UPDATING SUBJECTIVE PROBABILITY

BY

PERSI DIACONIS and SANDY ZABELL

TECHNICAL REPORT NO. 136
OCTOBER 1979

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
UPDATING SUBJECTIVE PROBABILITY

BY

PERSI DIACONIS and SANDY ZABELL

TECHNICAL REPORT NO. 136
OCTOBER 1979

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
UPDATING SUBJECTIVE PROBABILITY

by

Persi Diaconis
Bell Laboratories
and
Stanford University

and

Sandy Zabell
Rutgers The State University
and
University of California, Berkeley

ABSTRACT

We argue that "ordinary conditioning", \( P^*(A) = P(A|E) \), is not the only reasonable way to change a prior distribution \( P \) into a posterior distribution \( P^* \). We investigate a different rule proposed by Richard Jeffrey. Jeffrey's rule assumes there is a partition \( \{E_i\} \) such that \( P^*(A|E_i) = P(A|E_i) \) holds for all \( A \) and \( i \). Then \( P^*(A) = \sum P(A|E_i)P^*(E_i) \). Jeffrey's rule is connected to sufficiency and partial exchangeability. The final section discusses foundational aspects and the connection of Bayesian statistics and data analysis.
UPDATING SUBJECTIVE PROBABILITY

by

Persi Diaconis
Bell Laboratories
and
Stanford University

and

Sandy Zabell
Rutgers The State University
and
University of California, Berkeley

1. INTRODUCTION

We will argue that updating of subjective probability distributions can occur through other methods than conditioning. We present our current understanding of one of these methods - Richard Jeffrey's rule for updating subjective probability judgments in the light of newly acquired, but uncertain evidence. Our interpretation of probability follows de Finetti. In particular, we assume throughout that at any time we can quantify our probability judgments as a probability measure on a relevant algebra of events.

(1.1) Example. Suppose we are consulting statisticians in a medical trial. The experimenter says he has generated a random sequence of 100 a's and b's. Thinking
the experimenter is experienced, we believe the sequence will be exchangeably distributed and very much like 100 tosses of a fair coin. Given the first 10 a's and b's we will, under ordinary circumstances, make predictions concerning the remaining outcomes by updating our original probability $P$ to a new probability $P^*$ via

$$P^*(A) = P(A | \text{the results of the first 10 outcomes}).$$

Suppose that the first 10 outcomes are

$$a \ b \ a \ b \ a \ b \ a \ b \ a \ b.$$  

We might notice the perfect alternation of a's and b's and suspect that some element different from flips of a coin was operating. Surely, then, the new probability $P^*$ would have to be pretty much recomputed, taking into account whatever we know about the experimenter.

This example merits further discussion, which is given in Section 8. We argue there that much actual probability updating does not take place by using Bayes' Theorem. We may decide to update after an experience, thought, observation, or inspiration. In general, there is no reason that the new probability assessment need have any relation to the old. We will try to understand why, and in what way, relations can exist between new and old.

In general cases, the new probability $P^*$ has to be completely recomputed using the techniques for elicitation
of subjective probability as in Savage (1971), Winkler (1971),
or Hogarth (1975). In the situations in which Jeffrey's
rule is useful, much of the work in recomputing P* can be
saved. Suppose we have completely quantified P* after an
experience. There might be a partition \{E_i\} such that the
old probability P and the new probability P* satisfy

\[
(J) \quad P(A|E_i) = P^*(A|E_i) \quad \text{for all } A \text{ and all } i.
\]

Then we have

\[
P^*(A) = \sum_i P^*(A|E_i)P^*(E_i) = \sum_i P(A|E_i)P^*(E_i).
\]

The identity \(P^*(A) = \sum P(A|E_i)P^*(E_i)\) is called Jeffrey's
rule of conditioning. It is an algebraic fact which follows
from the rules of probability whenever condition \((J)\) holds.
In practice one will often proceed by thinking "the only
way the experience changed things was that \(P(E_i)\) changed to
\(P^*(E_i)\)". This last vague statement amounts to an approximate
check of the \((J)\) condition.

The next section presents more examples and points to
a source of natural partitions \(\{E_i\}\) arising from the ideas
of exchangeability and sufficiency. Section 3 points out the
connection between condition \((J)\) and the statistical notion
of sufficiency. We give an algorithm for finding a minimal
(coarsest) partition such that condition \((J)\) holds and give a
necessary and sufficient condition for two measures P and P*
to be related by conditioning. There is also a more general form of Jeffrey's rule which does not assume that \( P \) and \( P^* \) have the same support.

Sections 4, 5 and 6 describe what happens when two or more partitions are considered. In Section 4, we discuss commutativity of successive updating. In Section 5 we discuss methods for dealing with two partitions simultaneously, giving a necessary and sufficient condition for two probability measures on two algebras to have a common extension. In Section 6, we discuss some other motivations for Jeffrey updating when condition (J) has not been checked. We show that the Jeffrey updating is the "closest" measure to \( P \) which fixes \( P^*(E_i) \), and relate Jeffrey updating to the iterated proportional fitting procedure used in the statistical analysis of contingency tables.

For ease of exposition, most of this paper assumes a countable state space or a countable partition \( \{E_i\}_{i=1}^{\infty} \). In Section 7, we describe the mathematical machinery needed to extend the results to abstract probability spaces.

The final section, which may be read independently of Sections 2 through 7, presents a discussion of some philosophical issues connected with conditioning. The discussion includes comments on the disparity between Bayesian inference and exploratory data analysis.
Jeffrey's rule was introduced in Jeffrey (1957) and is further discussed in Jeffrey (1965, Chapter 11) and Jeffrey (1968). Isaac Levi (1967; 1970, pp. 147-152) is a vigorous critic of Jeffrey's version of probability kinematics, but has been thoroughly rebutted by Jeffrey (1970, especially at pp. 173-179). An early reference to a rule close to Jeffrey's is in Donkin (1851, p. 356). An independent proposal of Jeffrey's rule appears in Griffeath and Snell (1974). The last few years have seen a sudden upsurge of interest in Jeffrey conditionalization; papers have appeared by May and Harper (1976), Teller (1976), Field (1978), Williams (1979), van Fraassen (1979), Shafer (1979), and Domotor, Zanotti, and Graves (1979)!

The present work developed out of a seminar held by Richard Jeffrey during the Spring of 1979; we thank him for his comments, suggestions, and encouragement. We have also benefited from conversations with R. R. Bahadur, Peter Fishburn, David Freedman, Jim Pitman, Glenn Shafer, Amos Tversky, and Michael Woodroffe.
2. Examples

In the situations described below, Jeffrey's rule provides a natural tool for updating subjective probabilities.

(2.1) Example: Coin toss. Suppose we are thinking about 3 tosses of a coin. Under the usual circumstances a probability assignment is made on the eight possible outcomes $\Omega = \{(000), (001), (010), (011), (100), (101), (110), (111)\}$. Let us label the probability assignment as $\text{Prob} ((i,j,k)) = P(i,j,k)$. To give an example of "ordinary" conditioning, suppose that we are told that the sum of the number of ones in an outcome that has occurred is even. If the only impact of this information is to rule out those outcomes with an odd number of ones then our old probability will change to a new probability $P^*(i,j,k)$ on those triples with an even number of ones, this new probability being given by Bayes' Theorem. Thus

$$P^*(000) = \frac{P(000)}{P(000) + P(011) + P(110) + P(101)}.$$

Now imagine a situation where the information is of a less well defined sort. Suppose a friend, believed trustworthy, announces: "Oh, I see you're thinking about that coin. I just spun it 100 times in the other room and it came up heads 80 times." This is clearly relevant information and we will want to revise our opinion. The information cannot be put in terms of an event in the 8 point space $\Omega$. One possibility is to regard the 100 tosses
as interchangeable with the 3 unknown tosses but in many
cases this will not be a reasonable assumption. For example,
the friend may have spun the coin on a table and the present
experiment may involve tossing the coin in the air and
letting it land on the floor. In the most general setting
we will simply have to quantify eight new probabilities.
There is a wide class of situations where this requantification
can be simplified using Jeffrey's rule. Suppose that the
original probability assignment was exchangeable. That is,
two triples with the same number of heads were assigned the
same probability. Thus \( P(001) = P(010) = P(100) \) and \( P(110) = P(101) = P(011) \). In the situation described, the informa-
tion provided by the friend contains no information about
the order of the next 3 tosses and thus we may well believe
that our new probability assignment should still be exchange-
able. This amounts to providing a partition \( \{E_i\}_{i=0}^3 \) where
\[
E_0 = \{(000)\}, \quad E_1 = \{(001),(010),(100)\}, \\
E_2 = \{(110),(101),(011)\}, \quad E_3 = \{(111)\}
\]
and checking that, given \( E_1 \), the new and old probabilities
are the same. For example, under exchangeability of both
\( P \) and \( P^* \), consider the conditional probability of the
outcome \( (010) \) given \( E_1 \). This conditional probability is
zero given \( E_0, E_2, \) and \( E_3 \), and
\[
P((010)|E_1) = \frac{P((010))}{P(001) + P(010) + P(100)} = \frac{1}{3} = P^*((010)|E_1).
\]
Similarly, for any set $A$

$$P(A|E_1) = P^*(A|E_1).$$

Thus to quantify $P^*$, we must only quantify the $4$ numbers $P^*(E_i)$ and then define $P^*(A) = \sum_{i=1}^{4} P(A|E_i)P^*(E_i)$.

It is instructive to consider the problem of example 1 when the number of trials is much larger than 3.

For practical purposes one considers an infinite sequence of zero or one outcomes $X_1, X_2, X_3, \ldots$. Suppose that the joint distribution of $X_i$ is exchangeable and set

$$S_n = X_1 + \ldots + X_n.$$ Then, as shown by de Finetti, the limit

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

exists almost surely and

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

One consequence of de Finetti's Theorem is that one may decide on a subjective probability distribution for an infinite exchangeable sequence of coin-tosses by introspecting on the "prior distribution" $P(Z \in \mathcal{D}) \equiv \mu(p)$. In the story given above, the effect of the friend's
information could be taken into account by choosing a new prior \( d\mu^*(p) \) and Jeffrey's rule becomes:

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

This illustrates the use of Jeffrey updating via a continuous "sufficient statistic" rather than a discrete "sufficient partition" (see Section 3 below).

There are a number of cases in which the assumption of exchangeability is plausible. Some examples are described in Diaconis and Freedman (1979a). A very general notion of partial exchangeability which includes those situations where a sufficient statistic exists is given in Diaconis and Freedman (1979c). In any of these situations, when the information or experience is such that we believe that the new probability is still partially exchangeable, Jeffrey's rule can be used to do the updating.

(2.2) **Example:** Breaking a problem into pieces. Breaking a problem into pieces is a useful technique when forming probability judgments. For instance, one of the authors once had the practical task of deciding approximately how many components were in a typical mini-computer. It turns out that mini-computers are arranged by having their components on "boards" --
at the time there were about 5 or so boards with approximately 40 components per board in a typical mini-computer. A practical approach to the problem is to quantify a distribution of the number of boards \( B \), quantify a distribution of the number of components per board \( C \) and then use the distribution of the product \( B \times C \) as the distribution of the total number of components. This total number was, in turn, part of the larger problem of assessing the overall cost associated with adopting a particular mini-computer. At various times during the project, the opportunity of talking to experts about the numbers being estimated arose. This often resulted in information calling for wholesale revision of the distribution of one of the pieces. If the information was such that it didn't affect the other probability judgments -- this amounts to checking condition (J) -- then Jeffrey's rule says: quantify the new distribution for that piece and "plug it in" as a replacement for the old distribution of that piece. In this example, \( \Omega = \Omega_1 \times \Omega_2 \) is a product space and the partition is \( \Omega_1 \times \{ \omega_2 \}, \omega_2 \in \Omega_2 \).

(2.3) **Example:** 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 zero is recorded if the point of the tack is touching the floor, a one is recorded if the point is not touching the floor. This
example has been discussed by Mosteller, Rourke and Thomas (1970, pp. 113-115), Lindley and Phillips (1976) and Diaconis and Freedman (1979a p. 6). Most authors assume that the sequence of zeroes and ones thus obtained is exchangeable (the frequentists say independent with unknown probability). Suppose we initially assume exchangeability and quantify our opinion by specifying a measure \( \mu \) on \([0,1]\) as in Example 1. 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 (1979a) 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 zero and one. If the tack is initially point upward, so \( e_0 = 1 \), then

\[
P^*(e_0, e_1, \ldots, e_n) = \int \int t_{00}^* p_{00} t_{01}^* p_{01} t_{10}^* p_{10} t_{11}^* p_{11} \mu^* (dp_{00} dp_{11}).
\]

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 (1979b)).

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 are mutually singular. We now give a reasonable scenario which shows that much of the work that went into quantifying the measure \( \mu \) 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 \( \mu \), determined while considering the process as exchangeable, might be judged a reasonable quantification of the limiting relative frequency of ones in the non-sliding trials. One would also have to quantify the belief \( \alpha, \beta \) in the possibility that the tack slides without turning from both starting positions. Then \( \mu^* \) is a mixture of 3 measures on the unit square:

\[
\mu^* = \alpha \mu_{00} + \beta \mu_{11} + (1 - \alpha - \beta) \overline{\mu},
\]

where \( \mu_{00} \) and \( \mu_{11} \) are point masses at (0,0) and (1,1), while \( \overline{\mu} \) is the original prior \( \mu \) thought of as a measure on the line \( p_{00} = 1 - p_{11} \).
3. Jeffrey's Rule of Conditioning

In this section we develop some of the mathematics connected with Jeffrey's rule of conditioning. Throughout we work with a countable space Ω. We begin by giving a version of Jeffrey's rule which takes the support of the measures P and P* into account. Call a point ω ∈ Ω a support point of P if P(ω) > 0. Let supp(P) denote the set of support points of P. In general, P and P* will not have the same support — indeed with standard conditioning supp(P*) is strictly smaller than supp(P). Clearly P* will simply have to be quantified anew on supp(P*) − supp(P). This leads to the following generalized form of Jeffrey's rule:

(3.1) Suppose \{E_1\} is a partition of S = supp(P) ∩ supp(P*) such that

(J) P(ω|E_1) = P*(ω|E_1) for all ω ∈ S and all i.

Then for any set A,

P*(A) = \sum P(A|E_1) P*(E_1) + P*(A ∩ (supp(P*) − supp(P))).

In what follows, we will assume that supp(P*) = supp(P). Then Jeffrey's rule simplifies to the form P*(A) = \sum P(A|E_1) P*(E_1) as given in (1.3). All
the results we prove have straightforward modifications to the general situation (3.1) by restricting attention to $\text{supp}(P^*) \cap \text{supp}(P)$.

The partitions $\{E_i\}$ may be finite or countable. Sometimes partitions arise naturally as in examples (2.1) and (2.2). In other cases, the measures $P$ and $P^*$ can be used to define a suitable partition. To pursue this, we will rephrase Jeffrey's rule in terms of sufficient partitions. This allows us to use the well developed tools of sufficiency. We begin by recalling the definition of sufficiency. Blackwell and Girshick (1954, Chapter 8) or Lehmann (1959) contain further background.

(3.2) **Definition.** Let $\Omega$ be a countable set, $F$ a family of probability measures on $\Omega$ with the same support. A partition $\{E_i\}$ of $\Omega$ is **sufficient for $F$** if there is a function $\phi(\omega,i)$ such that for any $P \in F$

$$P(\omega|E_i) = \phi(\omega,i) \quad \text{for all } \omega \in \Omega \text{ and all } i.$$ 

The point is that $\phi$ does not depend on $P$. Sufficiency is often discussed in terms of sufficient statistics or sufficient algebras. The different versions are all easily equivalent. For instance, given a sufficient partition $\{E_i\}$, a statistic $T: \Omega \to \{1,2,3,\ldots\}$ can be defined by $T(\omega) = 1$ if $\omega = E_i$ and $T$ is called sufficient for $F$. Conversely, a sufficient statistic $T$ leads to a sufficient partition by defining $E_i = \{\omega: T(\omega) = i\}$. 
To apply Jeffrey's rule, it is required to find a partition \{E_i\} such that

\[ P(A|E_i) = P^*(A|E_i) \quad \text{for all } A \text{ and } i. \]

This is clearly the problem of finding a sufficient partition for the family \( F = \{P, P^*\} \). The next two Theorems are translations of the ideas of minimal sufficiency and likelihood ratio into the language of Jeffrey's rule.

Call a partition \{E_i\} coarser than a second partition \{E'_j\}, if every \( E_i \) is a union of sets in \( \{E'_j\} \).

For purposes of updating probability, a coarser partition has the advantage that \( P^* \) need be specified on fewer sets. A coarsest sufficient partition is said to be minimal sufficient. The next (well known) result shows that there is always a coarsest partition such that Jeffrey's rule is valid. This is also discussed by van Fraassen (1979).

\[(3.3) \quad \text{Theorem. Let } \Omega \text{ be a countable set, } P \text{ and } P^* \text{ probability measures on } \Omega \text{ with the same support. There exists a minimal sufficient partition for } P, P^*. \text{ If} \]

\[ L = \left\{ \frac{P(\omega)}{P^*(\omega)} \right\}_{\omega \in \Omega}, \text{ then the partition is } \{E_\ell\}_{\ell \in L} \]

where

\[ E_\ell = \{\omega: \frac{P(\omega)}{P^*(\omega)} = \ell\}. \]

\[ \text{Proof. We first argue that } \{E_\ell\}_{\ell \in L} \text{ is a sufficient partition: Fix } E_\ell \text{ and let } A \text{ be any subset of } \Omega. \text{ Then} \]

\[ P(A \mid E_\ell) = \frac{P(A \cap E_\ell)}{P(E_\ell)} = \frac{\sum_{\omega \in A \cap E_\ell} P(\omega)}{\sum_{\omega \in E_\ell} P(\omega)} = \frac{\sum_{\omega \in A \cap E_\ell} \ell P(\omega)}{\sum_{\omega \in E_\ell} \ell P(\omega)} = \frac{\sum_{\omega \in A \cap E_\ell} \ell P(\omega)}{\sum_{\omega \in E_\ell} \ell P(\omega)} = \ell P(A \mid E_\ell) = \ell P(A \mid E_\ell). \]

To show that \( E_\ell \) is the coarsest possible partition we now argue that if \( E \) is contained in a sufficient partition, then \( E \subset E_\ell \) for some \( \ell \). For if not, then \( E \) has non-empty intersection with say \( E_\ell \) and \( E_m \). Take \( x \in E \not\in E_\ell \) and \( y \in E \not\in E_m \). Then \( P(x \mid E) = P^*(x \mid E), P(y \mid E) = P^*(y \mid E) \) imply \( P(x)/P^*(x) = P(y)/P^*(y) \) which is a contradiction. \( \square \)

Remarks. 1. If \( \Omega \) is finite with \( n \) elements, Theorem 1 shows that there is a simple algorithm to determine the coarsest partition from the \( n \) pairs \((\omega, P(\omega)/P^*(\omega))\). Sort the pairs on their second component: \((\omega_1, P(\omega_1)/P^*(\omega_1)), (\omega_2, P(\omega_2)/P^*(\omega_2)), \ldots (\omega_n, P(\omega_n)/P^*(\omega_n))\) where if \( i < j \), \( P(\omega_i)/P^*(\omega_i) \leq P(\omega_j)/P^*(\omega_j) \). Finally, form the coarsest partition by letting \( E_1 = \{ \omega: P(\omega)/P^*(\omega) = P(\omega_1)/P(\omega_2) \} \) and so on. Since sorting can be done in approximately \( n \log n \) "operations" (Knuth (1973), Chapter 5), this gives an \( n \log n \) algorithm for determining the minimal partitions.

2. Theorem 1 can easily be adapted to the case when \( P \) and \( P^* \) have different support: just restrict attention to \( \omega \in \text{supp}(P) \cap \text{supp}(P^*) \) to get a partition of the common support.

Since \( \Omega \) is countable, \( P^* \) is specified by knowing \( P^*(\omega) \). The next result gives another way of writing Jeffrey's
rule and suggests the generalization of this rule to abstract spaces given in Section 7. The result is a version of the Fisher-Neyman factorization Theorem. The proof is implicit in the proof of (3.3) and is omitted.

(3.4) Theorem. Let $P, P^*$ be probability measures, with the same support, on the countable set $\Omega$. If $\{E_i\}$ is a partition of $\Omega$ such that condition (J) holds, then for each $\omega \in \Omega$

$$P^*(\omega) = \frac{P^*(E_i)}{P(E_i)} P(\omega) \text{ when } \omega \in E_i.$$  

Peter Fishburn has suggested an interesting interpretation for the results of this section. The measures $P$ and $P^*$ can be taken to be the probability assignments of two different people and we could be trying to quantify how they differ. In this context, there is no reason to restrict attention to two people; the family of probabilities $F$ can have any size. Finding a sufficient partition such that $\{E_i\}$ represents an interpretable parameter can help analyze the difference between the members of $F$: They differ on $\{E_i\}$ and given $E_i$ are the same. Theorem (3.3) has a direct extension to 3 (or more) probabilities by considering the ratios $P_1(\omega)/(P_1(\omega) + P_2(\omega) + P_3(\omega))$. For simplicity, we restrict the rest of this paper to two measures.

Jeffrey's rule of conditioning is a generalization of ordinary conditioning. We next consider conditions that
ensure that \( P^* \) could have arisen from \( P \) by conditioning.

Suppose \( P \) and \( P^* \) are measures on the countable space \( \Omega \).

We will say that \( P^* \) can be obtained from \( P \) by conditioning if there is a probability space \((\Omega', X, Q)\) and events

\[ \{E_\omega\}_{\omega \in \Omega}, \ E_\omega \in X \]

which can be thought of as "\( E_\omega = \omega \) occurred" in the sense that \( Q(E_\omega) = P(\omega) \). Also needed is an event \( E \in X \) such that \( Q(E_\omega | E) = P^*(\omega) \).

(3.5) **Theorem.** A necessary and sufficient condition that \( P^* \) be obtained from \( P \) by conditioning is that

\[
(3.6) \quad P^*(\omega) \leq B \cdot P(\omega) \quad \text{for some constant} \quad B \geq 1 \quad \text{and all} \quad \omega.
\]

**Proof.** If \( P^* \) can be obtained from \( P \) by conditioning, let \((\Omega', X, Q), \{E_\omega\}, E \) be given. Then for any \( \omega \in \Omega', \)

\[
P^*(\omega) = Q(E_\omega | E) \leq \frac{Q(E_\omega)}{Q(E)} = \frac{P(\omega)}{Q(E)}.
\]

This gives (3.6) with \( B = 1/Q(E) \).

Conversely, suppose (3.6) is satisfied. Let \( \Omega' = \Omega \times \{a, b\} = \{(\omega, a), (\omega, b)\} \omega \in \Omega \). Let \( X \) be the set of all subsets of \( \Omega' \). Let \( E_\omega = (\omega, a) \cup (\omega, b) \), and let \( E = \cup_{\omega \in \Omega} (\omega, a) \). Solving the problem of finding \( Q \) formally leads to introducing a parameter \( t \), \( 0 < t < 1 \) (\( t \) will be \( Q(E) \)), and setting

\[
Q((\omega, a)) = t \cdot P^*(\omega)
\]

\[
Q((\omega, b)) = P(\omega) - tP^*(\omega).
\]
Because (3.6) is satisfied, \( t \) can be chosen small enough so that \( Q((\omega,b)) > 0 \). It is then straightforward to check that \( Q \) is a probability on \( \mathcal{F} \) satisfying \( Q(E_\omega) = P(\omega) \) and \( Q(E_\omega | E) = Q((\omega,a))/\sum Q((\omega,a)) = P^*(\omega) \) as required. □

Condition (3.6) places a restriction on \( P, P^* \) when both have countable support but not when both have finite support and \( \text{supp}(P^*) \subseteq \text{supp}(P) \).
4. **Successive Updating**

In the usual applications of subjective probability, information builds up by successive conditioning. In this section we investigate successive Jeffrey updating. This is of direct interest and is also useful in understanding the combination of evidence discussed in Section 5.

Suppose that we initially have a probability P and, in the morning, decide to update it to a probability $P_E$ based on the partition $\{E_i\}_{i=1}^e$ and the new probability measure $P_E(E_i) = p_i \quad i = 1,2,\ldots,e$. In this section we will use $P_E$ instead of $P^*$. As always this means we have decided $P_E(A|E_i) = P(A|E_i)$ holds for our new opinion. In the evening, we then decide to update based on $\{F_j, q_j\}_{j=1}^f$. We will indicate the order of updating by $P_{EF}$. To use Jeffrey's rule at the second stage we must, of course, accept the Jeffrey condition $P_{EF}(A|F_j) = P_E(A|F_j)$.

(4.1) **Example.** A simple example has $E = F$. Thus, in the time between the two updatings our belief on the partition has changed from $P(E_i) = p_i$ to $P(E_i) = q_i$. The old and new opinions differ and we currently believe $P(E_i) = q_i$.

(4.2) **Example.** Suppose that in a criminal case we are trying to decide which of 4 defendants, called a, b, c, d, is a thief. We may initially think $P(a) = P(b) = P(c) = P(d) = 1/4$. In the morning evidence is
introduced to show that: "probably the thief was left-handed." The evidence does not demonstrate that the thief was definitely left-handed but says that $P(\text{thief left-handed}) = .8$. If $a$ and $b$ are the left-handed defendants, the partition becomes $E_1 = \{a,b\}$, $E_2 = \{c,d\}$ and $P_F(E_1) = .8$, $P_F(E_2) = .2$. If the only effect of the evidence was to alter the probability of left-handedness - note this amounts to checking the Jeffrey condition

$P(A|E_1) = P_F(A|E_1)$ - then $P$ is obtained from Jeffrey's rule as $P_F(a) = .4$, $P_F(b) = .4$, $P_F(c) = .1$, $P_F(d) = .1$.

Suppose next that evidence is presented that indicates that it is somewhat likely that the thief was a woman. If $a$ and $c$ are women, then $F_1 = \{a,c\}$, $F_2 = \{b,d\}$.

Say $P_{EF}(F_1) = .7$ and that Jeffrey updating is again judged acceptable, then the reader may check that

$P_{EF}(a) = .56$, $P_{EF}(b) = .24$, $P_{EF}(c) = .14$, $P_{EF}(d) = .06$

In the second example it is reasonable to ask if the order of updating affects the final answer. If the evidence $(F_1,.7)$, $(F_2,.3)$ was presented in the morning and $(E_1,.8)$, $(E_2,.2)$ was presented later and if at each stage Jeffrey updating was thought acceptable, would $P_{FE}$ equal $P_{EF}$? The first example, where $E = F$ and the update probabilities differ, shows that in general the order matters since the currently held opinion governs.
It turns out that in the second example order does not matter. The rest of this section investigates the effect of order.

There are two aspects to successive updating:

(4.3) The updating information at each stage:

\[
\{E_i, p_i\}_{i=1}^e, \quad \{F_j, q_j\}_{j=1}^f
\]

(4.4) The \textit{J} condition at each stage:

\[
P_E(A|E_1) = P(A|E_1) \quad \text{and} \quad P_{EF}(A|F_j) = P_E(A|F_j)
\]

or, if updating is being considered in the other order,

\[
P_F(A|F_j) = P(A|F_j) \quad \text{and} \quad P_{FE}(A|E_1) = P_F(A|E_1).
\]

The \textit{J} condition is an internal or psychological condition which must be checked or accepted at each stage. Mathematics has nothing to offer here.

Mathematics can be used to check if (4.3) is compatible with commutativity. Since Jeffrey updating fixes the probability on the partition (i.e., \(P_{EF}(F_j) = q_j\) and \(P_{FE}(E_1) = p_1\)), commutativity will be possible only if

(4.5) \(P_{EF}(E_1) = p_1\)

and

\(P_{FE}(F_j) = q_j\) for all \(i\) and \(j\).
It turns out that this condition is sufficient:

\((4.6)\) \textbf{Theorem.} If \((4.5)\) holds, then \(P_{EF} = P_{FE}\).

In other words, whenever \(P_{FE}\) and \(P_{EF}\) both incorporate \((4.3)\), they actually coincide. Theorem \((4.6)\) is an immediate consequence of Csiszár (1975, Theorem 3.2) and its proof will be omitted. Csiszár's theorem implies that the common measure \(P_{EF} = P_{FE}\) is an "I-projection" of the original measure \(P\) onto the set of measures which incorporate \((4.3)\). We discuss I-projections further in Section 6.

A second approach to the mathematical aspects of commutativity of successive Jeffrey updating uses independence. Two partitions \(E = \{E_i\}\) and \(F = \{F_j\}\), are \(P\)-independent if \(P(E_iF_j) = P(E_i)P(F_j)\) for all \(i\) and \(j\).

It is familiar that when all quantities are defined, \(E\) and \(F\) are independent if and only if

\[(4.7)\] \(P(E_i|F_j) = P(E_i)\) holds for all \(i\) and \(j\).

Of course \((4.7)\) holds if and only if \(P(F_j|E_i) = P(F_j)\) holds for all \(i\) and \(j\). Independence says that conditioning on \(F\) does not change the probabilities on \(E\). By analogy, we define:

\[(4.8)\] \(E\) and \(F\) are Jeffrey independent with respect to \(P,\{p_i\}\) and \(\{q_j\}\)

if

\(P_E(F_j) = P(F_j)\) and \(P_F(E_i) = P(E_i)\) holds for all \(i\) and \(j\).
Thus Jeffrey independence says that Jeffrey updating on $E$ with probability $p_i$ does not change the probability on $F$ and a similar statement with the roles of $E$ and $F$ interchanged. The next theorem shows the connection with commutativity.

(4.9) Theorem. Let $P, \{E_i, p_i\}$ and $\{F_j, q_j\}$ be given. Then $P_{EF} = P_{FE}$ if and only if $E$ and $F$ are Jeffrey independent with respect to $P, \{p_i\}, \{q_j\}$.

Theorem (4.9) is proved in the appendix.

The connection between $P$-independence and Jeffrey independence is seen in the following result which is also proved in the appendix. A closely related result is in Jeffrey (1957).

(4.10) Theorem. Two partitions $E$ and $F$ are $P$-independent if and only if $E$ and $F$ are Jeffrey independent with respect to any update probabilities $\{p_i\}$ and $\{q_j\}$.

Referring back to the second example at the start of this section, since the original probability $P(a) = P(b) = P(c) = P(d) = 1/4$ has $E$ and $F$ independent, any update probabilities can be used to get commutative successive updating.
Lest the reader think that commutativity is always possible, we conclude this section with an example which has \( P_{EF}(E_1) = p_1 \) (and of course \( P_{EF}(F_j) = q_j \)) but such that \( P_{FE}(F_j) \neq q_j \).

(4.11) **Example.** Let \( E = \{ E, \overline{E} \} \), \( F = \{ F, \overline{F} \} \), define \( P \) by \( P(EF) = 1/8 \), \( P(E\overline{F}) = 1/4 \), \( P(\overline{E}F) = 3/8 \), \( P(\overline{E}\overline{F}) = 1/4 \). Suppose \( p_1 = p_2 = 1/2 \) and \( q_1 = 7/15 \), \( q_2 = 8/15 \). Then a simple computation shows that \( P_{EF}(E) = 1/2 = P_{EF}(\overline{E}) \) but \( P_{FE}(F) \neq q_1 \).
5. Combining Several Bodies of Evidence

Suppose we have undergone a complex of experiences that result in our adopting new degrees of belief on two partitions \( E = \{E_i\} \) and \( F = \{F_j\} \), say

\[
(5.1) \quad P^*(E_i) = p_i \quad \text{and} \quad P^*(F_j) = q_j.
\]

How should we update our subjective probabilities so as to incorporate these new beliefs? In general, the theory put forth by de Finetti has no neat mathematical answer to this question - you just have to think about things and quantify your opinion as best you can. In this section we discuss several reasonable routes through this quantification process. The routes are reasonable in the same sense that exchangeability is a reasonable thing to consider when attempting to quantify probabilities on repeated events - the circumstances which make them subjectively acceptable occur frequently.

To begin with, if we are to adopt the degrees of belief \( P^* \) in (5.1), it is essential that they be coherent, i.e. that \( P^* \) be extendable to a probability measure. The next Theorem provides a simple necessary and sufficient condition for checking if such an extension exists. The proof, given in the appendix, contains an efficient algorithm when both partitions are finite.

\[
(5.2) \ \textbf{Theorem.} \ \text{If } \Omega \text{ is a countable set, } E = \{E_i\} \text{ and } F = \{F_j\} \text{ two partitions of } \Omega, \text{ and } \{p_i\}, \{q_j\}, \text{ two probability measures on } E \text{ and } F \text{ respectively, then there exists a probability measure}
\]
P* defined on all of \( \Omega \) such that \( P^*(E_i) = p_i \), \( P^*(F_j) = q_j \) holds for all \( i \) and \( j \), if and only if whenever \( A = \bigcup_{i \in I(A)} E_i \) and 
\[
B = \bigcup_{j \in I(B)} F_j \]

(5.3) \( A \cap B = \emptyset \) implies 
\[
\sum_{i \in I(A)} p_i + \sum_{j \in I(B)} q_j \leq 1.
\]

In the appendix we give an example devised by David Freedman and Jim Pitman to show that the simple condition (5.3) is not sufficient for Theorem (5.2) if \( \Omega \) is uncountable.

Assuming that (5.1) is coherent, we have two further tasks:

(5.4) Choose a probability \( P^* \) on the partition \( \{E_i \cap F_j\} \) which agrees with (5.1);

(5.5) Extend \( P^* \) to all of \( \Omega \), possibly by Jeffrey updating after checking the (J) condition

\[
P(A|E_i \cap F_j) = P^*(A|E_i \cap F_j).
\]

If judged valid, the easiest way of accomplishing (5.4) is to use independence: \( P^*(E_i \cap F_j) = P^*(E_i)P^*(F_j) = p_iq_j \).

Example 5.6. Suppose in the legal example (4.2) we were presented with the two items of evidence concerning left vs. right-handed, male vs. female in quick succession. We might accept the probabilities given -

\[
P^*(\text{left-handed}) = .8; P^*(\text{female}) = .7 - \text{and also judge that the evidence had not changed our assessment that being}
\]
left-handed and female were independent. In this case, the Jeffrey condition in (5.5) holds vacuously since \( \{E_i \cap F_j\} \) separates the four points \( a, b, c, d \). These assumptions fix the new probability which turns out to be the one arrived at in example (4.2).

For certain purposes we may want to consider the largest probability measure \( P^U \) consistent with (5.1), and the smallest probability measure \( P^L \), consistent with (5.1). These satisfy \( P^L(A) \leq P^*(A) \leq P^U(A) \) for any measure \( P^* \) satisfying (5.1). Such measures were first introduced by Hoeffding (1940) and Fréchet (1951), and are discussed by Kimeldorf and Sampson (1978), which contains further references.

Richard Jeffrey (1957, Chapter 4) has advocated another route from (5.1) to a final probability assignment: Use successive updating on the partitions \( E \) and \( F \). This raises two issues:

(5.7) When does successive updating satisfy (5.1)?

(5.8) When is successive updating reasonable?

Problem (5.7) arises because \( P_{EF} \) need not equal \( P_{FE} \). Indeed, example (4.11) provides a situation where (5.1) is coherent because \( P_{EF} \) satisfies (5.1), but \( P_{EF} \neq P_{FE} \). Since matters are simplified when \( P_{EF} = P_{FE} \) we note that the results of Section 4 imply that the following three conditions are equivalent:
(5.9a) \( P_{EF}(A) = P_{FE}(A) \) for all sets \( A \).

(5.9b) \( P_{EF}(E_i) = P_{FE}(E_i) \) and \( P_{FE}(F_j) = P_{EF}(F_j) \) for all \( i \) and \( j \).

(5.9c) \( P_{E}(E_i) = P(E_i) \) and \( P_{F}(F_j) = P(F_j) \) for all \( i \) and \( j \).

Even when the order doesn't matter, we still have the responsibility of justifying the use of successive updating, i.e. problem (5.8). One approach to this is through checking the two Jeffrey conditions at each stage of updating. We observe this is a somewhat unorthodox mental exercise in as much as we currently believe (5.1) which is a condition for both partitions \( E \) and \( F \). If we decide to update first on \( E \), then we must check \( P(A|E_1) = P_E(A|E_1) \), which amounts to thinking as if we don't know about \( F \) and are only thinking about \( E \). At the second stage, one must check \( P_E(A|F_j) = P_{EF}(A|F_j) \), which amounts to comparing one's opinion not knowing \( F \) to one's opinion knowing \( F \). Examples such as (4.11) show that this can be tricky. However, it is a possible route which is more general than the route, using independence, suggested before. It seems worthy of further study.

We note two final points. First, there is no reason to have \( P_{EF} = P_{FE} \) for successive updating to be useful and valid. If each of the (J) conditions is judged valid in forming \( P_{EF} \) and if \( P_{EF} \) satisfies (5.1), then \( P_{EF} \) is an adequate quantification of current belief. Second, (5.9) implies that it is not possible for both \( P_{FE} \) and \( P_{FE} \) to incorporate (5.1) and both be judged acceptable updates (in the
sense that both (J) conditions have been checked, and yet have $P_{EF} \neq P_{FE}$. Thus non-commutativity is not a real problem for successive Jeffrey updating.
6. Mechanical Updating

The approach we have taken thus far to justifying Jeffrey's rule is subjective - through checking condition (J). Several authors - Griffeth and Snell (1974), May and Harper (1976), Williams (1979), and van Fraassen (1979) - have pursued a different justification. Given a prior \( P \), partition \( \{E_i\} \), and given a new measure \( P^* \) on the partition, find the "closest" measure to \( P \) which agrees with \( P^* \) on the partition and take this as defining \( P^* \) on the whole space. Since this way of proceeding does not attempt to quantify one's degree of belief, we call this approach mechanical updating.

Our first group of results show that if "close" is defined in any of the usual ways, the closest measure is given by Jeffrey's rule. We work with three notions of closeness between measures \( P \) and \( Q \) on the countable set \( \Omega \):

(6.1) The variation distance between \( P \) and \( Q \) is

\[
||P - Q|| = \sup_{B \subseteq \Omega} |P(B) - Q(B)|
\]

Two measures are close in variation distance if they are uniformly close on all subsets.

(6.2) The Hellinger distance between \( P \) and \( Q \) is

\[
H(P, Q) = \sum_{\omega} (\sqrt{P(\omega)} - \sqrt{Q(\omega)})^2.
\]
(6.3) The Kullback-Leibler information of $Q$ with respect to $P$ is

$$I(Q,P) = \sum_{\omega} Q(\omega) \log \left( \frac{Q(\omega)}{P(\omega)} \right)$$

The variation and Hellinger distances are actual metrics on the space of probability distributions, the Kullback-Leibler information is not, being asymmetric in its arguments. Kailath (1967) gives a survey of distances with references and properties of (6.1), (6.2), and (6.3).

(6.4) **Theorem.** Let $\Omega$ be a countable set, $P$ a probability on $\Omega$, and $\{E_i\}$ be a partition of $\Omega$. Suppose $P^*(E_i) \geq 0$ are given numbers such that $\sum P^*(E_i) = 1$. Let $Q$ be a probability on $\Omega$ such that $Q(E_i) = P^*(E_i)$. Then

(6.5) \hspace{1cm} ||Q - P|| \geq \sum |P(E_i) - P^*(E_i)|

(6.6) \hspace{1cm} H(Q,P) \geq \sum (\sqrt{P(E_i)} - \sqrt{P^*(E_i)})^2

(6.7) \hspace{1cm} I(Q,P) \geq \sum P^*(E_i) \log \left( \frac{P^*(E_i)}{P(E_i)} \right)

In each of (6.5), (6.6), and (6.7), equality holds if and only if $Q(A) = \sum P(A|E_i) P^*(E_i)$.

**Remarks.** 1. In Theorem (6.4), the minimum distance between $P$ and $Q$ is the distance between $P$ and $Q$ viewed as measures on the partition $\{E_i\}$.

2. A result like (6.4) holds for several other notions of distance. We discuss this in Section 7 where we prove an abstract generalization of Theorem (6.4).
Mechanical updating allows the possibility of updating on more general collections of sets than partitions. Suppose we want to adapt new degrees of belief \( P^*(E_i) = p_i \), \( 1 \leq i \leq n \), where \( E = \{E_1, E_2, \ldots, E_n\} \) is not necessarily a partition of \( \Omega \). This situation is closely related to Jeffrey's proposal of updating simultaneously on several partitions, mentioned in Section 5, in as much as updating simultaneously on partitions \( E_1, E_2, \ldots, E_k \) is the same as updating on \( E = \bigcup_{i=1}^n E_i \). Conversely, updating on \( E = \{E_1, \ldots, E_n\} \) can be viewed as updating simultaneously on the partitions \( E_1 = \{E_1, E_1^c\}, E_2 = \{E_2, E_2^c\}, \ldots E_n = \{E_n, E_n^c\} \).

In general, the set \( C = \{Q: Q(E_i) = p_i \text{ for all } i\} \) is a convex set of probability measures on \( \Omega \) which can be empty, contain a single element, or contain many elements. In the first case \( P^* \) is incoherent, in the second, \( P^* \) is uniquely defined. When the third case holds, we can use Kullback-Leibler information as a notion of "distance" to pick a unique member of \( C \) closest to \( P \).

(6.8) **Theorem.** Let \( S(P, \infty) = \{Q: I(Q, P) < \infty\} \).

If \( S(P, \infty) \cap C \neq \emptyset \), then there exists a unique element \( Q_j \in C \) such that \( I(Q_j, P) = \inf \{I(Q, P): Q \in C\} \).

**Proof.** (6.8) is an immediate consequence of Csiszar's (1975) Theorem 2.1, since \( C \) is convex and closed with respect to variation distance. \( \square \)

In Csiszar's (1975, p. 147) terminology, \( Q_j \) is the \( I \)-projection of \( P \) onto \( C \). The term is meant to suggest the
projection of a vector in $\mathbb{R}^n$ onto a subspace. It turns out that $Q_J$ is closely related to a widely used technique in the statistical analysis of contingency tables.

A standard method of adjusting an $r \times c$ contingency table so that it has given marginal totals is the iterated proportional fitting procedure (IPFP). In this, one first adjusts the table to have (say) the right row sums (by dividing the numbers of a given row by the appropriate factor), then adjusts the new table to have the right column sums, and continues iteratively. It follows from Csiszár (1975, Theorem 3.2) that this procedure converges to the $I$-projection of the initial table onto the set of tables with the right row and column sums. The IPFP finds the "closest" table to the original table with the prescribed margins. This is essentially the same as finding the closest measure to an initial probability with prescribed values on two partitions.

The IPFP can be used to compute $Q_J$ of Theorem (6.8) by treating the problem as an $n$-dimensional contingency table with given margins $P^*(E_1), 1 - P^*(E_1), \ldots$

An amusing consequence of the IPFP can be said in the language of Section 4. Suppose we are considering two different aspects of a problem which are characterized by two partitions $E_1, E_2$. As people sometimes seem to do, we might oscillate back and forth, first updating on $E_1$, then updating on $E_2$, then updating on $E_1$ again, and so forth. If the update probabilities are fixed and coherent (so that the conditions of Theorem (5.2) are satisfied), the results for the IPFP show that our oscillating distribution converges to the $I$-projection of $P$. 
7. Abstract Probability Kinematics

In this section we briefly sketch the generalization of Jeffrey's rule of conditioning from the countable setting to an abstract one. The need for such a generalization is shown by the examples in Section 2 - each example there is naturally phrased in terms of sufficient statistics taking on uncountably many values. The generalization we use replaces partitions by \( \sigma \)-algebras. This section assumes familiarity with measure theoretic probability; Lehmann (1959) contains the necessary background.

Jeffrey's rule involves a probability space \((\Omega, \Theta, P)\) thought of as describing our current subjective beliefs about the \( \sigma \)-algebra of events \( \Theta \). Let \( P^* \) be the new probability and let \( A_0 \) be a sub-\( \sigma \)-algebra of \( \Theta \). Let \( C \) be an \( A_0 \) measurable set such that \( P(C) = 0 \) and \( \overline{P} \ll \overline{P}^* \) on \( \Omega - C \), where \( \overline{P}, \overline{P}^* \) are the restrictions of \( P, P^* \) to \( A_0 \). The appropriate version of Jeffrey's condition \((J)\) is:

\[
(J) \quad A_0 \text{ is sufficient for } \{P, P^*\}.
\]

When condition \((J)\) holds, Jeffrey's rule of conditioning is:

\[
(7.1) \quad P^*(A) = \int_{\Omega - C} P(A|A_0)P^*(d\omega) + P^*(A \cap C).
\]

where \( P(A|A_0) \) is the conditional probability of \( A \) with respect to \( A_0 \). If \( P^* \ll P \), we choose \( C = \phi \).
Much of the mathematical machinery for dealing with Jeffrey conditionalization in this generality has been developed (for a different purpose) by Csiszár (1967). His Lemma (2.2) translates into a likelihood ratio version of Jeffrey's rule (compare (3.3)): Let \( \lambda \) be a \( \sigma \)-finite measure which dominates \( P, P^* \). Let \( \lambda, \overline{P}, \overline{P}^* \) be the restrictions to \( A_0 \). Assume \( \lambda \) is \( \sigma \)-finite. Let \( \overline{p}(x), \overline{p}^*(x) \) be the densities of \( P, P^* \) with respect to \( \lambda \) and \( p^* \) the density of \( P^* \) with respect to \( \lambda \). If condition (J) holds, then:

\[
(7.2) \quad p^*(x) = \begin{cases} 
\frac{p^*(x)}{p(x)} & \text{if } p(x) > 0 \\
p^*(x) & \text{if } p = 0.
\end{cases}
\]

The identity (7.2) is, of course, a version of the Fisher-Neyman factorization theorem in the language of Jeffrey conditioning. In this generality (7.2) was first proved by Halmos and Savage (1949).

Csiszár's results allow us to give a single theorem which includes (6.4), showing that the closest measure to \( P \) which agrees with \( P^* \) on \( A_0 \) is the measure given by (7.1). Csiszár has introduced the notion of \( f \)-divergence, where \( f \) is a strictly convex function defined in the interval \((0, \infty)\). If \( \mu_1 \) and \( \mu_2 \) are two measures on \((\Omega, \theta)\), define the \( f \)-divergence by

\[
I_f(\mu_1, \mu_2) = \int_{\Omega} \frac{f(p_1(x)) \lambda}{p_2(x)} p_2(x) \, dx.
\]
where $\lambda$ dominates $\mu_1$ and $\mu_2$, and $p_1 = \frac{d\mu_1}{d\lambda}$.

Taking $f(u) = u \log u$ gives the Kullback-Leibler information. Taking $f(u) = (u^{1/2} - 1)^2$ gives the Hellinger distance and taking $f(u) = |u - 1|$ gives the variation distance. Csiszár shows that several other notions of distance are $f$-divergences for appropriate $f$.

(7.3) **Theorem.** Let $C$ be the set of probability measures on $(\Omega, \mathcal{F})$ which agree with $P^*$ on $A_0$. Then under condition (J),

$$I_f(P, P^*) = I_f(\overline{P}, P^*) = \inf\{I_f(P, Q) : Q \in C\}.$$ 

**Proof.** The first equality follows from sufficiency of $A_0$ for $P, P^*$ while the second follows from Csiszár's (1967, Section 3) version of the minimum discrimination distance Theorem of Kullback-Leibler. This states that $I_f(P, Q) \geq I_f(\overline{P}, \overline{Q})$. Since $I_f(\overline{P}, \overline{Q}) = I_f(\overline{P}, P^*)$, this completes the proof of (7.3) \(\square\)
8. 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 (1.1) is typical of what real scientists do all the time when confronting real data. The basis of the shift is noticing things and thinking; the shift 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 - they are useful things to consider when trying to quantify subjective probability.

From the subjectivist 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 \( P(A|E) = P(AE)/P(E) \) is not a definition, but follows as a theorem derived from the assumption of coherence (de Finetti, 1975, Chapter 4).

If we actually learn \( E \) to be true, it is conventional to adopt as one's new probability
(8.1) \[ P^*(A) = P(A|E). \]

Assumption (8.1) 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? Several authors have pointed out that (8.1) is an assumption. Hacking (1967, p. 314) refers to (8.1) 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 (8.1) 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) contains a particularly lucid discussion of the problem. A recent article with discussion and further references is Teller (1976). Gillies (1973, pp. 173-178) makes a number of similar points.

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

(8.2) Compute the new probability \( P^* \) after observing E. If

(8.3) \[ P^*(A) = P(A|E) \]

holds for each A,

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 (8.3), and is similar to the way condition (J) 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-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. Tukey (1972, pp. 61-63), Mallows (1970) and Savage (1970) have discussed in some detail that 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 the recognition 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. The difficulties are both practical and conceptual. We first examine some of the practical problems ((8.4),(8.5)) and then discuss a conceptual problem (8.6).
(8.4) "...the discovery of the irrelevance of past knowledge to the data before us can be one of the great triumphs of science" (Tukey, 1972, p. 63). Consider example (1.1). The effect of observing perfect alternation of a's and b's was to cause suspicion that the trials were not like normal coin tossing. In usual circumstances, none of the myriad possible options to exchangeability will have been formulated before the trial and built into the prior P. Hence (8.3) will not hold and simple conditioning is not the appropriate way to change our subjective distributions.

de Finetti (1975, § 11.3) in discussing a very similar point, suggests that the original assignment was "superficial". That is, in assigning the original distribution, we should have thought "what if the trials alternate" and explicitly built in the possibility of cheating, or malfunctioning of the generating process. This is a laudable aim, but seems practically impossible. After all, instead of alternating, the initial 10 trials could have a's at prime places only, or the string of 10 a's and b's could spell out some famous date base 2. As a practical matter, the subjectivist must make simplifying assumptions such as exchangeability.
Ordinarily, this will be tempered with a caveat about something suspicious happening. We cannot really hope to quantify an exhaustive list of what is suspicious and must be prepared to requantify our probabilities if our suspicions are aroused.

(8.5) "Bayesian techniques assume we know all the alternate possible states of nature" (Tukey 1972, p. 63). In practice, an observed event need not be an element of the algebra we have quantified over. For example, we may observe an event the possibility of which we were not aware of before. In principle, one might add such an event to one's algebra and retrospectively quantify probabilities on the assumption that the event had not yet been observed (de Finetti (1972, Chapter 8)). Often such an analysis seems unnatural; see Shafer (1979, § 3) for a number of examples. Shafer's theory of belief functions is one attempt to describe how new evidence should be incorporated in such cases.
Richard Jeffrey has advanced a fundamental
criticism of conditioning as the only way of learning from
experience:
(8.6) "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 that proposition to 1"
(Jeffrey 1968, p. 171). Jeffrey's point is not the jejune
one that nothing is known with absolute certainty, as much
as that there are large classes of observations which affect
our beliefs, but which cannot be captured as propositions
which we condition on. The effect of these observations is
to cause a change in our degree of belief in certain prop-
ositions which is less drastic than a change to zero or
one. This point of view is also taken by Shafer (1976).

As an example, suppose we are told we are about to
be played one of two recordings of Shakespeare, read by
either Olivier or Gielgud. Which of the two will be decided
by flipping a coin. It may be that this induces a prior
distribution with mass 1/2 on Olivier and 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;
all the integration of sensory stimuli has already taken
place at a subconscious level. To demand a list of objective
vocal features which we condition on in order to affect the
change would be a logician's parody of a complex psychological
process.
Isaac Levi has criticized Jeffrey's account of uncertain perceptions leading to changes in beliefs on the grounds that

Criticism is important...But our credal responses to sensory stimulation cannot be criticized if we make no claims about them...if we recognize a distinction between reliable and unreliable responses to sensory stimulation, as Jeffrey himself does, we must defend our credal responses by claiming that they are made under satisfactory conditions...One can explicitly describe some features of the conditions of observation. Moreover, the precision with which they are described is not limited in any a priori way; and if responses to sensory stimulation are to be critically evaluated, some description of the conditions of observation will be required. [Levi (1970, pp. 150-151)]

However, as our example with Olivier and Gielgud makes clear, Levi's invocation of the conditions under which our observations are made is not to the point. No matter how much information of this sort is conveyed (the quality of one's hearing, the loudness of the radio, were there distracting noises, etc.), we have no way of assessing the "reliability" of the resulting response (Olivier more likely than Gielgud). That is, although information as to the conditions may suggest upper bounds to the likelihood we may ascribe to one of two alternatives, it can say nothing as to our wisdom in actually opting for one as opposed to another.

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 situations where conditioning is not the 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 work involved in a total requantification can be saved. The justification for Jeffrey's rule is entirely similar to the justification of conditioning we have given in (8.2) and (8.3).
Appendix to Section 4

Proof of Theorem (4.9)

If $P_{EF}(A) = P_{FE}(A)$ for all events $A$ if and only if

\[
\sum_{ij} \frac{p_i q_j}{P_{E}(F_j) P(E_i)} P(AE_i F_j) 
= \sum_{ij} \frac{p_i q_j}{P_{F}(E_i) P(F_j)} P(AE_i F_j).
\]

Choose $A = E_{i_0} F_{j_0}$ to get

\[P_{E}(F_{j_0}) P(E_{i_0}) = P_{F}(E_{i_0}) P(F_{j_0}) \text{ for all pairs } i_0, j_0.\]

Keeping $i_0$ fixed and summing over $j_0$ yields

\[(4.13a) \quad P(E_{i_0}) = P_{F}(E_{i_0});\]

similarly, fixing $j_0$ and summing over $i_0$ yields

\[(4.13b) \quad P_{E}(F_{j_0}) = P(F_{j_0}).\]

Thus, $E$ and $F$ are Jeffrey independent with respect to $P$, \{p_i\}, \{q_j\}. Conversely, if (4.13) holds, then

\[P_{E}(F_j) P(E_i) = P(F_j) P(E_i) = P_{F}(E_i) P(F_j).\]
Using this equality shows that (4.12) holds and so
\[ P_{EF} = P_{FE}. \]

Proof of (4.10). First suppose \( E \) and \( F \) are independent. Then,

\[
(4.14) \quad P_E(F_j) = \sum_{i} P(F_j | E_i) p_i = \sum_{i} P(F_j) p_i = P(F_j).
\]

To see the converse, suppose \( E \) and \( F \) are not independent. Then there exists \( E_{10} \) and \( F_{j0} \) such that \( P(F_{j0} | E_{10}) \neq P(F_{j0}) \).

Pick \( p_{i0} \) sufficiently close to 1. Then (4.14) entails

\[
\sum_{i} P(F_{j0} | E_i) p_i \neq P(F_{j0}),
\]

and hence \( P_E(F_{j0}) \neq P(F_{j0}). \)
Appendix to Section 5.

We first prove a slight generalization of Theorem (5.2):

(5.10) **Theorem.** Let $\Omega$ be a countable set, $S$ a $\sigma$-algebra of subsets, $A$ and $B$ sub-$\sigma$-algebras of $S$, and $\mu$ and $\nu$ probability measures on $A$ and $B$ respectively. A necessary and sufficient condition for there to exist a probability measure $P$ on $(\Omega, S)$ such that $P$ equals $\mu$ on $A$ and $P$ equals $\nu$ on $B$ is

(5.11) for each $A \in A$ and $B \in B$ such that $A \cap B = \phi$,

$$\mu(A) + \nu(B) \leq 1.$$ 

**Proof.** The condition is clearly necessary. To prove sufficiency, let $\{A_i\}_{i=1}^\infty$ be the atoms of $A$ and $\{B_i\}_{i=1}^\infty$ be the atoms of $B$. Let $\Omega_a = \{A_i\}_{i=1}^\infty$ and $\Omega_b = \{B_i\}_{i=1}^\infty$ both thought of as discrete topological spaces. In $\Omega_a \times \Omega_b$, consider the set $F = \cup A_i \times B_j$. This is a closed set in $\Omega_a \times \Omega_b$ and according to Theorem 11 of Strassen (1964) a necessary and sufficient condition for there to exist a probability measure $\gamma$ on $F$ such that $\gamma$ has margins $\mu$ and $\nu$ is that for every $B \in B$,

(5.12) $$\nu(B) \leq \mu(\pi_a(F \cap S \times B))$$

where $\pi_a$ is the projection of a set into its first coordinate. Clearly $\pi_a(F \cap S \times B) = \sum_{A_1 \in A} A_1$ is the smallest $A$ measurable $A_1 \cap B = \phi$. 
set containing B. Thus Strassen's condition (5.12) is satisfied if and only if

\[(5.13) \quad \text{whenever } A \in A, \ B \in B \text{ and } B \cap A, \ \nu(B) \leq \mu(A).\]

Condition (5.13) is equivalent to (5.11). Hence Strassen's Theorem gives a measure \(\gamma\) which may be regarded as a measure on the partition \(\{A_i \cap B_j\}\). Since \(\Omega\) is countable, \(\gamma\) can clearly be extended to a measure on all of \(\Omega\) and then restricted to a measure \(P\) on \(S\) with the desired properties.

In the proof of Theorem (5.11), we have used Strassen's Theorem, which itself uses the Hahn-Banach Theorem. When the two partitions are both composed of a finite number of sets, Hansel and Troallic (1978) have shown that Strassen's Theorem follows from the max flow-min cut Theorem. There are efficient algorithms for finding maximum flows, and hence for checking (5.3), in Bondy and Murty (1976, Chapter 11).

As the final item of this section, we consider the extension of Theorem (5.11) to more general spaces.

Let \((\Omega, S)\) be a measurable space, let \(A, B\) be sub-\(\sigma\)-algebras of \(S\) and let \(\mu\) and \(\nu\) be probability measures on \(A\) and \(B\) respectively. When does there exist a probability measure \(P\) on \(\sigma(A, B)\) such that \(P\) restricts to \(\mu\) on \(A\) and \(\nu\) on \(B\)? We have proved above that if \(\Omega\) is countable then a necessary and sufficient condition is:

\[(5.14) \quad \forall A \in A, \ B \in B, \ A \cap B = \emptyset \implies \mu(A) + \nu(B) \leq 1.\]
It is easy to show that (5.14) is in fact necessary and sufficient for the existence of a \textit{finitely additive} measure $\hat{\mu}$ on the algebra generated by $A$ and $B$ (and hence on the algebra of all subsets of $\Omega$) which restricts to $\mu$ on $A$ and $\nu$ on $B$, even if $A$ and $B$ are merely algebras. Briefly, one considers the following linear subspace of bounded real valued functions from $\Omega$: $L = \{f+g: f \text{ is } A \text{ measurable and } g \text{ is } B \text{ measurable}\}$ and extends the positive linear functional $\ell(f+g) = \mu(f) + \nu(g)$ using the Hahn-Banach Theorem. Condition 5.14 is used to show $\ell$ is well defined.

We now present an example due to David Freedman and Jim Pitman which shows that condition (5.14) is not sufficient to ensure that a countably additive extension of $\mu$ and $\nu$ exists. The example shows a bit more: It shows that Theorem 11 of Strassen (1965) cannot be extended to give conditions for a measure on an $F_\sigma$ set of the unit square to have given margins.

(5.15) \textbf{Example} (D. Freedman, J. Pitman). There exists an $F_\sigma$ set $K$ in the unit square and a finitely additive probability $\pi$ on $K$ which has marginal projections equal to Lebesgue measure on each coordinate but such that $K$ supports no countably additive probability with these margins.

\textbf{Remark.} Taking $\Omega = K$

\begin{align*}
A &= \{(A \times [0,1]) \cap K: \ A \text{ is a Borel set of } [0,1]\}, \\
B &= \{([0,1] \times B) \cap K: \ B \text{ is a Borel set of } [0,1]\},
\end{align*}

with $\mu$ and $\nu$ as Lebesgue measure on $A$ and $B$ respectively,
gives a situation in which 5.14 is satisfied (since a
dinitely additive refinement exists) but no countably
additive refinement exists.

Proof. To construct $K$, let $a_n$ be a sequence of numbers in
$(0,1)$ with $a_n \uparrow 1$. Let $\ell_n$ be the line in the unit square
connecting $(0,0)$ to $(1,a_n)$. Let $K = \bigcup_n \ell_n$. Note that $K$
does not include the diagonal. To construct $\pi$, let $\pi_n$
be Lebesgue measure on the Borel sets of $\ell_n$. Let $\rho$ be any
finitely additive probability measure defined on all subsets
of the integers $\{1,2,3,\ldots\}$ such that $\rho$ is zero on finite
subsets. Let $\pi(s) = \int \pi_n(s) \rho(\mathrm{d}n)$. Each $\pi_n$, considered as
a probability on $K$, projects to Lebesgue measure on the
x-axis of the unit square. Further, the projection of $\pi_n$
onto the y-axis of the unit square gives Lebesgue measure
restricted to the set $\{(0,y) : 0 \leq y \leq a_n\}$. It follows
easily that the y-axis margin of $\pi$ is Lebesgue measure. It
only remains to argue that $K$ does not support a countably
additive probability measure $P$ which projects to Lebesgue
measure. If $X: K \to [0,1]$, and $Y: K \to [0,1]$ are the two
projections, then $E(X) = E(Y) = \frac{1}{2}$ because each of $X$ and
$Y$ are uniformly distributed by construction. Any countably
additive $P$ would have to put positive probability on some
line $\ell_n$ and since for all $(x,y) \in K, y < x$, this forces
$E(X) > E(Y)$. The contradiction shows $P$ cannot exist. □
Bibliography


