The Annals of Statistics
1982, Vol. 10, No. 4, 1075–1089

BAYES'S TWO ARGUMENTS FOR THE RULE OF CONDITIONING[1]

By Glenn Shafer

University of Kansas

The introductory section of Thomas Bayes's famous essay on probability contains two arguments for what we now call the rule of conditioning. The first argument, which leads to Bayes's third proposition, can be made rigorous if we use rooted trees to represent the step-by-step determination of events. The second argument, which leads to Bayes's fifth proposition, does not stand up to scrutiny.

1. Introduction. Compare the following statements, both of which appear in Thomas Bayes's famous essay on probability:

(I) . . . if of two subsequent events the probability of the 1st be a/N and the probability of both together be P/N, then the probability of the 2nd on supposition the 1st happens is P/a.

(II) If there be two subsequent events, the probability of the 2nd b/N and the probability of both together P/N, and it being first discovered that the 2nd event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is P/b.

Statement I is the corollary to Bayes's third proposition; statement II is his fifth proposition.

If we denote the first of two "subsequent events" (Bayes does not spell out what he means by this) by A and the second by B, then statement I becomes, when translated into modern notation,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \tag{1}$$

while statement II becomes

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2}$$

Most modern students of probability think of (1) and (2) as merely two instances of a single general rule. Yet Bayes evidently thought of I and II as quite distinct statements, requiring very different proofs.

I believe that modern students of probability can gain a great deal by trying to understand Bayes's viewpoint. Conditional probability is a pervasively important concept—it is the basis of all our modern approaches to inference. It is also an extremely subtle concept. Because we are so accustomed to using the concept, we can easily fall into the habit of thinking we understand it thoroughly. In fact, we are far from mastering all its subtleties. In reading Bayes's paper we have a unique opportunity—an opportunity to see an original mind grappling with the subtleties of conditional probability with neither the help nor the hindrance provided by the presumptions and clichés that have become attached to the concept during the past two centuries.

In order to understand the difference Bayes saw between statements I and II, we need a way of understanding his notion of subsequent events—we need, that is to say, a mathematical framework for probability that takes the timing of events into account. In this paper I use the simple and well-known device of the rooted tree to provide such a framework, and I translate the section of Bayes's essay that contains statements I and II

Received September 1981; revised May 1982.

[1] Research partially supported by National Science Foundation Grant MCS 800213.

AMS 1970 *subject classification.* Primary 60A05; secondary 62A15.

Key words and phrases. Bayes, conditional probability, exact events, rooted trees.

into this framework. It emerges from this exercise that most of Bayes's reasoning, including his argument for his third proposition (and its corollary, statement I), is mathematically sound. His argument for his fifth proposition (statement II) does not, however, stand up to scrutiny.

These results have profound implications for the modern concept of conditional probability. Twentieth-century writers, from Hausdorff (1901) onward, have insisted that the timing of events is irrelevant to the concept of conditional probability. And since Kolmogorov (1933) at least, they have treated (1) as a mere definition, applicable to any two events A and B. But our study of Bayes suggests that the timing of events may be crucial to (1); there is a justification for this definition that seems to be available only when A precedes B.

As I explain in Section 6 below, Bayes's third proposition is relevant to both objective and subjective probabilities. It gives a way of understanding why objective probabilities change as they do when a chance process unfolds step-by-step. And it gives a justification for changing subjective probabilities by conditioning in cases where the possibilities for the step-by-step development of our knowledge are built into our initial subjective probability model. It does not justify conditioning subjective probabilities on information that is acquired unexpectedly; this is what Bayes's unsuccessful fifth proposition sought to do.

Our current ways of thinking about conditional probability seem very deeply entrenched, and the thesis that these ways of thinking need to be or even can be changed will leave many readers incredulous. Appendix I below supplies an historical perspective that may help dispel this incredulity. As we see there, the very term "conditional probability" first appeared in print only in 1928, and the notation $P(A \mid B)$ is even younger. Some similar notations were used earlier, but before Markov (1900) and Hausdorff (1901), students of the mathematical theory of probability did not have any symbol for the probability of one event relative to another that they were willing to use for arbitrary pairs of events. Our current ideas are not as old and may not be as permanent as their constant repetition in textbook after textbook makes them seem.

It should be noted that the present paper is not concerned with the parts of Bayes's essay that are most often discussed. That essay consists of two sections. Section I, which contains the third and fifth propositions, is a brief treatment of the basic principles of probability. Section II, the heart of the essay, gives Bayes's solution to the problem of statistical inference for the binomial distribution. Most twentieth-century discussion of the essay has focused on Section II. Harold Jeffreys (1939, page 33) and R. A. Fisher (1973, page 14) discussed aspects of Section I, but most modern authors seem to agree with Issac Todhunter (1865, page 295) that Section I is "excessively obscure" and pass it by. The present paper, in contrast, is concerned only with Section I.

2. Bayes's first six propositions. Bayes's exposition is indeed very obscure, especially to the modern reader. His most important assumptions are left implicit, and his notation seems inadequate and cumbersome. But a brief glance at his own words will serve as a valuable preface to our discussion of his thinking.

Reproduced below are the first six propositions from Section I of Bayes's essay, together with the definitions on which they are based. His proofs are omitted. Notice, when reading Definitions 3, 4, and 5, how Bayes puts the question of when events happen at the very base of his concepts. In Definition 5, for example, he makes it clear that the "value of a thing" depends on what events have happened at the time the value is to be computed.

In order to understand Bayes's meaning when he writes of the "ratio compounded of" two probabilities, as in Propositions 3 and 6, the reader must bear in mind that Bayes thinks of probabilities as ratios (Definition 5). The "ratio compounded of" two ratios is the product of those ratios.

For the full text of Bayes's essay, the reader is referred to pages 293–315 of the 1958 volume of *Biometrika*, where it was reprinted together with a biographical note by G. A. Barnard, or to the later reproduction of this reprinting in Pearson and Kendall (1970).

DEFINITION 1. Several events are *inconsistent*, when if one of them happens, none of the rest can.

2. Two events are *contrary* when one, or other of them must; and both together cannot happen.

3. An event is said to *fail*, when it cannot happen; or, which comes to the same thing, when its contrary has happened.

4. An event is said to be determined when it has either happened or failed.

5. The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening.

6. By *chance* I mean the same as probability.

7. Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest.

PROPOSITION 1. *When several events are inconsistent the probability of the happening of one or other of them is the sum of the probabilities of each of them.*

PROPOSITION 2. *If a person has an expectation depending on the happening of an event, the probability of the event is to the probablity of its failure as his loss if it fails to his gain if it happens.*

PROPOSITION 3. *The probability that two subsequent events will both happen is a ratio compounded of the probablity of the 1st, and the probability of the 2nd on supposition the 1st happens.* COROLLARY. *Hence if of two subsequent events the probability of the 1st be a/N, and the probability of both together be P/N, then the probability of the 2nd on supposition the 1st happens is P/a.*

PROPOSITION 4. *If there be two subsequent events to be determined every day, and each day the probability of the 2nd is b/N and the probability of both P/N, and I am to receive N if both events happen the first day on which the 2nd does; I say, according to these conditions, the probability of my obtaining N is P/b.*

PROPOSITION 5. *If there be two subsequent events, the probability of the 2nd b/N and the probability of both together P/N, and it being first discovered that the 2nd event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is P/b.*

PROPOSITION 6. *The probability that several independent events shall all happen is a ratio compounded of the probabilities of each.*

3. A first look at the third proposition. Bayes defines the probability of an event as the ratio of the value of a contract[2] that awards a prize if the event happens to the value

[2] Bayes calls a contract that awards a prize if an event happens an *expectation* depending on that event. Abraham DeMoivre had used *expectation* in this same sense in *The Doctrine of Chances*. But DeMoivre's language is less meticulous than Bayes's; after initially distinguishing between expectations and their values, DeMoivre quickly slips into using the phrase "a player's expectation" to mean the value of that expectation. This latter usage has now become the standard one. DeMoivre's use of *expectation* to mean a contract in a game of chance derives, presumably, from the similar use of the Latin *expectatio* in the Latin translation (1657) of Christian Huygens's Dutch treatise *Van Rekeningh in Spelen van Geluck*. Freudenthal (1980, page 114) and Hacking (1975, page 93–95) differ on whether the introduction of the term *expectatio* should be ascribed to Huygens or to Van Schooten, who completed the translation.

of the prize. This means that the probability of A, $P(A)$ in our notation, is equal to the value of a contract that pays \$1 if A happens. And it means that the probability of B after A has happened, $P(B \mid A)$ in our notation, is the value after A has happened of a contract that pays \$1 if B happens. It also means (when we suppose that the prize to be awarded if A happens is a contract that pays \$1 if B happens) that

$$P(A) = \frac{\text{value of a contract that awards, if } A \text{ happens, a contract that pays \$1 if } B \text{ happens}}{\text{value after } A \text{ has happened of a contract that pays \$1 if } B \text{ happens}}$$
$$= \frac{\text{value of a contract that pays \$1 if } A \text{ and } B \text{ both happen}}{\text{value after } A \text{ has happened of a contract that pays \$1 if } B \text{ happens}},$$

or

$$P(A) = \frac{P(A \cap B)}{P(B \mid A)}. \tag{3}$$

And (3) is equivalent, of course, to (1). Bayes's proof of his third proposition boils down to this calculation.

Why does Bayes, in the statement of his third proposition, refer to the two events he is considering as "subsequent"?

Imagine a situation in which it is not known beforehand which of two events A and B will happen (or fail) first. In such a situation we cannot say beforehand what the probability of B will be immediately after A happens; that probability will be one if B has already happened by then, zero if B has already failed by then, and something in between if B has not yet happened or failed. Saying that B is subsequent to A may be an attempt on Bayes's part to resolve this ambiguity.

But the attempt is not fully successful. Once we admit that the timing of events may be contingent, we must guard not only against the possibility that B itself may or may not have happened by the time A happens but also against the possibility that other events affecting the probability of B may or may not have happened by then.

Once we have seen, in Section 5 below, how to take the order of events into account using rooted trees, it will become clear that what we need to assume in order to assure uniqueness for the probability of B immediately after A has happened is that there is only one way A can happen—i.e., only one possibility for what other events have happened at the point where A has just happened. This, it turns out, is equivalent to saying that when A happens it is the most specific event that happens. I call an event with this property *exact*. I conclude in Section 5 that Bayes's argument for his third proposition is sound if and only if A is an exact event.

4. The fifth proposition. In the argument leading up to his fifth proposition, Bayes introduces a new element into his reasoning. He is now concerned not just with the order in which events happen but also with the order in which we learn about their happening; the fifth proposition concerns the case where the second order reverses the first.

Let us again denote our two "subsequent events" by A and B. The first step in Bayes's argument for his fifth proposition is to embed A and B in an infinite sequence of similar pairs of events, say (A_1, B_1), (A_2, B_2), $\cdots$, where $A_1 = A$ and $B_1 = B$. Bayes assumes that the pairs are independent and that $P(B_i) = P(B)$ and $P(A_i \cap B_i) = P(A \cap B)$ for all i. In his fourth proposition, he deduces from these assumptions that

$$P(E) = \frac{P(A \cap B)}{P(B)},$$

where

$$E = \cup_{i=1}^{\infty} [(A_i \cap B_i) \cap (\cap_{j=1}^{i-1} \bar{B}_j)].$$

The event E is the event that A_i happens the first time B_i does; more precisely, it is the event that A_I happens, where I is the smallest value of i for which B_i happens. (The problem of proving Bayes's fourth proposition is discussed in Appendix II.)

Suppose now that we learn that B (which is the same as B_1) has happened, without yet learning whether A (which is the same as A_1) or any of the other A_i or B_i have happened or failed. In light of this new knowledge, A and E are equivalent and should be given the same probability. So if we do not change $P(E)$, it will become our new probability for A. In other words, (2) will hold.

Bayes's argument for keeping $P(E)$ unchanged after we learn B has happened is based on the idea that someone who follows a policy of always changing $P(E)$ after learning B_i has happened might suffer an infinite sequence of losses.

Consider a contract that pays \$1 if E happens—in Bayes's terminology, an expectation of receiving \$1 depending on E. The initial value or fair price of this contract is $\$P(E)$. Suppose I pay this price for it. And then I learn that B has happened. What can I infer from this new knowledge? According to Bayes,

> ... I can only infer that the event is determined on which my expectation depended,[3] and have no reason to esteem the value of my expectation either greater or less than it was before. For if I have reason to think it less, it would be reasonable for me to give something to be reinstated in my former circumstances, and this over and over again as often as I should be informed that the 2nd event [B_i] had happened, which is evidently absurd. And the like absurdity plainly follows if you say I ought to set a greater value on my expectation than before, for then it would be reasonable for me to refuse something if offered me upon condition I would relinquish it, and be reinstated in my former circumstances; and this likewise over and over again as often as (nothing being known concerning the 1st event) it should appear that the 2nd had happened.

It appears, from Bayes's use of a similar phrase in the proof of his Proposition 4, that when he writes about my paying to be "reinstated in my former circumstances," he means that I should pay to have my contract changed from one that pays \$1 if E happens to one that pays \$1 if

$$E_1 = \cup_{i=2}^{\infty} [(A_i \cap B_i) \cap (\cap_{j=2}^{i=1} \bar{B}_j)]$$

happens. (E_1 is the event that A_i happens the first time B_i does, starting with $i = 2$.) If I do this, and act similarly every time I learn that a B_i has happened before learning whether the corresponding A_i has happened, then I may end up paying over and over again. Bayes seems to feel that this possibility makes my willingness to pay absurd.

But why should I suppose, or even particularly fear, that I will always learn about a B_i happening before learning whether the corresponding A_i has happened? It makes just as much sense to worry that it is only when A_i has failed and B_i has happened that I shall be informed of B_i's happening before being informed whether A_i has happened.

Here, no doubt, is the crux of the matter. What rules govern my discovery of B's happening? Under what conditions will B's happening or something different be revealed to me? Has someone arranged, diabolically, that B's happening should be revealed to me only if B happens and A fails? Or, at the opposite extreme, is the discovery of whether B has happened or failed itself a chance event independent of the events A and B? Only if these questions are answered is there, it seems to me, any clear answer as to how A's probability should change when B is discovered to have happened.

It may clarify matters to reformulate Bayes's argument so as to eliminate the infinite sequence of similar pairs of events. This infinite sequence is of rhetorical value to Bayes; it enables him to raise the specter of an infinite sequence of futile payments. But it obscures the fundamental issues. And we can remove it from the argument easily enough; we can simply replace the prospect of "being reinstated in one's former circumstances" if B fails by the prospect of receiving an equivalent sum of money if B fails. Consider, indeed, a

[3] Recall that B is subsequent to A. So from knowledge that B has happened we can infer that A has been determined—i.e., either happened or failed.

contract that pays \$1 if both A and B happen and \$$x$ if B fails. Initially, the total expected value of this contract is \$$P(A \cap B) + \$xP(\bar{B})$. If $P(B) > 0$, then the value of x can be chosen so that this total is equal to x. Indeed, the equation

$$P(A \cap B) + xP(\bar{B}) = x$$

is satisfied by setting $x = P(A \cap B)/P(B)$. Suppose x does equal $P(A \cap B)/P(B)$, and suppose I have paid this much for the contract. When I learn B has happened without learning whether A has happened, my contract becomes, in effect, a contract that pays \$1 if A has happened, and so my new probability for A should equal its value. Bayes would contend, presumably, that this value should remain equal to $x = P(A \cap B)/P(B)$.

(If I pay \$$x$ for a contract that returns \$$x$ if $\bar{B}$ happens and \$1 if $A \cap B$ happens, then I have made what Bruno de Finetti calls a bet on A conditional on B. So in this revised form Bayes's argument is closely related to arguments presented by de Finetti (1964) and Ramsey (1931). For a more direct discussion of de Finetti's argument, see Section 3 of Shafer (1981).)

But why should the value of this particular contract be unchanged? We might argue for Bayes by pointing out that the value of our expectation before we found out B's happening was x, and that it would have continued to be x had things gone the other way—i.e., had we found out that B failed, for in that case we would know we are due to receive \$$x$. So symmetry would seem to demand that the value should still be x. But this argument is fallacious, for it is based on the unwarranted assumption that B's happening and B's failing were the only possibilities for what we might have found out.

So we come once again to the crux of the matter: Bayes's fifth proposition seems unconvincing unless we assume foreknowledge of the conditions under which the discovery of B's having happened will be made.

5. A mathematical formulation of the ideas involved in the third proposition. In this section we use rooted trees as a framework for describing the step-by-step determination of events. Within this framework we formulate explicit postulates about values, and we use these postulates to deduce versions of Bayes's first, second, third, and sixth propositions.

5.1. *Framework.* In order to describe mathematically an experiment whose outcome is determined and revealed all at once, it suffices to consider the set of all possible outcomes of the experiment, say Θ, and to call the subsets of Θ *events*. But this simple mathematical structure is not adequate to describe the step-by-step unfolding of events; it does not enable us, for example, to express mathematically the idea that an event A necessarily happens or fails before a second event B happens or fails.

Elementary graph theory provides a simple mathematical structure that can represent the step-by-step unfolding of events: the *rooted tree*. A rooted tree is a connected graph that has no cycles and one of whose nodes is singled out and called the root, or initial node. (See, for example, Marshall, 1971, page 20). There are many books on graph theory that discuss rooted trees, but there seems to be little consensus on terminology.) Such a tree is usually drawn "upside-down," with the initial node at the top, as in Figure 1. We can draw such a tree for any experiment or process whose outcome unfolds step-by-step; the unfolding corresponds to beginning at the initial node and moving down the tree step-by-step, deciding at each node which fork to take next.

Notice that every non-initial node in a rooted tree is the lower node of exactly one fork. In order to avoid trivialities, let us assume that every non-terminal node in the rooted tree with which we deal is the upper node of at least two forks.

Let us say that a node p in a rooted tree is *below* a node q if p comes after q on a path down the tree from the initial node. Given two distinct nodes p and q, exactly one of the following must be true: p is below q, q is below p, or there is no path down the tree that passes through both p and q.

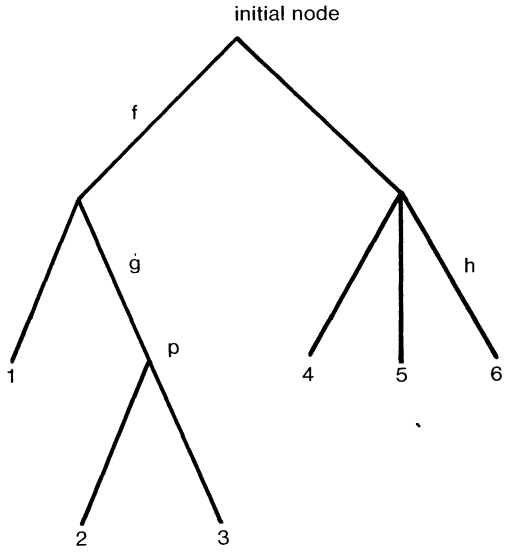

FIG 1. *A rooted tree with six terminal nodes.*

Let us assume, for simplicity, that we are dealing with a process of finite complexity, so that our rooted tree has only finitely many nodes. Then every path down the tree eventually arrives at a terminal node, and the terminal node finally arrived at uniquely identifies the path that was taken. So we may identify Θ, the set of possible outcomes of the process, with the set of all terminal nodes.

We may still call subsets of Θ *events*. And now we can talk about when events happen. Let us say that an event A *happens* when a fork is taken that forces the path down the tree to terminate in a node in A. Let us say that it *fails* when $\bar{A} = \Theta - A$ happens. And let us say that A is *determined* at a node n if either (i) A has happened or failed at some fork above n or (ii) $A = \Theta$ or $A = \varnothing$. (See Bayes's third and fourth definitions.) According to this terminology, the events Θ and $\varnothing$ never happen or fail, though they are always determined. If A is equal neither to Θ or to $\varnothing$, then A will have happened or failed by the time we complete a path down our rooted tree.

Usually several events will happen when a given fork is taken. When fork g in Figure 1 is taken, for example, the events $\{2, 3\}$, $\{2, 3, 4\}$, $\{2, 3, 5\}$, $\{2, 3, 6\}$, $\{2, 3, 4, 5\}$, $\{2, 3, 4, 6\}$, $\{2, 3, 5, 6\}$, and $\{2, 3, 4, 5, 6\}$ all happen. Notice that $\{1, 2, 3\}$ does not happen when g is taken; it happens earlier, when f is taken.

Given a node p other than the initial node, let us denote by E_p the event consisting of the terminal nodes of all the paths that go through p. (For the node labeled p in Figure 1, $E_p = \{2, 3\}$. If p is itself a terminal node, then $E_p = \{p\}$.) Let us call E_p the *exact event* that happens as the fork whose lower node is p is taken; E_p evidently does happen as this fork is taken, and any other event that happens at the same time is necessarily larger—i.e., less specific.

Two exact events E_p and E_q are always either nested or disjoint: if p is below q, then $E_p \subset E_q$; if q is below p, then $E_q \subset E_p$, and if there is no path down the tree that passes through both p and q, then $E_p \cap E_q = \varnothing$. It follows that if A is a proper non-empty subset of Θ, then the exact events corresponding to forks where A happens form a disjoint partition of A. (*Proof*: It is clear that every exact event corresponding to a fork where A happens is contained in A. And every element θ of A must be contained in one of these exact events, for A must happen somewhere on the path down to θ. So A is the union of these exact events. Since A cannot happen twice on the same path down the tree, these events must be disjoint.) We may call this partition the *canonical disjoint partition* of A.

An event is exact if and only if there is one and only one fork where it can happen. (If and only if, that is to say, its canonical disjoint partition has exactly one element—the event itself.) This is the precise meaning of the statement in Section 3 above that the exact events are those that can happen in only one way.

Let $[T, A]$ denote a contract whereby I am to receive the prize T if the event A happens and nothing if it fails. We shall suppose that if A happens then T is to be paid to me at a settling-up time that comes after the path down the rooted tree has been completed.

We will be interested primarily in contracts where the prize to be awarded is a sum of money—i.e., elements of the set

$$\xi = \{[\$r, B] \mid r \geq 0, B \subset \Theta\}.$$

Notice that a contract that awards an element of ξ as a prize if an event A happens is itself, in effect, an element of ξ. The contract $[[\$r, B], A]$, for example, amounts to a contract that pays $\$r$ if A and B both happen. In symbols $[[\$r, B], A] = [\$r, A \cap B]$.

(Readers who prefer more formal mathematical definitions are invited to think of ξ as the set of all real-valued functions on Θ that take at most two values, one zero and the other non-negative, and to interpret $[,]$ as a mapping from $(\xi \cup \mathbb{R}_0^+) \times 2^\Theta$ into ξ, where $\mathbb{R}_0^+$ denotes the set of all nonnegative real numbers.)

Contracts can be added together, and sometimes the sum of two elements of ξ is itself an element of ξ:

$$[\$r, A] + [\$s, A] = [\$(r + s), A]; \tag{4}$$

and if $A \cap B = \varnothing$, then

$$[\$r, A] + [\$r, B] = [\$r, A \cup B]. \tag{5}$$

Notice that (5) depends on the convention that the $\$r$ owed me under the contract $[\$r, A]$ if A happens is to be paid at a settling-up time after all events have been determined; it might not be true if the $\$r$ were to be paid immediately upon A's happening.

Let us assume that to every node n of our rooted tree and every element e of ξ there is assigned a non-negative real number $v_n(e)$, called the *value of e at n*. We shall abbreviate the symbol $v_n([e, A])$ to $v_n[e, A]$. Let us assume that the numbers $v_n(e)$ satisfy the following postulates:

I. If $r \geq 0$, then $v_n[\$r, \Theta] = r$.
II. If $E_n \cap A = \varnothing$, then $v_n[e, A] = 0$.
III. If e_1, e_2, and e are elements of ξ and $e_1 + e_2 = e$, then $v_n(e_1) + v_n(e_2) = v_n(e)$.
IV. If $e \in \xi$, $v_p(e) = r$, and n is above p, then $v_n[e, E_p] = v_n[\$r, E_p]$.

Postulate I merely says we are measuring value in dollars. Postulate II says that the prospect of getting nothing is worth nothing. Postulates III and IV spell out the idea that when a thing has a definite monetary value, that amount of money is fully equivalent to it; the thing can be replaced by that amount of money insofar as it contributes to a larger whole (Postulate III), and it can similarly be replaced in its role as the prize awarded by a contract (Postulate IV).

LEMMA. *If $r \geq 0$ and $s \geq 0$, then*

$$v_n[\$r, A] + v_n[\$s, A] = v_n[\$(r + s), A]. \tag{6}$$

If $r \geq 0$ and $A \cap B = \varnothing$, then

$$v_n[\$r, A] + v_n[\$r, B] = v_n[\$r, A \cup B]. \tag{7}$$

If $r \geq 0$, then

$$v_n[\$r, A] = r \cdot v_n[\$1, A]. \tag{8}$$

If $e \in \xi$ and n is above p, then

(9) $$v_n[e, E_p] = v_p(e)v_n[\$1, E_p].$$

If $e \in \xi$, $v_p(e) = r$ for every node p that is the lower node of a fork where A happens, and n is a node where A is not yet determined, then

(10) $$v_n[e, A] = r \cdot v_n[\$1, A].$$

PROOF. We obtain (6) and (7) when we apply Postulate III to (4) and (5). We can derive (8) from (6), and (9) follows from (8) together with postulate IV. Finally, we can derive (10) from (9) if we first apply (7) to express both sides in terms of the canonical disjoint partition of A. To carry out this last derivation, let us denote the lower nodes of the forks where A happens by $p_1, \cdots, p_k$. Since A is not yet determined at n, none of the p_i are above n. And if there is no path that passes through both p_i and n, then $v_n[e, E_{p_i}] = 0$ by Postulate II. So

$$v_n[e, A] = \sum_{i=1}^{k} v_n[e, E_{p_i}] = \sum \{v_n[e, E_{p_i}] \mid 1 \le i \le k;\ p_i \text{ is below } n\}$$

$$= r \sum \{v_n[\$1, E_{p_i}] \mid 1 \le i \le k;\ p_i \text{ is below } n\} = rv_n[1, A]. \qquad \square$$

5.2. *Probability.* Given a node n and an event A, let us define the *probability of A at n*, denoted by $P_n(A)$, by setting $P_n(A) = v_n[\$1, A]$. The following proposition, which follows immediately from our lemma, relates this definition to Bayes's definition of probability.

PROPOSITION. *If $r > 0$, then*

(11) $$P_n(A) = \frac{v_n[\$r, A]}{r}.$$

If $e \in \xi$, $A = E_p$, n is above p, and $v_p(e) > 0$, then

(12) $$P_n(A) = \frac{v_n[e, A]}{v_p(e)}.$$

If $e \in \xi$, $v_p(e) = r > 0$ for every node p that is the lower node of a fork where A happens, and n is a node where a is not yet determined, then

(13) $$P_n(A) = \frac{v_n[e, A]}{r}.$$

We are now in a position to state and prove versions of Bayes's first three propositions:

PROPOSITION 1. *If $A \cap B = \varnothing$, then $P_n(A \cup B) = P_n(A) + P_n(B)$.*

PROPOSITION 2. *If $r > 0$, then*

$$\frac{P_n(A)}{P_n(\bar{A})} = \frac{v_n[\$r, A]}{r - v_n[\$r, A]}.$$

PROPOSITION 3. *If A is exact and is not yet determined at n, then*

(14) $$P_n(A \cap B) = P_n(A) \cdot P_p(B),$$

where p is the lower node of the unique fork where A can happen.

COROLLARY. *If the hypotheses of Proposition 3 hold and $P_n(A) > 0$, then*

$$P_p(B) = \frac{P_n(A \cap B)}{P_n(A)}.$$

PROOF. Proposition 1 follows directly from (7). To prove Proposition 2, replace $P_n(\bar{A})$ by $1 - P_n(A)$ and then apply (11). To prove Proposition 3, notice that since A is not yet

determined at n, p must be below n. So we can apply (9), with $e = [\$1, B]$. Since $[[\$1, B], E_p] = [\$1, E_p \cap B]$, this yields

$$v_n[\$1, E_p \cap B] = v_p[\$1, B]v_n[\$1, E_p],$$

which is equivalent to (14). □

It is of some interest to note that (14) holds not only in the case where A is not yet determined at n but also in the case when A has already failed before n; in this case both sides of (14) are zero. But (14) may fail if A has already happened before n. In this case $P_n(A \cap B) = P_n(B)$ and $P_n(A) = 1$, so that (14) reduces to $P_n(B) = P_p(B)$, which may be false; the probability of B may change as we move from p down to n.

Let us say that an event B is *subsequent* to an event A if B becomes determined (i.e., happens or fails) after A does no matter what path down the tree is taken. In other words, B is still undetermined at the lower node of every fork where A becomes determined. In Figure 1, for example, the event $B = \{2, 4\}$ is subsequent to the event $A = \{1, 2, 3\}$.

Suppose the event B is subsequent to the event A. We shall say that A and B are *independent at n* if $P_p(B) = P_n(B)$ for every node p that is below n and is the lower node of a fork where A happens or fails.

PROPOSITION. *Suppose B is subsequent to A.*

(i) *If A is already determined at n, then A and B are independent at n.*

(ii) *If A and B are independent at n, then they remain independent at any lower node m; in fact, if A is not yet determined at m, then* $P_m(B) = P_n(B)$.

(iii) *If A and B are independent at n, then*

$$P_n(A \cap B) = P_n(A)P_n(B). \tag{15}$$

PROOF. (i) If A is already determined at n, then there are no forks below n where A happens or fails. (ii) Suppose A and B are independent at n, and suppose m is a node below n where A is not yet determined. Let $p_1, \cdots, p_k$ denote the nodes below m which are lower nodes of forks where A happens or fails. Then $P_{p_i}(B) = P_n(B)$ for $i = 1, \cdots, k$. And $E_{p_1}, \cdots, E_{p_k}$ is a disjoint partition of E_m. So

$$P_m(B) = P_m(B \cap E_m) = \sum P_m(B \cap E_{p_i}) = \sum P_m(E_{p_i})P_{p_i}(B)$$

$$= \sum P_m(E_{p_i})P_n(B) = P_m(E_m)P_n(B) = P_n(B).$$

(iii) If A is already determined at n, then (15) is obvious; both sides are equal to $P_n(B)$ if A has already happened, to zero if A has already failed. Suppose A is not yet determined at n. Then if A is exact, say $A = E_p$, the proof of (15) is simple: $P_n(A \cap B) = P_n(A)P_p(B)$ by (14), and $P_p(B) = P_n(B)$ by independence. If A is not exact, then we must consider the lower nodes $p_1, \cdots, p_k$ of forks below n where A can happen and write

$$P_n(A \cap B) = \sum P_n(E_{p_i} \cap B) = \sum P_n(E_{p_i})P_n(B) = P_n(A)P_n(B). \qquad \square$$

The following proposition generalizes (15) to the case of sequence of more than two events. This proposition can be thought of as a version of Bayes's Proposition 6.

PROPOSITION. *Consider a sequence* $A_1, \cdots, A_k$ *of events such that* A_j *is subsequent to and independent at n of* A_i *whenever* $1 \leq i < j \leq k$.

(i) A_i *is subsequent to and independent at n of* $A_1 \cap \cdots \cap A_{i-1}$, *for* $i = 2, \cdots, k$.

(ii) $P_n(A_1 \cap \cdots \cap A_k) = P_n(A_1) \cdots P_n(A_k)$.

PROOF. (i) is evident once we recognize that any fork where $A_1 \cap \cdots \cap A_{i-1}$ happens is a fork where A_{i-1} happens, and that any fork where $A_1 \cap \cdots \cap A_{i-1}$ fails is a fork where A_j fails, for some j less than or equal to $i - 1$. Given (i), (ii) follows from (15) by induction. □

5.3. *Discussion.* Bayes takes it for granted that there exists a numerical measure of value; he writes without elaboration about "an expectation of receiving N," N being a number. I have assumed, for simplicity, that value is measured in dollars. Many readers will not agree that values are additive (Postulate III) when measured in dollars but will accept postulates that imply the existence of an abstract scale on which values are additive. (See, for example, Savage, 1954, or Krantz et al., 1970, Chapter 5.) It would be interesting to see such postulates formulated within the framework of the preceding section and related to the theory developed there, but it seems wisest not to undertake this task in the present paper.

Bayes also uses the word "thing" in a general and imprecise way; he refers to the "value of a thing" without saying just what things have values. The simplest resolution of this ambiguity, and the one I have followed, is to assume only that elements of ξ (expectations of receiving sums of money) have values. It would do no harm to the mathematical development to assume the existence of values for a broader class of "things."

According to Bayes's definition of probability, the probability of an event A is the ratio of the value of an expectation depending on the happening of A to the "value of the thing expected upon its happening." Suppose the thing expected is an element e of ξ. Then in terms of our rooted tree, this definition says that

$$P_n(A) = \frac{v_n[e, A]}{\text{value of } e \text{ immediately after } A \text{ happens}}. \tag{16}$$

This makes sense if A is exact, for then the "value of e immediately after A happens" obviously means $v_p(e)$, where p is the lower node of the unique fork where A can happen, and so (16) is the same as (12). But if A is not exact there will be several forks where A can happen, and the value of e may not be the same at all the lower nodes of these forks. To make sense of Bayes's definition we must stipulate that e does have the same value at all these nodes, thus reducing (16) to (13). The simplest way to do this is to set $e = [\$r, \Theta]$, thus reducing (16) to (11).

In our version of Bayes's third proposition, it is assumed that the event A on which we are "conditioning" is exact. Is it possible to make sense of Proposition 3 and Bayes's own proof of it if we do not assume that A is exact? I think not. In his proof Bayes considers the expectation $[\$N, A \cap B]$. He denotes the present value of $[\$N, A \cap B]$ by P and the value of $[\$N, B]$ upon A's happening by b ("the value of my expectation ... will become b"). But can we take it for granted that there is a unique value b that the expectation $[\$N, B]$ can come to have upon A's happening? Only if A is exact. Unfortunately, Bayes did not formulate the idea of an exact event.

Notice that our proof of Proposition 3 makes no use of the idea that B is subsequent to A. And in fact the conclusion of the proposition does not depend on B being subsequent to A. Equation (14) is valid even if B is already determined at the unique p where the exact event A can happen. If B has already happened, both sides equal $P_n(A)$; if B has already failed, both sides equal zero.

When we turn to independence, we find an opposite contrast; Bayes's definition of independence says nothing about events being subsequent, but our translation of this definition into the language of rooted trees applies only to events that are subsequent. I am uncertain how successfully this definition can be generalized to apply to events that are not subsequent. But such a generalization does not appear necessary in the context of Bayes's essay. The independent events of interest to Bayes were successive trials in a binomial experiment, and these are certainly subsequent.

In his sixth proposition, Bayes asserts that the probability of several independent events all happening is the product of their probabilities. This means, in the case of three events A, B and C, for example, that

$$P_n(A \cap B \cap C) = P_n(A)P_n(B)P_n(C).$$

When we turn to Bayes's proof, we find that he tacitly assumes that these three events are subsequent; he labels them 1st, 2nd, and 3rd. It also turns out that his proof, since it

appeals to his third proposition, is valid only if A and $A \cap B$ are exact. But, as we saw in the preceding section, it is easy enough to extend the proof to cover the case where these events are not exact.

It is of some interest to note that Bayes uses Proposition 3 only in proving Proposition 6. So his restriction of Propositon 3 to subsequent events, though unnecessary, is not entirely inappropriate.

It is evident from this discussion that Bayes did not have clearly in mind the mathematical structure we have developed using rooted trees. But it is also clear that Bayes's reasoning cannot be made clear and rigorous without the use of some formulation that takes the order of events into account. And it seems likely that any successful formulation will be very much like the one we have developed.

6. The step-by-step development of our knowledge. From a mathematical point of view, the framework we have developed for representing the step-by-step determination of events is clear and simple. But there remains room for much discussion about what we mean by the "determination" of events.

One approach is to suppose that the step-by-step determination of events is entirely objective—entirely unrelated to the development of our knowledge. It is nature that moves down the rooted tree, and the probabilities that change as she does so are her probabilities, not ours. This approach is suggested by Bayes's statement of his fifth proposition, which seems to deny that the order in which events happen needs to be related in any particular way to the order in which we find out about their happening.

Another approach, suggested by examples where we watch the step-by-step playing out of a game of chance, is to suppose that we are indeed dealing with objective chance but that our knowledge keeps pace with events. Here we move down the rooted tree in tandem with nature, and the probabilities that change as we do so are both ours and nature's.

A third approach is to interpret the step-by-step determination of events in an entirely subjective way. This means requiring that the events represented by our rooted tree all be events *that we learn some given fact.* And it means that the tree itself must be part of our knowledge: at the initial node we assign probabilities not just to future possibilities as to what facts we might learn, but also to the various orders in which we might learn these facts. In this context, conditioning on an exact event means conditioning on all we have learned. And the justification for this conditioning is based on our assigning probabilities beforehand to the possibilities for how our knowledge might develop.

The subjective interpretation of the rooted tree is relevant, of course, to Bayes's fifth proposition. In Section 4 above we concluded that the fifth proposition is unconvincing precisely because it makes no assumptions about when and with what probabilities we might learn one thing rather than another. An attempt to repair this defect would likely lead to a subjective rooted tree. And the fifth proposition would then become merely a subjective version of the third.

APPENDIX I.

The Evolution of Conditional Probability. Abraham DeMoivre's work on games of chance seems to contain the first calculations of what we now call conditional probabilities. In DeMoivre's work, the timing of events is obviously important. DeMoivre was interested only in the question Bayes poses in his third proposition: how does the probability of a not-yet-determined event change when another event happens? He did not ask how our probability for one event might change when we learn of the happening of another that is not necessarily earlier in time.

(James Bernoulli, whose work on probability preceded DeMoivre's, was concerned with a different problem: the calculation of probabilities from the combination of arguments. Since the timing of events does not play a fundamental role in this problem, we might be tempted to look in Bernoulli's work for a concept of conditional probability that does not

require the event being conditioned on to have happened first. As it turns out, however, it is not possible to make much sense of Bernoulli's thinking in terms of conditional probability. The probabilities he calculated were non-additive, and his rules for combining arguments can be better understood by comparing them to Dempster's rule for combining belief functions than by comparing them to the rule of conditioning. See Shafer, 1978.)

The work of Bayes and Laplace brought DeMoivre's original idea a step closer to the modern concept of conditional probability. For after their contributions, mathematical probability was concerned not only with the probability of an event given an earlier event but also with the probability of an event given an hypothesis. For the most part, however, nineteenth century authors did not unify these two ideas. They did not regard them as special cases of a single concept of conditional probability, and they did not have a notation, like our $P(A \mid B)$, that encouraged them to do so.

A watershed is marked by the appearance in 1901 of a note by Felix Hausdorff. In this note, written in German, Hausdorff explicitly proposes the unification of disparate nineteenth-century ideas into a single concept of "relative probability" and proposed that the relative probability of E given F, for arbitrary events E and F, should be denoted by $p_F(E)$.

Hausdorff's proposal of the notation $p_F(E)$ was almost simultaneous with the publication, in 1900, of A. A. Markov's textbook on probability, *Ischislenie Veroyatnostii.* In this textbook, Markov denoted the probability of A given B by (A, B). Markov's textbook was quite influential; it was translated into French and German and went through several editions. We find the notation (A, B) as late as 1937, in J. V. Uspensky's *Introduction to Mathematical Probability.* Hausdorff's recommendation seems to have been more influential, however. Emanuel Czuber, who had followed the nineteenth-century tradition in his 1902 textbook, acknowledged Hausdorff's influence when he introduced the term "relative probability" and the notation $W_E(F)$ in the second edition, published in 1908. And we find $P_E(F)$ being used by Thornton C. Fry in 1928 and by A. N. Kolmogorov in 1933.

Though Markov and Hausdorff seem to have been the first to have used an explicit notation for the probability of one event given another in the literature on mathematical probability, such notations were used in the late nineteenth century by English-speaking logicians. According to John Maynard Keynes (1921, page 155), Hugh McColl was the first to use such notation. In 1880, McColl used x_a to denote the chance that x is true on the assumption that a is true; in 1897, he used $\frac{A}{B}$ to denote the chance that A is true on the assumption that B is true. Keynes himself used a notation similar to this second one; he used a/h for the probability of a given h. B. I. Gilman (1883) used $[a, b]_b$ for the "probability that an event of genus b will also be of the species a."

The current notation $P(A \mid B)$ seems to be due to Harold Jeffreys. In 1919, Dorothy Wrinch and Jeffreys used $P(p : q)$ for the probability of p given q; Jeffreys changed this to $P(p \mid q)$ in his *Scientific Inference* (1931, page 15).

The earliest use of the term "conditional probability" that I have seen is in Thornton C. Fry's 1928 textbook (see especially pages 43–44). A. N. Kolmogorov, in his *Grundbegriffe* (1933), used the German equivalent: "bedingte Wahrscheinlichkeit." Most nineteenth-century authors in English, French, German, and Russian found no need for such a term. Whitworth (1886, page 192), for example, refers simply to the probability of B "when A has happened." We do find the term "relative probability" in the French authors Lacroix (1816, page 20) and Liagre (1852, page 30). As I have already noted, Hausdorff recommended this term in 1901. Neyman used it in English as late as 1952.

Though nineteenth-century English writers seem not to have used the terms "conditional probability" or "relative probability," they did use the adjectives "conditional" and "relative" when they were making the general point that probability is relative to one's state of information. See Whitworth (1886, page 127) and Keynes (1921, pages 90–91).

APPENDIX II

The Mathematics of Bayes's Fourth Proposition. Consider an infinite sequence (A_1, B_1), (A_2, B_2), $\cdots$ of mutually independent pairs of events such that $P(B_i) = r$ and $P(A_i \cap B_i) = s$ for all i and j. Set

$$E = \cup_{i=1}^{\infty} ((A_i \cap B_i) \cap (\cap_{j=1}^{i-1} \bar{B}_i)).$$

Then

$$P(E) = \sum_{i=1}^{\infty} P(A_i \cap B_i \cap (\cap_{j=1}^{i-1} \bar{B}_i)) = \sum_{i=1}^{\infty} P(A_i \cap B_i) P(\cap_{j=1}^{i-1} \bar{B}_i)$$

$$= \sum_{i=1}^{\infty} s(1-r)^{i-1} = \frac{s}{r}.$$

This is Bayes's fourth proposition, as a modern student of the mathematical theory of probability might state and prove it.

The calculation in the preceding paragraph might seem to depend on the assumption of countable additivity for probabilities, but it is easy to see that its conclusion is valid even if the probability measure P is only finitely additive. Indeed, since

$$E \supset \cup_{i=1}^{k} [A_i \cap B_i \cap (\cap_{j=1}^{i-1} \bar{B}_j)],$$

the calculation yields $P(E) \geq (s/r)\ \{1-(1-r)^k\}$ for every positive integer k even if P is only finitely additive. So $P(E) \geq s/r$. And since

$$\bar{E} \supset \cup_{i=1}^{k} [\bar{A}_i \cap B_i \cap (\cap_{j=1}^{i-1} \bar{B}_j)],$$

a similar calculation yields $P(\bar{E}) \geq 1 - s/r$. So $P(E) = s/r$.

Bayes's own proof of his fourth proposition is based on a comparison of E with

$$E_1 = \cup_{i=2}^{\infty} [(A_i \cap B_i) \cap (\cap_{j=2}^{i-1} \bar{B}_j)].$$

Bayes takes it for granted that E_1 is independent of the pair (A_1, B_1) and that the probability of E_1 is the same as that of E. His argument amounts to a simple calculation; he puts it in terms of values, but it can easily be expressed in terms of probabilities:

$$E = (A_1 \cap B_1) \cap (\bar{B}_1 \cap E_1),$$

or

$$P(E) = P(A_1 \cap B_1) + P(\bar{B}_1 \cap E_1) = P(A_1 \cap B_1) + P(\bar{B}_1) P(E_1),$$

or

$$P(E) = P(A_1 \cap B_1) + \{1 - P(B_1)\} P(E),$$

whence

$$P(E) = \frac{P(A_1 \cap B_1)}{P(B_1)}.$$

In order to establish the fourth proposition in the context of rooted trees we would need, of course, to extend and modify our definitions to accommodate the case of an infinite rooted tree. There are no great difficulties involved in this extension, but it does not seem necessary to carry it out here.

Acknowledgments. I am indebted to Ian Hacking, David Krantz, Fred Mosteller, Steve Stigler, and Sandy Zabell for advice and references.

REFERENCES

BARNARD, G. A. (1958). Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika* **45** 293–315.

BAYES, THOMAS (1764). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London* for 1763 **53** 370–418. Reprinted by Deming (1940), by Barnard (1958), and by Kendall and Pearson (1970).

CZUBER, EMANUEL (1902). *Wahrscheinlichkeitsrechnung*. Teubner, Leipzig and Berlin.

CZUBER, EMANUEL (1908). *Wahrscheinlichkeitsrechnung*, 2nd ed., Vol. 1, Teubner, Leipzig and Berlin.

DEMING, W. EDWARDS, ed. (1940). *Facsimiles of two papers by Bayes*. The Graduate School, United States Department of Agriculture, Washington.

DEMOIVRE, ABRAHAM (1718). *The Doctrine of Chances*. Pearson, London.

DEFINETTI, BRUNO (1964). Foresight: its logical laws, its subjective sources. Pages 93–158 of *Studies in Subjective Probability*. Kyburg and Smokler (eds.), Wiley, New York.

FISHER, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd ed., Hafner, New York.

FREUDENTHAL, HANS (1980). Huygens' foundations of probability. *Historia Mathematica* **7** 113–117.

FRY, THORNTON C. (1928). *Probability and Its Engineering Uses*. Van Nostrand, New York.

GILMAN, B. I. (1883). Operations in relative number with applications to the theory of probabilities. Pages 107–125 of *Studies in Logic*, C. S. Peirce, ed., Little Brown, Boston.

HACKING, IAN (1975). *The Emergence of Probability*. Cambridge University Press.

HAUSDORFF, FELIX (1901). Beiträge zur wahrscheinlichkeitsrechnung. *Berichte der Königlich Sächischen Gesellschaft der Wissenschaften zu Leipzig, mathematisch-physische Classe* **53** 152–178.

HUYGENS, CHRISTIAN (1660). *Van Rekeningh in Spelen van Geluck*. Reprinted together with a French translation in Vol. XIV of Huygens's *Oeurves Complètes* (The Hague, 1888–1950).

JEFFREYS, HAROLD (1931). *Scientific Inference*. Cambridge University Press.

JEFFREYS, HAROLD (1939). *Theory of Probability*. The Clarendon Press, Oxford.

KEYNES, JOHN MAYNARD (1921). *A Treatise on Probability*. Macmillan, London.

KOLMOGOROV, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin.

KRANTZ, DAVID, et al. (1971). *Foundations of Measurement*. Academic Press, New York.

LACROIX, S. F. (1816). *Traité élémentaire du calcul des probabilités*. Courcier, Paris.

LIAGRE, J. B. J. (1852). *Calcul des probabilités*. Jaman, Brussels.

MCCOLL, HUGH (1880). The Calculus of Equivalent Statements (fourth paper). *Proceedings of the London Mathematical Society* **XI** 113–121.

MCCOLL, HUGH (1897). The calculus of equivalent statements (sixth paper). *Proceedings of the London Mathematical Society* **XXVIII** 555–579.

MARKOV, A. A. (1900). *Ischislenie Veroyatnostii*. Saint Petersburg, Academy of Sciences.

MARSHALL, CLIFFORD W. (1971). *Applied Graph Theory*. Wiley-Interscience, New York.

PEARSON, E. S., and KENDALL, M. G. (1970). *Studies in the History of Statistics and Probability*. Hafner, Darien, Connecticut.

RAMSEY, FRANK PLUMPTON (1931). *The Foundations of Mathematics and Other Logical Essays*. Routledge and Kegan Paul, London.

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

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

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

TODHUNTER, ISSAC (1865). *A History of the Mathematical Theory of Probability*. Cambridge. Reprinted in 1949 by Chelsea, New York.

USPENSKY, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.

WHITWORTH, WILLIAM ALLEN (1886). *Choice and Chance*, 4th ed., Stechart, New York.

WRINCH, DOROTHY AND HAROLD JEFFREYS (1919). On some aspects of the theory of probability. *The London, Edinburgh, and Dublin Philos. Mag. and J. of Sci. (Sixth Series)* **38** 715–731.

DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66045