SOME ALTERNATIVES TO BAYES'S RULE

Persi Diaconis and Sandy Zabell

1. INTRODUCTION

There are several different approaches to what might be called "the mathematics of changing one's mind." The most frequently discussed method, Bayes's rule, changes a prior or initial probability P to a posterior or final probability P*, based on the occurrence of an event E. It specifies that for any event A:

Bayes's rule $P^*(A) = P(A \text{ and } E)/P(E).$

Bayes's rule is *not* (at least directly) applicable if

- New information does not arrive in the form "event E occurred" (e.g., the murderer was a woman), but instead in the form "the odds on E have changed" (e.g., the murderer was likely to have been a woman). This is sometimes called the problem of *probable knowledge.*
- Even if "E occurs," we may not have thought about E beforehand. Thus we will not have previously assessed either P(A and E) or P(E)

and will therefore be unable to make direct use of Bayes's rule. We will call this the problem of *unanticipated knowledge.*

- Our (subjective) probabilities can change in the light of calculations or of pure thought without any change in the *empirical* data..." (Good, 1977, p. 140). I. J. Good terms such probabilities "evolving" or "dynamic" and has discussed them in a number of papers (e.g., 1950, p. 49; 1968; 1977). As noted by Savage (1967, p. 308) and others, such changes in belief are particularly difficult to model when they pertain to facts which are a logical or mathematical consequence of information already available (e.g., the 1,000th digit of π, or the evaluation of a chess position). We will call this the problem of *introspective knowledge.*

This review focuses on two proposed alternatives to Bayes's rule for revising probability assessments in the face of new information: Richard Jeffrey's rule of conditioning and Arthur Dempster's rule of combination. Section 2 describes Jeffrey's rule. Section 3 describes upper and lower probabilities and Dempster's rule for their combination. Section 5 shows that the two rules are in fact closely connected: Jeffrey's rule is the additive version of Dempster's rule in those situations where the two rules are comparable.

Our presentation is intended as an introduction to a growing and already sizeable literature on the limitations of Bayes's rule. For further information on Jeffrey's rule see Diaconis and Zabell (1982); for further references on upper and lower probability and Dempster's rule see Shafer (1976, 1982).

2. JEFFREY'S RULE OF CONDITIONING

While the mathematics of Bayes's rule presupposes some given event E, Jeffrey's rule assumes the existence of a partition $\{E_1, E_2, \ldots, E_n\}$ on which new probabilities $P^*(E_i)$ are given (the elements of a partition are, by definition, mutually exclusive and exhaustive). It specifies that for any event A:

Jeffrey's rule $$P^*(A) = \sum_{i=1}^{n} P(A|E_i)P^*(E_i).$$

Jeffrey's rule is mathematically equivalent to the judgment that the "J-condition"

(J) $$P^*(A|E_i) = P(A|E_i)$$

holds for all A and i. The J-condition can be interpreted as stating that the only impact of the new evidence was to change the probabilities on the elements of the partition; given an element of the partition, the new and old probabilities agree.

EXAMPLE 1 (*Bayes's Rule*). If (1) the partition consists of a set E and its complement E^c, and (2) if $P^*(E^c) = 0$, then Jeffrey's rule reduces to Bayes's rule $P^*(A) = P(A|E)$.

EXAMPLE 2 (*Uncertain Perception*). Suppose we are about to hear one of two recordings of Shakespeare on the radio, to be read by either Olivier (E) or Gielgud (E^c), but we are uncertain as to which, and we have a prior with mass $\frac{1}{2}$ on Olivier and $\frac{1}{2}$ on Gielgud. After hearing the recording, one might judge it fairly likely, but by no means certain, to be by Olivier. The change in belief takes place by direct recognition of the voice. If the *only* impact of hearing the recording is to change the odds on Olivier and Gielgud, in the sense that for any A, $P^*(A|E) = P(A|E)$ and $P^*(A|E^c) = P(A|E^c)$, then after assessing $P^*(E)$ we may proceed to apply Jeffrey's rule. (Of course, the former might well *not* be the case; for example the quality of the recording might convey additional information as to its date or manufacture.)

Jeffrey has argued that examples of this type are the norm: "it is rarely or never that there is a proposition for which the direct effect of an observation is to change the observer's degree of belief in a proposition to 1" (1968, p. 171).

EXAMPLE 3 (*Unanticipated Knowledge*). Suppose we are thinking about three possible trials of a new surgical procedure. Under the usual circumstances a probability assignment P is made on the eight possible outcomes $\Omega = \{000, 001, 010, 100, 011, 101, 110, 111\}$, where 1 denotes a successful outcome, 0 not. Suppose a colleague informs us that another hospital had performed this type of operation 100 times, with 80 successful outcomes. This is clearly relevant information, and we will obviously want to revise our opinion. The information cannot be put in terms of the occurrence of an event in the original eight-point space Ω, and Bayes's rule is not *directly* available.

Diaconis and Zabell (1982) discuss four possible approaches to the problem of forming P^*: complete reassessment, retrospective conditioning, the use of exchangeability, and Jeffrey's rule. We review here the use of Jeffrey's rule, as an example illustrating how natural partitions $\{E_i\}$ can arise.

Suppose that the original probability P was *exchangeable*; that is, $P(001) = P(010) = P(100)$, $P(110) = P(101) = P(011)$. In the situation described, the colleague's report says nothing about the order of the trials, and we may thus require the new P* to remain exchangeable. Consider the partition $\{E_0, E_1, E_2, E_3\}$, where E_i is the set of outcomes with i ones: $E_0 = \{000\}$, $E_1 = \{001, 010, 100\}$, $E_2 = \{110, 101, 011\}$, $E_3 = \{111\}$. The exchangeability of both P and P* is equivalent to Jeffrey's condition:

$$P(A|E_i) = P^*(A|E_i),$$

and so, to complete the assignment of P*, we need only undertake an assessment of $P^*(E_i)$. Then P* is determined by Jeffrey's rule: for any set A,

$$P^*(A) = \sum_{i=1}^{n} P(A|E_i)P^*(E_i).$$

EXAMPLE 4 (*Coin Tossing and Exchangeability*). It is instructive to consider the problem discussed in the preceding example when the number of trials can be very large (as in tossing a coin). In this case it is often assumed for simplicity that the sequence observed is the initial segment of an infinite sequence of 0–1 outcomes $X_1, X_2, X_3, \ldots$. Suppose that the X_i are exchangeable (so that all sequences of length n with the same number of heads are equiprobable), and let $S_n = X_1 + X_2 + \cdots + X_n$. Then, as first shown by de Finetti, the limiting relative frequency

$$Z = \lim_{n\to\infty} S_n/n$$

exists almost surely and, conditional on Z, S_n is binomially distributed with

$$P(S_n = k|Z = p) = \binom{n}{k} p^k(1-p)^{n-k}. \qquad (1)$$

It immediately follows that

$$P(S_n = k) = \int_0^1 p^k(1-p)^{n-k}\, d\mu(p),$$

where $d\mu(p) = P\{Z \in dp\}$. Thus, quantifying a subjective probability distribution for a sequence of coin tosses reduces to choosing a "prior" distribution $d\mu(p)$ for the limiting relative frequency of heads.

In many textbooks on Bayesian statistics the prior $d\mu$ is assumed to be unimodal, centered about $\frac{1}{2}$. The use of a unimodal prior is a reasonable assumption if the coin is being *tossed.* If, however, a coin is *spun* on its edge, it will typically exhibit a bias, often considerable, with the observed proportion of heads sometimes as great as $\frac{2}{3}$ or as small as $\frac{1}{3}$ (see Diaconis and Ylvisaker, 1983, Sec. 1). If the coin in question is being spun but, initially ignorant of the above, we chose our prior to be unimodal, we could take account of the new information by choosing a new bimodal prior $d\mu^*$, and then apply Jeffrey's rule to get

$$P^*(S_n = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k}\, d\mu^*(p).$$

This illustrates the use of Jeffrey conditioning via a continuous "sufficient statistic" rather than a discrete "sufficient partition"; since both P and P* are exchangeable, both satisfy (1), a continuous version of the J-condition.

EXAMPLE 5 (*Updating on an Event of Probability Zero*). Consider the following experiment: A tack is placed on the floor and given an energetic flick with the fingers. When it comes to rest, a 0 is recorded if the point of the tack is touching the floor, a 1 if not. This example has been discussed by Mosteller et al. (1970, pp. 113–115), Lindley and Phillips (1976), and Diaconis and Freedman (1980a). It is usually assumed that the sequence of zeros and ones thus obtained is exchangeable (or equivalently, independent with unknown probability). Suppose we initially assume exchangeability and quantify our opinion by specifying a measure μ on $[0,1]$ as in Example 4. Now consider the following reflections: If each successive flick of the tack is made from the position in which the tack just landed (as opposed to resetting the tack in a standard position), the outcome of a trial may well depend on the outcome of the previous trial. This thought led Diaconis and Freedman (1980a) to analyze the experiment, using partial exchangeability, as a mixture of Markov chains. Let $\mathbf{e} = (e_0, \ldots, e_n)$ be a sequence of zeros and ones. Let $t_{ij} = t_{ij}(\mathbf{e})$ denote the number of i to j transitions for i, j equal to 0 and 1. If the tack is initially point upward, so $e_0 = 1$, then

$$P^*(e_0, e_1, \ldots, e_n) = \iint p_{00}^{t_{00}} p_{01}^{t_{01}} p_{10}^{t_{10}} p_{11}^{t_{11}}\, d\mu^*(p_{00}, p_{11}).$$

Here p_{ij} can be thought of as the limiting relative frequency of i to j transitions (known to exist for recurrent partially exchangeable processes; see Diaconis and Freedman, 1980b).

Since the original exchangeable probability assigns probability zero to the Markovian aspects of the process, this is an example where the new and old probabilities can be mutually singular. We now give a reasonable scenario which shows that much of the work that went into quantifying the measure μ can be harnessed in quantifying the new probability.

It might be that the main reason for rejecting exchangeability was the thought that on a smooth floor the tack might just slide along without ever turning over. Then the outcome of the next trial would be identical with the last trial. If the tack didn't slide but tumbled irregularly, then the starting position might be judged as irrelevant. The measure μ, determined while considering the process as exchangeable, might be judged a reasonable quantification of the limiting relative frequency of ones in the nonsliding trials. One would also have to quantify the belief α, β in the possibility that the tack slides without turning from both starting positions. Then μ^* is a mixture of three measures on the unit square:

$$\mu^* = \alpha\mu_{00} + \beta\mu_{11} + (1 - \alpha - \beta)\bar{\mu},$$

where μ_{00} and μ_{11} are point masses at $(0, 0)$ and $(1, 1)$, while $\bar{\mu}$ is the original prior μ thought of as a measure on the line $p_{00} = 1 - p_{11}$.

Both exchangeability and Markov exchangeability are examples of the phenomenon of *partial exchangeability*; see Diaconis and Freedman (1980a). Whenever we consider sequences of outcomes initially thought partially exchangeable and receive new information which does not change this judgment, Jeffrey's rule is in effect being invoked.

Diaconis and Zabell (1982) link the J-condition with the statistical concept of sufficiency and show that Jeffrey's rule gives the closest probability P^* to P with prescribed values $P^*(E_i)$. Our 1982 paper also gives continuous versions of the rule, and an analysis of Jeffrey's rule when two or more sources of evidence are considered simultaneously.

3. UPPER AND LOWER PROBABILITIES, AND DEMPSTER'S RULE OF COMBINATION

We begin our discussion with an example drawn from Diaconis (1978). It concerns the well-known problem of the three prisoners:

> Of three prisoners, *a*, *b*, and *c*, two are to be executed but *a* does not know which. He therefore says to the jailer, "Since either *b* or *c* is certainly going to be executed, you will give me no information about my own chances if you give me the name of one man, either *b* or *c*, who is going to be executed." Accepting this argument, the jailer truthfully replies, "*b* will be executed." Thereupon *a* feels happier because before the jailer replied *a*'s own chance of execution was two-thirds but afterward there are only two people, himself and *c*, who could be the one not executed, and so his chance of execution is one-half.

Is *a* justified in believing that his chances of escaping execution have improved? Consider the set of possible outcomes

$$S = \{(a, b), (a, c), (b, c), (c, b)\},$$

where, for example, (a, b) means *a* will live and the jailer answers *b*. In the classical Bayesian solution of this problem (see, e.g., Gardner, 1961, Chap. 19), a, b, and c are assumed equally likely to be pardoned and if a is to be set free, it is assumed that the jailer will answer by choosing b or c with probability $\frac{1}{2}$. These assumptions translate into the probability P on S with $\mathrm{P}(a, b) = \mathrm{P}(a, c) = \frac{1}{6}$, $\mathrm{P}(b, c) = \mathrm{P}(c, b) = \frac{1}{3}$, and Bayes's rule gives

$$\mathrm{P}(a \text{ lives} | \text{jailer says } b) = \frac{\mathrm{P}(a, b)}{\mathrm{P}(a, b) + \mathrm{P}(c, b)} = \frac{1}{3},$$

that is, a's chances have *not* improved.

We will discuss three ways to model this problem using upper and lower probabilities P_*, P^*. Upper and lower probabilities are functions P_* and P^* defined on the subsets of a set S satisfying

$$\text{C1} \qquad \mathrm{P}_*(\varnothing) = 0, \qquad \mathrm{P}_*(\mathrm{S}) = 1,$$

$$\text{C2} \qquad \mathrm{P}^*(\mathrm{A}) = 1 - \mathrm{P}_*(\mathrm{A}^c),$$

and the inequalities

$$\text{C3} \qquad \mathrm{P}_*(\mathrm{A}_1 \cup \cdots \cup \mathrm{A}_n) \geq \sum \mathrm{P}_*(\mathrm{A}_i) - \sum_{i<j} \mathrm{P}_*(\mathrm{A}_i \cap \mathrm{A}_j) + \cdots + (-1)^{n+1} \mathrm{P}_*(\mathrm{A}_1 \cap \cdots \cap \mathrm{A}_n).$$

Conditions C1, C2, and C3 will be motivated later on. For the present we note that the definitions imply $\mathrm{P}_* \leq \mathrm{P}^*$, so that the upper-lower pair $(\mathrm{P}_*, \mathrm{P}^*)$ may be thought of as bounds on some "true probability" P, with $\mathrm{P}_* \leq \mathrm{P} \leq \mathrm{P}^*$. A simple example is the *vacuous* upper-lower pair defined by setting

$$\mathrm{P}_*(\mathrm{A}) = 0 \text{ if } \mathrm{A} \subsetneq \mathrm{S}, \qquad \mathrm{P}_*(\mathrm{S}) = 1.$$

The vacuous pair is often suggested as a way of quantifying a state of "no knowledge."

Arthur Dempster has suggested that, given the occurrence of an event E, the appropriate way of modifying an upper-lower pair to a new upper-lower pair incorporating the new information is via

$$\textbf{Dempster's rule} \qquad \mathrm{P}^*(\mathrm{A}|\mathrm{E}) = \mathrm{P}^*(\mathrm{A} \cap \mathrm{E}) / \mathrm{P}^*(\mathrm{E}).$$

A motivation for Dempster's rule will also be given later. First we return to the three-prisoner problem and show how it may be analyzed using different upper-lower pairs and Dempster's rule.

MODEL 1. Suppose that prisoner a models his (lack of) knowledge by putting the vacuous upper-lower pair on the four-point set S. Then the definitions imply $\mathrm{P}^*(a \text{ will live} | \text{jailer says } b) = 1$, $\mathrm{P}_*(a \text{ will live} | \text{jailer says } b) = 0$. Thus, with no assumptions on the problem, the jailer's information does not reduce his uncertainty, and the conditional upper-lower pair remains vacuous.

MODEL 2. Suppose that a assumes that the initial decision as to who will live is made at random but assumes nothing about how the jailer will act except that he will tell the truth. One way to model this is to consider the space $L = \{a, b, c\}$; the probability P on L corresponding to the random choice of who will live, that is, $P(a) = P(b) = P(c) = \frac{1}{3}$; and the multivalued map Γ, from L to the subsets of S, given by

$$\Gamma(a) = (a, b) \cup (a, c), \qquad \Gamma(b) = (b, c), \qquad \Gamma(c) = (c, b).$$

Thus Γ delineates the possible outcomes when a, b, and c are pardoned.

Dempster has described how an upper–lower pair can be constructed on S whenever a set L, probability P on L, and multivalued map $\Gamma: L \rightarrow$ subsets of S are given. Define $C = P\{\ell: \Gamma(\ell) \neq \varnothing\}$, and

$$P^*(A) = P\{\ell \in L: \Gamma(\ell) \cap A \neq \varnothing\}C \quad \text{and}$$

$$P_*(A) = P\{\ell: \Gamma(\ell) \neq \varnothing, \Gamma(\ell) \subset A\}/C.$$

P^* and P_* represent the largest and smallest probabilities that can be assigned to A consistent with Γ and P.

The French mathematician Gustave Choquet (1953) proved the following important result.

THEOREM. Every upper–lower pair constructed in this way from a multivalued map satisfies conditions C1, C2, and C3. Conversely, given an upper–lower pair satisfying C1, C2, and C3, there exists a set L, a probability P on L, and a multivalued map $\Gamma: L \rightarrow$ subsets of S which realizes the upper–lower pair.

This is the promised motivation for C1, C2, and C3. Any function P_* satisfying C1, C2, C3 is said to be a *capacity of infinite order.* The infinite system of inequalities C3 are known as the Block–Marschack inequalities in the psychology of choice; see the article by Batchelder in this volume.

Returning to the three-prisoner example, we have for the upper–lower pair P_*, P^* that arises from L, P, and Γ:

$$P^*(\text{jailer says } b) = P^*\{(a, b) \cup (c, b)\} = P\{a \cup c\} = \tfrac{2}{3};$$

$$\begin{aligned} P_*(\text{jailer says } b) &= 1 - P^*\{\text{jailer does not say } b\} \\ &= 1 - P^*\{(a, c) \cup (b, c)\} = \tfrac{1}{3}. \end{aligned}$$

This result is intuitively reasonable: If the jailer said b when he truthfully could, he would say b two-thirds of the time. If the jailer avoided saying b whenever he truthfully could, he would say b one-third of the time. Dempster's rule of conditioning then gives

$$P^*(a \text{ will live}|\text{jailer says } b) = P_*(a \text{ will live}|\text{jailer says } b) = \tfrac{1}{2}.$$

Thus, with this set of assumptions a is justified in reasoning exactly as described in the original version of the problem. Observe that after Dempster conditioning the two members of the upper-lower pair are actually equal, coalescing to a bona fide probability.

A "lazy Bayesian" could regard the formation of an upper-lower pair based on a multivalued mapping as a way of proceeding without quantifying belief within the elements of $\Gamma(\ell)$. The calculations result in bounds which would be useful in checking a more refined quantification.

Here is Dempster's motivation for his rule of conditioning, via multivalued mappings. Consider a pair of probability spaces and multivalued mappings:

$$(L_1, P_1) \xrightarrow{\Gamma_1} \mathcal{S}, \qquad (L_2, P_2) \xrightarrow{\Gamma_2} \mathcal{S},$$

where $\mathcal{S}$ denotes the subsets of S. Define a product space $(L_1 \times L_2, P_1 \times P_2)$ and $\Gamma_1 \times \Gamma_2: L_1 \times L_2 \to \mathcal{S}$ by

$$\Gamma_1 \times \Gamma_2(\ell_1, \ell_2) = \Gamma_1(\ell_1) \cap \Gamma_2(\ell_2).$$

It is easy to show the following:

D1 If $\Gamma_1(\ell) = S$, then the upper-lower pair associated with Γ_1 is vacuous and the upper-lower pair associated with the product $\Gamma_1 \times \Gamma_2$ is identical to the upper-lower pair associated with Γ_2.

D2 If either of the component upper-lower pairs is a probability, then the product is a probability.

D3 If $\Gamma_1(\ell) = E$, then the product yields Dempster's rule of conditioning.

For further discussion of this motivation for Dempster's rule, see Dempster (1968).

To us, the multivalued mapping approach to upper-lower pairs seems preferable to their direct use and interpretation (as favored by Shafer, 1976).

4. FURTHER ANALYSIS OF THE PRISONER PARADOX

To discuss Model 2 and the example further, consider the general Bayesian solution to the three-prisoner problem: let π_a, π_b, and π_c be the prior probabilities that a, b, and c are pardoned, and let p be the probability that the jailer names b when a is pardoned. Then

$$P(a, b) = \pi_a p, \qquad P(a, c) = \pi_a(1 - p),$$
$$P(b, c) = \pi_b, \qquad P(c, b) = \pi_c,$$

and

$$\mathrm{P}(a \text{ lives}|\text{jailer says } b) = \frac{\pi_a \mathrm{p}}{\pi_a \mathrm{p} + \pi_c}.$$

For Model 2 we have $\pi_a = \pi_b = \pi_c = \frac{1}{3}$, with p remaining a free parameter. The resulting family $\mathscr{P}$ of possible probability measures can be used to define another type of upper–lower probability, say U and L, defined by

$$\mathrm{U(A)} = \max\{\mathrm{P(A)}: \mathrm{P} \in \mathscr{P}\} \quad \text{and} \quad \mathrm{L(A)} = \min\{\mathrm{P(A)}: \mathrm{P} \in \mathscr{P}\}.$$

In this case, it is easy to check that U and L are exactly the same as those derived via the multivalued map L. In general, however, upper–lower pairs defined by sups and infs will *not* be capacities of infinite order, nor even capacities of order 2; see Huber and Strassen (1973) for a simple counterexample.

Note that the *conditional* probabilities generated by $\mathscr{P}$ range from 0 to $\frac{1}{2}$, while Dempster's rule of conditioning picks out the unique value $\frac{1}{2}$. This is a disturbing result for a Bayesian, since it calls into question both the interpretation of and justification for Dempster's rule. *Either* Dempster's rule contains further hidden, implicit assumptions, here responsible for narrowing down the range of possible conditional probabilities to but one, *or* it appears in a manner very different from ordinary, Bayesian conditioning, in which case we would wish some further guidance as to its interpretation and meaning. Mere surface plausibility is insufficient, for it is possible to suggest at least one equally plausible alternative to Dempster's rule, namely

$$\mathrm{P}_*(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}_*(\mathrm{A \text{ and } B})}{\mathrm{P}_*(\mathrm{B})} \quad \text{and} \quad \mathrm{P}^*(\mathrm{A}|\mathrm{B}) = 1 - \mathrm{P}_*(\mathrm{A}^c|\mathrm{B}).$$

This yields a rule of conditioning different from Dempster's; yet the resulting conditional set functions are capacities. In what sense is one of them right? (Note that for this method of conditioning the upper–lower pair for Model 2 of the three-prisoner problem yields upper and lower conditional probabilities of 0!)

MODEL 3. Suppose a knows nothing about the selection process for who will live but assumes (or is told) that if he lives, the jailer will choose randomly between answering b or c. (Of course, if the jailer knows b is to live, he will answer c, and vice versa.) This problem can be modeled by assuming that three different probability measures are given on the set $\mathrm{W} = \{b, c\}$ of the jailer's possible answers: $\mathrm{P}_a(b) = \mathrm{P}_a(c) = \frac{1}{2}$; $\mathrm{P}_b(c) = 1$; $\mathrm{P}_c(b) = 1$. Given the jailer's answer, Chapter 11 of Shafer (1976) proposes a method related to direct use of likelihood for deriving an upper–lower pair on the parameter set $L = \{a, b, c\}$. This yields P_*(a will live|jailer says b) = 0, P^*(a will live|jailer says b) = $\frac{1}{2}$. In this model, before questioning the jailer, a might have expressed his ignorance by P_*(a lives) = 0, P^*(a lives) = 1. After learning b will die, a can no longer be so optimistic.

Again, the comparison with the Bayesian analysis is instructive. Now π_a, π_b, π_c are arbitrary and $p = \frac{1}{2}$, so that the resulting conditional probabilities for P(a will live|jailer says b) range *from* 0 *to* 1. Thus while Shafer's method does not suffer in this case from the defect of picking out a unique conditional probability, the range spanned by his resulting upper-lower pair differs markedly from that arising from the Bayesian analysis, again calling into question both the interpretation of and justification for the method.

Dempster (1966) has proposed a different approach to this problem. In general, the two methods do not agree, but in this simple example they do, and hence the objection just voiced to Shafer's analysis applies with equal force to Dempster's.

5. RELATIONSHIPS BETWEEN JEFFREY'S RULE AND DEMPSTER'S RULE

Shafer has observed that Jeffrey's rule and Dempster's rule agree in certain cases. This is an easy consequence of the three properties D1–D3 of Dempster's rule given at the end of Section 3. To be precise, let P_1 be a probability on a set S, let $\{E_i\}_{i=1}^{n}$ be a partition of S, and suppose that $P_2(E_i)$ are positive numbers summing to 1. Define multivalued mappings Γ_i from $L_i \rightarrow$ subsets of S as follows:

$$L_1 = S, \qquad \Gamma_1(s) = s,$$

$$L_2 = \{1, \ldots, n\}, \qquad \Gamma_2(i) = E_i.$$

The product of (P_1, L_1, Γ_1) and (P_2, L_2, Γ_2) combine to give a probability on S because of property D2. Shafer (1981b, Sec. 7) shows that this is precisely the probability given by Jeffrey's rule.

Thus Dempster's rule may be viewed as a generalization of Jeffrey's rule. The difference between them may be summarized as follows:

1. Jeffrey's rule works with ordinary probabilities which have a well-understood interpretation in a variety of real-world situations. Dempster's rule works with upper and lower probabilities which presently lack an operational interpretation, objective or subjective.

2. Dempster's rule is a way to pool fairly general types of information. If one is willing to work outside the world of well-defined probabilities, upper–lower pairs representing information from very general sources can

be combined. An additive approach to the combination of different types of evidence is given in Secs. 3, 4, 5 of Diaconis and Zabell (1982). The comparison of the two approaches is instructive: Dempster's rule is based on an intuitive notion of independence; the method using Jeffrey's rule that we suggest is not tied to such independence.

Finally, it is worth considering a problem that neither theory claims to know how to treat. Suppose we have a probability P defined on a class $\mathscr{F}$ of subsets of a space S. After observation or reflection we decide that we need to work with a richer collection of sets $\mathscr{F}^*$, perhaps even a larger basic space S^*. For example, new data may force us to consider outcomes previously thought impossible or unimportant. How should we proceed to extend P, changing it as little as possible? Several procedures are available under special circumstances, but any semblance of a general theory is presently lacking.

6. SOME FOUNDATIONAL QUESTIONS

Changes of opinion, through conditioning, Jeffrey's rule, or other means, raise substantive foundational questions. The kind of shift in opinion exemplified in the examples of Section 2 are typical of what people do all the time when confronting real data. The basis of the shifts were sensory experiences, imprecise information and reflection; the shifts did not occur via conditioning.

Our view is that conditioning is not the only reasonable way to justify probability shifts. Conditioning and Jeffrey's rule are similar to exchangeability: they cannot be justified on the basis of coherence but are useful things to consider when trying to quantify subjective probability.

From the subjectivistic perspective, the conditional probability P(A|E) is the probability we *currently* would attribute to an event A if in addition to our present information we were also to learn E. In the language of betting, it is "the probability that we would regard as fair for a bet on A to be made immediately but to become operative only if E occurs" (de Finetti 1972, p. 193). In this formulation, the equality P(A|E) = P(A and E)/P(E) is not a definition but follows as a theorem derived from the assumption of coherence (de Finetti, 1975, Chap. 4). It gives a subjectivist's interpretation for the conditional probability routinely used in combinatorial computations.

If we actually *learn* E to be true, it is conventional to adopt as one's new probability

$$P^*(A) = P(A|E). \tag{2}$$

Assumption (2) seems entirely plausible—what else should our probability of A be, given that we have learned E, and nothing else, other than the probability which we were willing to attribute to A if we were subsequently to learn E? But, as Ramsey himself pointed out (1931, p. 180);

> [The degree of belief in *p* given *q*] is not the same as the degree to which [a subject] would believe *p*, if he believed *q* for certain; for knowledge of *q* might for psychological reasons profoundly alter his whole system of beliefs.

Several authors have pointed out that (2) is in fact an *assumption.* Hacking (1967, p. 314) refers to (2) as the *dynamic assumption of personalism,* to contrast it with the static nature of the assumption of coherence. Hacking (1967, pp. 315–316) points out that coherence in its usual sense does not entail (2), and de Finetti concedes as much when he refers to an unexplained "criterion of temporal coherency" (de Finetti, 1972, p. 150). De Finetti (1972, pp. 193–194) gives a particularly lucid discussion of the problem; see also Teller (1976), Gillies (1973, pp. 173–178).

In any case, coherence is a *consistency* condition and does not indicate the direction in which inconsistencies are to be resolved. Indeed, de Finetti goes so far as to describe equation (2) as an invitation to compare its two sides when attempting to calculate a conditional probability. I. J. Good's useful "device of imaginary results" (Good, 1950, pp. 35, 70; 1965, pp. 19–20, 45), where prior probabilities are elicited by first assessing posterior probabilities conditional on imaginary outcomes, carries this invitation out to its logical extreme.

In our view, the only way to justify Bayes's rule of conditioning is through the following tautology:

Compute the new probability P^* after observing E. (3)

If $P^*(A) = P(A|E)$ holds for each A, (4)

then assign $P^*(A) = P(A|E)$.

In practice, we will not usually recompute probabilities. We observe E and ask: "Did observing E cause my opinion to differ from P(A|E)?" This amounts to an approximate check of (4) and is similar to the way the J-condition is checked when applying Jeffrey's rule. We do not, contra Jeffrey (1970, pp. 178–179), regard this as entirely circular or question begging.

We think that our view that Bayesian conditioning is not the only way that probabilities shift goes a long way toward explaining why Bayesian methods are not generally useful in exploratory data analysis. Mallows (1970), Savage (1977) and others (including John Tukey) have discussed why the orthodox Bayesian view of statistics—incorporating new information via conditioning a prior distribution—does not come close to describing what a scientist does when confronting a real data set. Our explanation for this is that the impact of new data in the exploratory phases of scientific work is quite different from the impact of new data in later stages of scientific work. Ordinary conditioning will be judged useful or valid less often in the first case. For further discussion, see Diaconis (1985).

We summarize our views about conditioning as follows. Conditioning is a special, frequently useful way of changing probability assessments. Because of practical limitations, we are often in a situation where conditioning is not an appropriate way to quantify our current beliefs. We can always proceed by totally requantifying a new probability distribution. Jeffrey's rule points to a class of situations where some of the effort involved in a total requantification can be avoided. The justification for employing Jeffrey's rule parallels the justification of conditioning given in (3) and (4), above.

Since the writing of this survey, Shafer has published considerably more material on upper and lower probabilities (see, for example, Shafer and Tversky, *Cognitive Sciences*, 1985). In this last paper Shafer implicity replaces the model criticized in our Section 4 with another, but we do not find the new model any more convincing than the old. We have always found his writings valuable for their critical analysis of the assumptions underlying different models of probability and belief, but we remain troubled by the absence of "real" examples in which upper and lower probabilities perform tasks that more classical tools cannot handle.