Journal of Statistical Planning and Inference 25 (1990) 329-345
North-Holland

RANDOMIZATION IN A BAYESIAN PERSPECTIVE

Joseph B. KADANE and Teddy SEIDENFELD

Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A.

Received August 1987; revised manuscript received February 1989
Recommended by K. Hinkelmann

"Applying the theory (of personal probability) naively one quickly comes to the conclusion that randomization is without value for statistics. This conclusion does not sound right; and it is not right. Closer examination of the road to this untenable conclusion does lead to new insights into the role and limitations of randomization but does by no means deprive randomization of its important function in statistics." L.J. Savage (1961)

"Though we all feel sure that randomization is an important invention, the theory of subjective probability reminds us that we have not fully understood randomization... The need for randomization presumably lies in the imperfection of actual people and, perhaps, in the fact that more than one person is ordinarily concerned with an investigation." L.J. Savage (1962)

Randomization has thus been a puzzle for Bayesian theory for many years. In this paper, we give our current views on this subject.

There are two principal arguments for randomization that we are familiar with. The first is to support a randomization-analysis of the data. This notion goes back to Fisher, and is exposited in a series of papers by Kempthorne (1955, 1966, 1977). It asks whether what is observed is surprising given all the other designs that might have been randomly selected and data that might have been observed, but were not. By its appeal to what did not occur, such an analysis violates the likelihood principle; hence, it is not compatible with Bayesian ideas. Many have criticized randomization-analysis for failing the likelihood principle: see Basu (1981), and Bunke and Bunke (1978), for illustrations. This reply to Fisher, Kempthorne, and others who defend randomization-analysis is, we think, what Savage means by the 'naive' Bayesian rejection of randomization.

We explore the argument for randomization-analysis in detail in Section 1. It is our purpose there to distinguish two cases. In one (illustrated with Example A), the randomized inference depends upon what is wholly irrelevant evidence according to the likelihood principle, evidence that we define as 'suppressible'. In the second case (Example B), the randomized inference ignores ancillary information which is *not*

suppressible. The two cases are different ways of violating the likelihood principle with a randomization-analysis. Case 2, the theme of validating randomization by ignoring ancillary but not suppressible evidence, is central to our subsequent discussion of how randomized experimental designs are justified (Example C).

The second of the two arguments we know for randomization is that, in design, it is thought to provide methodological insurance against a variety of observer 'biases'. In Section 2 we explain what bias is about, why, and *for whom* it is a problem. The latter question can be motivated with the aid of the following:

Randomization amounts to deciding by some random device, whose outcomes are out of the control of the researcher-analyst, which units to study, say which people to interview in a survey, or which treatments to assign to each patient in a clinical trial. Thought of as a decision, each such choice or assignment has some expected utility (see Lindley (1972) for a clear exposition of how that expected utility is calculated). If one choice has higher expected utility than all others, why not choose it? If many are equally good, choosing randomly among them is optimal, but so is choosing with certainty any of the optimal choices. Thus, it would appear that randomization is always unnecessary and sometimes suboptimal. Without loss of utility, why then cannot decisions be chosen without randomization?

The hallmark of this reasoning, in our opinion, is that it applies to what we call 'experiments to learn'. That is, the objective of the researcher is to inform himself or herself. In such a case, which is typical for example of pilot studies in many disciplines, it is indeed unnecessary and quite possibly suboptimal to randomize. *When only one decision-maker is relevant*, we accept this analysis and would not randomize.

Observer bias, however, seems to us to concern a different goal for experiments, which we call 'experiments to prove'. This, we think, is what Savage intends as the 'closer examination'. With 'experiments to prove', several decision-makers are involved. In Section 3 we discuss some Bayesian facts about decisions involving more than one decision-maker. And, in Section 4, we consider randomization in this light. We conclude that randomization has no particular merit in creating 'evidence to prove'. Alternative, non-randomized designs are available for this purpose. In Section 5 we illustrate this for clinical trials, where ethical considerations can be given priority over unrestricted randomization without creating a 'biased' design.

1. The 'naive' Bayesian Theory opposes the suppression of evidence

The 'naive' application of the theory of personal probability to which Savage refers is, we suspect, the appeal to 'ancillarity' as a refutation of randomization. That argument proceeds as follows.

Suppose hypotheses of interest are indexed by a parameter θ and there are new data d. If we use Bayes' rule (conditionalization) to update our 'prior' opinion about θ, $p(\theta)$, to a 'posterior' probability given the new data d, $p_d(\theta)$, then the

'posterior' is proportional to the product of the 'likelihood' and 'prior':

$$p_d(\theta) \propto p(\mathrm{d} \mid \theta) p(\theta).$$

A statistic t is called *ancillary for* θ if its likelihood is constant, i.e., if

$$p(t \mid \theta) = p(t),$$

independent of θ. Then, by Bayes' rule, ancillary data are irrelevant:

$$p_t(\theta) \propto p(t \mid \theta) p(\theta) = p(t) p(\theta) \propto p(\theta).$$

If the new data d can be written as a conjunction of the two statistics s and t, $d = (s, t)$, with t ancillary for θ, then (by Bayes' rule) the updated probability for θ depends only on the (conditional) likelihood for s, given t and θ:

$$\begin{aligned} p_d(\theta) \propto p(d \mid \theta) p(\theta) &= p(s \mid t, \theta) p(t \mid \theta) p(\theta) \\ &= p(s \mid t, \theta) p(t) p(\theta) \\ &\propto p(s \mid t, \theta) p(\theta). \end{aligned} \qquad (*)$$

Thus, whatever there is of relevance in data d to hypotheses θ is contained in the statistic s, given the ancillary data t.

Define t to be *suppressible* in the presence of s for θ if t is ancillary and also

$$p(s \mid t, \theta) = p(s \mid \theta), \quad \text{independent of } t.$$

Then we have that the data d may be contracted to s, without loss of relevant evidence about θ since

$$\begin{aligned} p(\theta \mid s, t) &\propto p(s \mid t, \theta) p(\theta) && \text{as } t \text{ is ancillary} \\ &\propto p(s \mid \theta) p(\theta) && \text{as } t \text{ is suppressible} \\ &\propto p(\theta \mid s) && \text{by Bayes' theorem.} \end{aligned}$$

That is, when t is suppressible in the presence of s, s is sufficient for θ:

$$\begin{aligned} p(t \mid s, \theta) &= p(s \mid t, \theta) p(t \mid \theta) / p(s \mid \theta) && \text{by Bayes' theorem} \\ &= p(s \mid t, \theta) p(t) / p(s \mid \theta) && \text{since } t \text{ is ancillary} \\ &= p(t) && \text{as } t \text{ is suppressible.} \end{aligned}$$

Thus, $p(d \mid s, \theta) = p(d \mid s)$ independent of θ, as required for sufficiency.

We may apply this directly both (i) to overturn 'orthodox' procedures based on randomized data analysis, and (ii) to refute the the 'classic' (Fisher's) argument: that randomization in experimental design offers methodological insurance against a biased sample. Let us illustrate each of these criticisms. In Example A below, the random digit y is suppressible. However in Example B, the random pairing is ancillary but not suppressible.

Randomized statistical decisions

There are two, familiar uses of randomization in 'orthodox' (Neyman–Pearson) testing and interval estimation. (A) It can be the basis of a 'most powerful' test. That is, the 'best' test (at a fixed size, $\alpha = \alpha_0$) may be a mixed test. (B) Second, a randomized test may have an exact size whereas no non-randomized one does; hence, in such problems, only 'mixed' confidence intervals have exact 'coverage' probabilities.

(A) Randomization may improve the power of a test by simulating continuity for discrete random variables

Example A. Let x be a random quantity with $x \in \{0, 1, \ldots, 10\}$. Consider tests of a simple null hypothesis h_0 versus a simple rival hypothesis h_1. Under the null hypothesis the distribution of x, $p(x \mid h_0)$, is:

$$p_0(0) = 0.05; \qquad p_0(n) = n(0.019)/1.1 \quad (n = 1, \ldots, 10).$$

Under the rival hypothesis the distribution of x, $p(x \mid h_1)$, is:

$$p_1(0) = 0.5; \qquad p_1(n) = n(0.01)/1.1.$$

When $\alpha = 0.02$ (the probability of a 'type 1' error), the only 'pure' test (based on an observation of x) meeting the size restriction rejects h_0 if $x = 1$. Its 'power' is a paltry 0.009, so that this pure test is 'biased'. (One is more likely to reject the null hypothesis when it is true than when it is false!)

Suppose, instead, you can augment the observation, x, with a randomly chosen digit $y \in \{0, \ldots, 9\}$. Then there is a mixed test ($\alpha = 0.02$), with 'power' 0.2: reject h_0 provided $x = 0$ and $y \in \{0, 1, 2, 3\}$. (This is the best $\alpha = 0.02$ test, mixed or pure, based on x.) The increase in power results from using the randomizer to simulate a continuous random quantity z with the same likelihood as x. Then, by the Neyman–Pearson lemma, the best test with z is a likelihood ratio test, i.e. let the rejection region consist of z-outcomes with lowest likelihood ratio, $p(z \mid h_0)/p(z \mid h_1) < d$. In the example, this region corresponds to $x = 0$.

Of course, with data (x, y), the random digit y is ancillary to the hypotheses of interest. By the result (*) above, from a Bayesian point of view only the conditional likelihood ratio matters:

$$p(x \mid y, h_0)/p(x \mid y, h_1) = p(x \mid h_0)/p(x \mid h_1).$$

Thus, the random digit y is suppressible in the presence of x.

(B) Randomization may provide a basis for exact confidence levels by bypassing 'nuisance' factors

Example B. Let x_j be a sequence of n (identically, independently) normally distributed random quantities, $N(\mu_x, \sigma_x^2)$. Likewise, let y_j be a sequence of n (i.i.d.) normal $N(\mu_y, \sigma_y^2)$ quantities. The parameter of interest is the difference in population means, $\delta = \mu_x - \mu_y$, and the (unknown) population variances (σ_x^2, σ_y^2) are 'nuisance' factors. Use random numbers from the set $\{1, \ldots, n\}$ to randomly match the x's and y's. This forms n-pairs (x_j, y_j) $(j = 1, \ldots, n)$. Define the n-differences $z_j = x_j - y_j$. Then the random variable z is normally distributed, $N(\delta, \sigma_z^2)$, where $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$. An exact confidence interval for δ is created from the familiar t-test based on the data z_j. (This analysis corresponds to a permutation-test solution of the Behrens–Fisher problem.) Thus, with the aid of randomization, the 'orthodox' analysis bypasses the 'nuisance' factor σ_x/σ_y.

This randomized t-test is not valid from the Bayesian point of view and for two reasons. One objection, which is not germane to our discussion of randomization, is the fact that the reduction of the data (x_j, y_j) to the paired differnces z_j fails to preserve all the relevant information about δ.

For instance, assume the (x_j, y_j) have a bivariate normal distribution ($\varrho = 0$) and the familiar 'improper' prior ($\propto 1/\sigma_x\sigma_y$) is used in order to reproduce the t-test analysis. Still, we find that the posterior distribution based on all the data differs from the posterior distribution based on just the paired differences, $p(\delta \mid x_j, y_j) \neq p(\delta \mid z_j)$. The difference amounts to a reduction by $\frac{1}{2}(n-1)$ in the degrees of freedom of the estimate for $\sigma_x^2 + \sigma_y^2$ as provided by the z_j-data. Put another way, $p(\delta \mid x_j, y_j)$ is a function of both the paired sums $w_j = x_j + y_j$ and the paired differences, z_j. (See DeGroot (1980) for an interesting analysis of the inference problem about an unknown ϱ, given 'broken' pairs.)

Let us sidestep this objection and, in the fashion of Lindley's (1965, Vol 2, p. 84) analysis, consider the reduction from data (x_j, y_j) to data z_j an acceptable approximation – as it becomes with increasing n. The objection to the randomized t-test that we focus upon is concerned with the *suppression* of the ancillary pairing that results when the data are contracted to the paired differences, z_j, alone.

The t-test using the z_j takes the form $(n-1)^{1/2}(\bar{z} - \delta)/s_z$. That is, the t-test uses the further reduction of the data to the two statistics $(\bar{z}, s_z^2)$. Alternative (random) pairings of the original data leave $\bar{z}$ unchanged but alter the variance, s_z^2. There are $n!$ pairings of the data and thus (barring ties) there are $n!$ possible variances as rank ordered by the relative magnitude of the resulting s_z^2. Randomization picks out one of these, with probability $1/n!$, yielding the ancillary data $t = \text{rank}(s_z^2)$. That is, the evidence available includes both $s = (\bar{z}, s_z^2)$ and the ancillary statistic, t, of which one among the $n!$ possible s_z^2-ranks was selected.

According to the result (*) above,

$$p(\delta \mid s, t) \propto p(s \mid \delta, t)\, p(\delta).$$

However, $p(s \mid \delta, t) \neq p(s \mid \delta)$, and t is not suppressible. Specifically, $p(s \mid \delta, t)$, but not $p(s \mid \delta)$, involves the nuisance factor σ_x/σ_y. That is, the randomized t-test analysis, based on $p(s \mid \delta)$, is invalid because it mistakenly takes the ancillary, ran-

domized choice of t to be suppressible. Hence, using the randomization to eliminate the nuisance parameter is unwarranted from this Bayesian point of view.

Let us illustrate our analysis in the simple case of two observations from each population, $n = 2$. There are two, alternative pairings of the x's and y's. In the one case we get the smaller of the two possible s_z^2, to wit, with $\text{rank}(s_z^2) = 1$, we have the 't-statistic':

$$[(x_1 + x_2) - (y_1 + y_2)] / ||x_1 - x_2| - |y_1 - y_2||.$$

And with $\text{rank}(s_z^2) = 2$, using the larger of the two possible s_z^2, we have:

$$[(x_1 + x_2) - (y_1 + y_2)] / [|x_1 - x_2| + |y_1 - y_2|].$$

Unfortunately, the distributions of these statistics depend upon the nuisance factor σ_x/σ_y. In particular, the second of these corresponds to the Behrens-Fisher statistic. See, for example, Fisher's discussion (1973, pp. 97-103) for its exact distribution.

Thus, given the ancillary statistic t, the posterior distribution $p(\delta \mid s, t)$ remains a function of the agent's beliefs about the unknown ratio σ_x/σ_y. From a Bayesian point of view, the randomization does not succeed in eliminating the nuisance parameter. A Bayesian can use the permutation test as a solution to the problem of Example B only by censoring the (ancillary) datum t. We find it is counter-intuitive to adopt a strategy to willfully censor evidence which, though ancillary, is not suppressible.

I.J. Good (1971, #679) observes that often a Bayesian can make sense of 'orthodox' statistical procedures by avoiding parts of the data. In this case Good could employ his 'Statistician's Stooge' to carry out the random pairing for him, but to report (to Good) only the values z_j. (A related idea is found in Levi's (1980, Chapter 17) use of 'data as input', rather than 'data as evidence'.) That would allow Good to respect the total evidence principle while offering him the convenience of the permutation test analysis. Next, we explore the censoring of ancillary but not suppressible data in connection with our discussion of the role of randomization for sound experimental design.

(C) Randomized designs as methodological insurance against a 'biased' sample

In his pedagogically influential discussion of an experiment for testing an hypothesis that a Lady can taste the difference between tea made with milk added first (rather than second), Fisher (1971, Chapter 2) raises three methodological concerns to which randomization is the supposed remedy:

(1) By presenting the Lady with 8 cups (4 prepared in each way) *in a random order*, the researcher is (supposed to be) assured that the Lady's success is reasonably attributed to her professed ability to taste differences and is not the result, instead, of her ability to anticipate the pattern of cups when that is left to the experimenter's imagination.

(2) By randomly assigning each preparation to four of the eight cups, the study is (supposed to be) protected against a confounding of uncontrolled factors with what is tested for. For example, suppose the Lady reacts to small differences in the cups rather than to the preparation of the tea. Then not knowing which are the relevant cup characteristics is of no concern since, when randomized, the chance is negligible that just those cups to which she would react 'milk-first' get milk-first.

(3) By randomly assigning treatments to cups the experimenter is (supposed to be) warranted in adopting a particular statistical distribution corresponding to the 'null' hypothesis. And for problems where the 'alternative' hypothesis is well-formed, the randomization justifies adopting a particular statistical model for the experimental outcomes. Accordingly, the probability is 1/70 that the Lady correctly identifies all four 'milk-first' of the eight randomly prepared cups, given the null hypothesis that she cannot discriminate (and that she knows four of eight were so made).

Observe that, according to Fisher's randomization solution, the three problems are treated exactly the same in their first- and third-person interpretations.

Under randomization, the researcher and the reader each believes the Lady's success at discerning the two kinds of tea infusions is separated from her ability to anticipate subconsciously preferred patterns – for the simple reason that the allocation scheme used is known to be a mechanism without preferences. Of course, when there can be no game between researcher and subject, e.g., when the responses are involuntary (because the subjects are plots of land with no interest in the outcomes), this problem doesn't even arise.

Similarly, the researcher and the reader are to agree that, if the cups are randomized between the two preparations and if the Lady has no real ability to discriminate 'milk-first' tea, then:

– 'uncontrolled' factors are uncorrelated with her accuracy – 'nuisance' factors are eliminated from consideration;

– and there is an established chance of 1/70 that she correctly identifies all four cups prepared with milk-first – there is consensus on a simple statistical model for her responses.

Unfortunately, these arguments are invalid once the researcher (or the reader) is made aware of the (ancillary) outcome of the randomization. That is, just as in Example B, these analyses are cogent (only) prior to observing the particular allocation arrived at by randomization. If the random allocation directs serving all four cups of one kind first, or directs putting the 'milk-first' tea into the cups which are seen to be the brightest in color, is there any ground for continued agreement on the (conditional) odds of 1/70 that the Lady correctly identifies the tea (given this allocation and the null hypothesis)? Even if the researcher can convince himself these factors are irrelevant to the Lady's performance, what features of the randomization lend credence to this judgment? What is it concerning randomization that makes such a judgment (of irrelevance of the allocation to the test outcomes) compelling for the

reader? We can find none and suspect that randomization has little to do with whatever grounds there are for the belief that the allocation is irrelevant to the test results.

This, then, is the 'naive' Bayesian criticism of randomization. Of course, if neither the researcher nor the reader learns the (ancillary) details of the allocation, if they know only that allocations are made 'at random' and they learn the frequency data of 'favorable' test responses, then there is a Bayesian interpretation of the 'orthodox' claims. Then there is a Bayesian reconstruction of how randomization supports a consensus of low probability that allocations are 'biased', etc. Rubin's (1978) Bayesian account of randomization is based on this fact. By *failing* to make a record of the ancillary information in a randomized allocation, both the 'assignment' and 'recording' mechanisms are 'ignorable' (in Rubin's terms); what we call a 'transparent' allocation. Also, then there is consensus on a statistical distribution for the recorded data, given the null hypothesis, since they are simplified by the elimination of 'nuisance' factors – exactly as in Example B.

This is shown, as follows. Suppose the Lady will specify one of the 70 possible sets of four cups in response to the test question, "Which four are made 'milk-first'?" Let the Lady's response that it is the i-th quadruple be denoted by Q_i and let T_j designate the allocation of 'milk-first' to the j-th quadruple of cups. Under the 'null' hypothesis (that the Lady has no ability to identify which are 'milk-first' cups), these are independent,

$$p(Q_i \mid T_j) = p(Q_i);$$

where p is 'subjective'. Then, with a randomized allocation of the two treatments, four cups prepared each way,

$$\begin{aligned} p(\text{"the lady gets all four right"}) &= \sum_i p(T_i \& Q_i) \\ &= \sum_i p(Q_i \mid T_i)\, p(T_i) \\ &= \frac{1}{70} \sum_i p(Q_i \mid T_i) \\ &= \frac{1}{70} \sum_i p(Q_i) \\ &= 1/70, \end{aligned}$$

independent of the 'subjective' probability, $p(Q_i)$. A similar analysis shows that

$$p(\text{"the lady gets just 3 right"}) = 16/70$$

and

$$p(\text{"the lady gets just 2 right"}) = 36/70,$$

all in accord with the 'usual' distribution for the recorded data under the 'null' hypothesis.

Thus, if the recorded data include only the number correctly identified while the (ancillary) statistic of which cups were treated 'milk first' is censored then, under the 'null' hypothesis, there is a consensus about a simple statistical distribution for the experimental outcomes. Nuisance factors, such as the effect of the thickness or color of the cups on the Lady's response, are eliminated – just as σ_x/σ_y is eliminated in Example B.

In this example, the randomization-plus-censoring does not produce a consensus model under the alternative hypothesis, given that the Lady has some ability to identify 'milk-first' tea. The difficulty is that the claim 'some ability' remains too vague. In other settings, however, for instance in sampling problems where the quantities of interest are population frequencies for particular characteristics, randomized-sampling-plus-censoring (of the ancillary information of which items were sampled) justifies a Multinomial statistical model for the recorded data of the sample frequencies.

The 'closer examination', called for by Savage is, we suppose, based on the idea that sometimes it may be efficient to seek interpersonal agreement about what the data show by using a simple randomized design (with the concomitant censoring of ancillary data) instead of an experiment which maximizes what the researcher (alone) can expect to learn. In an experiment that optimizes what the researcher expects to learn, there is an obstacle to consensus (a limitation on what he can expect to prove). Then there is no *robust* Bayesian analysis of the experimental data: different 'prior' opinions about 'nuisance' factors (themselves unaddressed by the experiment) can interfere with a shared interpretation of experimental results. Also, if the investigator attends to the cost of calculation and deliberation then it may be efficient *to learn* (1st-person) from a randomized experiment (in which the ancillary data are unrecorded). The non-randomized design may require costly calculations to analyze its data. Rubin (1978, p. 54) put the point this way: "A comparable non-randomized design would generally be substantially more difficult to execute and analyze because of the need to deal explicitly with all covariates being balanced."

Reliance on a method that requires and encourages censoring of evidence is a problem not just because of the lost opportunity to squeeze a bit more out of the data. Such a method actively encourages researchers to avoid reporting their data and beliefs fully, for fear that the additional evidence will destroy the inference they wish to make. Our discomfort with this incentive for data censoring should not be taken to mean that we think that all details of a study must, or should, or can be reported. The details of a study that are omitted should be those that don't matter, those that are suppressible. Here, to the contrary, details are omitted because they do or might matter.

However, there exist other strategies for designing experiments that promote a consensus of posterior probabilities and which avoid randomization and censoring data: we discuss one of these in Section 5. Moreover, the alternative designs are

ethically superior in the setting of clinical trials, as we argue below. Thus, even a sophisticated (rather than naive) Bayesian defense of randomization, one which emphasizes the goal of experiments 'to prove' rather than 'to learn', fails to establish it as a sine qua non of sound experimental methods.

2. Observer bias

Suppose that you are the reader of a paper I have written about the comparison of two treatments for a psychological ailment. Suppose also that I am the inventor and proponent of treatment A, which, in my study, did better than treatment B (I am pleased to report). Suppose also that patients who have the ailment severely are known to be very difficult to treat, but that patients who have the ailment only mildly are quite tractable. What might you make of my results? How can I make a study that will be persuasive to you? If I could decide which patient went into each treatment, the possibilities for mischief are huge. Even if I assigned patients to whatever treatment I think will be best for them (a not uncommon, nor unworthy thing to do), how can I (or, what is even more difficult, you), sort out what happened afterward, especially in the circumstance that a perfect measure of severity of illness is unavailable? How can I (or, what is even more difficult, you) be sure that I have not essentially built the apparent superiority of treatment A into the study by how the patients were chosen? This is the claimed role for randomization, and it is one that deserves to be taken seriously.

We have built into the example above the motive of being a proponent of treatment A. But this is really unnecessary to the argument. If it is, as the Bible tells us, the truth that will make us free, as researchers it is the surprising truth that will make us famous. Thus each researcher may have, or may be suspected of having by a skeptical and critical reader, a motive to exaggerate the truth. We are not writing here of deliberate and flagrant falsification, but of the much more subtle ways in which we can fool ourselves, and try to fool others.

Thus to leave the assignment of patients to treatments to the researcher seems perilous. One method around this is randomization, which would assign patients to treatments with a probability mechanism outside the control of the researcher. And this is what is meant, as we understand it, by using randomization to avoid bias. Nothing in this argument suggests that randomization is unique in this; for example designs that maximize balance might also deal with observer bias.

There are two ways that we see to view this argument. One concentrates on the researcher, and says that because he or she may not be aware of the implicit biases inherent in the experimental procedures used, randomization, by relieving the experimenter of the the responsibility, can aid good design. This argument is within 'experiments to learn', but relies on the inability of the researcher to know his or her own mind. A second view of observer bias involves more than one decision-maker, at least an experimenter and a reader, who may have different beliefs. Consequently this argument is in the 'experimentation to prove' category. In order to analyze it in a Bayesian perspective, we first review some literature on the Bayesian analysis of problems involving more than one decision-maker.

3. Bayesian analysis of decisions involving more than one decision-maker

The Bayesian literature on problems that involve two or more decision-makers stems from the same principles that animate the analysis of one-decision-maker problems. Given that I must make a decision among various alternatives, my outcome, that is, the utility of my action, depends on what you do. I do not know what you will do, but, as a Bayesian, I do have a probability distribution over your possible acts. Then, not surprisingly, the Bayesian norm is that I should decide in such a way that I maximize my subjective expected utility, where here the expectation is over the probability distribution generated by my uncertainty about your behavior.

It is important that I do not have to model your behavior as if you were a Bayesian. As is usual in the Bayesian paradigm, I am entitled to think whatever I think about you. I might think that you are a Bayesian, with a known probability distribution and a utility function uncertain to me, but in my opinion, with probability $\frac{1}{3}$ it is absolute error and with probability $\frac{2}{3}$ it is squared error. Then I am saying that with probability $\frac{1}{3}$ you will choose the median of this known probability distribution, and with probability $\frac{2}{3}$ you will choose the mean. Of course, I may be quite wrong about how you will decide, but this is an ever-present risk to a Bayesian. Adherence to the axioms only guarantees coherence (Lindley, 1972), a kind of consistency, but not correctness. But I may, equally as coherently, believe that you do not make your decisions in a Bayesian way. I must still have a probability distribution over your actions, but that is the only constraint. In fact, for the purpose of making my decision optimally, all that matters in my probability distribution, and not the underlying theory that led me to it (and of course, my utility function).

While seemingly innocuous, this principle has interesting consequences in the social sciences. Many social scientists apparently believe that uncertainty about the actions of other people is fundamentally different from other kinds of uncertainty, and that other rules of optimal behavior ought to apply. For some flavor of this debate, see Kadane and Larkey (1982, 1983), Harsanyi (1982) and Aumann (1987). The debate has been particularly sharp in game theory. One of the games that most sharply raises the issue is the prisoner's dilemma. In this two-player game, each player faces a choice between two alternatives, a cooperative choice and a competitive choice. If both choose to cooperate, both are rewarded. If both choose to compete, both are punished. But if one chooses to compete and the other to cooperate, the former is greatly rewarded while the latter is severely punished. In a single play of this game, both traditional game theory as inherited from von Neumann and Morgenstern and Bayesian game theory prescribe the competitive response. However in multiple play games, the traditional theory only recommends

competition, while Bayesian game theory, taking account as it must of the likely effect of my aciton now on your action later, can recommend either move, depending on the prior employed. See Wilson (1986) for more on the Bayesian approach to the prisoner's dilemma. A very different game analyzed in the Bayesian framework can be found in DeGroot and Kadane (1983). Important discussion has also been undertaken in the economics literature under the label of the principal-agent problem (Pratt and Zeckhauser, 1985). Actual modelling of real decision-making in a clinical trial, for example, is enormously complicated, and is still in its infancy.

4. Bayesian analysis of observer bias

In an 'experiment to prove', there are at least two actors: the reader and the author. Applying the theory of Section 3, the reader may model the author as a Bayesian whose utility reflects a desire to prove a hypothesis, and who knows more about the phenomenon at hand, having studied it more closely, the experimental design becomes a legitimate source of worry to the reader. Consequently the author must explain just how the sample was chosen, or just how patients were assigned treatments. A method that allows the author's judgment to enter the design, then imposes on the reader a substantial cost of elicitation and analysis. A reader who is sufficiently interested could do such an analysis, but it is wise of an author to try to arrange things so that such an effort is not required. This would reduce the entry cost of reading the work, and consequently encourage readership. Being read is also an element in becoming famous.

All that is required to achieve this simplicity of analysis, however, is that the author not be able to control the design once it is set in motion; that is, the author should not be able to achieve the goal of being able to assign the less sick patients to the favored treatment. Randomization is one available method to do this, but it is not unique in this respect. There are other designs that can do this.

We have been preceded by Stone (1969) in finding that a Bayesian analysis of randomization requires more than one decision-maker. However, Stone's analysis and conclusions differ from ours in several important respects. Stone presupposes a Bayesian B who is doing the study, and another Bayesian A who is reading it, in a finite sampling context. Bayesian B reports his prior π_B, as well as the results of his sampling experiment. Stone then wonders how Bayesian A is to use the information contained in π_B to analyze the data. Stone argues that only random sampling performs this function:

> "If B selects his n units by simple random sampling and A knows this, A's problems are resolved. A then can and should use his own prior π_A in conjunction with the ... likelihood function. Note that simple random sampling is the only sampling scheme that performs this resolution perfectly. If B used a random sampling scheme specified by probabilities $\{P(S) \mid S \text{ any sample of size } n\}$ with $P(S)$ dependent on S, A would be back in a position of uncertainty about how to deal with the information in the different $P(S)$ values, no matter how close these values were to constancy. Such is the argument for simple random sampling, requested by A and understood by B."

Within this argument, the sample size n chosen by B may well carry information about his prior and utility on the parameters of interest. Consequently even a randomized trial with the sample size chosen by B in an unknown way leaves the reader A in the very position of uncertainty that Stone claims randomization avoids.

By contrast, our argument is that any mechanism that is transparent to A in that it makes B's choice of sample depend only on measured covariates in a functional or probabilistic way will suffice to remove both π_B and B's utility function from A's inference problem. Thus, we argue that many other designs beyond simple random sampling achieve the same protection from observer bias as does randomization. This would be of only theoretical interest were it not for the existence of practical situations in which this consideration is important, one of which is described in Section 5.

5. Design of a more ethical clinical trial

While randomization maybe have the advantage of familiarity to statisticians, it can have other, counterbalancing disadvantages. One area where those disadvantages are especially acute occurs in clinical trials of new medical treatments on people. Here, the proposition is to be put to patients that their treatment is to be decided by a random device. In the United States, a patient must be informed of all information that reasonably bears on the decision to participate in a clinical trial. Many trials now make clear how the assignment of treatments is to be done, and it is quite possible that all informed consent statements will be required to reveal this information. What might a knowledgeable patient make of such information?

Many potential patients who have been trained in statistics and the making of optimal decisions under uncertainty might respond to such a proposal by saying in effect, "Doctor, you know about me and about my disease. You must have a hunch about which treatment would be better for me. Please give me that treatment and forget about flipping coins." This line of argument, which is essentially the same as the experiments to learn argument against randomization, indicates that such randomization is actively harmful to the patient except in that rare case in which the expected utilities of the two treatments are exactly balanced. To make this point is not to advance the wise treatment of patients, however, if the upshot is to prevent carefully designed clinical trials from being conducted. Thus the burden of proof is on the critic to propose a better system.

A system that purports to be an improvement was proposed in Kadane and Sedransk (1980). It involves the following steps:

(1) the appointment of a small number of experts on the disease and treatments under study;

(2) agreement in that group on the single indicator of outcome most reasonably of concern to a patient in the trial;

(3) agreement on a few measurable diagnostic indicators possibly linked to outcome as measured in (2) above;

(4) elicitation of each expert's opinion about the outcome indicator as a (probabilistic) function of the diagnostic indicators and treatment;

(5) agreement on a likelihood function to update the expert's opinions as data become available; and

(6) actual updating of those opinions in a computer.

Thus when a new patient is to be assigned, an updated opinion is available that arguably represents each expert's current opinion. We call a treatment recommended by at least one (updated) expert admissible. Different sets of treatments might be admissible at any given time for patients that differ in the values of their diagnostic variables, and of course the set of admissible treatments changes over time as data are collected. The fundamental ethical stance proposed is that patients be assigned only admissible treatments. Within this constraint, treatments could be assigned in many ways, including randomly. However, the restriction to admissible treatments, which depends on the expert's priors (as updated by the data collected to date), is different from classical randomization which allows no such dependence. An exposition of the reasoning behind these ideas, with stress on the ethical side, can be found in Kadane (1986).

A group of doctors, lawyers, philosophers and statisticians have been meeting over the past few years to examine and refine these ideas. The results of these meetings are to be published in a volume now in preparation. Due to a fortunate happenstance, it became possible to do a test case of this method at Johns Hopkins Hospital. The trial compares verapamil and nitroprusside infusions as treatment for hypertension after separation from cardiopulmonary bypass during cardiac surgery. There are five experts from a variety of relevant specialties, only one of whom is from Johns Hopkins. Four independent variables were used to predict a function of the patient's blood pressure just after open heart surgery. As a function of these four independent variables, a normal linear model was proposed. The expert's priors were elicited using the methods of Kadane et al. (1980). The design, after taking account of the restriction to admissible treatments, stresses balance among the independent variables (Sedransk, 1973).

The nature of this design is that it is a compromise between the desire to respect better the right of a patient to the best available treatment, and the need for interpretable data. Since it is a compromise, it is liable to attack from both sides, those who feel that patients should have the right to choose their own treatment, and those who feel that only an unconstrained randomization can be allowed scientifically. This paper is aimed at the second group, and bases its appeal on the fact that unlike patient or physician choice, the design used here is a known, albeit complex and

dynamic, function of measured patient characteristics. Consequently, if the analysis respects the necessary conditioning on the four independent variables representing those measured patient characteristics, the treatment assigned carries no information about the parameters of interest, which measure the effectiveness of the treatments.

Stone asserts that only randomization permits the reader (Bayesian A) to ignore the experimenter's (Bayesian B's) priors and utility functions. We argue that it is a useful property of a design that the Bayesian reader need not model the author's prior and utility; but, we disagree with Stone that only classically randomized designs achieve this goal. The clinical trial, discussed above, does not require such modelling on the reader's part, whether or not it uses randomization to assign admissible treatments to patients.

6. Conclusion

In view of these comments, what are we to make of Savage's conjecture that the Bayesian explanation of randomization must lie in human imperfection and that more than one decision-maker is involved? With the latter, we heartily agree. However, the former seems to us too harsh. The desire that the analysis of an experiment not involve a judgment of the motivation of the experimenter seems natural for science. All too much outside of science rests on ad hominem arguments, which science tries to avoid when it can. Of course, in judging research proposals, the likely payoffs from proposed experiments are exactly the issue, and hence judgments of the people involved are inevitable. But is is certainly a comfort that the results of an experiment can be analyzed without having to impugn the experimeter. This consideration independent of the prior and utility function of the experimeter can be conducted under any design that puts the experiment on 'automatic pilot' once it is started, and hence does not allow the experimenter to get a finger on the scales. Randomization is one way to accomplish this, but it is not unique in having this virtue. We join with Savage and many others, however, in his respect for randomization as a statistical tool for enhancing interpersonal communication.

Acknowledgement

This research was sponsored in part by the Office of Naval Research under Contract N00014-85-K-0539 and in part by the Ethics and Values in Science and Technology Program of the National Science Foundation and the National Endowment for the Humanities under Grant ISP-8116810. Other scholars working on the latter grant are: David Kairys, Ken Schaffner, and Nell Sedransk, advised by Thomas J.J. Blanck, Eugenie S. Casella, Jack Coulehan, Alan Meisel, Preston Covey, A. John Popp, Jerome J. DeCosse, John C. Ruckdeschel, Arvin S. Glicksman, Katheryn D.

Katz and Rachelle Hollander. The test case at Johns Hopkins University is being conducted by Drs. Thomas J.J. Blanck and Eugenie S. Casella. The authors are grateful for the helpful comments of Morris DeGroot, Persi Diaconis and Dennis Lindley. While all these have contributed ideas that may appear here, none but the authors should be held responsible.

References

Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* **55**, 1-18.

Basu, D. (1981). Randomization analysis of experimental data: The Fisher randomization test (with discussion). *J. Amer. Statist. Assoc.* **75**, 575-595.

Bunke, H. and O. Bunke (1978). Randomization. Pro and contra. *Math. Operationsforsch. Statist. Ser. Statist.* **9**, 607-623.

DeGroot, M. (1980). Estimation of the correlation coefficient from a broken random sample. *Ann. Statist.* **8**, 264-278.

DeGroot, M. and J.B. Kadane (1983). Optimal sequential decisions involving more than one decision maker. In: H. Rizvi, J.S. Rustagi and D. Siegmund, Eds., *Recent Advances in Statistics-Papers Submitted in Honor of Herman Chernoff's Sixtieth Birthday*. Academic Press, New York, 197-210.

Fisher, R.A. (1971). *The Design of Experiments*, 9th ed., Hafner Press, New York.

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

Good, I.J. (1971). Twenty-seven principles of rationality. In: V.P. Godambe and D.A. Sprott, Eds., *Foundations of Statistical Inference*. Holt, Rinehart, and Winston, Toronto, 124-127.

Good, I.J. (1974). Random thoughts about randomness. In: K. Schaffner and R. Cohen, Eds., *PSA 1972*. Reidel, Dordrecht, 117-135.

Harsanyi, J.C. (1982). Subjective probability and the theory of games: Comments on Kadane and Larkey's paper. *Management Sci.* **28**, 121-124.

Kadane, J.B. (1986). Toward a more ethical clinical trial. *J. Medicine and Philos.* **11**, 385-404.

Kadane, J.B., J. Dickey, R. Winkler, W. Smith and S. Peters (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75**, 845-854.

Kadane, J.B. and P.D. Larkey (1982). Subjective probability and the theory of games. *Management Sci.* **28**, 113-120.

Kadane, J.B. and P.D. Larkey (1983). The confusion of is and ought in game theoretic contexts, *Management Sci.* **29**, 1365-1379.

Kadane, J.B. and N. Sedransk (1980). Toward a more ethical clinical trial. In: J. Bernardo, M. DeGroot, D. Lindley and A. Smith, Eds., *Bayesian Statistics*. University Press, Valencia, 329-338.

Kempthorne, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50**, 946-967.

Kempthorne, O. (1966). Some aspects of experimental inference. *J. Amer. Statist. Assoc.* **16**, 11-34.

Kempthorne, O. (1977). Why Randomize? *J. Statist. Plann. Inference* **1**, 1-25.

Levi, I. (1983). Direct inference and randomization. In: P. Asquith and T. Nickles, Eds., *PSA 1982*, Vol. 2. Edward Brothers, Ann Arbor, MI. 447-463.

Lindley, D.V. (1969). *Introduction to Probability and Statistics from a Bayesian Viewpoint* (2 Vols.) Cambridge University Press, Cambridge.

Lindley, D.V. (1972). *Bayesian Statistics, A Review*. SIAM, Philadelphia, PA.

Lindley, D.V. (1983). The role of randomization in inference. In: P. Asquith and T. Nickles, Eds., *PSA 1982*, Vol. 2. Edward Brothers, Ann Arbor, MI, 431-446.

Lindley, D.V. and M.R. Novick (1981). The role of exchangeability in inference. *Ann. Statist.* **9**, 45-58.

Pratt, J.W. and R.J. Zeckhauser (Eds.) (1985). *Principals and Agents: The Structure of Business*. Harvard School of Business Press, Boston, MA.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6**, 34-58.

Savage, L.J. (1961). The foundations of statistics reconsidered. In: J. Neyman, Ed., *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I. University of California Press, Berkeley, CA, 575-586.

Savage, L.J. (1962). Subjective probability and statistical practice. In: M.S. Bartlett, Ed., *The Foundations of Statistical Inference*. Methuen, London, 33-34.

Sedransk, N. (1973). Allocation of sequentially available units to treatment groups. *Internat. Statist. Inst. Proc.* **2**, 393-400.

Seidenfeld, T. (1981). Levi on the dogma of randomization in experiments. In: R. Bogdan, Ed., *Henry E. Kyburg, Jr. and Isaac Levi*. Reidel, Dordrecht, 263-291.

Stone, M. (1969). The role of experimental randomization in Bayesian statistics: Finite sampling and two bayesians. *Biometrika* **56**, 681-683.

Suppes, P. (1983). Arguments for randomizing. In: P. Asquith and T. Tickles, Eds., *PSA 1982*, Vol. 2. Edward Brothers, Ann Arbor, MI, 464-475.

Swijtink, Z. (1982). A Bayesian argument in favor of randomization. In: P. Asquith and T. Nickles, Eds., *PSA 1982*, Vol. I. Edward Brothers, Ann Arbor, MI, 159-168.

Wilson, J.G. (1986). Subjective probability and the prisoner's dilemma. *Management Sci.* **32**, 45-55.