001} + p'_{011}\,. \end{aligned} \tag{3.3}$$

These relations may be solved to yield

$$\begin{aligned} p'_{100} &= P_*\{s_1\}, && \text{and similarly for } p'_{010} \text{ and } p'_{001}\,; \\ p'_{110} &= P_*\{s_1, s_2\} - P_*\{s_1\} - P_*\{s_2\} && \text{and similarly for } p'_{101} \text{ and } p'_{011}\,; \\ p'_{111} &= 1(= P_*\{s_1, s_2, s_3\}) - P_*\{s_1, s_2\} - P_*\{s_1, s_3\} - P_*\{s_2, s_3\} \\ &\quad + P_*\{s_1\} + P_*\{s_2\} + P_*\{s_3\}. \end{aligned} \tag{3.4}$$

The obvious extension of (3.4) to general m is easily proved by induction, and is omitted here.

The relations (3.4) may be applied to any member of Ω_1 using on the right side the bounding planes of support for that member. The result is a set of $p'_{\delta_1\delta_2\cdots\delta_m}$ which may be used as in (3.3) to determine the bounds of probabilities and hence give back the member of Ω_1. It also follows from (3.4) that the $p'_{\delta_1\delta_2\cdots\delta_m}$ sum to unity, but a difference between Ω_1 and Ω_2 arises because the $p'_{\delta_1\delta_2\cdots\delta_m}$ need not all be non-negative in Ω_1. A simple example of the latter when $m = 3$ is pictured in case (b) of Figure 1. For this example, $P_*\{s_1\} = P_*\{s_2\} = P_*\{s_3\} = 0$ while $P_*\{s_1, s_2\} = P_*\{s_1, s_3\} = P_*\{s_2, s_3\} = \frac{1}{2}$ and (3.4) yields $p'_{100} = p'_{010} = p'_{001} = 0$, $p'_{110} = p'_{101} = p'_{011} = \frac{1}{2}$ and $p'_{111} = -\frac{1}{2}$. Thus there are closed convex subsets in Ω_1 which are not in Ω_2, where Ω_2 is the class of primary interest in this paper.

Many of the basic relationships of ordinary probability theory have analogues for systems of upper and lower probabilities. For example, in Ω one has

$$P_*(\emptyset) = P^*(\emptyset) = 0, \qquad P_*(S) = P^*(S) = 1; \tag{3.5}$$

if the complement of T is denoted by $\bar{T}$, then

$$P_*(T) + P^*(\bar{T}) = 1; \tag{3.6}$$

if T and R are mutually exclusive, then

$$\begin{aligned} P_*(T) + P_*(R) &\leqq P_*(T \cup R) \leqq P_*(T) + P^*(R) \\ &\leqq P^*(T \cup R) \leqq P^*(T) + P^*(R); \end{aligned} \tag{3.7}$$

if $E_*(V) = \inf_{P \varepsilon \mathcal{C}} E(V)$ and $E^*(V) = \sup_{P \varepsilon \mathcal{C}} E(V)$ are used to define upper and lower expectations, then

$$E_*(V) = -E^*(-V) \tag{3.8}$$

or more generally

$$\begin{aligned} E_*(a + bV) &= a + bE_*(V) \quad \text{if} \quad b \geqq 0 \\ &= a + bE^*(V) \quad \text{if} \quad b \leqq 0 \end{aligned} \tag{3.9}$$

together with a similar formula for $E^*(a + bV)$; for any pair of variates V and W,

$$\begin{aligned} E_*(V) + E_*(W) &\leqq E_*(V + W) \leqq E_*(V) + E^*(W) \\ &\leqq E^*(V + W) \leqq E^*(V) + E^*(W). \end{aligned} \tag{3.10}$$

Note that (3.10) and (3.8) generalize (3.7) and (3.6), respectively. To prove (3.10), for example, note that there exists a measure in $\mathcal{C}$ such that $E_*(V + W) = E(V + W) = E(V) + E(W) \geqq E_*(V) + E_*(W)$. The remaining parts of (3.10) follow from the first part together with (3.8).

It is interesting to note that the definitions (1.12) and (2.8) do not coincide in Ω as they do in Ω_2, i.e., for general convex sets it can happen that

$$\int_{-\infty}^{\infty} v\, dF^*(v) < \inf_{P \varepsilon \mathcal{C}} E(V). \tag{3.11}$$

This comes about because there is in general no measure P which simultaneously minimizes each of the terms in (2.10).

Another relation which holds in Ω_2 but not in general in Ω is

$$\begin{aligned} P_*(T) + P_*(R) &\leqq P_*(T \cup R) + P_*(T \cap R) \leqq P_*(T) + P^*(R) \\ &\leqq P^*(T \cup R) + P^*(T \cap R) \leqq P^*(T) + P^*(R) \end{aligned} \tag{3.12}$$

for any $T, R \subset S$. Simple counterexamples may be found in Ω even for $m = 3$.

To prove (3.12) in Ω_2 define $T_1 = T \cap R$, $T_2 = T - T_1$, $T_3 = R - T_1$, and $T_4 = S - (T_1 \cup T_2 \cup T_3)$. Then, analogous to (3.4) define

$$\begin{aligned} t_{1000} &= P_*(T_1), \quad \text{etc.}, \\ t_{1100} &= P_*(T_1 \cup T_2) - P_*(T_1) - P_*(T_2), \quad \text{etc.}, \\ t_{1110} &= P_*(T_1 \cup T_2 \cup T_3) - P_*(T_1 \cup T_2) - P_*(T_1 \cup T_3) - P_*(T_2 \cup T_3) \\ &\quad + P_*(T_1) + P_*(T_2) + P_*(T_3), \quad \text{etc.}, \\ t_{1111} &= 1 - P_*(T_1 \cup T_2 \cup T_3) - \cdots + P_*(T_1 \cup T_2) + \cdots - P_*(T_1) \\ &\quad - \cdots - P_*(T_4). \end{aligned} \tag{3.13}$$

By a simple argument of inclusion and exclusion, these $2^4 - 1$ quantities are nonnegative and sum to unity. Like (3.4), the relations (3.13) may be solved to yield lower probabilities and thence upper probabilities in terms of $t_{\delta_1\delta_2\delta_3\delta_4}$ for every event determined by T_1, T_2, T_3, and T_4 or equivalently by T and R. The relations (3.12) follow simply by replacing each quantity with its expression in terms of the $t_{\delta_1\delta_2\delta_3\delta_4}$ and using the fact that each $t_{\delta_1\delta_2\delta_3\delta_4} \geqq 0$.

I do not know whether (3.11) can happen or whether (3.12) can fail in Ω_1.

The literature on upper and lower probabilities is to my knowledge quite small. Good (1962) has presented an axiomatic approach which he believes simplifies but does not necessarily agree with an earlier axiomatic approach of Koopman (1940a), (1940b). I have not attempted to produce a compact set of probability axioms sufficient to characterize Ω, Ω_1 or Ω_2, as the case may be. Nor does Good appear to discuss models with the mathematical concreteness of the families Ω, Ω_1 or Ω_2. Smith (1961), (1965) has also discussed upper and lower probabilities, largely in the context of upper and lower betting odds. Since upper and lower odds for any bet are equivalent to a pair of planes of support for a convex set $\mathcal{C}$ of measures P, it appears that Smith is considering the family Ω. Fishburn (1964) considers upper and lower probabilities and their corresponding expectations apparently in the framework Ω_1. There appears to be no hint of the family Ω_2 in any of the work referred to.

From the viewpoint of a reader to whom probabilities are essentially determinants of bets or rational decisions, it may seem undesirable to restrict the class of convex subsets $\mathcal{C}$ to Ω_1 or even less to Ω_2, since any member of Ω would seem to be a defensible position for a rational consistent man. On the other hand, when upper and lower probabilities can be traced back to a single measure μ, a more stringent kind of logic can be introduced in the area of conditioning. This concept of conditioning and its generalization to the concept of combining independent sources of information are the crux of this paper and, I believe, the most attractive feature of restriction to Ω_2.

4. Upper and lower conditional probabilities. Given a system of upper and lower probabilities defined over subsets $T \subset S$ by $(X, \mathfrak{F}, \mu)$ and Γ, what are the appropriate upper and lower conditional probabilities of T given R, i.e., probabilities appropriate when $S - R$ is ruled impossible? The obvious answer is to use the same $(X, \mathfrak{F}, \mu)$ and Γ except restricting Γ to subsets of R, or more precisely, using the multivalued mapping Γ' from X to R defined by

$$\Gamma' x = \Gamma x \cap R. \tag{4.1}$$

The upper and lower conditional probabilities defined by Γ' may be expressed simply in terms of the unconditional upper and lower probabilities defined by Γ, i.e.,

$$\begin{aligned} P^*(T \mid R) &= P^*(T \cap R)/P^*(R); \\ P_*(T \mid R) &= 1 - P^*(\bar{T} \mid R) = 1 - P^*(\bar{T} \cap R)/P^*(R). \end{aligned} \tag{4.2}$$

The first line of (4.2) is an application of (1.3) with Γ' in place of Γ, and the second line of (4.2) follows from (3.6) and the first line of (4.2). Note that upper and lower conditional probabilities given R are undefined unless $P^*(R) > 0$, i.e., unless the range of Γ' includes more than $\emptyset$.

The following lemma is a consequence of the above definitions.

LEMMA. *If T_1 and T_2 are mutually exclusive subsets of R, then*

$$\begin{aligned} P_*(T_1)/P^*(T_2) &\leqq P_*(T_1 \mid R)/P^*(T_2 \mid R) \leqq P^*(T_1 \mid R)/P_*(T_2 \mid R) \\ &\leqq P^*(T_1)/P_*(T_2). \end{aligned} \tag{4.3}$$

Only the first inequality need be proved, since the second is obvious and the third follows from the first. To prove the first write

$$\begin{aligned} P_*(T_1 \mid R)/P^*(T_2 \mid R) &= (1 - P^*(R - T_1 \mid R))/P^*(T_2 \mid R) \\ &= (P^*(R) - P^*(R - T_1))/P^*(T_2) \\ &\geqq P_*(T_1)/P^*(T_2) \end{aligned} \tag{4.4}$$

where the inequality between the last two numerators follows from (3.7). Relations (4.3) assert that the elimination of possibilities extraneous to a given bet serves to tighten the upper and lower betting odds appropriate to that bet. Note that these upper and lower betting odds do not in general come together, even when $R = T_1 \cup T_2$.

The definition of upper and lower conditional probabilities given above relies for its motivation on the structure of Ω_2. In Ω or Ω_1 one could use (4.2) but it would no longer appear natural; instead, one might regard

$$P^{**}(T \mid R) = \sup_{P \varepsilon \mathfrak{C}} P(T \mid R), \qquad P_{**}(T \mid R) = \inf_{P \varepsilon \mathfrak{C}} P(T \mid R) \tag{4.5}$$

as the natural definitions of upper and lower conditional probabilities. The relationship between (4.2) and (4.5) as alternatives in Ω_2 may be clarified as follows:

Define $T_1 = T \cap R$, $T_2 = R - T$, and $T_3 = S - R$. Analogous to (3.13), define

$$t_{100} = P_*(T_1), \quad \text{etc.,}$$

$$(4.6)\quad t_{110} = P_*(T_1 \cup T_2) - P_*(T_1) - P_*(T_2), \quad \text{etc.},$$
$$t_{111} = 1 - P_*(T_1 \cup T_2) - P_*(T_1 \cup T_3) - P_*(T_2 \cup T_3) + P_*(T_1) + P_*(T_2) + P_*(T_3),$$

which in Ω_2 are seven non-negative quantities summing to unity. From (4.2) and (4.6) it follows that

$$(4.7)\quad \begin{aligned} P^*(T \mid R) &= (t_{100} + t_{110} + t_{101} + t_{111})/(t_{100} + t_{010} + t_{110} + t_{101} + t_{011} + t_{111}), \\ P_*(T \mid R) &= t_{100}/(t_{100} + t_{010} + t_{110} + t_{101} + t_{011} + t_{111}). \end{aligned}$$

On the other hand, the maximum and minimum of $P(T \mid R) = P(T_1)/(P(T_1) + P(T_2))$ are found by distributing the pieces (4.6) appropriately among T_1, T_2 and T_3 where t_{100} must go to T_1, while t_{110} may go to T_1 or T_2, and so on. Thus

$$(4.8)\quad \begin{aligned} P^{**}(T \mid R) &= (t_{100} + t_{110} + t_{101} + t_{111})/(t_{100} + t_{110} + t_{101} + t_{111} + t_{010}), \\ P_{**}(T \mid R) &= t_{100}/(t_{100} + t_{010} + t_{110} + t_{011} + t_{111}). \end{aligned}$$

From (4.7) and (4.8)

$$(4.9)\quad P^{**}(T \mid R) \geqq P^*(T \mid R) \geqq P_*(T \mid R) \geqq P_{**}(T \mid R).$$

Thus the additional structure used in the definitions (4.2) serves to pull the upper and lower probabilities inward relative to the less structured definitions (4.5).

In Section 5 the definitions (4.2) will be seen as a very special case of a method of assimilating new information into a system of upper and lower probabilities

5. Combination of independent sources of information. A probability measure may be regarded as defining degrees of belief which quantify a state of partial knowledge. Any such measure arises in some way from a limited range of human experience which will be called a source of information. A mechanism for combining such sources of information is a virtual necessity for a theory of probability oriented to statistical inference. The mechanism adopted here assumes *independence* of the sources, a concept whose real world meaning is not so easily described as its mathematical definition. Opinions of different people based on overlapping experiences could not be regarded as independent sources. Different measurements by different observers on different equipment would often be regarded as independent, but so would different measurements by one observer on one piece of equipment: here the question concerns independence of errors. In the application referred to in Section 6, the independent sources are taken to be non-overlapping random samples from a population, together with prior information which may be regarded as a distillation of previous samples or experiences.

The sources considered here are mathematically defined by their basic probability spaces $(X_i, \mathfrak{F}_i, \mu_i)$ and multivalued mappings Γ_i, where i indexes the source. The space S into which Γ_i maps is the same for each i, i.e., the different sources are giving information about the same uncertain outcome in S. If the n sources $i = 1, 2, \cdots, n$ are assumed independent, then the combined source $(X, \mathfrak{F}, \mu)$ and Γ is defined from

$$\begin{aligned} X &= X_1 \times X_2 \times \cdots \times X_n , \\ \mathfrak{F} &= \mathfrak{F}_1 \times \mathfrak{F}_2 \times \cdots \times \mathfrak{F}_n , \\ \mu &= \mu_1 \times \mu_2 \times \cdots \times \mu_n , \\ \Gamma x &= \Gamma_1 x \cap \Gamma_2 x \cap \cdots \cap \Gamma_n x \end{aligned} \tag{5.1}$$

for all $x \,\varepsilon\, X$. The product measure space $(X, \mathfrak{F}, \mu)$ is motivated by the usual definition of statistical independence. The definition of Γ reflects the idea that $x_i \,\varepsilon\, X_i$ is consistent with a particular $s \,\varepsilon\, S$ if and only if $s \,\varepsilon\, \Gamma_i x_i$, for $i = 1, 2, \cdots, n$, and consequently $x = (x_1, x_2, \cdots, x_n) \,\varepsilon\, X$ is consistent with that s if and only if s belongs to all of the $\Gamma_i x_i$ simultaneously.

It is a characteristic of the above combination rule that neither upper probabilities, nor lower probabilities nor probabilities of the type $p_{\delta_1\delta_2\cdots\delta_m}$ have a simple product rule of combination. A set of probabilities $q_i(T)$ which do obey a simple product rule is defined as follows: For the systems $(X, \mathfrak{F}, \mu)$ and Γ defined by (5.1) from the systems $(X_i, \mathfrak{F}_i, \mu_i)$ and Γ_i, and for any $T \subset S$, set

$$\tilde{T} = \{x \,\varepsilon\, X, \Gamma x \supset T\} \quad \text{and} \quad \tilde{T}_i = \{x_i \,\varepsilon\, X_i , \Gamma_i x_i \supset T\} \tag{5.2}$$

and set

$$q(T) = \mu(\tilde{T}) \quad \text{and} \quad q_i(T) = \mu_i(\tilde{T}_i). \tag{5.3}$$

It follows immediately that

$$\tilde{T} = \tilde{T}_1 \times \tilde{T}_2 \times \cdots \times \tilde{T}_n \tag{5.4}$$

and hence that

$$q(T) = q_1(T) \times q_2(T) \times \cdots \times q_n(T). \tag{5.5}$$

It will be seen shortly that, at least for finite S, the probabilities $q(T)$ are sufficient to determine all upper and lower probabilities defined by a system $(X, \mathfrak{F}, \mu)$ and Γ and hence from (5.5) they provide a convenient form, ready for further combination, for storing the information in a given source. Note also that, if T consists of a single element $s \,\varepsilon\, S$, then $\tilde{T} = T^*$ so that $q(T) = \mu(T^*)$ whence from (1.3) the $q\{s\}$ as s ranges over S are proportional to $P^*\{s\}$.

The foregoing ideas will now be concretely illustrated using a finite S, beginning with $m = 3$. A source is characterized here by p_{000}, p_{100}, p_{010}, p_{001}, p_{110}, p_{101}, p_{011} and p_{111}. If the $q(T)$ corresponding to the $T = \emptyset, \{s_1\}, \{s_2\}, \{s_3\}, \{s_1, s_2\}, \cdots, \{s_1, s_2, s_3\}$ are denoted by $q_{000}, q_{100}, q_{010}, q_{001}, q_{110}, \cdots, q_{111}$, it follows directly that

$$q_{000} = 1 = p_{100} + p_{010} + p_{001} + p_{110} + p_{101} + p_{011} + p_{111} ,$$

$$(5.6)\quad \begin{aligned} q_{100} &= p_{100} + p_{110} + p_{101} + p_{111}, && \text{and similarly for } q_{010} \text{ and } q_{001},\\ q_{110} &= p_{110} + p_{111}, && \text{and similarly for } q_{101} \text{ and } q_{011}, \text{ and}\\ q_{111} &= p_{111}. \end{aligned}$$

The relations (5.6) may be solved to yield

$$(5.7)\quad \begin{aligned} p_{000} &= 1 - q_{100} - q_{010} - q_{001} + q_{110} + q_{101} + q_{011} - q_{111},\\ p_{100} &= q_{100} - q_{110} - q_{101} + q_{111}, && \text{and similarly for } p_{010} \text{ and } p_{001},\\ p_{110} &= q_{110} - q_{111}, && \text{and similarly for } p_{101} \text{ and } p_{011}, \text{ and}\\ p_{111} &= q_{111}, \end{aligned}$$

thus showing that the set of a $q(T)$ determine the $p_{\delta_1\delta_2\delta_3}$. Note also that the extensions of (5.6) and (5.7) from $m = 3$ to general m are evident and easily proved.

A pair of sources $i = 1, 2$ may be characterized by their $p^{[i]}_{\delta_1\delta_2\delta_3}$ or by their $q^{[i]}_{\delta_1\delta_2\delta_3}$. The relations $q_{\delta_1\delta_2\delta_3} = q^{[1]}_{\delta_1\delta_2\delta_3}q^{[2]}_{\delta_1\delta_2\delta_3}$ from (5.5) together with the relations (5.6) and (5.7) applied to the two sources and their combination yield

$$(5.8)\quad \begin{aligned} p_{000} &= p^{[1]}_{100}p^{[2]}_{010} + p^{[1]}_{100}p^{[2]}_{001} + p^{[1]}_{100}p^{[2]}_{011} + p^{[1]}_{010}p^{[2]}_{100} + p^{[1]}_{010}p^{[2]}_{001} + p^{[1]}_{010}p^{[2]}_{101}\\ &\quad + p^{[1]}_{001}p^{[2]}_{100} + p^{[1]}_{001}p^{[2]}_{010} + p^{[1]}_{001}p^{[2]}_{110} + p^{[1]}_{110}p^{[2]}_{100} + p^{[1]}_{101}p^{[2]}_{010} + p^{[1]}_{011}p^{[2]}_{100},\\ p_{100} &= p^{[1]}_{100}p^{[2]}_{100} + p^{[1]}_{100}p^{[2]}_{110} + p^{[1]}_{100}p^{[2]}_{101} + p^{[1]}_{100}p^{[2]}_{111}\\ &\quad + p^{[1]}_{110}p^{[2]}_{100} + p^{[1]}_{110}p^{[2]}_{101} + p^{[1]}_{101}p^{[2]}_{100} + p^{[1]}_{101}p^{[2]}_{110} + p^{[1]}_{111}p^{[2]}_{100},\\ &\qquad\qquad \text{and similarly for } p_{010} \text{ and } p_{001},\\ p_{110} &= p^{[1]}_{110}p^{[2]}_{110} + p^{[1]}_{110}p^{[2]}_{111} + p^{[1]}_{111}p^{[2]}_{110},\\ &\qquad\qquad \text{and similarly for } p_{101} \text{ and } p_{011}, \text{ and}\\ p_{111} &= p^{[1]}_{111}p^{[2]}_{111}. \end{aligned}$$

The general rule here, extended to any m, is that

$$(5.9)\quad p_{\delta_1\delta_2\cdots\delta_m} = \sum p^{[1]}_{\delta_1'\delta_2'\cdots\delta_m'}p^{[2]}_{\delta_1''\delta_2''\cdots\delta_m''}$$

with summation over all $(\delta_1', \delta_2', \cdots, \delta_m', \delta_1'', \delta_2'', \cdots, \delta_m'')$ such that $\delta_i = \delta_i'\delta_i''$ for $i = 1, 2, \cdots, m$. It is clear, however, that combining sources directly in terms of $p_{\delta_1\delta_2\cdots\delta_m}$ is awkward, and by referring from (5.8) back to Table 1 one sees that the situation is no better in terms of upper and lower probabilities.

This section concludes with two important properties of the combination rule. To introduce the first of these, note that, if a source $(X_1, \mathfrak{F}_1, \mu_1)$ and Γ_1 is combined with an informationless source, then the result is again the original source $(X_1, \mathfrak{F}_1, \mu_1)$ and Γ_1. By an informationless source is meant an $(X_2, \mathfrak{F}_2, \mu_2)$ and Γ_2 such that $\Gamma_2 x_2 = S$ for all $x_2 \in X_2$, i.e., a source for which $P^*(T) = 1$ and $P_*(T) = 0$ for every T other than $\varnothing$ and S. The more general version of the first property asserts that, if a source $(X_1, \mathfrak{F}_1, \mu_1)$ and Γ_1 is combined with a source $(X_2, \mathfrak{F}_2, \mu_2)$ and Γ_2 where $\Gamma_2 x_2 = R \subset S$ for all $x_2 \varepsilon X_2$, then

$$(5.10) \qquad P^*(T) = P_1^*(T \mid R), \qquad P_*(T) = P_{1*}(T \mid R)$$

for $T \subset S$, where $P_1^*(T \mid R)$ and $P_{1*}(T \mid R)$ are upper and lower conditional probabilities for the system $(X_1, \mathfrak{F}_1, \mu_1)$ and Γ_1 according to the definitions (4.2). In other words, the rule of this section is sufficiently general to include the definition of conditioning as a special case. The relations (5.10) are immediate consequences of the definitions adopted.

The second property concerns sharp sources. A source will be called *sharp* if it is sharp with respect to T for all $T \varepsilon \mathcal{E}$, and will be called *sharp with respect to* T *if* $P^*(T) = P_*(T)$. Thus a sharp source is an ordinary probability measure over the events $T \varepsilon \mathcal{E}$. Assuming finite S, it will be shown that *a source which is sharp with respect to a given* T *remains sharp with respect to* T *after combination with any other source.* A similar property therefore holds for sharpness with respect to all T. Thus, if sharpness is once achieved by a user of this theory, it remains a characteristic of all subsequent states of knowledge of the user.

The demonstration depends on a simple lemma:

LEMMA. *A source is sharp with respect to* $T \varepsilon \mathcal{E}$ *if and only if* $q(R) = 0$ *for every* R *such that* $R \cap T \neq \emptyset$ *and* $R \cap \bar{T} \neq \emptyset$. Clearly, $P^*(T) - P_*(T) \geqq q(R)$ for any R such that $R \cap T \neq \emptyset$ and $R \cap \bar{T} \neq \emptyset$, so that $P^*(T) = P_*(T)$ implies $q(R) = 0$ for all such R. Conversely, if $q(R) = 0$ for all such R, then no Γx which intersects both T and $\bar{T}$ may have positive probability, which implies that $P^*(T) = P_*(T)$. In view of the simple combination property of the $q(R)$ function, the sharpness property of the preceding paragraph follows immediately from the above lemma.

A new kind of limit theorem becomes possible in discussions of upper and lower probability, namely results about convergence to sharpness. For example, in view of (5.1) one would expect the combination of n sources to be sharper than a typical member of the sources combined. Thus rates of convergence to sharpness deserve definition and study. An illustration of this may be found in equation (4.18) of Dempster (1966).

6. An application. The theory proposed in this paper has been implicitly applied to statistical inference in an earlier paper (Dempster (1966)). The nature of the application will be sketched briefly. Individual sample observations may be regarded as sources whose information may be combined according to the rule of Section 5. In such an individual source, the role of X is played by a space representing the possible sample individuals, and the role of S is represented by a parameter space or more generally by the product of a parameter space and a space of future observations. Before a particular sample observation is recorded, the source defined by that sample individual is informationless, but after conditioning by the sample observation one generally gets a non-trivial system of upper and lower probabilities referring to the parameters or to the parameters and future observations jointly.

The combination of many sample individuals appears to lead to sharp inferences which agree with standard asymptotic inferences given either by Bayesian methods or by confidence methods. It is also suggested as valid to treat a prior distribution as a source of information independent of sample data. If source 1 is taken to be combined sample data and source 2 is taken to be the prior information, and if this prior information is sharp and has a density, then (5.13) applies and in fact reduces to the familiar formula for a Bayes posterior distribution. The reason for this is that, in the particular models defined for the inferential situation, $\mu_1^*\{s\}$ turns out to be the familiar likelihood function.

Further concrete examples of these applications to inference will be forthcoming soon.

7. Acknowledgment. Many helpful comments were provided by L. J. Savage who read an earlier version of this paper.

REFERENCES

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

FISHBURN, PETER C. (1964). *Decision and Value Theory*. Wiley, New York.

GOOD, I. J. (1962). The measure of a non-measurable set. *Logic, Methodology and Philosophy of Science* (edited by Ernest Nagel, Patrick Suppes, and Alfred Tarski). Stanford Univ. Press. 319–329.

KOOPMAN, B. O. (1940a). The axioms and algebra of intuitive probability. *Ann. Math.* **41** 269–292.

KOOPMAN, B. O. (1940b). The bases of probability. *Bull. Amer. Math. Soc.* **46** 763–774.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision, (with discussion). *J. Roy. Statist. Soc. Ser. B* **23** 1–25.

SMITH, C. A. B. (1965). Personal probability and statistical analysis, (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 469–499.