Reasoning to a Foregone Conclusion

Joseph B. KADANE, Mark J. SCHERVISH, and Teddy SEIDENFELD

When can a Bayesian select an hypothesis H and design an experiment (or a sequence of experiments) to make certain that, given the experimental outcome(s), the posterior probability of H will be greater than its prior probability? We discuss an elementary result that establishes sufficient conditions under which this reasoning to a foregone conclusion cannot occur. We illustrate how when the sufficient conditions fail, because probability is finitely but not countably additive, it may be that a Bayesian can design an experiment to lead his/her posterior probability into a foregone conclusion. The problem has a decision theoretic version in which a Bayesian might rationally pay not to see the outcome of certain cost-free experiments, which we discuss from several perspectives. Also, we relate this issue in Bayesian hypothesis testing to various concerns about "optional stopping."

KEY WORDS: Coherence; Finite additivity; Sequential tests; Stopping rules; Value of information.

1. INTRODUCTION

In a lively (1962) discussion of some foundational issues, several noted statisticians, especially L. J. Savage, focused on the controversy of whether an experimenter's stopping rule is relevant to the analysis of his or her experimental data. Savage wrote (1962, p. 18).

> The [likelihood] principle has important implications in connection with optional stopping. Suppose the experimenter admitted that he had seen 6 red-eyed flies in 100 and had then stopped because he felt that he had thereby accumulated enough data to overthrow some popular theory that there should be about 1 per cent red-eyed flies. Does this affect the interpretation of 6 out of 100? Statistical tradition emphasizes, in connection with this question, that if the sequential properties of his experimental programme are ignored, the persistent experimenter can arrive at data that nominally reject any null hypothesis at any significance level, when the null hypothesis is in fact true. These truths are usually misinterpreted to suggest that the data of such a persistent experimenter are worthless or at least need special interpretation; see, for example, Anscombe (1954), Feller (1940), Robbins (1952). The likelihood principle, however, affirms that the experimenter's intention to persist does not change the import of his experience.

The tradition Savage refers to is, in a paraphrase of Anscombe (1954), Feller (1940), Robbins (1952), and Cornfield (1970), captured by the following imaginary circumstance. Suppose that a statistician has his designs on rejecting the null hypothesis H_0: $\theta = 0$, that the mean of iid normal data is zero. The data have known unit variance. Fix k_α so that $k_\alpha/\sqrt{n}$ corresponds to the nominal rejection point in an α-level uniformly most powerful unbiased (UMPU) test of H_0 versus the composite alternative hypothesis H_0^c: $\theta \neq 0$, based on a sample of size n. Let H_t denote the simple hypothesis that $\theta = t$. Continue observing data until the sample average, $\bar{x}_n = (x_1 + \cdots + x_n)/n$, satisfies the inequality (1), then halt:

$$|\bar{x}_n| > k_\alpha/\sqrt{n}. \tag{1}$$

The likelihood principle entails that the statistician's intent to stop only when (1) obtains is irrelevant to the "evidential import" of the data for hypotheses about θ. (See Berger 1985, sec. 7.7 for a recent view.) If, contrary to traditional theory, significance is calculated independent of the stopping rule for the experiment—a mistake by traditional theory—then when the inquiry halts, H_0 has achieved an observed significance of α, or less. Moreover, given the truth of H_0, by the law of the iterated logarithm, with probability 1 the experiment terminates; that is, almost surely the inequality (1) is eventually satisfied.

In response to this tradition, Savage asserted (1962, p. 18)

> The true moral of the facts about optional stopping is that significance level is not really a good guide to "level of significance" in the sense of "degree of import," for the degree of import does depend on the likelihood alone, a theme to which I must return later in the lecture.

And later in the discussion (1962, p. 72),

> It is impossible to be sure of sampling until the data justifies an unjustifiable conclusion, just as surely as it is impossible to build a perpetual motion machine. After all, whatever we may disagree about, we are surely agreed that Bayes's theorem is true where it applies.

We begin our inquiry by reviewing some of the details for what we guess Savage meant as the evident import of Bayes's theorem for solving the problem of sampling to a foregone conclusion. Because we find that the Bayesian position on forgone conclusions is complicated by mathematical conditions that are important for statistics, we rehearse the following arguments, even though their conclusions have been in the literature before (see, e.g., Kerridge 1963).

Let $(S, \mathcal{A}, P)$ be a (countably additive) probability space, which we think of as the underlying joint space for all quantities of interest. Expectations are with respect to the probability P. Unconditional expectation is denoted by $E(\cdot)$, and conditional expectation given a random variable X is denoted by $E(\cdot|X)$. Let $(\mathcal{X}, \mathcal{B})$ and (Ω, τ) be measurable spaces where

X: $S \to \mathcal{X}$ is a random quantity to be learned,

Θ: $S \to \Omega$ is any random quantity,

and h: $\Omega \to \Re^*$ is an (extended) real-valued function whose expectation $E(h)$ exists.

Joseph B. Kadane is Leonard J. Savage Professor of Statistics and Social Sciences, Department of Statistics; Mark J. Shervish is Professor of Statistics, Department of Statistics; and Teddy Seidenfeld is Professor of Philosophy and Statistics, Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213. This research was supported by National Science Foundation Grants SES-9123370, DMS-9005858, DMS-9302557, and SES-9208942 and by Office of Naval Research Contracts N00014-89-J-1851 and N00014-91-J-1024.

Then the familiar law of total probability implies that

$$E[E(h(\Theta)|X)] = E(h(\Theta)) \tag{2}$$

(see, e.g., Ash 1972, T.6.5.4, p. 257).

This says, in short, that there can be no experiment with outcome X designed (almost surely with respect to P) to drive up or drive down the conditional expectation of h, given X. Of course, Equation (2) has no special logical dependence on Bayes's theorem, except that non-Bayesian statistical methods often begin with the claim that neither the "prior" expectation $E(h(\Theta))$ nor the "posterior" expectation $E(h(\Theta)|X)$ has objective status.

As an example, suppose that h is the indicator for an hypothesis H (an unobserved "event") in Ω; that is, $h(\Theta) = 1$ if $\Theta \in H, h(\Theta) = 0$ otherwise. Thus $E(h(\Theta))$ is the "prior" probability of H, denoted by $P(H) = p$. Let $X_1, X_2, \ldots$ be observations that become available sequentially. To consider experimental designs that mandate a minimum sample size, $k \geq 0$, define

$$N = \inf\{n \geq k\colon P(H|X_1, \ldots, X_n) \geq q\},$$

where $N = \infty$ if the set is empty. That is, N identifies the first point after the kth in the sequence of X_i observations when the "posterior" probability of H reaches q, at least. The event $N = \infty$ obtains when, after k-many observations, the sequence of conditional probabilities, $P(H|X_1, \ldots, X_n)$ $(n \geq k)$, all remain below q. We assume that $q > p$.

Let $F_{X_1,\ldots,X_n|N}(\cdot|N = n)$ denote the conditional cdf of $(X_1, \ldots, X_n)$, given that $N = n$. Then,

$$\begin{aligned} P(H) &\geq P(H, N < \infty) \\ &= \sum_{n=k}^{\infty} P(N = n)P(H|N = n) \\ &= \sum_{n=k}^{\infty} P(N = n) \int P(H|X_1 = x_1, \ldots, X_n = x_n) \\ &\quad \times dF_{X_1,\ldots,X_n|N}(x_1, \ldots, x_n|n) \\ &\geq \sum_{n=k}^{\infty} P(N = n) \int q\, dF_{X_1,\ldots,X_n|N}(x_1, \ldots, x_n|n) \\ &= qP(N < \infty). \end{aligned}$$

Hence

$$P(N < \infty) \leq p/q < 1. \tag{3}$$

Thus, when $p < q$, the prior probability is less than 1 that a Bayesian will halt the sequence of experiments and conclude that the posterior probability of H has risen to q (at least). Moreover, by Doob's martingale convergence theorem (thm. 7.4.3, 1953), $\lim_{n\to\infty}\{P(H|X_1, \ldots, X_n)\}$ converges (P almost surely). Thus for (P almost all) infinite sequences $(x_1, \ldots) \in (N = \infty)$, we have that $P(H|(x_1, \ldots)) \leq q$. Hence the prior probability is less than p/q a Bayesian will conclude that the posterior probability of H is more than q. This result obtains no matter which experimental design is adopted and allows for "infinite" samples, but it relies on Bayes's rule for updating a prior to a posterior.

The same argument provides a bound for the conditional probability of terminating the experiment in finite time, given that H is false. Assume that $P(H) < 1$ and write

$$P(N < \infty|\neg H) = \frac{P(\neg H|N < \infty)P(N < \infty)}{P(\neg H)}.$$

Note three facts: that $P(\neg H|N < \infty) \leq 1 - q$ by the design of the experiment, that $P(N < \infty) \leq p/q$ as just shown, and that $P(\neg H) = 1 - p$ by assumption. Then,

$$P(N < \infty|\neg H) \leq \frac{p(1-q)}{q(1-p)}. \tag{4}$$

For example, assume that $0 < p \leq .1$ and let $q/p = 10$. That is, by the choice of the stopping rule, if the experiment terminates, then the posterior probability for H increases tenfold (at least). Then the inequality (4) asserts the following: Given that H is false, the conditional probability of terminating the experiment is no more than $(.1 - p)/(1 - p) < .1$.

Savage (1962, pp. 72–73) offered a similar conclusion and illustrated with a simple case of two binomial hypotheses. Kerridge (1963) derived the same bounds as in (4) for the case of a uniform prior ($p = .5$). Cornfield (1970, pp. 20–21) used Kerridge's inequality to argue that as a Bayesian, you cannot be sure to defeat a true "null" hypothesis.

However, with extended Bayesian methods based on so-called "improper" priors, the situation is complicated. The following example reveals how one type of improper prior leads to a violation of (3) through the formal application of Bayes's theorem. Examples of this type exist with other sorts of data distributions as well.

Example 1.1. Let X and Y be independent Poisson random variables with X having mean θ and Y having mean $\lambda\theta$. Let the prior be the product of the "flat" improper prior, the Lebesgue measure for λ, and the proper prior gamma (2, 1) distribution for θ. Suppose that we observe $Y = y$ first. The product of the prior and likelihood densities is

$$f(y, \theta, \lambda) = \exp[-\theta(\lambda + 1)](\theta\lambda)^y\theta/y!.$$

The integral of $f(y, \theta, \lambda)$ with respect to λ is $g(y, \theta) = \exp[-\theta]$, and the integral of $g(y, \theta)$ with respect to θ is 1 for all y. Thus a formal application of Bayes's theorem with this improper prior for λ yields the gamma (1, 1) posterior (marginal) density of θ, given $Y = y$. That is, the posterior density $h(\theta|Y = y) = g(y, \theta) = \exp[-\theta]$, for all possible values y. Hence we know before observing Y that the posterior on θ will be different from the prior on θ and will not depend on the value of Y in fact observed. This is an illustration of a foregone conclusion.

Now consider the distribution of X. Prior to observing $Y = y$, we can calculate

$$P(X = 2) = \int_0^{\infty} \left(\frac{\theta^2}{2}\exp[-\theta]\right)(\theta\exp[-\theta])\, d\theta = \frac{3}{16}.$$

After observing $Y = y$, because X and Y are independent given the parameters, we have

$$P(X = 2|Y = y) = \int_0^\infty \left(\frac{\theta^2}{2}\exp[-\theta]\right)(\exp[-\theta])\,d\theta = \frac{1}{8},$$

no matter what value y is. That is, we know now that after we observe $Y = y$ (and assuming that we use the formal Bayes's theorem to update our beliefs), regardless of what y is, we will assign $\{X = 2\}$ a probability that is 2/3 the size of its prior probability. Thus we have reasoned to a foregone conclusion about the observable event $\{X = 2\}$ in addition to a foregone conclusion about the unobserved parameter θ.

Next, we begin an explanation of how improper priors can admit such reasoning to a foregone conclusion.

2. THE ROLE OF FINITE ADDITIVITY IN FOREGONE CONCLUSIONS

2.1 Finite Additivity Allows Experimentation to Foregone Conclusions

If the probability P is merely finitely additive and not countably additive, then (3) and (4) are not valid. That is, with a finitely additive P, it is possible for a Bayesian agent to design an experiment that surely terminates in foregone conclusions. The following illustrates the key phenomenon, what deFinetti (1974) called "non-conglomerability."

Let P be a (finitely additive) probability defined on the algebra $\mathcal{A}$. Denote expectation with respect to P by $\mathbf{E}_P[\cdot]$. Subject to the usual account of finitely additive conditional probability given a σ algebra, denote the P expectation given X by $\mathbf{E}_P[\cdot|X]$. (See Dubins 1975, sec. 3, and Schervish, Seidenfeld, and Kadane 1984, p. 213 for some discussion of existence of coherent conditional finitely additive probabilities.)

Definition (deFinetti). P is conglomerable in the (denumerable) partition $X = \{x_1, \ldots\}$ if for every bounded random variable Y and constants k_1 and k_2,

$$k_1 \le \mathbf{E}_P[Y] \le k_2 \quad \text{whenever} \quad k_1 \le \mathbf{E}_P[Y|X = x_i] \le k_2 \quad (i = 1, \ldots).$$

Example 2.1 (Dubins 1975). Let the set S be the product of a binary event $\{E, E^c\}$ and the countably infinite space of natural numbers, $X = \{1, 2, \ldots\}$. (We identify the events with their indicator functions.) Let the algebra $\mathcal{A}$ be the σ field of all subsets of S. Define a finitely additive P by $P(E) = P(E^c) = .5; P(X = i|E) = 2^{-i}$; and $P(X = i|E^c) = 0$. That is, given E, X is the random variable corresponding to the selection of an integer by flipping a fair coin until the first "head" results, assuming that eventually the coin lands heads up. Given E^c, X is a "uniformly" chosen positive integer. Then, in violation of conglomerability, $P(E) = .5$ yet $P(E|X = i) = 1$, regardless of the observed value $(i = 1, \ldots)$. The experimenter who has the personal probability P recognizes that, given each finite sequence of coin flips, E is practically certain.

When probability is updated by Bayes's rule of conditional probability, the experiment X makes E a foregone conclusion, contrary to the findings in Section 1. In explanation of what goes wrong with the reasoning from Section 1 applied to merely finitely additive probabilities, nonconglomerability creates a failure of the equality (2) with respect to the conditional probability $P(E|X = i)$. Specifically, $\mathbf{E}_P[E] = .5$, but $\mathbf{E}_P[E|X = x_i] = 1\,(i = 1, \ldots)$. Evidently, the door to foregone conclusions is opened whenever P is not conglomerable. Moreover, in this example the conditioning events all have positive probability, $P(X = i) = 2^{-(i+1)} > 0$. There is no issue of conditional probability given events of zero measure to bother with. Bayes's theorem applies without any extra conditions; nor is a special sequential argument needed. The "foregone conclusion" E is arrived at after a single trial of the experiment X.

In earlier work (Schervish et al. 1984) we investigated when, and by how much, finitely additive probabilities are nonconglomerable for events. In short, unless P is countably additive, there is a denumerable partition X and event E for which the conditional probability $P(E|X)$ is to one side of $P(E)$ and bounded away from it. If the data to be observed specify uniquely a member of that partition, then the anomalous behavior of Dubins' example is recreated. (Of course, when P is countably additive, conglomerability is satisfied in every denumerable partition.) Then, as merely finitely additive probabilities display nonconglomerability in predictable ways, are agents whose personal probabilities are not countably additive open to the criticism that they accept sampling to foregone conclusions? Contrary to the views of deFinetti (1974) and Savage (1954), is countably additivity a requirement for a coherent personal probability? Is countable additivity justified to avoid sampling to a foregone conclusion? In Sections 2.2, 2.3, and 2.4 we consider alternative views about finitely additive probability to determine whether they endorse reasoning to foregone conclusions.

2.2 Finite Additivity and the Value of Experimentation

Can it be that a Bayesian would rationally pay not to see the results of a cost-free experiment before making a decision? A classic result of Bayesian decision theory (Good 1967; Raiffa and Schlaifer, 1961, sec. 4.5.2; Ramsey 1990) is that cost-free evidence is worth waiting for in advance of making a terminal decision. If it does not cost anything to postpone a decision to conduct a (free) experiment, then, ex ante, delaying the decision to learn something new carries higher expected utility than choosing to act at once. The following is a formalization of this claim.

Suppose that a Bayesian agent has the opportunity now to postpone a terminal decision $D \in \mathcal{D}$, without penalty. Suppose also that the Bayesian agent has the chance to acquire, cost-free, new evidence X from a (finite) experiment before choosing among the same terminal decisions in $\mathcal{D}$. Assume that the agent uses Bayes's rule to update and knows this of himself or herself. (As is shown in Sec. 2.3, this turns out

to be an important premise for the conclusion that follows.) Then, the current value of deferring the terminal decision until after the experiment is observed is not less than the current value of making a terminal decision at once.

Example 2.2. Consider a binary decision problem: two terminal decisions, D_1 and D_2, on a binary state space E and its complement E^c. The terminal acts are functions from states to outcomes, defined in the usual way: $D_i(E) = o_{i1}$ and $D_i(E^c) = o_{i2}\,(i = 1, 2)$. The Bayesian agent carries one personal probability P over the events and one utility $U(o_{ij}) = u_{ij}$ over outcomes and seeks to maximize (conditional) expected utility. Assume, for convenience, that the agent judges the acts and states are probabilistically independent, $P(s_j|D_i) = P(s_j)$. Thus denote by $\mathbf{E}_P(u_{ij})$ the agent's expected utility for choosing $D_i, \Sigma_j P(s_j)u_{ij}$. In a terminal decision, now, the agent maximizes expected utility. Say that $\max_{\mathcal{D}} \mathbf{E}_P(u_{ij}) = k$.

Allow a binary experiment, $X = \{x_1, x_2\}$, followed by a choice between D_1 and D_2. (Because the outcome space is finite, there is no issue of countable additivity of P.) Assume that there is no cost for postponing the terminal decision—the utility of outcomes is unaffected by the experiment. The extensive form decision is depicted in Figure 1.

Denote by $\max_{\mathcal{D}} \mathbf{E}_{P(\cdot|x)} u_{ij}$ the maximum conditional expected utility of choosing between D_1 and D_2, given $X = x$. The following result was reported by Good (1967), Raiffa and Schlaifer (1961), and Ramsey (1990).

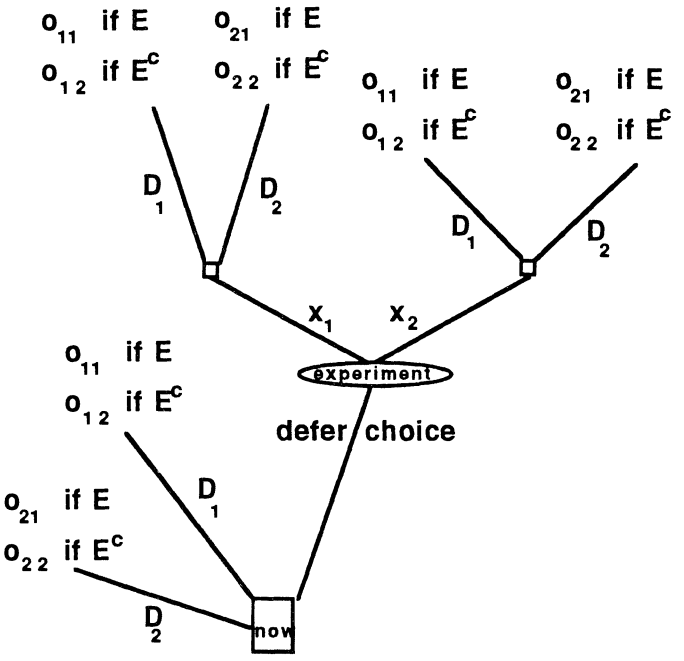

Figure 1. Extensive Form for the Sequential Decision Whether or not to Experiment and Observe, Cost Free, the Binary Random Variable X Prior to Choosing Between the Terminal Decisions D_1 and D_2. Following standard notation, boxes denote choice nodes and ovals denote chance nodes. The payoff structure of each terminal option is provided at the end of the corresponding branch of the decision tree. The first line gives outcome when event E occurs; the second line gives outcome when event E^c occurs.

Theorem 2.1

(a) $k \leq \mathbf{E}_P[\max_{\mathcal{D}} \mathbf{E}_{p(\cdot|x)} u_{ij}]$.

(b) Provided that no single act in $\mathcal{D}$ maximizes the conditional expected utility, $\mathbf{E}_{p(\cdot|x)} u_{ij}$ for each outcome x, then $k < \mathbf{E}_p[\max_{\mathcal{D}} \mathbf{E}_{p(\cdot|x)} u_{ij}]$.

Part (a) asserts that, ex ante, the value of waiting for the outcome X is not less than that of maximizing utility by choosing without first learning X. Part (b) asserts that, provided that the choice given X depends on how the experiment turns out, it is strictly better to await the new evidence before deciding between D_1 and D_2.

Of course, because X has only finitely many possible outcomes and because the state space associated with the terminal options is finite (E, E^c), the law of total probability obtains. Let i be fixed, corresponding to choosing a particular act D_i. Then $\mathbf{E}_P[\mathbf{E}_{P(\cdot|x)} u_{ij}] = \mathbf{E}_P u_{ij}$. In other words, there is no utility difference for the Bayesian between a simple sequential decision problem expressed in normal or in extensive form. However, when the experiment X is not simple and P is merely finitely additive, the law of total probability may fail and there is the possibility that the agent will prefer, strictly to decide without first learning X. In fact, the experiment X may carry negative value.

Example 2.2 (continued, itself a continuation of Example 2.1). Suppose that D_1 is a 2:1 bet on event E with stake 3 utiles, $u_{11} = 1, u_{12} = -2$. Suppose that D_2 is an even-odds bet against E with stake 2 utiles, $u_{21} = -1, u_{22} = 1$. What if the agent has the opportunity of postponing a terminal choice between these two options to learn the outcome of the experiment X? Then, ex ante, D_1 carries negative expected utility $(E_P(D_1) = -.5)$, whereas D_2 has expected utility zero. That is, D_2 maximizes expected utility, given what the agent knows now. However, given $X = x_i$ (because $P(E|x_i) = 1, i = 1, 2$), the conditional expected utility of D_1 is 1, whereas the conditional expected utility of D_2 is -1.

Thus the agent knows that if the choice to maximize expected utility is deferred until X is learned, then D_1 will be chosen for certain. But the current value (the value now) of deferring this sure choice of D_1 is $-.5$. Hence, ex ante, the value of waiting to see X and then choosing is $-.5$, whereas the value of choosing now (the value of choosing D_2) is zero. When finitely additive probability is updated by Bayes's rule, it may be better to avoid learning X than to postpone and face a foregone conclusion. The foregone conclusion leads to a choice now judged to be inferior to a current option.

There have been two important earlier attempts to mitigate the disruption to Bayesian theory caused by finite additivity. The next two sections take up each of these attempts in turn.

2.3 Some Consequences of Goldstein's Theory for Finitely Additive Probability

Here, we consider what Goldstein's (1983) theory has to say about the problem of reasoning to a foregone conclu-

sion with finitely additive probability. Specifically, we apply his theory to Example 2.1. We think that it is important to review Goldstein's work because his result, (5), appears to deny what we assert about the possibility of reasoning to a foregone conclusion when probability is only finitely additive. We show that Goldstein's theory requires giving up Bayes's conditioning as the rule for updating opinion and, in fact, any rule for updating that is a function of the data.

Let the current prevision for a quantity Z be denoted by $P_N(Z)$, where the subscript N indexes the time now. Also, let $P_F(Z)$ denote the future prevision for Z with respect to some well-defined future time F, later than N. Goldstein used sequential decision reasoning to argue that so-called "G coherence" of the previsions entails the equality

$$P_N(Z) = P_N(P_F(Z)). \tag{5}$$

That is, without regard for what one might learn between now and later, and without assuming how one updates future opinions, if each of the current and future previsions is G-coherent, then the current prevision of Z must equal the current prevision of future prevision of Z.

In short, Goldstein's result, (5), establishes that one cannot (now) expect that the future previsions will suffer a foregone conclusion, regardless of how one plans to stop experimenting. But Goldstein maintained that his theory is consistent with an agent holding (merely) finitely additive previsions, just as deFinetti allowed. How can this be when, as in Example 2.1, finitely additive probabilities open the door to foregone conclusions? The answer in short, we think, is that Goldstein's theory *prohibits* updating by Bayes's rule (in Eq. 2.1). Here is our analysis.

Suppose that we know now that we will observe X before we need to express $P_F(E)$, but that we are unsure what the total future evidence might be. We might also learn other information, but we are certain by the future time F to learn at least X and, we know this now, at N. In this case we can extend (5) to include some conditional previsions. Specifically,

$$P_N(E|X = k) = P_N(P_F(E)|X = k) \qquad (k = 1, \ldots) \tag{6}$$

follows within his theory, by the same reasoning that he uses to derive (5). Recall that in Example 2.1, the current probability for the event E is nonconglomerable in the X partition is $P_N(E) = .5$, yet $P_N(E|X = k) = 1\,(k = 1, \ldots)$. Then, by (6),

$$P_N(P_F(E)|X = k) = 1 \qquad (k = 1, 2, \ldots). \tag{7}$$

Because $0 \leq P_F(E) \leq 1$, (7) asserts that

$$(\forall k)(\forall \varepsilon > 0) \quad P_N(1 \geq P_F(E) > 1 - \varepsilon|X = k) = 1. \tag{8}$$

From (5) and (8), it follows that one now believes with probability .5 that $P_F(E)$ will be near zero.

Claim.

$$(\forall \varepsilon > 0) \quad P_N(P_F(E) < \varepsilon) = .5. \tag{9}$$

Proof. Note that

$$\begin{aligned}(\forall \varepsilon > 0) \quad & P_N(P_F(E) > 1 - \varepsilon) \\ & \geq \Sigma_k(P_N(P_F(E) > 1 \\ & \qquad - \varepsilon|X = k)P_N(X = k)) \\ & = \Sigma_k P_N(X = k) = .5.\end{aligned}$$

The inequality follows by subadditivity of finitely additive probability. The first equality follows from (8). But, by (5), $P_N(P_F(E)) = .5$; hence (for all $\varepsilon > 0$) $P_N(0 \leq P_F(E) < \varepsilon) = .5$.

To discuss the variety of rules permitted in Goldstein's theory for updating probabilities, suppose that between now, N, and the future, F, one learns *only* the new information X. This simplification avoids unnecessary complications about what later is the totality of the new evidence learned.

It is evident that, because of (5), in Goldstein's theory Bayes's rule is *not* (now) an authorized rule for updating your previsions at the later time F. Specifically, Bayes' rule mandates that when X is all the new evidence acquired between N and F,

$$P_F(E) = P_N(E|X). \tag{10}$$

But as $P_N(E|X = k) = 1\,(k = 1, \ldots)$, under Bayes's rule for updating, $P_F(E) = 1$. If one knows now that (10) applies, then $.5 = P_N(E) \neq P_N(P_F(E)) = 1$, contradicting (5). Hence, in Example 2.1, Bayes' rule is G incoherent for updating in light of the new information X. In fact, for this example, we show shortly that every rule that is a function of X is also G incoherent for updating.

Recall that we are examining the special case where one currently believes that X is *all* the relevant information that one will learn between now and the future time F when one must assess the prevision for event E. In that case, one currently believes (9); namely, that with probability .5, later one will depart in an extreme way from using Bayes's rule to update the prevision for E. Later, one will assign to E a prevision that is maximally far from what Bayes's rule prescribes.

If we suppose that there is a rule for determining $P_F(E)$ that is a function of X alone, then the event $\{P_F(E) < \varepsilon\}$ is also an event of the form $\{X \in B\}$, where B is a set of integers. (This is just what is ordinarily meant by saying that $P_F(E)$ is a function of X alone.) Because $P_N(X = k) > 0$ for all k, then for each $k \in B, P_N(P_F(E) < \varepsilon \,\&\, X = k) = P_N(X \in B \,\&\, X = k) = P_N(X = k) > 0$. But this last inequality implies that $P_N(P_F(E) < \varepsilon|X = k) = 1$, which contradicts (8).

For example, let D_m be the event $\{X > m\}$. It is easy to calculate that $P_N(E|D_m) = (1 + 2^m)^{-1}$. One might contemplate choosing some large integer m and setting $P_F(E) = P_N(E|D_m)$ if D_m occurs and setting $P_F(E) = P_N(E|D_m^c) = 1$ if the complement of $D_m, (D_m^c = \{X \leq m\})$ occurs. If $(1 + 2^m)^{-1} < \varepsilon$, then this rule corresponds to $\{P_F(E) < \varepsilon\} = \{X > m\}$, and (8) is violated for each value $k > m$. One might wish to take the (pointwise) limit of these rules as m goes to infinity. It is easy to see that

amounts to using Bayes's rule, which is also G incoherent. How strange it is that G coherence compels one to anticipate changing one's previsions in a manner that is not expressible as a function of the new evidence X that one will acquire between N and F. If it comes to pass that one does choose $P_F(E)$ close to zero, then we would be interested to hear an explanation of why one made such a choice.

The preceding example is typical of what happens in Goldstein's theory when a finitely additive probability is nonconglomerable in the margin of the data X. It is straightforward to show that if G coherence is to accommodate the simple situation where the data have positive probability, then either G coherence requires that finitely additive probabilities be conglomerable in the margin of the data to be observed, or it requires that future previsions not be a function of the data that we know now that we will observe. Next we turn to a theory of finitely additive probability that incorporates the first of these two alternatives.

2.4 Heath–Sudderth Coherence and Reasoning With Finitely Additive Probability

Heath and Sudderth (1978) proposed a theory of finitely additive probability based on their own account of (what we call) H–S coherence. Theirs is a *local* (rather than *global*) decision criterion involving wagers on select parameters of interest Θ and quantities to be observed X. An agent is required to indicate conditional fair odds over the sample space $\mathcal{X}$, given each $\theta \in \Theta$, and conditional fair odds over Θ, given each possible observation $x \in \mathcal{X}$. Usually, a statistical "model" fixes conditional probability over the sample space, $P(X|\theta)$, which determines the fair odds for X given Θ. Thus only the agent's "posterior" odds, $P(\Theta|x)$, are open for specification.

Let the function $sI_{X=x}[I_A - P(A|x)]$ be a (called-off) bet on event A, given $X = x$, at the agent's conditional odds $P(A|x)$, with total stake s. Thus the agent wins $s[1 - P(A|x)]$ if both $X = x$ and event A obtains, the agent loses $sP(A|x)$ if both $X = x$ and event A^c obtains, and the bet is annulled if $X \neq x$. For each possible outcome $x \in \mathcal{X}$, let $\mathcal{B}_x$ be a finite selection of bets on events involving Θ (i.e., finitely many bets on subsets of Θ) using the agent's posterior odds $P(\Theta|x)$, and let $\mathcal{B}$ be the system of these bets taken together. H–S coherence requires that there be no such system of bets $\mathcal{B}$ that have a negative *risk;* that is, no $\mathcal{B}$ with $\sup_{\theta\in\Theta} \mathbf{E}_P(\mathcal{B}|\theta) < 0$.

Heath and Sudderth showed (1978, Theorem 1) that H–S coherence is equivalent to the existence of a finitely additive probability P over the joint space, (X, Θ), such that for each bounded (measurable) function $h(X, \Theta)$,

$$\mathbf{E}_P[h] = \iint hP(dx|\theta)P(d\theta) = \iint hP(d\theta|x)P(dx). \quad (11)$$

That is, for the agent to be H–S coherent, the finitely additive odds must be conglomerable in the X partition and in the Θ partition. More precisely, (11) requires that the finitely additive probability p be disintegrable in both partitions. Dubins (1975 Theorem 1) established equivalence of disintegrability in a partition and conglomerability in that partition.

Heath and Sudderth's sense of "coherence" is to be contrasted with, for instance, deFinetti's criterion of avoiding a *sure loss.* deFinetti's criterion requires only that there be no finite set of fair unconditional, or called-off bets that lead to a negative outcome (rather than negative expectation), bounded below zero, no matter which state (say, no matter which θ) obtains. To see the difference, Heath and Sudderth's system of bets $\mathcal{B}$ includes betting on Θ given X, whereas deFinetti's indexes these as infinitely many called-off bets if X takes on more than finitely many different values. A probability P may support "coherent" fair odds in deFinetti's sense but fail to be H–S coherent—the finitely additive probability of Example 2.1 is one such P. It is evident that if a system of "fair" odds is incoherent in deFinetti's sense, then it is H–S incoherent too. In Example 2.1, event E is the subject of wagering and X is the quantity to be observed. The finitely additive probability P has the property that $.5 = \mathbf{E}_P[E] \neq \mathbf{E}_P[\mathbf{E}_P[E|X]] = 1$; hence that P is H–S incoherent though coherent in deFinetti's sense.

When a finitely additive probability is H–S coherent, there can be no experiment the agent designs to lead to a foregone conclusion. This obtains provided, of course, that the *combined* set of quantities to be observed through experimentation falls within the scope of the Heath–Sudderth criterion. It may be that P is H–S coherent with bets on event E for observations X; likewise, P may be H–S coherent with bets on E for observations Z, though P is H–S incoherent with bets on E for the combined observations (X, Z). This is illustrated next.

Example 2.3. This example is based on example 4.2 in earlier work (Schervish et al. 1984), that a mixture of nonconglomerable distributions may be conglomerable. Let Q be a finitely additive probability on the (power set) algebra $\mathcal{A} = \{E, E^c\} \times \mathcal{X} \times \{F, F^c\}$ where E and F are (binary) events and $\mathcal{X} = \{1, 2, \ldots\}$. Let $Q(E) = Q(F) = .5$, and assume that $Q(E\,\&\,F) = Q(E)Q(F) = .25$, so that E and F are independent under Q. Let $Q(E, X|F)$ be as $P(E, X)$ in Example 2.1; that is, $Q(E\,\&\,X = k|F) = 2^{-k+1}$ and $Q(E^c\,\&\,X = k|F) = 0$. Finally, let $Q(E\,\&\,X = k|F^c) = 0$ and $Q(E^c\,\&\,X = k|F^c) = 2^{-k+1}$.

Observe that because E and F are events, there is no issue of additivity associated with Q on the four atom subalgebra $\{E, E^c\} \times \{F, F^c\}$. That is, trivially, Q is H–S coherent over this subalgebra. Also, note that the probability Q is H–S coherent over the subalgebra $\{E, E^c\} \times \mathcal{X}$. Specifically, $Q(E\,\&\,X = k) = Q(E^c\,\&\,X = k) = 2^{-k+2}$, and thus $Q(E|X = k) = Q(E) = .5\,(k = 1, \ldots)$. That is, Q is conglomerable in the X partition. Similarly, $Q(F) = Q(F|X = k) = Q(F^c|X = k) = Q(F^c) = .5\,(k = 1, \ldots)$, and Q is H–S coherent over the subalgebra $\mathcal{X} \times \{F, F^c\}$.

However, Q is not conglomerable in the (X, F) partition. Note that $Q(E|X = k\,\&\,F) = 1 = Q(E^c|X = k\,\&\,F^c)$ (for $k = 1, \ldots$), yet $Q(E|F) = Q(E^c|F^c) = .5$. Thus a bettor might display the H–S incoherence in Q by wagering with the agent using the agent's conditional "posterior" odds, once X is revealed. After X is revealed, the bettor makes

a "called-off" gamble $\mathfrak{b}_1$ on E^c, given F, with the amount \$1 contributed to the stake by the agent against $\$\varepsilon$ wagered by the bettor. The agent wins $\$\varepsilon$ in the event that $E \& F$ occurs. The bettor wins \$1 in the event that $E^c \& F$ occurs. The gamble is called-off in the event that F^c occurs—in which case there is no exchange of funds. The bet $\mathfrak{b}_1$ is a favorable (conditional) wager from the standpoint of the agent's posterior Q odds after X is revealed, because, given $X, Q(E|F \& X = k) = 1\,(k = 1, \ldots)$. For a second bet, $\mathfrak{b}_2$, after X is revealed, the bettor wins the agent's \$1 if $E \& F^c$ and the agent wins the bettor's $\$\varepsilon$ if $E^c \& F^c$. The bet $\mathfrak{b}_2$ is favorable for the agent because $Q(E^c|F^c \& X = k) = 1\,(k = 1, \ldots)$. Let $\mathcal{B} = \{\mathfrak{b}_1 + \mathfrak{b}_2\}$ be the conjunction of these two bets. B has negative risk regardless whether E or E^c obtains:

$$\mathbf{E}_Q[\mathcal{B}|E] = \$\varepsilon Q(F|E) + \$(-1)Q(F^c|E) = \$(\varepsilon - 1)/2 < 0$$

and

$$\begin{aligned}\mathbf{E}_Q[\mathcal{B}|E^c] &= \$(-1)Q(F|E^c) + \$\varepsilon Q(F^c|E^c) \\ &= \$(\varepsilon - 1)/2 < 0\end{aligned}$$

Thus Q is H–S incoherent over $\mathcal{A}$ when both X and F are learned, though Q is H–S coherent when either one of X or F alone is observed.

The problem that we see with Heath and Sudderth's approach is that, though the agent may use the finitely additive probability Q when betting on E given X or when betting on E given F, what shall that agent do when the offer is made to learn about both X and F when only one is known? Knowing X, how does the agent assess the value of conducting an experiment to learn which of $\{F, F^c\}$ obtains? Knowing F, how does the agent assess the value of conducting an experiment to learn X? It seems difficult to insulate finitely additive probabilities against the threat of nonconglomerability by trying to determine which quantities may be H–S-coherently observed. As the example shows, nonconglomerability may surface when several quantities are observed, though no proper subset of these quantities causes H–S incoherence on its own. An H–S-coherent agent may be confronted with evidence for which his or her beliefs no longer satisfy H–S coherence.

3. CLASSICAL TESTING AND BAYESIAN INFERENCE TO FOREGONE CONCLUSIONS

In Section 1 we recalled the familiar concern that with classical hypothesis testing, optional stopping opens the door to foregone conclusions when classical (i.e., fixed sample size) significance levels are used to report evidential import. We have seen how the fear of sampling to a foregone conclusion can be alleviated within the (countably additive) Bayesian paradigm. In short, there are bounds on how high the probability can be of sampling until the posterior probability reaches a specified level.

At first, these two results may seem to conflict with the observation that many fixed sample size classical procedures are Bayes or nearly Bayes procedures. For example, with normal data $X \sim N(\theta, 1)$ and the usual improper prior, the posterior probability that $\Theta \leq \theta_0$ given X is the same as the significance level for testing the hypothesis that $\Theta \leq \theta_0$ against the alternative $\Theta > \theta_0$. If sequential significance testing with a fixed sample size tests leads to foregone conclusions, is not the same true for the Bayes procedures?

A naive response might be that the improper prior corresponds to a finitely additive prior; hence foregone conclusions are not ruled out in that case, as we discussed in Section 2. However, in this particular case, the posterior inferences are coherent in the sense of Heath and Sudderth (1978), so no foregone conclusions are possible. That is, the apparent conflict is not attributable to the mere finite additivity of the probability as represented by the improper prior.

To sort our what is happening, note that we have stated the problem as one in which sampling continues until the posterior probability of the null hypothesis H: $\Theta \leq \theta_0$ rises above a specified level. It can be shown that given $\Theta = \theta \leq \theta_0$, the probability is 1 that such a sampling scheme will stop and assign a high posterior probability to the true null hypothesis. This is true whether one uses the usual improper (finitely additive) prior or a conjugate proper prior. Also, it can be shown that for these same priors, given $\Theta = \theta > \theta_0$, the probability is less than 1 that the sampling scheme will stop and assign a high posterior probability to the false null hypothesis. But if we modify the hypotheses and move the endpoint θ_0 into the alternative hypothesis, then, conditional on $\Theta = \theta_0$, there is probability 1 that the sampling scheme will stop and declare that the probability is high for the false null hypothesis $\Theta < \theta_0$. This case allows a foregone conclusion. Of course, each of the prior distributions used in this analysis puts zero probability on the event $\Theta = \theta_0$.

If one is particularly concerned about the possibility that $\Theta = \theta_0$ (as opposed to Θ being very close to θ_0), then one can avoid the foregone conclusion just mentioned by using a countably additive prior probability that puts sufficient mass at $\Theta = \theta_0$; for example, let $P(\Theta = \theta_0) > p(1 - q)/q$. This will ensure that the conditional probability given $\Theta = \theta_0$ is strictly less than 1 for stopping and declaring that the posterior probability of the false hypothesis $\Theta < \theta_0$ is high. Of course, with this changed prior, the stopping rule (based on the posterior) no longer is to sample until $\Theta < \theta_0$ achieves a preassigned classical significance level. That is our resolution to the earlier puzzle.

Note also that a similar problem does not arise in the two-sided case, because the usual fixed sample size significance levels are not posterior probabilities, as Jeffreys discussed many years ago (1939, sec. 5.1). (See Lindley 1957 for additional commentary.) And for the special case of a point-null hypothesis with a two sided alternative, when improper priors are used for the adjustable parameter of the alternative, it can happen that there is a foregone conclusion for the null hypothesis with only one observation. (See Jeffreys' 1939, p. 251, for an analysis.)

4. SUMMARY

In Section 1 we displayed some elementary reasoning, based on the law of total probability, about the protection

simple Bayesian theory affords against sampling to a foregone conclusion. However, Example 1.1 alerts us to the possibility of foregone conclusions when inference is based on an improper prior. In Section 2 we explored limitations on the results from Section 1 that arise from using a finitely, but not countably additive, probability corresponding to an improper prior. We considered three perspectives on finitely additive probabilities:

- where foregone conclusions are avoided by strictly preferring; that is, paying to choose "now" rather than to collect "cost-free" evidence
- where foregone conclusions are avoided by mandating against the use of Bayes's rule for updating probabilities, as in Goldstein's theory
- where foregone conclusions are avoided by trying to circumscribe the partitions in which nonconglomerability appears, as in Heath–Sudderth coherence.

Each of these viewpoints carries a price for relaxing the principle of countably additivity.

Of course, there is a fourth perspective which also avoids foregone conclusions. This is to proscribe the use of merely finitely additive probabilities altogether. The cost here would be an inability to use improper priors. These have been found to be useful for various purposes, including reconstructing some basic "classical" inferences, affording "minimax" solutions in statistical decisions when the parameter space is infinite, approximating "ignorance" when the improper distribution is a limit of natural conjugate priors, and modeling what appear to be natural states of belief (Kadane and O'Hagan 1995).

Whether in each case the price is too high is a question for further investigation and serious debate. We need to discuss the extent to which Bayesian methodology affords or ought to afford protection against reasoning to a foregone conclusion.

[Received August 1994. Revised October 1995.]

REFERENCES

Anscombe, F. J. (1954), "Fixed Sample Size Analysis of Sequential Observations," *Biometrics*, 10, 89–100.

Ash, R. B. (1972), *Real Analysis and Probability*, New York: Academic Press.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Cornfield, J. (1970), "The Frequency Theory of Probability, Bayes's Theorem, and Sequential Clinical Trials," in *Bayesian Statistics*, eds. D. L. Meyer and R. O. Collier, Jr., Itasca, IL: Peacock Publishing, pp. 1–28.

De Finetti, B. (1974), *The Theory of Probability*, New York: John Wiley.

Doob, J. L. (1953), *Stochastic Processes*, New York: John Wiley.

Dubins, L. E. (1975), "Finitely Additive Conditional Probabilities, Conglomerability, and Disintegrations," *Annals of Probability*, 3, 89–99.

Feller, W. (1940), "Statistical Aspects of E.S.P." *Journal of Parapsychology*, 4, 271–298.

Goldstein, M. (1983), "The Prevision of a Prevision," *Journal of the American Statistical Association*, 78, 817–819.

Good, I. J. (1967), "On the Principle of Total Evidence," *The British Journal for the Philosophy of Science*, 17, 319–321.

Heath, D., and Sudderth, W. (1978), "On Finitely Additive Priors, Coherence, and Extended Admissibility," *The Annals of Statistics*, 6, 333–345.

Jeffreys, H. (1939), *Theory of Probability*, Oxford, U.K.: Oxford University Press.

Kadane, J. B., and O'Hagan, A. (1995), "Using Finitely Additive Probability: Uniform Distributions on the Natural Numbers," *Journal of the American Statistical Association*, 90, 626–631.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1986), "Statistical Implications of Finitely Additive Probability," *Bayesian Inference and Decision Techniques With Applications*, eds. Goel, P. K. and A. Zellner, New York: Elsevier Science Publishers, pp. 59–76.

Kerridge, D. (1963), "Bounds for the Frequency of Misleading Bayes Inferences," *Annals of Mathematical Statistics*, 34, 1109–1110.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge, MA: Harvard University Press.

Ramsey, F. P. (1990), "Weight or the Value of Knowledge," *The British Journal for the Philosophy of Science*, 41, 1–4.

Robbins, H. (1952), "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 58, 527–535.

Savage, L. J. (1954), *The Foundations of Statistics*, New York: John Wiley.

Savage, L. J. with prepared contributions from Bartlett, M. S., Barnard, G. A., Cox, D. R., Pearson, E. S., and Smith, C. A. B. (1962), *The Foundations of Statistical Inference*, London: Methuen.

Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (1984), "The Extent of Non-Conglomerability of Finitely Additive Probabilities," *Zeitschrift fur Wahrscheinlictkeitstheorie und verwandte Gebiete*, 66, 205–226.