The Annals of Statistics
1978, Vol. 6, No. 1, 34-58

BAYESIAN INFERENCE FOR CAUSAL EFFECTS: THE ROLE OF RANDOMIZATION

BY DONALD B. RUBIN

Educational Testing Service, Princeton, New Jersey

Causal effects are comparisons among values that would have been observed under all possible assignments of treatments to experimental units. In an experiment, one assignment of treatments is chosen and only the values under that assignment can be observed. Bayesian inference for causal effects follows from finding the predictive distribution of the values under the other assignments of treatments. This perspective makes clear the role of mechanisms that sample experimental units, assign treatments and record data. Unless these mechanisms are ignorable (known probabilistic functions of recorded values), the Bayesian must model them in the data analysis and, consequently, confront inferences for causal effects that are sensitive to the specification of the prior distribution of the data. Moreover, not all ignorable mechanisms can yield data from which inferences for causal effects are insensitive to prior specifications. Classical randomized designs stand out as especially appealing assignment mechanisms designed to make inference for causal effects straightforward by limiting the sensitivity of a valid Bayesian analysis.

1. Introduction and overview. Discussion of the role of randomization in the search for effective treatments is commonplace in the social and health sciences. See, for example, Campbell and Erlebacher (1970), Gilbert (1975), Gilbert, Light and Mosteller (1974), and Weinstein (1974). The rules of randomization imply that treatment assignment be made by an objectively defined random mechanism and not according to ad hoc decisions of the experimenters or the subjects of the experiments. Since human subjects may be randomized to treatments that some believe are less efficacious than other treatments under study, there is increasing interest in designs that reduce or eliminate randomization.

Some opponents of randomization turn to Bayesian statistics as a conceptual foundation for their position. However, careful development of the Bayesian framework for drawing inferences about causal effects of treatments explicates the steps required to analyze randomized and nonrandomized studies, and demonstrates that randomized studies are in general substantially easier to analyze than comparable nonrandomized studies. Therefore, we argue that randomization plays a central role in Bayesian inference for causal effects.

Intuitively, the causal effect of one treatment relative to another for a particular experimental unit is the difference between the result if the unit had been exposed to the first treatment and the result if, instead, the unit had been

Received October 1975; revised January 1976.

AMS 1970 *subject classifications*. Primary 62A15, 62B15, 62C10, 62F15, 62K99.

Key words and phrases. Bayesian, inference, randomization, causality, missing data, experimentation.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

exposed to the second treatment (Rubin, 1974): "If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone," or "Because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone." The problem is, of course, that we cannot return in time to give the other treatment, and so must compare an observed result and an unobserved result.

Section 2 formalizes this idea by conceptualizing inference for causal effects as inference about values that would have been observed under all possible assignments of treatments. Section 3 describes Bayesian models for (1) the prior distribution of the potentially observable data, (2) the mechanism that selects experimental units for exposure to treatments and assigns treatments, and (3) the mechanism that chooses values to record for data analysis.

Section 4 shows that the Bayesian statistician[1] generally needs all three models in order to draw inferences about causal effects. The models for the mechanisms used to select experimental units, assign treatments and record data can be ignored only in special cases when they are "ignorable," which means that they specify rules based on known, possibly probabilistic, functions of recorded values, as with randomization. For example, we show that if the experimenter assigns treatments so that the design "promised to tell him the most" (Savage, 1962), the assignment mechanism is ignorable only if the values used for making the assignment decisions are recorded and modelled in the data analysis. If these values are not recorded, the assignment mechanism is not ignorable and must itself be modelled in order to account for possible correlations between recorded values and unrecorded values used for assignment. In this case, inferences for causal effects are sensitive to prior specifications because recorded values cannot directly estimate correlations between recorded and unrecorded values. Inferences for causal effects are also sensitive to prior specifications when ignorable assignment mechanisms yield data poorly balanced with respect to recorded covariates.

Section 5 discusses the role of randomization. Ignorable mechanisms incorporating some randomization guard against data poorly balanced with respect to recorded covariates. More importantly, we show that classical randomized designs allow the Bayesian to achieve approximate balance with respect to many blocking variables, and to draw inferences for causal effects without having to formalize these blocking variables, record their values, or model their joint prior distribution with outcome variables. This freedom from having to deal explicitly with many blocking variables results in analyses for causal effects that are relatively straightforward and insensitive to prior specifications.

[1] By the terms "Bayesian statistician" or "Bayesian" we mean, more precisely, "the statistician who analyzes data by calculating, via Bayes theorem, the conditional distribution of unknowns given knowns."

2. Assumptions and notation. Consider a study of T *treatments* and a population P of N *experimental units* for which we wish to estimate the causal effects of the treatments. By a treatment we mean a series of well-defined actions that can be applied to a unit of study. Typical examples of treatments are medical or surgical interventions on patients with coronary artery disease. The treatment consisting of no active intervention is often of interest when the efficacies of proposed interventions are unclear. By an experimental unit we mean a particular unit of study (e.g., a person) at a particular time, since the effect of a treatment on a unit may depend on when the treatment is applied. Only a finite number of experimental units need be considered since no treatment will be applied into the infinite future.

The rows of the matrix in Figure 1 represent the N experimental units in P. The entries of the matrix correspond to all values that one might record in a study of the T treatments. The matrix is not a standard "units by variables" data matrix of observed values since in any study only some of the values represented are actually recorded. This explicit representation of all potentially observable values leads to substantial notation, but once established, the notation permits important conclusions to be drawn almost immediately.

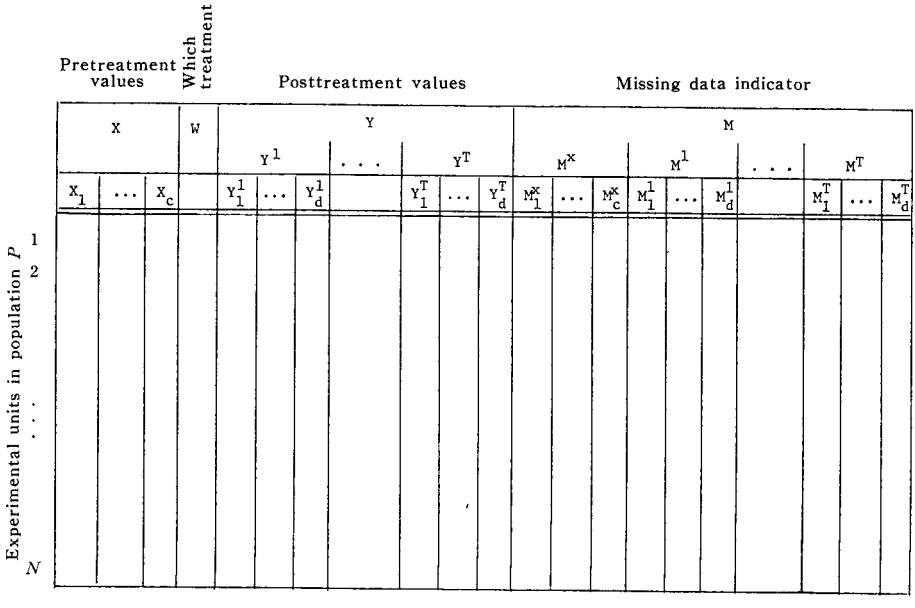

FIG. 1. All values in a study of T treatments.

2.1 *Pretreatment values—covariates.* The collection of c columns labelled $X = (X_1, X_2, \cdots, X_c)$ refers to the values of all variables describing experimental units that might be recorded *before* treatments are assigned. Each column of X refers to a particular aspect of experimental units such as preoperative blood

pressure, order of entry into experiment, age, or doctor's assessment of preoperative health. When recorded and used in the data analysis, a particular column of X is often called a covariate, concomitant variable, background variable, blocking factor or group indicator if used to define subpopulations of P. We assume that all pretreatment variables that might be used to distinguish between experimental units are included in X whether or not any of their values are actually recorded for data analysis.

2.2 *Assignment to treatment condition.* The column labelled W indicates which experimental units were selected for exposure to a treatment and which treatment each selected experimental unit received. Specifically, the elements of W take one of the $T + 1$ values $0, 1, \cdots, T$: $W_i = 0$ indicates that the ith experi mental unit was not selected and so not exposed to any of the T treatments being studied; $W_i = t\,(> 0)$ indicates that the experimental unit was selected and received treatment t.

In controlled experiments, W reflects two mechanisms, the sampling mechanism which determines the experimental units to be studied (i.e., to be exposed to a treatment) and the treatment assignment mechanism which determines the sampled experimental units to receive each treatment. In observational studies, W reflects the mechanism, beyond the control of the experimenter, that determines the experimental units to be exposed to each treatment; in such studies, the sampling mechanism that selects which treated experimental units are to be studied is reflected by the indicator variable discussed in Section 2.4.

2.3 *Posttreatment values—dependent variables.* The T collections of columns labelled $Y = (Y^1, Y^2, \cdots, Y^T)$ refer to the values of all variables describing experimental units that might be recorded *after* the assignment of treatments. Specifically, the collection of d columns labelled $Y^1 = (Y_1^1, \cdots, Y_d^1)$ refers to the observable values of Y if all experimental units were exposed to treatment 1 (i.e., if $W = (1, \cdots, 1)'$), $Y^2 = (Y_1^2, \cdots, Y_d^2)$ refers to the observable values of Y if all experimental units were exposed to treatment 2, and so on. If an experimental unit was not selected for exposure to one of the treatments, we assume that no Y values will be observed for the experimental unit. (Hence there is no need for a Y^0.) Each collection Y^t includes the same d aspects of experimental units, a particular column in Y^t referring for example to postoperative blood pressure, or doctor's assessment of postoperative health. The two columns Y_k^1 and Y_k^2 thus refer to the same aspect of the experimental units but given exposure to different treatments. In a data analysis, a particular $Y_k = (Y_k^1, Y_k^2, \cdots, Y_k^T)$ is usually called an outcome variable, a dependent variable, a response variable, or a criterion variable. There are T columns representing each response variable because the observable value of the aspect represented by Y_k would generally differ under different treatments.

In order for this representation using T columns for each Y_k to be adequate, we must assume that if the ith experimental unit is selected for treatment

exposure and assigned treatment t (> 0), the observable value of Y is the same for all assignments of treatments to the other experimental units. That is, for each k, Y_{ki}^t represents the ith experimental unit's observable value of Y_k for all values of W such that $W_i = t$ (> 0). Without this assumption, we would need more than T versions of each Y_k, since we require a different version for each value of W that leads to a potentially different value of Y_k. This assumption, called "no 'interference' between different units" by Cox (1958, page 19), is usually made in practice. A common exception is cross-over designs with additive carry-over effects assumed. Our model can be extended to include more general assumptions about interference effects between units, but for notational and descriptive simplicity we assume T versions of each Y_k are adequate. The general results of this paper do not rely on the no-interference assumption.

2.4 *Defining causal effects.* The causal effects of the treatments are comparisons among the Y^t values. For example, it is common to define the causal effect of treatment 1 vs. treatment 2 on Y_k for the ith experimental unit to be the ith component of $Y_k^1 - Y_k^2$, $Y_{ki}^1 - Y_{ki}^2$: the difference for the ith experimental unit between Y_k given exposure to treatment 1 and Y_k given exposure to treatment 2. The fundamental problem facing inference for causal effects is that if treatment t is assigned to the ith experimental unit (i.e., if $W_i = t$), only values in Y^t can be observed, Y^j values for $j \neq t$ being unobservable (or missing).

Without the no-interference assumption discussed in Section 2.3, more complicated definitions of causal effects are needed. However, such definitions still involve comparisons of Y_k values only some of which could be observed in a particular study.

2.5 *Missing-data indicator.* The $c + dT$ columns labelled M in Figure 1 indicate recorded and unrecorded values in (X, Y) at the time of the data analysis. The c columns labelled $M_1^x, \cdots, M_c^x$ indicate recorded and unrecorded values in $X_1, \cdots, X_c$: if $M_{ki}^x = 1$, X_{ki} is recorded for the data analysis; if $M_{ki}^x = 0$, X_{ki} is not recorded for the data analysis. Similarly, the d columns labelled $M_1^t, \cdots, M_d^t$ indicate recorded and unrecorded values in $Y^t = (Y_1^t, \cdots, Y_d^t)$, $t = 1, \cdots, T$. For notational simplicity we assume that W is always observed.

The fundamental problem facing inference for causal effects is reflected by the restriction that $W_i = t$ implies $M_{ki}^j = 0$ for all $j \neq t$ and $k = 1, \cdots, d$. That is, one cannot record values under a treatment not assigned.

Sometimes an element of M is itself unrecorded (e.g., two measurements of blood pressure are taken on each patient, but for some patients only one value is recorded and it is not clear whether it is the first or second measurement). As with unrecorded W_i, this can be handled by using more indicator variables, but the extra generality adds little insight and clutters the notation. Hence, we assume M is known at the time of the data analysis.

In practice, the indicator M reflects the mechanism that chooses which values to record for the data analysis, mechanisms that create unexpected (accidental)

missing values, as well as the sampling mechanism and treatment assignment mechanism (via the restrictions placed on M by W).

2.6 *A specific study.* In our model, we assume that any values capable of distinguishing between experimental units are elements in (X, Y, W, M) with W and M fully observed and (X, Y) partially observed. In a specific study of T treatments let $\tilde{W}$ and $\tilde{M}$ be the actual observed values of W and M, and for notational convenience write $\tilde{X} = (X_{(0)}, \tilde{X}_{(1)})$ and $\tilde{Y} = (Y_{(0)}, \tilde{Y}_{(1)})$ where each element of $(X_{(0)}, Y_{(0)})$ corresponds to an element of $\tilde{M}$ that is 0, and each element of $(\tilde{X}_{(1)}, \tilde{Y}_{(1)})$ corresponds to an element of $\tilde{M}$ that is 1. Thus $(\tilde{X}_{(1)}, \tilde{Y}_{(1)}, \tilde{W}, \tilde{M})$ is comprised of known numbers and $(X_{(0)}, Y_{(0)})$ is comprised of unknowns. Notice that these partitions of $\tilde{X}$ and $\tilde{Y}$ are defined by the observed value of M, $\tilde{M}$.

Within the structure developed, problems of inference for causal effects of treatments on individual experimental units or collections of experimental units are equivalent to problems of inference about values of missing data. That is, most of $(\tilde{X}, \tilde{Y})$ is missing, as indicated by $\tilde{M}$. If all of $(\tilde{X}, \tilde{Y})$ were observed (which is impossible), we would simply calculate causal effects. Because of missing data, we must turn to a method of statistical inference in order to estimate causal effects. The concern here is with Bayesian inference. See Rubin (1976) for general discussion of inference when confronted with missing data and Rubin (1977a) for brief discussion of sampling distribution analogues for the Bayesian results presented here.

2.7 *Relating the model to the real world.* Several aspects of our model for causal effects should be clearly understood before it is applied to real world problems. The common theme throughout this section is the need for clear definition and understanding of the actions to be performed on those experimental units selected for treatment.

First, within our model, each of the T treatments must consist of a series of actions that could be applied to each experimental unit. This requirement may seem obvious, but some colloquial uses of "cause" specify treatments that either cannot be applied or are so ambiguous that no series of actions can be inferred from the description of the treatment; such questions have no causal answer within our framework. For example, consider the causal effect of sex (male—female) on intelligence. What are the actions to be applied to each experimental unit that define the treatments? Are we to give hormone shots beginning at birth and surgically perform a "sex-change" operation, or at conception "change" Y-chromosomes and X-chromosomes? Even if an "at-conception X-for-Y chromosome change" becomes possible, presumably there will be several techniques developed for effecting the change with potentially different causal effects. Without treatment definitions that specify actions to be performed on experimental units, we cannot unambiguously discuss causal effects of treatments.

Second, in our model, the causal effects being estimated reflect the pretreatment manipulations carried out on sampled experimental units as well as the

different actions that define the treatments. For example, suppose in a medical experiment all sampled patients are exposed to extensive medical examination before treatments are assigned. The causal effects being estimated in such a study assume that these examinations are given to experimental units before exposure to treatment. If the pretreatment manipulations performed on the sampled experimental units are quite different from the pretreatment manipulations that are likely to be used with future experimental units exposed to the treatments, it may be wise to consider studying more realistic pretreatment manipulations since there may exist interactions between the treatments and the pretreatment manipulations.

Finally, in our model, we cannot attribute cause to one particular action in the series of actions that define a treatment. Thus treatments that appear similar because of a common salient action are not the same treatment and may not have similar causal effects. An important practical implication is that treatments given under double-blind conditions are different treatments than those given under blind or simple conditions and may have different causal effects; see Rosenthal (1976) for a summary of studies of interactions between experimenters and their subjects. Double-blind versions of treatments are generally of more scientific interest, although simple versions of treatments may in fact be of more immediate applied interest. For example, in an experiment to compare aspirin and a prescription drug, simple versions of the treatments may be important because in practice a patient will usually know if his doctor is recommending aspirin rather than a prescription drug. In some cases then, it may be wise to consider studying both scientific (double-blind) and applied (simple) versions of treatments, since different versions of treatments, being different treatments, may have different causal effects.

3. The distribution of the random variables. Bayesian inference considers the observed values $\tilde{X}_{(1)}$, $\tilde{Y}_{(1)}$, $\tilde{W}$, $\tilde{M}$ to be realizations of random variables and the missing values $X_{(0)}$, $Y_{(0)}$ to be unobserved random variables. For a study of T treatments, the random variables are thus (X, Y, W, M). There is a specification for the distribution of the random variables given an *unknown* (vector) parameter π which is itself a random variable having a known prior (or marginal) distribution. For reasons to be given shortly, we let $f(X, Y|\pi)k(W|X, Y, \pi)g(M|X, Y, W, \pi)$ be the joint probability density function for the random variables (X, Y, W, M) given π, and let $p(\pi)$ be the prior distribution of π.[2] The distribution $f(X, Y|\pi)$ is the marginal density of the potentially observable data (X, Y)

[2] Actually $p(\pi)$ is the conditional distribution of π given the choice of the families of models f, k, g. Writing $\pi_{Y,X}$ for the function of π that appears in f, the distribution of $\pi_{Y,X}$ given the family of models f, g, k may or may not be the same as the distribution of $\pi_{Y,X}$ given f, depending on the distribution placed on the process of choosing the models f, k, g. The distribution of $\pi_{Y,X}$ given f (not given f, k, g) is presumably what a Bayesian means by the prior distribution of the parameter of the data. This distinction is usually unimportant in practice when prior distributions are chosen to be relatively diffuse.

given the parameter π. The distribution $k(W|X, Y, \pi)$ is the probability of the assignment W of treatments given the value (X, Y) for the data and the parameter π; we call $k(W|X, Y, \pi)$ the *assignment mechanism* since it reflects the mechanisms that select experimental units to be assigned treatments and assign treatments to the selected experimental units. The distribution $g(M|X, Y, W, \pi)$ is the probability of the pattern M of recorded and unrecorded values in X, Y given: the value (X, Y) for the data, the value W for the assignment of treatments and the parameter π; we call $g(M|X, Y, W, \pi)$ the *recording mechanism* since it reflects the mechanisms that determine which values are recorded for the data analysis.

The purpose of formulating all of these distributions is simply to tie unobserved values to observed values so that the values we see tell us something about the the values we do not see. This perspective holds that the models are only used to draw inferences about the unobserved values $X_{(0)}$, $Y_{(0)}$, and thereby the causal effects of the treatments.

3.1 *Restrictions on the prior distribution of the data.* By assumption (Section 2.6), all values capable of distinguishing between experimental units are included in (X, Y, W, M). Hence, without loss of generality we can assume that the indices of the experimental units are assigned as a random permutation of the integers $1, \cdots, N$. Doing so, the distribution $f(X, Y|\pi)k(W|X, Y, \pi)g(M|X, Y, W, \pi)$ must remain constant under permutation of the row indices of (X, Y, W, M). Similarly the distribution $f(X, Y|\pi)$ must remain constant under permutation of the row indices of (X, Y). It follows from standard results on exchangeable random variables with infinite N (de Finetti, 1964; Hewitt and Savage, 1955; Feller, 1965, pages 225–226) and recent extensions to finite N (Diaconis, 1976), that we can assume with nearly no loss of generality the rows of (X, Y) to be independent and identically distributed (i.i.d) given π:

$$f(X, Y|\pi) = \prod_{i=1}^{N} f_*((X, Y)_i|\pi) \tag{3.1}$$

where $(X, Y)_i$ is the ith row of (X, Y).

In practice, other restrictions besides equation (3.1) are made on $f(X, Y|\pi)$. Already discussed are the restrictions on Y that follow from the usual assumption of no-interference between experimental units. Less commonly, a particular Y_k may be assumed to have the same value no matter which treatment is applied: $Y_k^1 = Y_k^2 = \cdots = Y_k^T$ (when studying whether an additive in the diet of patients decreases cholesterol in the blood, the precision of the experiment might be improved by using "the amount of saturated fat eaten" as a covariate, claiming that it is unaffected by the additive).

Since in any practical problem, any Y or any X can take only a finite number of distinct values, with no loss of generality, $f(X, Y|\pi)$ can be assumed to be an unrestricted multinomial over the $c + Td$-dimensional contingency table of possible values of X, Y. Often, in order to reduce the number of parameters

appearing in this contingency table, $p(\pi)$ places severe restrictions on these parameters, and so smooths the multinomial probabilities. Prior restrictions that model each row of (X, Y) as, for example, multivariate normal, reflect the usual efforts of model building.

3.2 *Ignorable assignment and recording mechanisms.* The explicit inclusion of W and M as random variables is not standard but is central to understanding Bayesian inference for causal effects. Moreover, the factorization of the joint distribution of (X, Y, W, M) used above isolates essential differences between the random variables (X, Y), W and M. The model $f(X, Y|\pi)$ is never totally under the experimenter's control since the conditional distribution of Y given X reflects the state of nature; the marginal distribution of X is to some extent under the experimenter's control since he defines P by choice of experimental units. The assignment mechanism $k(W|X, Y, \pi)$ can be under the experimenter's control since he can assign treatments to experimental units. The recording mechanism $g(M|X, Y, W, \pi)$ is mainly under the experimenter's control, since subject to constraints of treatment assignment and aside from unexpected missing data, he controls which values to record for data analysis. Thus the assignment and recording mechanisms differ from the model for the distribution of the data in that they can be "known" a priori in the following sense.

DEFINITION 1. The assignment mechanism $k(W|X, Y, \pi)$ is ignorable at $(\tilde{X}_{(1)}, \tilde{Y}_{(1)}, \tilde{W}, \tilde{M})$ if the probability of the observed pattern of treatment assignments given $\tilde{X}$, $\tilde{Y}$ and π, $k(\tilde{W}|\tilde{X}, \tilde{Y}, \pi)$, takes the same known value for all values of the unknowns $X_{(0)}$, $Y_{(0)}$, π.

DEFINITION 2. The recording mechanism $g(M|X, Y, W, \pi)$ is ignorable at $(\tilde{X}_{(1)}, \tilde{Y}_{(1)}, \tilde{W}, \tilde{M})$ if the probability of the observed pattern of recorded values given $\tilde{X}$, $\tilde{Y}$, $\tilde{W}$ and π, $g(\tilde{M}|\tilde{X}, \tilde{Y}, \tilde{W}, \pi)$ takes the same known value for all values of the unknowns $X_{(0)}$, $Y_{(0)}$, π.

If the assignment mechanism depends on (X, Y) values, then those values must be recorded by the recording mechanism if the assignment mechanism is to be ignorable. For example, consider a two-treatment study and the "play-the-winner" sequential $k(W|X, Y, \pi)$ where the next patient receives the treatment that past data suggest is better; all values used in making these decisions, such as order of entry into the study, must be recorded for data analysis if the assignment mechanism is to be ignorable. Similarly, if a doctor assigns patients to treatments so as to balance the distribution of background variables such as age and sex, these variables must be recorded if the assignment mechanism is to be ignorable. If a doctor assigns treatments according to his unrecorded judgments about the health of patients, the assignment mechanism is not ignorable. Or if patients select the treatments themselves on the basis of their unrecorded opinions of their health, the assignment mechanism is not ignorable. These examples illustrate that for a particular assignment mechanism, one can choose a recording mechanism that makes the assignment mechanism not ignorable

except when $k(W|X, Y, \pi) = k(W|\pi)$ (e.g., simple random sampling followed by a completely randomized experiment). The more involved the assignment mechanism (in the sense of depending on more values), the more complete must be the recording mechanism if the assignment mechanism is to be ignorable.

In addition, the assignment mechanism cannot depend on (unknown) parameters if it is to be ignorable. For example, if in the play-the-winner study the stopping rule for ending the study (letting further $W_i = 0$) is determined by the experimenter's unrecorded criterion of "enough evidence," the assignment mechanism is not ignorable since the criterion is unknown and thus some function of the parameter π. In observational studies (Cochran and Rubin, 1973), such as those of heart disease and smoking, the assignment mechanism is beyond the control of the experimenter, and thus generally it is not reasonable to model it as being ignorable.

The recording mechanism $g(M|X, Y, W, \pi)$ is largely under the control of the experimenter/data analyst. In fact, in almost any real data analysis, some values that could be recorded are not. For example, often unit labels, times of initiation of treatments and other aspects thought a priori to be uninteresting (e.g., length of fingers of patients) are not recorded for data analysis. Such a priori decisions are completely specified and so imply recording mechanisms that are ignorable. However, if there is the possibility of unintended missing values, the recording mechanism would not be ignorable (except by assumption) even if there are no unintended missing values since the probability of such an occurrence may be a function of values in $(X_{(0)}, Y_{(0)})$ or π (e.g., perhaps Y^1_{1i} is observed because it is less than a function of π or because Y^2_{1i} would have been greater than 0).

4. Bayesian inference for causal effects. Using the distributions defined in Section 3, Bayesian inference for causal effects proceeds by calculating the predictive distribution of $Y_{(0)}$ given the observed values, $\tilde{X}_{(1)}$, $\tilde{Y}_{(1)}$, $\tilde{W}$, $\tilde{M}$:

$$\text{(4.1)}\qquad \Pr(Y_{(0)}|\tilde{X}_{(1)}, \tilde{Y}_{(1)}, \tilde{W}, \tilde{M}) = \frac{\int\int k(\tilde{W}|\tilde{X}, \tilde{Y}, \pi)g(\tilde{M}|\tilde{X}, \tilde{Y}, \tilde{W}, \pi)f(\tilde{X}, \tilde{Y}|\pi)p(\pi)\,d\pi\,dX_{(0)}}{\int\int\int k(W|\tilde{X}, \tilde{Y}, \pi)g(M|\tilde{X}, \tilde{Y}, W, \pi)f(\tilde{X}, \tilde{Y}|\pi)p(\pi)\,d\pi\,dX_{(0)}\,dY_{(0)}}.$$

The predictive density (4.1), in conjunction with the observed $\tilde{Y}_{(1)}$ values, yields the Bayesian inferences for the causal effects of the T treatments for the N experimental units in the population P. Commonly, the predictive distribution of the column means of Y is calculated, the difference between each pair of means being the average causal effect in P of one treatment relative to another. It is also common to calculate predictive distributions for average causal effects in various subgroups of P (e.g., males and females).

4.1 *The role of ignorable assignment and recording mechanisms.*

THEOREM. *If the assignment and recording mechanisms are ignorable, then Bayesian inference for causal effects is completely determined by*

(a) *the observed values* $\tilde{X}_{(1)}$, $\tilde{Y}_{(1)}$, $\tilde{M}$,

(b) *the specification for the conditional distribution of* Y *given* $X_{(1)}$ *and* π:

$$h(Y \mid X_{(1)}, \pi) = \int f(X, Y \mid \pi)\, dX_{(0)} / \int\int f(X, Y \mid \pi)\, dX_{(0)}\, dY\,, \tag{4.2}$$

and

(c) *the specification for the conditional distribution of* π *given* $X_{(1)}$:

$$q(\pi \mid X_{(1)}) = p(\pi) \int\int f(X, Y \mid \pi)\, dX_{(0)}\, dY / \int\int\int p(\pi) f(X, Y \mid \pi)\, dX_{(0)}\, dY\, d\pi\,. \tag{4.3}$$

PROOF. This result is immediate because, from the definitions of ignorable mechanisms, equation (4.1) can be rewitten as

$$\text{Pre}\,(Y_{(0)} \mid \tilde{X}_{(1)}, \tilde{Y}_{(1)}, \tilde{W}, \tilde{M}) = \frac{\int h(\tilde{Y} \mid \tilde{X}_{(1)}, \pi) q(\pi \mid \tilde{X}_{(1)})\, d\pi}{\int\int h(\tilde{Y} \mid \tilde{X}_{(1)}, \pi) q(\pi \mid \tilde{X}_{(1)})\, d\pi\, dY_{(0)}}\,. \tag{4.4}$$

This theorem shows that, given ignorable assignment and recording mechanisms, inference for causal effects follows from the observed values and the usual data specification ignoring the assignment and recording mechanisms.

Further simplification often occurs in practice in the case of ignorable assignment and recording mechanisms. It is common to parameterize $f(X, Y \mid \pi)$ so that the conditional distribution of Y given X depends on one function of π say $\pi_{Y \mid X}(\pi)$, and the marginal distribution of X depends on another function of π, say $\pi_X(\pi)$, where $\pi_{Y \mid X}$ and π_X are a priori independent. For example, in normal regression models, $\pi_{Y \mid X}$ includes the regression coefficients of Y on X and the conditional covariance of Y given X, and these are often declared a priori independent of the parameter of the marginal distribution of X, π_X. With such models and ignorable assignment and recording mechanisms, inferences for causal effects are determined simply by the observed values $\tilde{X}_{(1)}$, $\tilde{Y}_{(1)}$, $\tilde{M}$, and the prior specifications for the conditional distribution of Y given $X_{(1)}$ and the marginal distribution of $\pi_{Y \mid X}$.

If either the assignment mechanism or the recording mechanism is not ignorable, a valid Bayesian analysis follows from (4.1); using (4.4) in place of (4.1) does not yield a valid Bayesian analysis. However, in practice, mechanisms that are not ignorable because of dependence on $(X_{(0)}, Y_{(0)})$ usually pose a greater problem than those that are not ignorable solely because of dependence on π. The reason is that in many practical problems, it is common to choose prior distributions to be relatively diffuse and to reflect weak prior dependencies between parameters. In a notation analogous to that used above, if $\pi_{X,Y}(\pi)$ and $\pi_{W,M \mid X,Y}(\pi)$ are a priori independent, then if neither $k(\tilde{W} \mid \tilde{X}, \tilde{Y}, \pi)$ nor $g(\tilde{M} \mid \tilde{X}, \tilde{Y}, \tilde{W}, \pi)$ depends on $(X_{(0)}, Y_{(0)})$, inferences for causal effects follow from (4.4). Of course if $\pi_{X,Y}$ and $\pi_{W,M \mid X,Y}$ are not independent a priori, inferences for causal effects will vary with the specification of the a priori relationship between them.

If either $k(\tilde{W} \mid \tilde{X}, \tilde{Y}, \pi)$ or $g(\tilde{M} \mid \tilde{X}, \tilde{Y}, \tilde{W}, \pi)$ depends on $(X_{(0)}, Y_{(0)})$, inference for causal effects (equation (4.1)) depends on relationships between the observed values $(\tilde{X}_{(1)}, \tilde{Y}_{(1)})$ and missing values used to assign treatments and/or record values. Since observed values cannot directly estimate these relationships, causal

inferences vary with their prior specification, and it is usually inappropriate to assume that given π the unobserved values used to assign treatments and/or record data are uncorrelated with observed values $(X_{(1)}, Y_{(1)})$.

4.2 *Illustrating the sensitivity of inferences for causal effects.* Suppose $T = 2$ and n experimental units (e.g., patients) are randomly sampled from P for exposure to treatments (e.g., operations). Also suppose the recording mechanism records a Y_1 value for each of the n experimental units exposed to a treatment but records no other values; Y_1 is dichotomous: 1 indicates success (e.g., of the operation) and 0 indicates failure. Let $\bar{Y}_1^t = \sum_{i=1}^N Y_{1i}^t/N$ be the proportion of experimental units in P for whom treatment t is successful, and suppose for simplicity that interest focuses on the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$.

First assume that each sampled experimental unit is assigned with probability $\frac{1}{2}$ to treatment 1 and probability $\frac{1}{2}$ to treatment 2. Thus, the assignment and recording mechanisms are ignorable. By the theorem of Section 4.1, inferences for causal effects are determined by the observed values $(\tilde{Y}_{(1)}, \tilde{M})$ and by the specifications for the joint distribution of (Y_1^1, Y_1^2) given π and the prior distribution of π. Let the rows of (Y_1^1, Y_1^2) given π be i.i.d. with $\pi^t = \pi^t(\pi)$ the prior probability that $Y_{1i}^t = 1$, $t = 1, 2$. Thus, the model for (Y_1^1, Y_1^2) is a 2×2 contingency table, Y_{1i}^t being binomial with probability π^t of success, $t = 1, 2$; see Figure 2a. We do not explicitly parameterize the joint distribution of (Y_1^1, Y_1^2) for the following reason. In most practical cases, the population of experimental units that might be exposed to the treatments being studied is considered to be much larger than the sample of experimental units exposed to treatments. Since as $N/n \to \infty$ the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ converges to the posterior distribution of (π^1, π^2), interest often focuses on this posterior distribution, with other parameters being considered nuisance parameters.

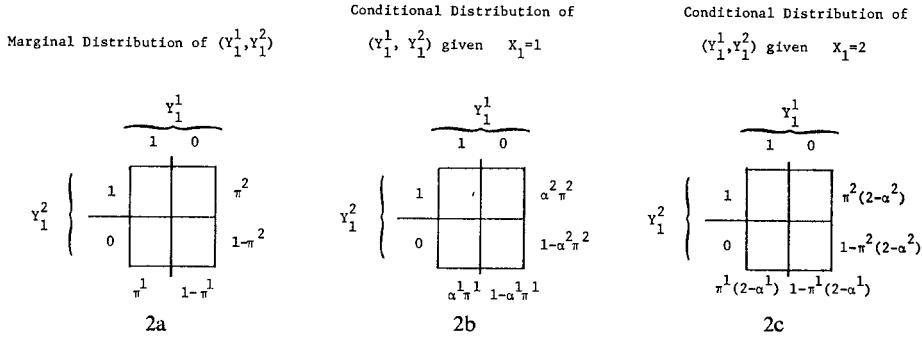

FIG. 2. Specification of distribution of (Y_1^1, Y_1^2, X_1) for examples.

Let $\bar{y}^t = \sum_1^N \tilde{Y}_{1i}^t \tilde{M}_{1i}^t/n^t$ (where $n^t = \sum_1^N \tilde{M}_{1i}^t$), the observed proportion of experimental units exposed to operation t for which operation is successful, $t = 1, 2$. Then the posterior distribution of π is proportional to

$$p(\pi) \prod_{t=1}^2 [\pi^t]^{n^t \bar{y}^t} [1 - \pi^t]^{n^t(1-\bar{y}^t)} . \tag{4.5}$$

In order to dramatize the reduced sensitivity to model specification that is possible when ignorable mechanisms are used, consider the large sample case with $n \to \infty$ and $N/n \to \infty$. The limiting posterior distribution of π with a $p(\pi)$ giving positive density to all $(\pi^1, \pi^2) \in [0, 1] \times [0, 1]$ is then proportional to

$$p(\pi) \prod_{t=1}^{2} \delta(\pi^t - \bar{y}^t) \tag{4.6}$$

where

$$\begin{aligned} \delta(a) &= 1 \quad \text{if} \quad a = 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Thus, in this limiting case, the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ converges to point mass at $(\bar{y}^1, \bar{y}^2)$.

If the assignment and/or recording mechanisms are not ignorable, inferences for $(\bar{Y}_1^1, \bar{Y}_1^2)$ are sensitive to prior specifications even when $n \to \infty$ and $N/n \to \infty$. An extension of the above example will illustrate this fact. As before, suppose that (a) the recording mechanism records only Y_1 values, (b) the n experimental units exposed to treatments are a random sample from P, and (c) the prior distribution of (Y_1^1, Y_1^2) is the 2×2 contingency table in Figure 2a. However, now suppose that the unrecorded dichotomous variable X_1 (e.g., doctor's assessment of health status prior to operation) is used to assign treatments to sampled experimental units: if $X_{1i} = 1$ (good health), with probability $\theta = \theta(\pi)$ the ith experimetal unit (if sampled) is exposed to treatment 1 and with probability $(1 - \theta)$ is exposed to treatment 2; if $X_{1i} = 2$ (poor health), with probability $(1 - \theta)$ the ith experimental unit (if sampled) is exposed to treatment 1 and with probability θ is exposed to treatment 2. Let the rows of (Y_1, X_1) be i.i.d. given π, where the prior probability that $X_{1i} = k$ is $\frac{1}{2}$, $k = 1, 2$, and $\alpha^t \pi^t \leqq 1$ is the prior probability that $Y_{1i}^t = 1$ given $X_{1i} = 1$, $\alpha^t = \alpha^t(\pi) \leqq 2$. Thus, the model for (Y_1^1, Y_1^2, X_1) is a $2 \times 2 \times 2$ contingency table where X_{1i} is binomial 0.5, and conditional on $X_{1i} = k$, Y_{1i}^t is binomial with probability of success $\alpha^t \pi^t$ if $k = 1$ and $\pi^t(2 - \alpha^t)$ if $k = 2$; see Figure 2.

In this nonignorable case it is easy to show[3] that

$$\begin{aligned} \Pr(Y_{1i}^t = 1 \mid W_i = t, \pi) = \xi^t = \xi^t(\pi) & \\ = \pi^1[2(1 - \theta) + \alpha^1(2\theta - 1)] \quad & \text{if} \quad t = 1 \\ = \pi^2[2\theta + \alpha^2(1 - 2\theta)] \quad & \text{if} \quad t = 2\,. \end{aligned} \tag{4.7}$$

---

[3] $\Pr(Y_{1i}^t = 1 \mid W_i = t, \pi) = \Pr(W_i = t \mid \pi)^{-1} \Pr(Y_{1i}^t = 1, W_i = t \mid \pi)$
$= 2 \sum_{k=1}^{2} \Pr(X_i = k \mid \pi) \Pr(Y_{1i}^t = 1, W_i = t \mid X_i = k, \pi)$
$= \sum_{k=1}^{2} \Pr(W_i = t \mid Y_{1i}^t = 1, X_i = k, \pi) \Pr(Y_{1i}^t = 1 \mid X_i = k, \pi)$

where

$$\begin{aligned} \Pr(W_i = t \mid Y_{1i}^t = 1, X_i = k, \pi) &= \theta \quad \text{if} \quad t = k \\ &= 1 - \theta \quad \text{if} \quad t \neq k \end{aligned}$$

and

$$\begin{aligned} \Pr(Y_{1i}^t = 1 \mid X_i = k, \pi) &= \pi^t \alpha^t \quad \text{if} \quad k = 1 \\ &= \pi^t(2 - \alpha^t) \quad \text{if} \quad k = 2\,. \end{aligned}$$

Hence the posterior distribution of π is proportional to

$$p(\pi) \prod_{t=1}^{2} [\xi^t]^{n^t \bar{y}^t} [1 - \xi^t]^{n^t(1-\bar{y}^t)}, \tag{4.8}$$

Suppose $p(\pi)$ assigns positive density to all $(\xi^1, \xi^2) \in [0, 1] \times [0, 1]$. Then as $n \to \infty$ and $N/n \to \infty$, the limiting posterior distribution of π is proportional to

$$p(\pi) \prod_{t=1}^{2} \delta(\xi^t - \bar{y}^t). \tag{4.9}$$

Suppose further that conditional on each $(\theta, \alpha^1, \alpha^2)$ having positive prior density, $p(\pi)$ assigns positive prior density to all $(\pi^1, \pi^2) \in [0, 1] \times [0, 1]$. Then conditional on a $(\theta, \alpha^1, \alpha^2)$ that has positive posterior density, the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ converges to point mass at

$$(\bar{y}^1/[2(1 - \theta) + \alpha^1(2\theta - 1)], \bar{y}^2/[2\theta + \alpha^2(1 - 2\theta)]). \tag{4.10}$$

Therefore, the limiting predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ is expression (4.10) averaged over the posterior distribution of $(\theta, \alpha^1, \alpha^2)$, which is simply expression (4.10) averaged over the prior distribution of $(\theta, \alpha^1, \alpha^2)$ restricted to values of $(\theta, \alpha^1, \alpha^2)$ such that for $t = 1, 2$, (a) the tth component of (4.10) is $\leqq 1$ and (b) α^t times the tth component of (4.10) is $\leqq 1$. Consequently, even in this limiting case, the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ will be quite sensitive to the specification of the prior distribution of $(\theta, \alpha^1, \alpha^2)$.

If $\theta = \frac{1}{2}$, each experimental unit was assigned with probability $\frac{1}{2}$ to each treatment condition, the assignment mechanism is ignorable, and the predictive distribution converges to point mass at $(\bar{y}^1, \bar{y}^2)$. Or if $\alpha^1 = \alpha^2 = 1$, X_1 is a priori independent of (Y_1^1, Y_1^2) (and so equivalent to a binary random number), the assignment mechanism is equivalent to an ignorable mechanism, and the predictive distribution converges to point mass at $(\bar{y}^1, \bar{y}^2)$. In general however, the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ does not converge to a single point, since the possible values of the components of (4.10) range between $\bar{y}^t/2$ and $\bar{y}^t/2 + .5$.

Since the likelihood function in (4.5) has a unique maximizing value of (π^1, π^2) equal to $(\bar{y}^1, \bar{y}^2)$ while the likelihood in (4.8) has a ridge of maximizing values of (π^1, π^2) defined by equating (π^1, π^2) to expression (4.10), in finite samples the nonignorable assignment mechanism will lead to inferences for causal effects more sensitive to $p(\pi)$ than the ignorable assignment. In real data problems with moderate sample sizes and assignment and/or recording mechanisms that are not ignorable, the consideration of realistic models for $f(X, Y|\pi)p(\pi)$ may easily lead to such a variety of inferences under the different models, or such a large posterior variance when a specific prior distribution is placed on the models, that the Bayesian may consider his data worthless for inference for causal effects.

4.3 *Ignorable mechanisms poorly suited for causal inference.* The example of Section 4.2 indicated the difficult problems of inference for causal effects that can arise when the assignment mechanism is not ignorable: a valid Bayesian analysis can easily be very sensitive to the data specification $f(X, Y|\pi)p(\pi)$. This does *not* imply that all combinations of ignorable assignment and recording

mechanisms lead to data from which causal inferences are insensitive to the data specification. By the theorem of Section 3.1, even with ignorable assignment and recording mechanisms and $\pi_{Y|X}$ a priori independent of π_X, we still must estimate the conditional distribution of Y given $\tilde{X}_{(1)}$, and this can be a problem. If the experimental units exposed to one treatment have different values of $\tilde{X}_{(1)}$ than the experimental units exposed to another treatment, the causal inferences will be quite sensitive to the specifications $h(Y \mid X_{(1)}, \pi)$ and $p(\pi)$.

As an example, Weinstein (1974, page 7) in some cases seems to recommend assigning patients to operations on the basis of their preferences and tries to justify the procedure by a Bayesian decision-theoretic argument. Suppose the only X variable recorded is X_1 = patient's preference for operation 1 or operation 2. Since all patients receive the operation they preferred, resultant inferences about the causal effects of the operations are very sensitive to prior specifications about associations between patient preference and Y given the nonpreferred operation. That is, for $X_{1i} = 1$ we can observe only Y^1, and for $X_{1i} = 2$ we can observe only Y^2, yet we need to estimate the conditional distributions of Y^1 given $X_1 = 2$ and Y^2 given $X_1 = 1$. For a specific example with a dichotomous Y, consider the example of Section 4.2 with $\theta = 0$, noting that $\theta = 0$ implies X_{1i} is known; in large samples, the predictive distribution of $(\bar{Y}_1^1, \bar{Y}_1^2)$ then converges to $(\bar{y}^1/(2 - \alpha^1), \bar{y}^2/\alpha^2)$ averaged over the prior distribution of (α^1, α^2) restricted to those (α^1, α^2) for which $\alpha^1 \leqq 2 - \bar{y}^1$ and $\alpha^2 \geqq \bar{y}^2$. Although there are nonstatistical justifications for assignment to patient-preferred treatments, the Bayesian statistician must consider the design inappropriate for estimating the causal effects of the operations unless strong prior information exists about relationships between patient preference for operation and the outcome of each operation.

Other obvious examples of ignorable assignment and recording mechanisms that lead to data poorly suited for causal inference include cases with Y recorded only for experimental units exposed to treatment 1, or assuming the average causal effect in P is of interest, cases with Y recorded only for experimental units with $X_1 = 1$.

4.4 *Ignorable mechanisms relatively well suited for causal inference.* The implications of Sections 4.2 and 4.3 are rather simple. Inferences for causal effects will be relatively insensitive to prior specifications when both the assignment and recording mechanisms are ignorable, and, for each distinct value of recorded covariates, there are experimental units with recorded Y values within each treatment condition.

A simple modification of the previous examples further illustrates this rather obvious point. Suppose (a) n sampled units are exposed to one of two treatments, (b) Y_1 is dichotomous (0, 1) and either Y_1^1 or Y_1^2 is recorded for each sampled unit, (c) X_1 is dichotomous (1, 2) and recorded for each sampled unit, (d) the

assignment and recording mechanisms are ignorable, and (e) the specification for (Y_1^1, Y_1^2, X_1) is the $2 \times 2 \times 2$ contingency table given in Figure 2.

Let $n_k^t = \sum_1^N \tilde{M}_{1i}^t \delta(\tilde{X}_{1i} - k)$ (the number of units exposed to treatment t with $\tilde{X}_{1i} = k$) and $\bar{y}_k^t = \sum_1^N \tilde{Y}_{1i}^t \tilde{M}_{1i}^t \delta(\tilde{X}_{1i} - k)/n_k^t$ (the average observed value of $\tilde{Y}_{1i}^t$ when $\tilde{X}_{1i} = k$). The posterior distribution of π is proportional to

$$(4.11) \qquad p(\pi) \prod_{t=1}^2 [\alpha^t \pi^t]^{n_1^t \bar{y}_1^t} [1 - \alpha^t \pi^t]^{n_1^t (1 - \bar{y}_1^t)} \times \prod_{t=1}^2 [\pi^t (2 - \alpha^t)]^{n_2^t \bar{y}_2^t} [1 - \pi^t (2 - \alpha^t)]^{n_2^t (1 - \bar{y}_2^t)} .$$

Suppose for $\tilde{X}_{1i} = 1$ and $\tilde{X}_{1i} = 2$ there are experimental units in both treatment conditions (i.e., $n_k^t > 0$, $t = 1, 2$, $k = 1, 2$). Then the likelihood in (4.11) has a unique maximizing value of (π^1, π^2) given by $((\bar{y}_1^1 + \bar{y}_2^1)/2, (\bar{y}_1^2 + \bar{y}_2^2)/2)$, and a unique maximizing value of (α^1, α^2) given by

$$\left(\frac{2\bar{y}_1^1}{\bar{y}_1^1 + \bar{y}_2^1}, \frac{2\bar{y}_1^2}{\bar{y}_1^2 + \bar{y}_2^2} \right).$$

Hence the data are informative about the causal effects of the treatments in P and the two subpopulations of P defined by $X_{1i} = 1$ and $X_{1i} = 2$. In this sense, inferences for causal effects are relatively insensitive to prior specifications. If some $n_k^t = 0$, then the inferences for causal effects will be quite sensitive to prior specifications because there will exist a ridge in the likelihood in (4.11); for example, if $n_2^2 = 0$, then π^1 and α^1 have unique maximizing values as given above, but there exists a ridge of maximizing (α^2, π^2) defined by the equation $\alpha^2 \pi^2 = \bar{y}_1^2$.

We consider these examples with a dichotomous Y and a dichotomous X representative of the most general practical case. This is because in any practical problem any Y can take only a finite number of possible values and any X can take only a finite number of possible values. The modelling of the rows of (Y, X) as i.i.d. multinomial is thus completely general, models such as a normal linear regression of Y on X being viewed as restrictions in $p(\pi)$ that smooth conditional multinomial probabilities in special ways.

Thus, letting $f(X, Y | \pi)$ be the unrestricted multinomial, we see that given ignorable assignment and recording mechanisms, inferences for causal effects in P and in subpopulations of P defined by each observed value of X are relatively insensitive to prior specifications only if for each observed value of X there exist experimental units with Y recorded in each treatment condition.

5. The role of randomization. There were four main messages in Section 4. First, when the assignment and/or recording mechanism is nonignorable, a valid Bayesian analysis requires the incorporation of models for the nonignorable mechanisms, as well as a model relating observed variables to unobserved variables used in the nonignorable mechanisms. Second, inference for causal effects is generally very sensitive to prior specifications when the assignment and/or recording mechanisms are nonignorable even with no imbalance in the distribution

of recorded covariates across treatment groups. Third, inference for causal effects can also be very sensitive to prior specifications with combinations of ignorable assignment and recording mechanisms yielding unbalanced distributions of recorded covariates across treatment groups. Fourth, inferences for causal effects in P and in subpopulations of P defined by observed values of covariates can be insensitive to prior specifications only when for each distinct value of the recorded covariates there are experimental units in each treatment condition.

A standard justification for randomization is that it has prophylactic effect, guarding against data unbalanced with respect to recorded covariates. A Bayesian interpretation of this statement is given by the result that, although randomized designs are generally not optimal, they satisfy certain minimax properties over choices of $p(\pi)\, f(X, Y \mid \pi)$ (see Savage, 1972; Stone, 1973). These results, however, do not imply that classical randomized designs should be used. Of critical importance, they do not address the following question: given a particular treatment assignment $\tilde{W}$, is there any advantage in knowing that the assignment was obtained by a randomized rule rather than a deterministic rule?

The framework we have developed shows, however, that classical randomized designs can markedly reduce the sensitivity of a valid Bayesian analysis, because only a randomized assignment mechanism can be ignorable and yield data having more than one treatment condition represented for a distinct value of recorded covariates.[4] Furthermore, we argue that in many practical problems, classical randomized designs can achieve this reduced sensitivity to prior specifications and still balance covariates used to form blocks almost as well as, and perhaps better than "optimal" designs. We will demonstrate the advantages of classical randomized designs by first considering the problems of execution and analysis that face a study using a nonrandomized (deterministic) assignment mechanism to balance the distribution of many covariates across treatment groups and then showing how these problems of execution and analysis can be obviated by a comparable randomized design.

5.1 *Deterministically balancing many covariates.* In any study, the assignment mechanism should be described explicitly in order to constrain the class of models $k(W \mid X, Y, \pi)$ that needs to be considered in the data analysis. Rules such as "a large number of allocations were tried and one chosen that seemed to exhibit the most balanced distribution of relevant characteristics across treatment groups" or "the study was continued until conclusive evidence was collected" correspond to models for $k(W \mid X, Y, \pi)$ that depend on π and possibly unobserved (X, Y) values. Deterministic rules that try to balance many covariates may be

[4] Two experimental units with identical values of $\tilde{X}_{(1)}$ appear identical (except for their randomly assigned indices) to an ignorable assignment mechanism because an experimental unit's Y-values cannot be recorded until after treatment assignment. Hence, if two such units receive different treatments under an ignorable assignment mechanism, some randomization must have been employed.

hard to formalize especially when the covariates being balanced are based on personal assessments about the experimental units (e.g., medical judgments about the health of the patients). The effort required to formalize covariates and/or deterministic rules balancing many covariates may be great and delay the assignment of treatments.

Even if deterministic rules and the covariates which they attempt to balance are successfully formalized, in practice the rules may be difficult to follow because they are intricate, or easy to avoid following because once the covariates' values are known, the treatment assignment is known (e.g., by changing the value of a covariate, perhaps based on a personal assessment, treatment assignment may be changed—this means that variables reflecting judgments about preferred treatments are also being used to assign treatments).[5] Thus, it may be difficult (and, if some values used to make assignment decisions are not recorded, virtually impossible) to verify that the proposed assignment rules were actually used. Bailar (1976) makes similar points when discussing the infrequent use of intricate patient assignment algorithms in medical research. The practical implication is that when the experimental design proposes a deterministic assignment mechanism, a model for $k(W \mid X, Y, \pi)$ that reflects the assignment mechanism actually used may have to be more complicated than the proposed mechanism. Hence, even if an "optimal" deterministic design is proposed, the actual assignment mechanism used may be neither optimal nor ignorable.

Furthermore, using a deterministic assignment mechanism to balance many covariates may lead to a nonnegligible possibility of unintended missing values because of the bookkeeping required to record the covariates' values. Consequently, in order to reflect the recording mechanism actually used in this case, a complicated nonignorable model for $g(M \mid X, Y, W, \pi)$ may have to be considered in the data analysis (remember that even if there exist no unintended missing values, the possibility that there might have been missing values makes the recording mechanism nonignorable).

In spite of the real possibility of nonignorable assignment and/or recording mechanisms when attempting to use a deterministic assignment mechanism to balance many covariates, suppose that the experimenter has successfully used ignorable assignment and recording mechanisms. It now may be difficult to perform a valid Bayesian analysis because the observed data set is so highly multivariate. Even if the parameter $\pi_{Y|X}$ is a priori independent of the parameter π_X, the data analysis must still specify a model $h(Y \mid X_{(1)}, \pi_{Y|X})$ and a prior distribution for $\pi_{Y|X}$. With many recorded X variables, the sensitivity of causal inferences to a broad class of models for $h(Y \mid X_{(1)}, \pi_{Y|X})$ is so great that a Bayesian

[5] Results in Blackwell and Hodges (1957) and Stigler (1969) address this issue in a simple sequential experiment: if the experimental unit is to be assigned a treatment (i.e., if $W_i > 0$), the proposed assignment rule must be followed, but the experimenter can try to bias results by choosing which experimental units to study (i.e., he is allowed to choose $W_i = 0$ or $W_i > 0$ by some unknown rule).

analysis capable of yielding sharp causal inferences requires strong prior restrictions on these models; with a continuous Y, simply consider polynomial regression models of order N with many covariates (independent variables). Hence, if a valid Bayesian analysis is to yield worthwhile causal inferences, the statistician must expend the effort to formalize "reasonable" prior restrictions on the specification for the distribution of the data. As more and more covariates are recorded, even an approximately valid Bayesian analysis becomes practically impossible since an enormous commitment of time and resources is needed in order to perform the mental contemplation, mathematical and/or Monte Carlo analyses, and numerical computations necessary to understand a highly multivariate data set. The results of such efforts may be of interest for covariates that are recorded because their relationships with Y are of central scientific or practical importance but of dubious interest for covariates that are recorded principally because they were used to balance treatment groups.

Of course, one could decide to avoid a highly multivariate data set in this case by not recording some covariates' values that were used to assign treatments.[6] However, then the assignment mechanism is not ignorable and the data analysis must explicitly incorporate the model $k(W \mid X, Y, \pi)$ as well as still model the joint distribution of Y and the covariates used to assign treatments. A Bayesian cannot simply assert that incorporating these models will not affect causal inferences and have a valid Bayesian analysis. Section 4.2 showed that a valid Bayesian analysis with an assignment mechanism that is not ignorable can be very sensitive to the specification $f(X, Y \mid \pi)p(\pi)$ even with no imbalance in recorded covariates. The inferences for causal effects may be especially sensitive if the models for $k(W \mid X, Y, \pi)$ and $g(M \mid X, Y, W, \pi)$ reflect mechanisms that might actually be operating when a complicated deterministic assignment mechanism is proposed.

In sum, we argue that deterministically balancing many covariates in practice generally leads to a study that is difficult to execute properly and hard to analyze.

5.2 *Using blocking and randomization to balance many covariates.* Classical randomized designs utilize blocking and randomization to balance the distribution of covariates that were used for blocking but not necessarily recorded for data analysis. These designs thus offer alternatives to nonrandomized, deterministic balancing of these covariates' distributions and the subsequent need to model their joint distribution with Y variables and either (a) record their values for data analysis, or (b) explicitly incorporate models for assignment and recording mechanisms.

The advantages of randomized designs are most dramatic when there are many covariates to be balanced, including many based on personal assessments.

---

[6] This appears to be the recommendation in "biased coin designs" (Efron, 1971) with "time of entry into experiment" and "block" used to assign treatments but the former not recorded for data analysis.

Specifically, suppose early in the study indices were assigned randomly to the experimental units; then the experimenter blocked the experimental units in such a way that on the basis of all covariates he wanted to control, the experimental units within each block appeared similar to him; a final randomization assigned treatments to experimental units within blocks (e.g., for $T = 2$ and 12 experimental units in a block, the 6 experimental units with lowest indices received treatment t, $t = 1$ or 2 being decided by a coin toss). Since the assignment rule is simple, it should be easy to follow. Also, it is difficult to avoid following because within each block, the decision as to the experimental units to receive each treatment is not made until the final coin toss, so that neither blocking nor the initial assignment of indices (nor both) can be used to determine treatment assignment. The initial random assignment of indices guarded against the experimenter grouping experimental units within a block and assigning "preferred" treatments to the groups on the basis of some unrecorded variables; since the experimental units with high and low indices are randomly grouped, there is little to be gained by avoiding the coin toss. In order to verify that the assignment rule is being followed, one need only verify that indices were in fact randomly assigned and that the final assignment of treatments was also random. Although the experimenter can "cheat," the possibilities are far less than when using a deterministic rule where a specific change in a covariate value is known to correspond to a specific change in treatment asssignment. Thus when using the randomized design, it is more likely that the proposed assignment rule is being followed and consequently more likely that complicated models for $k(W|X, Y, \pi)$ need not be considered.

Since assignment is on the basis of the covariate X_1 = block number, the recording mechanism need not record any X values other than X_1 in order for the assignment mechanism to be ignorable, and thus covariates used for blocking do not have to be formalized. Since the recording mechanism need record only Y and X_1, it too can be ignorable more easily than if a complicated deterministic assignment rule were used. Also, in the data analysis, $f(X, Y|\pi)p(\pi)$ need only specify an acceptable model for the randomized block experiment with no covariates recorded (e.g., an analysis of variance model). Since within each block we have a completely randomized experiment, the rows of Y in a block are exchangeable and thus may be modelled as i.i.d. given π. Although inferences for causal effects will vary somewhat with prior specifications about the distributions of Y given each block, a valid Bayesian analysis is far more straightforward and insensitive to prior specifications than if many covariates were recorded and this extensive exchangeability given recorded covariates could not be invoked.

If some X variables are especially important a priori (in the sense that that a clear relationship to Y is likely), they should be recorded and modelled (e.g., an analysis of covariance model), or many blocks should be used to represent their values. If some important X variables can be used to make future assignment decisions they should be recorded so that different treatment effects can be

estimated for subpopulations defined by their values. For example, the relative effectiveness of medical treatments may be different for different age groups, and age can usually be measured before a treatment decision is made. Ideally, in each subpopulation defined by the values of these X variables, there should be a classical randomized experiment. Of course for a fixed total sample size, increasing the number of subpopulations increases the sensitivity of the Bayesian analysis for causal effects to the prior specifications since the exchangeability is reduced. Consequently, if many subpopulations are of interest we may be faced with inferences for the causal effects in each subpopulation that are quite sensitive to prior specifications.

We are not claiming that classical randomized designs make Bayesian inference for causal effects trivial, but rather that they make it simple relative to Bayesian inference for causal effects using data obtained by a comparable deterministic rule balancing many covariates. The objection that the randomized design, although yielding treatment groups balanced with respect to the covariates used to form blocks, has not "optimally" balanced these covariates under any specific model, seems irrelevant in real world problems having no accepted specific model relating Y to X. This view agrees with some experimenters' practical experience suggesting that the gain from using an "optimal" design rather than a good classical randomized design is usually trivial.[7] Furthermore we have argued that even if an optimal deterministic design is proposed, in many practical problems it is unlikely that it will be followed.

In sum, when using the randomized design, the assignment rule is easy to follow and difficult to avoid following, the recording mechanism can be quite simple, a valid Bayesian analysis for causal effects is relatively straightforward, and treatment groups are balanced with respect to covariates used to form blocks almost as well as if the experimenter found the one allocation most satisfying to him, and yet there was no need to formalize and record these covariates' values. A comparable nonrandomized design would generally be substantially more difficult to execute and analyze because of the need to deal explicitly with all covariates being balanced.

5.3 *Deterministic sampling of experimental units in randomized experiments.* In practice, most randomized studies use deterministic sampling of experimental units since the actual population of interest, P, usually consists of experimental units from different geographical areas and from the future (the objective of most studies being to determine the potential efficacy of widely applicable treatments such as medical operations). Suppose that by proper definition of the

[7] W. G. Cochran, in personal communications, has suggested that in his experience, experimenters often feel many covariates are important and should be explicitly controlled, but the data analysis surprises them by showing that the relationship of many of these covariates with Y is actually quite weak. D. R. Cox, in a personal communication relating unpublished work comparing optimal designs and randomized designs in some agricultural studies, concludes that in practice any advantage to the optimal designs appears to be quite small.

covariates used to select experimental units for study, all selected experimental units have the same values of those covariates. If we focus attention on the experimental units sampled for study, considering them to constitute the population of interest, a valid Bayesian analysis for the causal effects of the treatments is straightforward since the covariates used to determine participation in the study are constant in this subpopulation and thus need not be recorded for data analysis.

When interest focuses on P, the actual population of interest, covariates used to determine participation in the study must be recorded for data analysis if the assignment mechanism is to be ignorable. In some cases such as when sampling from the present but generalizing to the near future, it may be reasonable to believe that the covariate used for selection (time) is unrelated to recorded variables and so model the sampled units as a simple random sample from P. In other cases such as when using prisoners for medical experiments, it may not be reasonable to believe that the covariates used to select experimental units are unrelated to recorded variables. In any case, the classical randomized experiment on the sampled experimental units has led to a straightforward, valid Bayesian analysis of causal effects for these experimental units. In many observational studies, the sampled experimental units may be more nearly a random sample from P than in many controlled experiments; however, the rule for assigning treatments is unknown. Therefore, in these observational studies it is generally difficult to estimate causal effects for any collection of experimental units, so that the ability to generalize results to the intended population may be of limited practical use.

5.4 *Conclusions and extensions.* In some cases with strong prior knowledge, randomization may not be important. For example, in an industrial experiment comparing manufacturing procedures, the relevant covariates may be easily recorded and their relationships to dependent variables well-understood; hence, a design that is optimal under restrictive prior specifications for the data may be appropriate. In other cases, such as when operating a familiar piece of equipment (e.g., driving a car), causal effects of treatments (turning the steering wheel left vs. right) may be so dominant that any formal design may be superfluous.

However, when causal effects of treatments are subtle enough to warrant the attention of a statistician for the design and/or analysis of the study, some relevant covariates will be difficult to record and their relationships with dependent variables poorly understood. In such cases, which abound in health and social research, classical randomized designs must be considered vital tools for the Bayesian statistician since they can dramatically reduce the sensitivity of valid Bayesian analyses to prior specifications and greatly simplify the computation of such analyses.

Nevertheless, it is possible in the age of the computer that other designs may also be of interest. For example, consider "biased coin designs" (Efron, 1971)

and "finite selection models" (Morris, 1975). These are randomized (and "unbiased" in some sense) but use covariates to assign treatments without requiring these covariates to be recorded for data analysis. Thus they correspond to randomized nonignorable assignment mechanisms. Suppose the recording mechanism is ignorable; then equations (4.1)—(4.4) imply that the predictive distribution of $Y_{(0)}$ is proportional to

$$[\int h(\tilde{Y}|\tilde{X}_{(1)}, \pi)q(\pi|\tilde{X}_{(1)})\,d\pi]S(Y_{(0)}) \tag{5.1}$$

where

$$S(Y_{(0)}) = \frac{\int\int k(\tilde{W}|\tilde{X}, \tilde{Y}, \pi)f(\tilde{X}, \tilde{Y}|\pi)p(\pi)\,d\pi\,dX_{(0)}}{\int\int f(\tilde{X}, \tilde{Y}|\pi)p(\pi)\,d\pi\,dX_{(0)}}\,.$$

Note that $S(Y_{(0)})$ is the expectation of $k(\tilde{W}|\tilde{X}, \tilde{Y}, \pi)$ over the conditional distribution of $(\pi, X_{(0)})$ given $(\tilde{X}_{(1)}, \tilde{Y}_{(1)}, Y_{(0)})$ determined by the prior specification $f(X, Y|\pi)p(\pi)$. For a particular $f(X, Y|\pi)p(\pi)$, a necessary and sufficient condition such that inferences for causal effects will be the same as if the assignment mechanism were ignored is that $S(Y_{(0)})$ takes the same value for all $Y_{(0)}$. Thus if $S(Y_{(0)})$ is functionally independent of $Y_{(0)}$ we may say that the assignment mechanism is ignorable given that particular prior specification of the data.

In the biased coin design and the finite selection model

$$k(\tilde{W}|\tilde{X}, \tilde{Y}, \pi) = k(\tilde{W}|\tilde{X}_{(1)}, X_{(0)})\,. \tag{5.2}$$

Hence, if Y and the unrecorded X's used to assign treatments are assumed independent given $\tilde{X}_{(1)}$, then $S(Y_{(0)})$ takes the same value for all $Y_{(0)}$ and the assignment mechanism (5.2) is ignorable given that prior specification. Furthermore, for these assignment mechanisms, $S(Y_{(0)})$ can be nearly functionally independent of $Y_{(0)}$ for a broad range of prior specifications. As an example, consider the biased coin design where $X_{(1)}$ represents blocks and $X_{(0)}$ represents "time of treatment assignment." For fixed $(\tilde{X}_{(1)}, \tilde{Y}_{(1)}, \tilde{W})$ and any fixed distribution of $X_{(0)}$, as the bias of the coin becomes smaller, the distribution of the values of $k(\tilde{W}|\tilde{X}_{(1)}, X_{(0)})$ generated by the distribution of $X_{(0)}$ becomes more concentrated about some fixed value (e.g., $(1/2)^n$ for the biased coin design without blocks and n experimental units having received treatments). Thus, $S(Y_{(0)})$ becomes "nearly" independent of $Y_{(0)}$ as the coin becomes fair.

Consequently, there are nonclassical randomized designs that can be "nearly ignorable" in the sense that in practice we may be able to approximate them as being ignorable. Of course, before making such approximations, we should be convinced that $S(Y_{(0)})$ really is nearly independent of $Y_{(0)}$ for the full range of reasonable data specifications. This checking may be possible in many cases using a computer or may be possible to do analytically.[8]

[8] Sections 5 and 6 of Efron (1971) provide justifications for the recommendation to ignore the covariate "time of treatment assignment" in some biased coin designs; the blanket recommendation was implicitly criticized here in footnote 6. These types of arguments suggest that

(continue to next page)

If the assignment mechanism is not ignorable or nearly so, then models incorporating the assignment mechanism should be used. Observational studies are especially difficult to analyze properly because the form of the assignment mechanism is itself unknown. Models appropriate for prospective and retrospective observational studies need to be developed. In many observational studies, no sharp inferences for causal effects will be possible, while in others, reasonable models for nonignorable assignment mechanisms may lead to consistent conclusions. Clearly, analyses like these require more computation and demand more attention than analyses of comparable data obtained by ignorable mechanisms; an example in the simpler but related context of nonresponse in sample surveys is given in Rubin (1977b).

Further extensions of this work include application to data where nonstandard definitions of treatment effects may be useful, e.g., data typically analyzed using competing risks models. Since, within our framework, causal effects are defined without reference to the parametric structure of particular models, it is conceptually straightforward to evaluate the sensitivity of inferences for causal effects to prior specifications (e.g., nonindependence of competing risks), even if the parametric structure of the specifications change.

In conclusion, we feel that our framework not only provides theoretical justification for classical methods of estimating causal effects, but also suggests new approaches to drawing inferences for causal effects in nonstandard problems.

Acknowledgments. I wish to thank P. W. Holland, C. Morris, and M. R. Novick for many helpful comments on earlier drafts of this work. I also want to thank A. P. Dempster for numerous stimulating and influential conversations which have greatly improved the presentation of ideas in this paper.

REFERENCES

BAILAR, J. C. (1976). Patient assignment algorithms—an overview. *Proc. 9th International Biometric Conf.* **1** 189-206. The Biometric Society, Raleigh.

BLACKWELL, D. and HODGES, J. L. (1957). Design for the control of selection bias. *Ann. Math. Statist.* **28** 449-460.

CAMPBELL, D. T. and ERLEBACHER, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In *The Disadvataged Child, Compensatory Education: A National Debate*, **3** (J. Hellmuth, ed.). Brunner/Mazel, New York.

COCHRAN, W. G. and RUBIN, D. B. (1974). Controlling bias in observational studies: a review. *Sankhyā A* **35** 417-446.

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York.

---

weaker definitions of nearly ignorable may be of interest. For example, in some cases we may be able to show that under mild prior restrictions the predictive distribution of some particular function of $Y_{(0)}$ (e.g., $\bar{Y}^1 - \bar{Y}^2$) is equal or nearly equal to the predictive distribution of that function ignoring the assignment mechanism (i.e., ignoring the factor $S(Y_{(0)})$ in equation (5.1)); or perhaps we can show this equality holds in large samples; or weaker still, we may be able to show that the means of the predictive distributions are equal or nearly equal.

DE FINETTI, B. (1963). Foresight: its logical laws, its subjective sources. *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds.). Wiley, New York.

DIACONIS, P. (1976). Finite forms of de Finetti's theorem on exchangeability. Technical Report No. 84, Stanford Univ.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 3, 403–417.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications, Volume II.* Wiley, New York.

GILBERT, J. P. (1974). Randomization of human subjects. *New England Journal of Medicine*, **291** 24, 1305–1306.

GILBERT, J. P., LIGHT, R. J. and MOSTELLER, F. (1975). Assessing social innovations: an empirical base for policy. *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs* (A. R. Lumsdaine and C. A. Bennett, eds.). Academic Press, New York.

HEWITT, E. and SAVAGE, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80** 470–501.

MORRIS, C. (1975). A finite selection model for experimental design of the health insurance study. *Proc. Social Statist. Sect., Amer. Statist. Assoc.* 78–85.

ROSENTHAL, R. L. (1976). *Experimenter Effects in Behavioral Research.* Irvington, New York.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology* **66** 5, 688–701.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 3, 581–592.

RUBIN, D. B. (1977a). Assignment to treatment group on the basis of a covariate. *J. Educational Statist.* **2** 1, 1–26.

RUBIN, D. B. (1977b). Formalizing subjective notions about the effect of non-response in sample surveys. *J. Amer. Statist. Assoc.* **72** 359, 538–543.

SAVAGE, L. J. (1962). *The Foundations of Statistical Inference* (M. S. Bartlett, ed.). 33–34. Wiley, New York.

SAVAGE, L. J. (1972) *The Foundations of Statistics.* Dover, New York.

STIGLER, S. (1969). The use of random allocation for the control of selection bias. *Biometrika* **56** 3, 553–560.

STONE, M. (1973). Role of experimental randomization in Bayesian statistics: An asymptotic theory for a single Bayesian. *Metrika* **20** 170–176.

WEINSTEIN, M. C. (1974). Allocation of subjects in medical experiments. *New England Journal of Medicine* **291** 1278–1285.

OFFICE OF DATA ANALYSIS RESEARCH
EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08540