BAYESIAN INFERENCE FOR CAUSALITY: THE IMPORTANCE OF RANDOMIZATION

Donald B. Rubin

Abstract

It is often held that experimental randomization is irrelevant to a Bayesian since he conditions on all the observed values of the data in any case. By considering randomization as a special process for creating missing data that allow inferences for causality, it can be shown that randomization is as important to a Bayesian as to a frequentist. A subjective Bayesian cannot ignore the method used to assign treatments unless it is a known (possibly probabilistic) function of recorded values of covariates. A Bayesian who considers objective priors can generally ignore the method used to assign treatments only if it is complete randomization.

1. Introduction

Discussion of the role of randomization in the search for effective treatments is becoming common in the social and medical sciences. See for example, Campbell and Erlebacher (1970), Gilbert (1975) and Gilbert, Light, Mosteller (1974). The basic problem concerns assigning human subjects to treatments that some suspect are less efficacious than other treatments under study. The rules of randomization imply that this assignment be made by a random mechanism and not by the subjects themselves or with their consent. Perhaps we can do away with randomization and still reach valid inferences about the causal effects of treatments?

Bayesian statisticians often claim that randomization is irrelevant or at least of secondary importance for inference of any kind. Thus, Savage (1954, p.66) suggests that randomization may mean little more to a Bayesian than some sort of haphazard assignment, and neither de Finneti (1975) nor Lindley (1970) even list "randomization" or "randomized" in their indexes.

The implication is that a statistician faced with the results of a study and looking for causal effects of treatments should not care if the treatments were assigned randomly or by some other mechanism, since his analysis will simply condition on the observed values of the dependent variables and covariates in either case. This position is incorrect: the resultant analyses are, in general, not correct Bayesian analyses of the data. The reason is that they do not condition on the observed value of the random variable that indicates how treatments were assigned.

Understanding the process that assigned treatments is as important to a Bayesian as to a frequentist. Randomization stands out as the only commonly used method for assigning treatments that allows the Bayesian statistician to ignore the assignment mechanism when making causal inferences.

This argument will be developed here. The theoretical foundations lie in the conceptualization of inference for causality as a special case of inference when faced with missing data as developed in Rubin (1975). Derived issues of Bayesian experimental design are mentioned but not explored, as are the sampling distribution analogues for the conclusions presented. These results for causal inference are more complicated versions of similar results for descriptive inference where the utility of random sampling is the central issue.

2. Defining Causal Effect

Often, discussion of how to estimate causal effects seems to be rather confused, ambiguous, and commonly consists of semantic arguments rather than statistical comparisons. By carefully defining causal effect to be the difference between an observed value and an unobserved value, we hope to avoid this.

Intuitively, the causal effect of one treatment over another for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to the first treatment initiated at t_1 and what would have happened at t_2 if instead the unit had been exposed to the second treatment initiated at t_1: "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." Our definition of the causal effect of one treatment versus another treatment will reflect this intuitive meaning (Rubin, 1974).

We begin with some simple definitions. These lead to the definition of causal effect.

(.) A unit is an object of study (e.g., a person, a rat, a block of copper).

(.) Y is some aspect of units as recorded by some particular measuring instrument (e.g., the score on an achievement test, blood pressure in inches of mercury as recorded by a specific instrument).

(.) A treatment is a series of well-defined actions performed on a unit (e.g., the injection of one ounce of a drug, exposure to a special compensatory reading program as taught by a specially trained teacher).

(.) A trial is a triple: a unit, a time of initiation of treatment, and a time of recording of aspect.

(.) P is a population of trials (generally a hypothetical collection of those units and times in the future to which the treatments under study may be applied).

(.) y_{ij} is the random variable representing the value of the aspect Y for the ith trial given exposure to the jth treatment.

(.) The causal effect on Y of treatment 1 vs. treatment 2 for the ith trial is the expectation of $y_{i1} - y_{i2}$, say $\mu_{i1} - \mu_{i2}$; similarly for all other pairs of treatments.

(.) The typical causal effect on Y of treatment 1 vs. treatment 2 for the population P is the average value of $\mu_{i1} - \mu_{i2}$ over all trials in P, say $\mu_1 - \mu_2$; similarly for all other pairs of treatments.

Any question that cannot be formulated in the above framework has no "causal" answer. For example, questions of the causal effect of sex or race as sometimes discussed are totally ambiguous with respect to what the treatment and its time of initiation mean. Do we mean at birth we "dye" the child's skin a different color, or at conception we "change" all Y chromosomes to be X chromosomes? The point is that we need well-defined treatments and times at which we are to initiate them before we can discuss their causal effects.

Sometimes, we can get away with rather sloppy definitions of treatments if the implied causal effects are gross enough. For example, in the statement "the sun causes the planets to travel in orbits around it" the implication is that if one were to destroy or remove the sun in any way at any time, the planets would no longer travel in their orbits. However, in many practical cases, we have to be far more precise in order to be clear. In some cases, we have to be terribly precise in the definitions of treatments in order to avoid deceiving ourselves, especially when dealing with humans who may know they are part of a study (e.g., treatments given under double-blind conditions are different treatments than those given under simple conditions because the actions performed on the units are different).

3. A Study of T Treatments

Consider a study of T treatments. We will be very explicit in listing all the random variables one might record because it is essential that the Bayesian conditions on the observed values of all random variables. Figure 1 presents a matrix of random variables. This matrix is not the usual "units by variables" matrix of observed values: in any study only a few values of the random variables in this matrix are actually recorded. Each row of the matrix refers to one trial in the population P.

The T columns under Y labelled 1,..., T refer to the values of the aspect Y given exposure to the various treatments, i.e., the random variables y_{ij}, as described in Section 2. Similarly, the T columns under Z labelled 1,...,T refer to the values of another aspect Z (vector-valued) given exposure to the various treatments. Every characteristic of the trials that is recorded after the initiation of treatments (except Y) is included in Z. Because both aspects Y and Z are recorded after the initiation of treatments, each generates T columns in the matrix of random variables. The column labelled X refers to an aspect (vector-valued) that is recorded before the initiation of treatments and thus generates only one column of random variables in the matrix. Every characteristic of the trials that is recorded before assignment of treatments is included in X.

The columns labelled Y are commonly referred to as the dependent variable, the column labelled X as the covariate or concomitant variable. The columns labelled Z are, for example, dependent variables of secondary importance or

measures of how the treatment was actually carried out. The important point is that every aspect of the trials that is to be recorded is included in this matrix of all the random variables which we call U; U is the potentially observable data.

Within the ith row of the matrix U (i.e., the ith trial), at most one of the y_{ij} is actually observed, and at most one of the z_{ij} is actually observed. That is, if the ith trial in P was not included in the study, values for $y_{i1},\ldots,y_{iT}$, $z_{i1},\ldots,z_{iT}$, are not observed. If the ith trial was in the study and "complete" data were obtained, the value of exactly one of $y_{i1},\ldots,y_{iT}$ is recorded and the value of exactly one of $z_{i1},\ldots,z_{iT}$ is recorded; which values are recorded depends on which treatment the ith trial received.

The columns labelled M in Figure 1 are the same in number as the columns labelled U; M is simply an indicator random variable: if $m_{ij}=1$, u_{ij} is observed while if $m_{ij}=0$, u_{ij} is not observed. The potentially observable random variable in a study of T treatments is thus (U,M), not U alone.

Notice that within the structure we have developed, the problem of inference for causality is equivalent to the problem of inference given missing data. That is, most of U is missing, as indicated by M. If all of U were observed (which is impossible) standard methods of inference could be applied; given the presence of missing data, we have to be more careful. We have developed this structure because it enables us to use existing results on inference with missing data presented in Rubin (1975). Before using these results, we must specify the distribution of the random variable (U,M).

4. The Distribution of the Random Variables

Let $f_\theta(U)\, g_\phi(M|U)$ be the joint probability density function for the random variables, where θ and ϕ are vector parameters; $f_\theta(U)$ is the marginal density of the potentially observable data U, and $g_\phi(M|U)$ is the conditional density of the indicator M given U. Because the causal effects are differences of the expectations of the Y columns of U, the causal effects are functions of θ and have nothing to do with the assignment process g_ϕ.

Under $f_\theta\, g_\phi$, the probability that the ith trial was exposed to the jth treatment is positive for all i,j. This is true because if for some trial the probability of some treatment is zero, then that trial really is not a member of the population of trials to which we want to generalize the results of the T treatment study; that is, if it is impossible that some trial will ever get some treatment, then that trial does not belong to the population of trials that might be exposed to that treatment. Hence, M is a non-degenerate random variable taking values other than the observed value.

The following paragraph is an aside on model building for f_θ. It is not central to our argument but is relevant in Section 7. In many cases, the interval of time from application of treatment to measurement of dependent variable is roughly the same length for all trials in P; that is, it makes sense to consider the effects of treatments at a fixed length of time after application. Then, each trial is distinguished by the time of application and the unit. Often the times of application are considered nearly constant or a priori irrelevant to Y for all trials in the population P; then the population of trials may be considered to be identical to the population of units to which we want to generalize the results -- each trial is a distinct unit. Since it is often reasonable to assume that given the aspects X,Y,Z, the labelling of the units is irrelevant, in many cases one is naturally led to assuming that under f_θ the rows of U are exchangeably distributed. Often in fact, the rows of U are in addition assumed independent under f_θ (and thus are i.i.d). The justification for this last assumption is that if the number of trials in P is extremely large, given knowledge of θ, knowledge of the value of U for one trial does not really restrict the values of U for any other trial.

Example: Suppose Y is a measure of health and each trial represents a different person. Also suppose T=2 and that a priori all other aspects that are recorded are irrelevant to Y. Then an obvious choice for f_θ would be

$$(y_{i1},y_{i2}) \sim \text{i.i.d.}\ N\left((\mu_1,\mu_2),\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right) .$$

Notice that y_{i1} and y_{i2} can never both be observed. Thus ρ cannot be estimated, e.g., given a prior distribution on ρ that is independent of the prior distribution for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, the posterior distribution of ρ equals the prior distribution of ρ. In general, only the marginal distributions of Y are estimable.

5. Naive Bayesian Inference for Causality

Section 2 defined causal effect; the formulation and the explicit inclusion of the population P of trials is not standard. Section 3 described in detail the random variables in a study of T treatments; the explicit inclusion of M is not standard. Similarly in Section 4 when defining the distributions of the random variables, the explicit inclusion of g_ϕ is not standard. The standard Bayesian analysis of the data establishes a prior distribution for θ, say $p(\theta)$, and proceeds to find the posterior distribution of θ (and thus of the treatment effects which are functions of θ) from f_θ and the observed values in U. Variously expressed, this naive Bayesian analysis for causal effects:

(.) Ignores the assignment process

(.) Ignores the process that causes missing data

(.) Fixes M at its observed value without conditioning on this value

(.) Bases inferences for θ solely on f_θ

(.) Bases inferences for θ solely on the marginal likelihood of the observed data.

More specifically, let $\tilde{M}$ be the observed value of M. The definition of $\tilde{U}$ is more subtle. Let $U = \{u_{ij}\}$ be defined by: $\tilde{u}_{ij}$ = the observed value of u_{ij} (i.e., a real number) if $\tilde{m}_{ij} = 1$, and $\tilde{u}_{ij} = u_{ij}$ (i.e., an argument representing a random variable that will be integrated out) if $\tilde{m}_{ij} = 0$. Then naive Bayesian analysis lets the posterior distribution of θ be proportional to

$$(5.1) \qquad p(\theta) \int f_\theta(\tilde{U}) \; dU_0^{\tilde{M}}$$

where the integral is over those random variables in U that are not observed (i.e., those u_{ij} for which $\tilde{m}_{ij} = 0$). For instance, in the example of Section 4, if both $N_1 = \Sigma \tilde{m}_{i1}$ and $N_2 = \Sigma \tilde{m}_{i2}$ are large, the posterior distribution of the treatment effect, $(\mu_1 - \mu_2)$, converges to $(\bar{y}_1 - \bar{y}_2)$ where $\bar{y}_i = \Sigma_j \tilde{m}_{ij} \tilde{y}_{ij}/N_i$, the average observed value of y_{ij}.

6. Proper Bayesian Analysis for Causality

The problem with the naive Bayesian approach to causal inference as outlined in Section 5 is that the indicator M is a random variable whose value is always observed, and thus a proper Bayesian analysis must condition on it as well as the observed values in U. Conditioning on both the observed value of M and the observed values in U leads to the joint posterior distribution of θ and ϕ, which is proportional to:

$$(6.1) \qquad p(\theta)\, p(\phi|\theta) \int f_\theta(\tilde{U}) g_\phi(\tilde{M}|\tilde{U}) \; dU_0^{\tilde{M}}$$

where $p(\phi|\theta)$ is the prior of ϕ given θ.

The question of primary interest is: when does the naive Bayesian approach (i.e., based on 5.1) produce the correct Bayesian answer (i.e., based on 6.1)? The answer is:

> If (a) The missing data are missing at random (Rubin, 1975)
>
> and (b) ϕ is *a priori* independent of θ,
>
> Then all Bayesian inferences for θ based on (5.1) will be correct Bayesian inferences.

The missing data are missing at random if the probability of the observed pattern of missing data is the same for all possible values of the missing data, i.e., $g_\phi(\tilde{M}|U)$ does not depend on any u_{ij} such that $\tilde{m}_{ij} = 0$. If the missing data are missing at random and $p(\phi|\theta) = p(\phi)$, equation (6.1) becomes

$$\left[p(\theta) \int f_\theta(\tilde{U}) \; dU_0^{\tilde{M}}\right]\left[p(\phi)\; g_\phi(\tilde{M}|\tilde{U})\right]$$

so that θ and ϕ are *a posteriori* independent and the posterior of θ is proportional to equation (5.1). These two conditions are also necessary for naive Bayesian inferences to be correct except in rather artificial examples, see Rubin (1975).

If the missing data are not known to be missing at random, a proper Bayesian analysis has to specify the range of possible assignment processes given by g_ϕ and the prior distribution $p(\phi|\theta)$. In practical cases in which the assignment process is not known, this specification, if taken seriously by the Bayesian statistician, should require much thought and mental anguish. For instance, in the example of Section 3 it might be possible that the healthy (with respect to Y) patients received treatment 1 and the sick patients received treatment 2; clearly such an assignment would alter the posterior mean of the causal effect of treatment 1 vs. treatment 2. Or it might also be possible that the healthy patients received treatment 2 and the sick patients received treatment 1. Since the posterior distribution of the causal effect may change substantially as the various assignment procedures are considered, it is clear that the posterior variance of the causal effect may be much larger than in the naive Bayesian analysis.

More specifically, suppose in this example that *a priori* $\rho=0$ and $g_\phi(M|U)$ specifies

$$(6.2) \qquad (m_{i1},m_{i2}) = \begin{cases} (1,0) \text{ if } y_{i2} \geq \mu_2 \\ (0,1) \text{ if } y_{i2} < \mu_2 \end{cases}.$$

Thus, the patients with above average health under the second (e.g., control) treatment receive the first treatment, while patients with below average health receive the second treatment. Then with large samples the posterior distribution of $(\mu_1 - \mu_2)$ converges to $\bar{y}_1 - \max(\tilde{y}_{2i})$ rather than $\bar{y}_1 - \bar{y}_2$ as when one ignores g_ϕ, where $\max(\tilde{y}_{2j})$ is the maximum observed value of y_{2j}.

Practically, the only assignment procedures the Bayesian statistician need not be concerned with are those that are known (possibly probabilistic) functions of the observed values of the covariates X, such as (probability) sampling, (restricted) randomization, and assignment on the basis of a covariate. The fact that the process is known assures that ϕ is independent of θ since there is no ϕ to estimate; the fact that the process depends only on values known at the time of treatment assignment assures that the missing data are missing at random.

If the Bayesian statistician does not know the assignment process is some combination of randomization and assignment using the known values of a covariate, he must be prepared to perform the proper Bayesian analysis explicitly incorporating the assignment process.

Notice that there are good and bad assignment processes that satisfy conditions (a) and (b) above, i.e., processes that lead to small and large posterior variance and/or are robust or not to ranges of particular distributions f_θ. These issues are important for Bayesian experimental design but are not discussed here since the essential point is simply that the assignment process generally cannot be ignored.

7. The Objective Bayesian and Sampling Distribution Inferences

Thus far, our discussion of the Bayesian position has been very much from the subjective Bayesian point of view in the sense that priors were to be found by inner contemplation. Many statisticians are what might be called objective Bayesian in that they use the Bayesian framework but tend to strive for priors that lead to posteriors having good sampling distribution properties (e.g., a posterior mean that has small mean squared error) or that express ignorance relative to the expected information (in the sampling distribution sense) available from the experiment being contemplated. For example, consider Box and Tiao (1973 p. 46) in their discussion of non-informative priors and binomial sampling:

> This says that when we sample till the number of successes reaches a certain value some downward adjustment of probability is needed relative to sampling with fixed n. We find this result much less surprising than the claim that they ought to agree.
>
> In general we feel that it is sensible to choose a non-informative prior which expresses ignorance relative to information which can be supplied by a particular experiment. If the experiment is changed, then the expression of relative ignorance can be expected to change correspondingly.

Thus an objective Bayesian does not have a fixed prior found by contemplation but rather uses a prior determined from the sampling distribution of statistics.

The naive objective Bayesian analysis calculates the sampling distribution of these statistics from the density

$$(7.1) \qquad \int f_\theta(U)\, dU_0^{\tilde{M}}$$

i.e., he implicitly assumes $\tilde{M}$ is the only possible value of M. The correct sampling distribution of the statistic is found from the density

$$(7.2) \qquad \int f_\theta(U)\ g_\phi(M|U)\ dU^M_0 \quad .$$

It is clear that in general, the resulting sampling distributions of a statistic are not the same under densities (7.1) and (7.2) (Rubin, 1975).

However, as discussed in Section 3, it is common to assume f_θ is exchangeable in the rows of U. Given this assumption, it is sensible to restrict attention to statistics that for each value of M are exchangeable in the rows of U. The following result is immediate.

> If (a) f_θ is exchangeable in the rows of U,
>
> (b) M and U are independently distributed,
>
> and (c) all possible values of M are permutations of the rows of $\tilde{M}$,
>
> Then for every statistic S(U,M) that is exchangeable in the rows of U for each M, the sampling distribution of S(U,M) ignoring the assignment process (i.e., calculated from density (7.1)) is the correct sampling distribution of S(U,M) (i.e., as calculated from density (7.2)).

Also consider the necessity of the conditions (b) and (c) above. If condition (b) does not hold, in general the density (7.1) is not even the correct conditional density given that M takes the value $\tilde{M}$ (see Rubin, 1975). If condition (c) does not hold, then there exists an M* such that an exchangeable statistic has a different conditional distribution given M* than given $\tilde{M}$, and thus, different distributions under densities (7.1) and (7.2).

Conditions (b) and (c) hold only for simple random sampling without replacement followed by complete randomization. Probability sampling and randomization within blocks are excluded because the assignment depends on recorded variables (i.e., sampling weights and/or block indicators). Hence, not surprisingly, an objective prior for a completely randomized experiment is not necessarily the same as an objective prior for a randomized blocks experiment.

This raises the issue of what are good objective priors within the experimental design framework. This issue appears to be interesting but is not explored here since the point is simply that to an objective Bayesian, the only assignment process that allows him to ignore the assignment process when calculating objective priors is simple random sampling followed by complete randomization, and even this holds only in cases in which f_θ is exchangeable in the rows of U.

8. Some Verse

The points of this paper may be summarized fairly simply in the following doggerel which is given only for its simplicity of presentation and not for its literary merit.

> There exists no causation without manipulation;
>
> There exists no generalization without a population;
>
> There does exist experimentation without randomization,
>
> But
>
> The Bayesian must consider the manipulation in order to obtain the proper generalization.

9. Acknowledgments

I wish to thank P.W. Holland and the discussants of this paper, A.P. Dempster and M.R. Novick, for their helpful comments.

10. References

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley.

Campbell, D.T., and Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education lock harmful. In J. Hellmuth (Ed.), The Disadvantaged Child, Vol. 3. Compensatory Education: A National Debate. New York: Brunner/Mazel.

de Finneti, Bruno (1975). Theory of Probability. Vol. 1 (1974) & Vol. 2. John Wiley.

Gilbert, J.P. (1975). Randomization of human subjects. New England Journal of Medicine.

Gilbert, J.P., Light, R.J., Mosteller, F. (1974). Assessing social innovations: an empirical base for policy. To appear in Evaluation and Experiment: Some Critical Issues in Assessing Social Programs. A.R. Lumsdaine and C.A. Bennett (Eds).

Lindley, D.V. (1965). Introduction to Probability and Statistics from a Bayesian Viewpoint, Vols. 1 & 2. Cambridge University Press.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology Vol. 66, No. 5, 688-701.

Rubin, D.B. (1975). Inference and missing data. Educational Testing Service Research Bulletin 75-14. To appear in Biometrika.

Savage, L.J. The Foundations of Statistics. New York: Wiley, 1954.

Figure 1: The Matrix of Random Variables in a Causal Effect Study of T Treatments