Handbook of Statistics, Vol. 25
ISSN: 0169-7161

DOI 10.1016/S0169-7161(05)25001-0

Bayesian Inference for Causal Effects

Donald B. Rubin

Abstract

A central problem in statistics is how to draw inferences about the causal effects of treatments (i.e., interventions) from randomized and nonrandomized data. For example, does the new job-training program really improve the quality of jobs for those trained, or does exposure to that chemical in drinking water increase cancer rates? This presentation provides a brief overview of the Bayesian approach to the estimation of such causal effects based on the concept of potential outcomes.

1. Causal inference primitives

Although this chapter concerns Bayesian inference for causal effects, the basic conceptual framework is the same as that for frequentist inference. Therefore, we begin with the description of that framework. This framework with the associated inferential approaches, randomization-based frequentist or Bayesian, and its application to both randomized experiments and observational studies, is now commonly referred to as "Rubin's Causal Model" (RCM, Holland, 1986). Other approaches to Bayesian causal inference, such as graphical ones (e.g., Pearl, 2000), I find conceptually less satisfying, as discussed, for instance, in Rubin (2004b). The presentation here is essentially a simplified and refined version of the perspective presented in Rubin (1978).

1.1. Units, treatments, potential outcomes

For causal inference, there are several primitives – concepts that are basic and on which we must build. A "unit" is a physical object, e.g., a person, at a particular point in time. A "treatment" is an action that can be applied or withheld from that unit. We focus on the case of two treatments, although the extension to more than two treatments is simple in principle although not necessarily so with real data.

Associated with each unit are two "potential outcomes": the value of an outcome variable Y at a point in time when the active treatment is applied and the value of that outcome variable at the same point in time when the active treatment is withheld. The

objective is to learn about the causal effect of the application of the active treatment relative to the control (active treatment withheld) on Y.

For example, the unit could be "you now" with your headache, the active treatment could be taking aspirin for your headache, and the control could be not taking aspirin. The outcome Y could be the intensity of your headache pain in two hours, with the potential outcomes being the headache intensity if you take aspirin and if you do not take aspirin.

Notationally, let W indicate which treatment the unit, you, received: $W = 1$ the active treatment, $W = 0$ the control treatment. Also let $Y(1)$ be the value of the potential outcome if the unit received the active version, and $Y(0)$ the value if the unit received the control version. The causal effect of the active treatment relative to its control version is the comparison of $Y(1)$ and $Y(0)$ – typically the difference, $Y(1) - Y(0)$, or perhaps the difference in logs, $\log[Y(1)] - \log[Y(0)]$, or some other comparison, possibly the ratio.

We can observe only one or the other of $Y(1)$ and $Y(0)$ as indicated by W. The key problem for causal inference is that, for any individual unit, we observe the value of the potential outcome under only one of the possible treatments, namely the treatment actually assigned, and the potential outcome under the other treatment is missing. Thus, inference for causal effects is a missing-data problem – the "other" value is missing.

How do we learn about causal effects? The answer is replication, more units. The way we personally learn from our own experience is replication involving the same physical object (ourselves) with more units in time. That is, if I want to learn about the effect of taking aspirin on headaches for me, I learn from replications in time when I do and do not take aspirin to relieve my headache, thereby having some observations of $Y(0)$ and some of $Y(1)$. When we want to generalize to units other than ourselves, we typically use more objects.

1.2. Replication and the Stable Unit Treatment Value Assumption – SUTVA

Suppose instead of only one unit we have two. Now in general we have at least four potential outcomes for each unit: the outcome for unit 1 if unit 1 and unit 2 received control, $Y_1(0, 0)$; the outcome for unit 1 if both units received the active treatment, $Y_1(1, 1)$; the outcome for unit 1 if unit 1 received control and unit 2 received active, $Y_1(0, 1)$, and the outcome for unit 1 if unit 1 received active and unit 2 received control, $Y_1(1, 0)$; and analogously for unit 2 with values $Y_2(0, 0)$, etc. In fact, there are even more potential outcomes because there have to be at least two "doses" of the active treatment available to contemplate all assignments, and it could make a difference which one was taken. For example, in the aspirin case, one tablet may be very effective and the other quite ineffective.

Clearly, replication does not help unless we can restrict the explosion of potential outcomes. As in all theoretical work, simplifying assumptions are crucial. The most straightforward assumption to make is the "stable unit treatment value assumption" (SUTVA – Rubin, 1980, 1990) under which the potential outcomes for the ith unit just depend on the treatment the ith unit received. That is, there is "no interference between units" and there are "no versions of treatments". Then, all potential outcomes for N units with two possible treatments can be represented by an array with N rows and two columns, the ith unit having a row with two potential outcomes, $Y_i(0)$ and $Y_i(1)$.

There is no assumption-free causal inference, and nothing is wrong with this. It is the quality of the assumptions that matters, not their existence or even their absolute correctness. Good researchers attempt to make assumptions plausible by the design of their studies. For example, SUTVA becomes more plausible when units are isolated from each other, as when using, for the units, schools rather than students in the schools when studying an educational intervention.

The stability assumption (SUTVA) is very commonly made, even though it is not always appropriate. For example, consider a study of the effect of vaccination on a contagious disease. The greater the proportion of the population that gets vaccinated, the less any unit's chance of contracting the disease, even if not vaccinated, an example of interference. Throughout this discussion, we assume SUTVA, although there are other assumptions that could be made to restrict the exploding number of potential outcomes with replication.

1.3. Covariates

In addition to (1) the vector indicator of treatments for each unit in the study, $W = \{W_i\}$, (2) the array of potential outcomes when exposed to the treatment, $Y(1) = \{Y_i(1)\}$, and (3) the array of potential outcomes when not exposed, $Y(0) = \{Y_i(0)\}$, we have (4) the array of covariates $X = \{X_i\}$, which are, by definition, unaffected by treatment. Covariates (such as age, race and sex) play a particularly important role in observational studies for causal effects where they are variously known as potential "confounders" or "risk factors". In some studies, the units exposed to the active treatment differ on their distribution of covariates in important ways from the units not exposed. To see how this can arise in a formal framework, we must define the "assignment mechanism", the probabilistic mechanism that determines which units get the active version of the treatment and which units get the control version.

In general, the N units may not all be assigned treatment 1 or treatment 0. For example, some of the units may be in the future, as when we want to generalize to a future population. Then formally W_i must take on a third value, but for the moment, we avoid this complication.

1.4. Assignment mechanisms – unconfounded and strongly ignorable

A model for the assignment mechanism is needed for all forms of statistical inference for causal effects, including Bayesian. The assignment mechanism gives the conditional probability of each vector of assignments given the covariates and potential outcomes:

$$\Pr\big(W|X, Y(0), Y(1)\big). \tag{1}$$

Here W is a N by 1 vector and X, $Y(1)$ and $Y(0)$ are all matrices with N rows. An example of an assignment mechanism is a completely randomized experiment with N units, with $n < N$ assigned to the active treatment.

$$\Pr\big(W|X, Y(0), Y(1)\big) = \begin{cases} 1/C_n^N & \text{if } \sum W_i = n, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

An "unconfounded assignment mechanism" is free of dependence on either $Y(0)$ or $Y(1)$:

$$\Pr\big(W|X, Y(0), Y(1)\big) = \Pr(W|X). \tag{3}$$

With an unconfounded assignment mechanism, at each set of values of X_i that has a distinct probability of $W_i = 1$, there is effectively a completely randomized experiment. That is, if X_i indicates sex, with males having probability 0.2 of receiving the active treatment and females probability 0.5, then essentially one randomized experiment is described for males and another for females.

The assignment mechanism is "probabilistic" if each unit has a positive probability of receiving either treatment:

$$0 < \Pr\big(W_i = 1|X, Y(0), Y(1)\big) < 1. \tag{4}$$

A "strongly ignorable" assignment mechanism (Rosenbaum and Rubin, 1983) satisfies both (2) and (3): it is unconfounded and probabilistic. A nonprobabilistic assignment mechanism fails to satisfy (4) for some units.

The assignment mechanism is fundamental to causal inference because it tells us how we got to see what we saw. Because causal inference is basically a missing data problem with at least half of the potential outcomes not observed, without understanding the process that creates missing data, we have no hope of inferring anything about the missing values. Without a model for how treatments are assigned to individuals, formal causal inference, at least using probabilistic statements, is impossible. This does not mean that we need to know the assignment mechanism, but rather that without positing one, we cannot make any statistical claims about causal effects, such as the coverage of Bayesian posterior intervals.

Randomization, as in (2), is an unconfounded probabilistic assignment mechanism that allows particularly straightforward estimation of causal effects, as we see in Section 3. Therefore, randomized experiments form the basis for inference for causal effects in more complicated situations, such as when assignment probabilities depend on covariates or when there is noncompliance with the assigned treatment. Unconfounded assignment mechanisms, which essentially are collections of distinct completely randomized experiments at each distinct value of X_i, form the basis for the analysis of observational nonrandomized studies.

1.5. Confounded and ignorable assignment mechanisms

A confounded assignment mechanism is one that depends on the potential outcomes:

$$\Pr\big(W|X, Y(0), Y(1)\big) \neq \Pr(W|X). \tag{5}$$

A special class of possibly confounded assignment mechanisms are particularly important to Bayesian inference: ignorable assignment mechanisms (Rubin, 1978). Ignorable assignment mechanisms are defined by their freedom from dependence on any missing potential outcomes:

$$\Pr\big(W|X, Y(0), Y(1)\big) = \Pr(W|X, Y_{\text{obs}}), \tag{6}$$

$$\text{where } Y_{\text{obs}} = \{Y_{\text{obs},i}\}$$
$$\text{with } Y_{\text{obs},i} = W_i Y_i(1) + (1 - W_i) Y_i(0).$$

Ignorable assignment mechanisms do arise in practice, especially in sequential experiments. Here, the next unit's probability of being exposed to the active treatment depends on the success rate of those previously exposed to the active treatment versus the success rate of those exposed to the control treatment, as in "play-the-winner" designs (e.g., see Efron, 1971).

All unconfounded assignment mechanisms are ignorable, but not all ignorable assignment mechanisms are unconfounded (e.g., play-the-winner designs). Seeing why ignorable assignment mechanisms play an important role in Bayesian inference requires us to present the full Bayesian approach. Before doing so, we place the framework presented thus far in an historical perspective.

2. A brief history of the potential outcomes framework

2.1. Before 1923

The basic idea that causal effects are the comparisons of potential outcomes seems so direct that it must have ancient roots, and we can find elements of this definition of causal effects among both experimenters and philosophers. For example, Cochran (1978), when discussing Arthur Young, an English agronomist, stated:

> A single comparison or trial was conducted on large plots – an acre or a half acre in a field split into halves – one drilled, one broadcast. Of the two halves, Young (1771) writes: "The soil is exactly the same; the time of culture, and in a word every circumstance equal in both."

It seems clear in this description that Young viewed the ideal pair of plots as being identical, so that the outcome on one plot of drilling would be the same as the outcome on the other of drilling, $Y_1(\text{Drill}) = Y_2(\text{Drill})$, and likewise for broadcasting, $Y_1(\text{Broad}) = Y_2(\text{Broad})$. Now the difference between drilling and broadcasting on each plot are the causal effects: $Y_1(\text{Drill}) - Y_1(\text{Broad})$ for plot 1 and $Y_2(\text{Drill}) - Y_2(\text{Broad})$ for plot 2. As a result of Young's assumptions, these two causal effects are equal to each other and moreover, are equal to the two possible observed differences when one plot is drilled and the other is broadcast: $Y_1(\text{Drill}) - Y_2(\text{Broad})$ and $Y_1(\text{Broad}) - Y_2(\text{Drill})$.

Nearly a century later, Claude Bernard, an experimental scientist and medical researcher wrote (Wallace, 1974, p. 144):

> The experiment is always the termination of a process of reasoning, whose premises are observation. Example: if the face has movement, what is the nerve? I suppose it is the facial; I cut it. I cut others, leaving the facial intact – the control experiment.

In the late nineteenth century, the philosopher John Stuart Mill, when discussing Hume's views offers (Mill, 1973, p. 327):

> If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the source of his death.

And Fisher (1918, p. 214) wrote:

> If we say, "This boy has grown tall because he has been well fed," we are not merely tracing out the cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter.

Despite the insights evident in these quotations, there was no formal notation for potential outcomes until 1923, and even then, and for half a century thereafter, its application was limited to randomized experiments, apparently until Rubin (1974). Also, before 1923 there was no formal discussion of any assignment mechanism.

2.2. *Neyman's (*1923*) notation for causal effects in randomized experiments and Fisher's (*1925*) proposal to actually randomize treatments to units*

Neyman (1923) appears to have been the first to provide a mathematical analysis for a randomized experiment with explicit notation for the potential outcomes, implicitly making the stability assumption. This notation became standard for work in randomized experiments from the randomization-based perspective (e.g., Pitman, 1937; Welch, 1937; McCarthy, 1939; Anscombe, 1948; Kempthorne, 1952; Brillinger et al., 1978; Hodges and Lehmann, 1970, Section 9.4). The subsequent literature often assumed constant treatment effects as in Cox (1958), and sometimes was used quite informally, as in Freedman et al. (1978, pp. 456–458).

Neyman's formalism was a major advance because it allowed explicit frequentistic probabilistic causal inferences to be drawn from data obtained by a randomized experiment, where the probabilities were explicitly defined by the randomized assignment mechanism. Neyman defined unbiased estimates and asymptotic confidence intervals from the frequentist perspective, where all the probabilities were generated by the randomized assignment mechanism.

Independently and nearly simultaneously, Fisher (1925) created a somewhat different method of inference for randomized experiments, also based on the special class of randomized assignment mechanisms. Fisher's resulting "significance levels" (i.e., based on tests of sharp null hypotheses), remained the accepted rigorous standard for the analysis of randomized clinical trials at the end of the twentieth century. The notions of the central role of randomized experiments seems to have been "in the air" in the 1920's, but Fisher was apparently the first to recommend the actual physical randomization of treatments to units and then use this randomization to justify theoretically an analysis of the resultant data.

Despite the almost immediate acceptance of randomized experiments, Fisher's significance levels, and Neyman's notation for potential outcomes in randomized experiments in the late 1920's, this same framework was not used outside randomized experiments for a half century thereafter, and these insights were entirely limited to randomization-based frequency inference.

2.3. *The observed outcome notation*

The approach in nonrandomized settings, during the half century following the introduction of Neyman's seminal notation for randomized experiments, was to build mathematical models relating the observed value of the outcome variable $Y_{\text{obs}} = \{Y_{\text{obs},i}\}$

to covariates and indicators for treatment received, and then to define causal effects as parameters in these models. The same statistician would simultaneously use Neyman's potential outcomes to define causal effects in randomized experiments and the observed outcome setup in observational studies. This led to substantial confusion because the role of randomization cannot even be stated using observed outcome notation. That is, Eq. (3) does not imply that $\Pr(W|X, Y_{\text{obs}})$ is free of Y_{obs}, except under special conditions, i.e., when $Y(0) \equiv Y(1) \equiv Y_{\text{obs}}$, so the formal benefits of randomization could not even be formally stated using the collapsed observed outcome notation.

2.4. The Rubin causal model

The framework that we describe here, using potential outcomes to define causal effects and a general assignment mechanism, has been called the "Rubin Causal Model" – RCM by Holland (1986) for work initiated in the 1970's (Rubin, 1974, 1977, 1978). This perspective conceives of all problems of statistical inference for causal effects as missing data problems with a mechanism for creating missing data (Rubin, 1976).

The RCM has the following salient features for causal inference: (1) Causal effects are defined as comparisons of a priori observable potential outcomes without regard to the choice of assignment mechanism that allows the investigator to observe particular values; as a result, interference between units and variability in efficacy of treatments can be incorporated in the notation so that the commonly used "stability" assumption can be formalized, as can deviations from it; (2) Models for the assignment mechanism are viewed as methods for creating missing data, thereby allowing nonrandomized studies to be considered using the same notation as used for randomized experiments, and therefore the role of randomization can be formally stated; (3) The underlying data, that is, the potential outcomes and covariates, can be given a joint distribution, thereby allowing both randomization-based methods, traditionally used for randomized experiments, and model-based Bayesian methods, traditionally used for observational studies, to be applied to both kinds of studies. The Bayesian aspect of this third point is the one we turn to in the next section.

This framework seems to have been basically accepted and adopted by most workers by the end of the twentieth century. Sometimes the move was made explicitly, as with Pratt and Schlaifer (1984) who moved from the "observed outcome" to the potential outcomes framework in Pratt and Schlaifer (1988). Sometimes it was made less explicitly as with those who were still trying to make a version of the observed outcome notation work in the late 1980's (e.g., see Heckman and Hotz, 1989), before fully accepting the RCM in subsequent work (e.g., Heckman, 1989, after discussion by Holland, 1989). But the movement to use potential outcomes to define causal inference problems seems to be the dominant one at the start of the 21st century and is totally compatible with Bayesian inference.

3. Models for the underlying data – Bayesian inference

Bayesian causal inference requires a model for the underlying data, $\Pr(X, Y(0), Y(1))$, and this is where science enters. But a virtue of the framework we are presenting is that

it separates science – a model for the underlying data, from what we do to learn about science – the assignment mechanism, $\Pr(W|X_1 Y(0), Y(1))$. Notice that together, these two models specify a joint distribution for all observables.

3.1. The posterior distribution of causal effects

Bayesian inference for causal effects directly confronts the explicit missing potential outcomes, $Y_{\text{mis}} = \{Y_{\text{mis},i}\}$ where $Y_{\text{mis},i} = W_i Y_i(0) + (1 - W_i) Y_i(1)$. The perspective simply takes the specifications for the assignment mechanism and the underlying data (= science), and derives the posterior predictive distribution of Y_{mis}, that is, the distribution of Y_{mis} given all observed values,

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, W). \tag{7}$$

From this distribution and the observed values of the potential outcomes, Y_{obs}, and covariates, the posterior distribution of any causal effect can, in principle, be calculated.

This conclusion is immediate if we view the posterior predictive distribution in (7) as specifying how to take a random draw of Y_{mis}. Once a value of Y_{mis} is drawn, any causal effect can be directly calculated from the drawn values of Y_{mis} and the observed values of X and Y_{obs}, e.g., the median causal effect for males: $\text{med}\{Y_i(1) - Y_i(0)|X_i$ indicate males$\}$. Repeatedly drawing values of Y_{mis} and calculating the causal effect for each draw generates the posterior distribution of the desired causal effect. Thus, we can view causal inference completely as a missing data problem, where we multiply-impute (Rubin, 1987, 2004a) the missing potential outcomes to generate a posterior distribution for the causal effects. We have not yet described how to generate these imputations, however.

3.2. The posterior predictive distribution of Y_{mis} under ignorable treatment assignment

First consider how to create the posterior predictive distribution of Y_{mis} when the treatment assignment mechanism is ignorable (i.e., when (6) holds). In general:

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, W) = \frac{\Pr(X, Y(0), Y(1)) \Pr(W|X, Y(0), Y(1))}{\int \Pr(X, Y(0), Y(1)) \Pr(W|X, Y(0), Y(1)) \, \mathrm{d}Y_{\text{mis}}}. \tag{8}$$

With ignorable treatment assignment, Eqs. (3), (6) becomes:

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, W) = \frac{\Pr(X, Y(0), Y(1))}{\int \Pr(X, Y(0), X(1)) \, \mathrm{d}Y_{\text{mis}}}. \tag{9}$$

Eq. (9) reveals that under ignorability, all that needs to be modelled is the science $\Pr(X, Y(0), Y(1))$.

Because all information is in the underlying data, the unit labels are effectively just random numbers, and hence the array $(X, Y(0), Y(1))$ is row exchangeable. With essentially no loss of generality, therefore, by de Finetti's (1963) theorem we have that the distribution of $(X, Y(0), Y(1))$ may be taken to be i.i.d. (independent and identically

distributed) given some parameter θ:

$$\Pr\big(X, Y(0), Y(1)\big) = \int \left[\prod_{i=1}^{N} f\big(X_i, Y_i(0), Y_i(1)|\theta\big) \right] p(\theta)\, \mathrm{d}(\theta) \tag{10}$$

for some prior distribution $p(\theta)$. Eq. (10) provides the bridge between fundamental theory and the practice of using i.i.d. models. A simple example illustrates what is required to apply Eq. (10).

3.3. Simple normal example – analytic solution

Suppose we have a completely randomized experiment with no covariates, and a scalar outcome variable. Also, assume plots were randomly sampled from a field of N plots and the causal estimand is the mean difference between $Y(1)$ and $Y(0)$ across all N plots, say $\overline{Y}_1 - \overline{Y}_0$. Then

$$\Pr(Y) = \int \prod_{i=1}^{N} f\big(Y_i(0), Y_i(1)|\theta\big) p(\theta)\, \mathrm{d}\theta$$

for some bivariate density $f(\cdot|\theta)$ indexed by parameter θ with prior distribution $p(\theta)$. Suppose $f(\cdot|\theta)$ is normal with means $\mu = (\mu_1, \mu_0)$, variances (σ_1^2, σ_0^2) and correlation ρ. Then conditional on (a) θ, (b) the observed values of Y, Y_{obs}, and (c) the observed value of the treatment assignment, where the number of units with $W_i = K$ is n_K $(K = 0, 1)$, we have that when $n_0 + n_1 = N$ the joint distribution of $(\overline{Y}_1, \overline{Y}_0)$ is normal with means

$$\frac{1}{2}\left[\bar{y}_1 + \mu_1 + \rho\frac{\sigma_1}{\sigma_0}(\bar{y}_0 - \mu_0)\right],$$
$$\frac{1}{2}\left[\bar{y}_0 + \mu_0 + \rho\frac{\sigma_0}{\sigma_1}(\bar{y}_1 - \mu_1)\right],$$

variances $\sigma_1^2(1-\rho^2)/4n_0$, $\sigma_0^2(1-\rho^2)/4n_1$, and zero correlation, where $\bar{y}_1$ and $\bar{y}_0$ are the observed sample means of Y in the two treatment groups. To simplify comparison with standard answers, now assume large N and a relatively diffuse prior distribution for $(\mu_1, \mu_0, \sigma_1^2, \sigma_0^2)$ given ρ. Then the conditional posterior distribution of $\overline{Y}_1 - \overline{Y}_0$ given ρ is normal with mean

$$E\big[\overline{Y}_1 - \overline{Y}_0|Y_{\text{obs}}, W, \rho\big] = \bar{y}_1 - \bar{y}_0 \tag{11}$$

and variance

$$V\big[\overline{Y}_1 - \overline{Y}_0|Y_{\text{obs}}, W, \rho\big] = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0} - \frac{1}{N}\sigma_{(1-0)}^2, \tag{12}$$

where $\sigma_{(1-0)}^2$ is the prior variance of the differences $Y_i(1) - Y_i(0)$, $\sigma_1^2 + \sigma_0^2 - 2\sigma_1\sigma_0\rho$. Section 2.5 in Rubin (1987, 2004a) provides details of this derivation. The answer given by (11) and (12) is remarkably similar to the one derived by Neyman (1923) from the randomization-based perspective, as pointed out in the discussion by Rubin (1990).

There is no information in the observed data about ρ, the correlation between the potential outcomes, because they are never jointly observed. A conservative inference for $\overline{Y}_1 - \overline{Y}_0$ is obtained by taking $\sigma^2_{(1-0)} = 0$.

The analytic solution in (11) and (12) could have been obtained by simulation, as described in general in Section 3.2. Simulation is a much more generally applicable tool than closed-form analysis because it can be applied in much more complicated situations. In fact, the real advantage of Bayesian inference for causal effects is only revealed in situations with complications. In standard situations, the Bayesian answer often looks remarkably similar to the standard frequentist answer, as it does in the simple example of this section:

$$(\bar{y}_1 - \bar{y}_0) \pm 2\left(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\right)^{1/2}$$

is a conservative 95% interval for $\overline{Y}_1 - \overline{Y}_0$, at least in relatively large samples.

3.4. Simple normal example – simulation approach

The intuition for simulation is especially direct in this example of Section 3.3 if we assume $\rho = 0$; suppose we do so. The units with $W_i = 1$ have $Y_i(1)$ observed and are missing $Y_i(0)$, and so their $Y_i(0)$ values need to be imputed. To impute $Y_i(0)$ values for them, we need to find units with $Y_i(0)$ observed who are exchangeable with the $W_i = 1$ units, but these units are the units with $W_i = 0$. Therefore, we estimate (in a Bayesian way) the distribution of $Y_i(0)$ from the units with $W_i = 0$, and use this estimated distribution to impute $Y_i(0)$ for the units missing $Y_i(0)$.

Since the n_0 observed values of $Y_i(0)$ are a simple random sample of the N values of $Y(0)$, and are normally distributed with mean μ_0 and variance σ_0^2, with the standard independent noninformative prior distributions on (μ_0, σ_0^2), we have for the posterior of σ_0^2:

$$\sigma_0^2/s_0^2 \sim \text{ inverted } X^2_{n_0-1}/(n_0 - 1);$$

and for the posterior distribution of μ_0 given σ_0:

$$\mu_0 \sim N\big(\bar{y}_0, s_0^2/n_0\big);$$

and for the missing $Y_i(0)$ given μ_0 and σ_0:

$$Y_i(0) \ni W_i \neq 0 \overset{\text{i.i.d.}}{\sim} N\big(\mu_0, s_0^2\big).$$

The missing values of $Y_i(1)$ are analogously imputed using the observed values of $Y_i(1)$.

When there are covariates observed, these are used to help predict the missing potential outcomes using one regression model for the observed $Y_i(1)$ given the covariates, and another regression model for the observed $Y_i(0)$ given the covariates.

3.5. Simple normal example with covariate – numerical example

For a specific example with a covariate, suppose we have a large population of people with a covariate X_i indicating baseline cholesterol. Suppose the observed X_i is

dichotomous, *HI* versus *LO*, split at the median in the population. Suppose that a random sample of 100 with $X_0 = HI$ is taken, and 90 are randomly assigned to the active treatment, a statin, and 10 are randomly assigned to the control treatment, a placebo. Further suppose that a random sample of 100 with $X_i = LO$ is taken, and 10 are randomly assigned to the statin and 90 are assigned to the placebo. The outcome Y is cholesterol a year after baseline, with $Y_{i,\text{obs}}$ and X_i observed for all 200 units; X_i is effectively observed in the population because we know the proportion of X_i that are *HI* and *LO*.

Suppose the hypothetical observed data are as displayed in Table 1.

Table 1
Final cholesterol in artificial example

Baseline	$\bar{y}_1$	n_1	$\bar{y}_0$	n_0	$s_1 = s_0$
HI	200	90	300	10	60
LO	100	10	200	90	60

Then the inferences based on the normal-model are as follows:

Table 2
Inferences for example in Table 1

	HI	*LO*	Population $= \frac{1}{2}HI + \frac{1}{2}LO$
$E(\overline{Y}_1 - \overline{Y}_0 \mid X, Y_{\text{obs}}, W)$	−100	−100	−100
$V(\overline{Y}_1 - \overline{Y}_0 \mid X, Y_{\text{obs}}, W)^{1/2}$	20	20	$10\sqrt{2}$

Here the notation is being slightly abused because the first entry in Table 2 really should be labelled $E(\overline{Y}_1 - \overline{Y}_0 \ni X_i = HI \mid X, Y_{\text{obs}}, W)$ and so forth.

The obvious conclusion in this artificial example is that the statin reduces final cholesterol for both those with *HI* and *LO* baseline cholesterol, and thus for the population which is a 50%/50% mixture of these two subpopulations. In this sort of situation, the final inference is insensitive to the assumed normality of $Y_i(1)$ given X_i and of $Y_i(0)$ given X_i; see Pratt (1965) or Rubin (1987, 2004a, Section 2.5) for the argument.

3.6. Nonignorable treatment assignment

With nonignorable treatment assignment, the above simplifications in Sections 3.2–3.5, which follow from ignoring the specification for $\Pr(W \mid X, Y(0), Y(1))$, do not follow in general, and analysis typically becomes far more difficult and uncertain. As a simple illustration, take the example in Section 3.5 and assume that everything is the same except that only Y_{obs} is recorded, so that we do not know whether baseline is *HI* or *LO* for anyone. The actually assignment mechanism is now

$$\Pr\bigl(W \mid Y(0), Y(1)\bigr) = \int \Pr\bigl(W \mid X, Y(0), Y(1)\bigr)\, \mathrm{d}P(X)$$

because X itself is missing, and so treatment assignment depends explicitly on the potential outcomes, both observed and missing, which are both correlated with the missing X_i.

Inference for causal effects, assuming the identical model for the science, now depends on the implied normal mixture model for the observed Y data within each treatment arm, because the population Y values are a 50%/50% a mixture of those with *LO* and *HI* baseline cholesterol, and these subpopulations have different probabilities of treatment assignment. Here the inference for causal effects is sensitive to the propriety of the assumed normality and/or the assumption of a 50%/50% mixture, as well as to the prior distributions on μ_1, μ_0, σ_1 and σ_0.

If we mistakenly ignore the nonignorable treatment assignment and simply compare the sample means of all treated with all controls, we have $\bar{y}_1 = .9(200) + .1(100) = 190$ versus $\bar{y}_0 = .1(300) + .9(200) = 210$; doing so, we reach the incorrect conclusion that the statin is bad for final cholesterol in the population. This sort of example is known as "Simpson's Paradox" (Simpson, 1951) and can easily arise with incorrect analyzes of nonignorable treatment assignment mechanisms, and thus indicates why such assignment mechanisms are to be avoided whenever possible.

Randomized experiments are the most direct way of avoiding nonignorable treatment assignments. Other alternatives are ignorable designs with nonprobabilistic features so that all units with some specific value of covariates are assigned the same treatment. With such assignment mechanisms, randomization-based inference is impossible for those units since their treatment does not change over the various possible assignments.

4. Complications

There are many complications that occur in real world studies for causal effects, many of which can be handled much more flexibly with the Bayesian approach than with standard frequency methods. Of course, the models involved, including associated prior distributions, can be very demanding to formulate in a practically reliable manner. Here I simply list some of these complications with some admittedly idiosyncratically personal references to current work from the Bayesian perspective. Gelman et al. (2003), especially starting with Chapter 7, is a good reference for some of these complications and the computational methods for dealing with them.

4.1. Multiple treatments

When there are more than two treatments, the notation becomes more complex but is still straightforward under SUTVA. Without SUTVA, however, both the notation and the analysis can become very involved. The exploding number of potential outcomes can become especially serious in studies where the units are exposed to a sequence of repeated treatments in time, each distinct sequence corresponding to a possibly distinct treatment. Most of the field of classical experiment design is devoted to issues that arise with more than two treatment conditions (e.g., Kempthorne, 1952; Cochran and Cox, 1957, 1992).

4.2. Unintended missing data

Missing data, due perhaps to patient dropout or machine failure, can complicate analyzes more than one would expect based on a cursory examination of the problem. Fortunately, Bayesian/likelihood tools for addressing missing data such as multiple imputation (Rubin, 1987, 2004a) or the EM algorithm (Dempster et al., 1977) and its relatives, including data augmentation (Tanner and Wong, 1987) and the Gibbs sampler (Geman and Geman, 1984) are fully compatible with the Bayesian approach to causal inference outlined in Section 3. Gelman et al. (2003), Parts III and IV provide, guidance on many of these issues from the Bayesian perspective.

4.3. Noncompliance with assigned treatment

Another complication, common when the units are people, is noncompliance. For example, some of the subjects assigned to take the active treatment take the control treatment instead, and some assigned to take the control manage to take the active treatment. Initial interest focuses on the effect of the treatment for the subset of people who will comply with their treatment assignments. Much progress has been made in recent years on this topic from the Bayesian perspective, e.g., Imbens and Rubin (1997), Hirano et al. (2000). In this case, sensitivity of inference to prior assumptions can be severe, and the Bayesian approach is ideally suited to not only revealing this sensitivity but also to formulating reasonable prior restrictions.

4.4. Truncation of outcomes due to death

In other cases, the unit may "die" before the final outcome can be measured. For example, in an experiment with new fertilizers, a plant may die before the crops are harvested and interest may focus on both the effect of the fertilizer on plant survival and the effect of the fertilizer on plant yield when the plant survives. This problem is far more subtle than it may at first appear to be, and valid Bayesian approaches to it have only recently been formulated following the proposal in (Rubin, 2000); see (Zhang and Rubin, 2003) for simple large sample bounds. It is interesting that the models also have applications in economics (Zhang et al., 2004).

4.5. Direct and indirect causal effects

Another topic that is far more subtle than it first appears to be is the one involving direct and indirect causal effects. For example, the separation of the "direct" effect of a vaccination on disease from the "indirect" effect of the vaccination that is due solely to its effect on blood antibodies and the "direct" effect of the antibodies on disease. This language turns out to be too imprecise to be useful within our formal causal effect framework. This problem is ripe for Bayesian modelling as briefly outlined in Rubin (2004b).

4.6. Principal stratification

All the examples in Sections 4.3–4.5 can be viewed as special cases of "principal stratification" (Frangakis and Rubin, 2002), where the principal strata are defined by partially

unobserved intermediate potential outcomes, namely in our examples: compliance behavior under both treatment assignments, survival under both treatment assignments, and antibody level under both treatment assignments. This appears to be an extremely fertile area for research and application of Bayesian methods for causal inference, especially using modern simulation methods such as MCMC (Markov Chain Monte Carlo); see, for example, Gilks et al. (1995).

4.7. Combinations of complications

In the real world, such complications typically do not appear simply one at a time. For example, a randomized experiment in education evaluating "school choice" suffered from missing data in both covariates and longitudinal outcomes; also, the outcome was multicomponent as each point in time; in addition, it suffered from noncompliance that took several levels because of the years of school. Some of these combinations of complications are discussed in Barnard et al. (2003) in the context of the school choice example, and in Mealli and Rubin (2003) in the context of a medical experiment.

Despite the fact that Bayesian analysis is quite difficult when confronted with these combinations of complications, it is still a far more satisfactory attack on the real scientific problems than the vast majority of ad hoc frequentist approaches in common use today.

It is an exciting time for Bayesian inference for causal effects.

References

Anscombe, F.J. (1948). The validity of comparative experiments. *J. Roy. Statist. Soc., Ser. A* **61**, 181–211.

Barnard, J., Hill, J., Frangakis, C., Rubin, D. (2003). School choice in NY city: A Bayesian analysis of an imperfect randomized experiment. In: Gatsonis, C., Carlin, B., Carriquiry, A. (Eds.), *Case Studies in Bayesian Statistics, vol. V*. Springer-Verlag, New York, pp. 3–97. (With discussion and rejoinder.)

Brillinger, D.R., Jones, L.V., Tukey, J.W. (1978). Report of the statistical task force for the weather modification advisory board. In: *The Management of Western Resources, vol. II: The Role of Statistics on Weather Resources Management. Stock No. 003-018-00091-1*. Government Printing Office, Washington, DC.

Cochran, W.G. (1978). Early development of techniques in comparative experimentation. In: Owen, D. (Ed.), *On the History of Statistics and Probability*. Dekker, New York, pp. 2–25.

Cochran, W.G., Cox, G.M. (1957). *Experimental Designs*, second ed. Wiley, New York.

Cochran, W.G., Cox, G.M. (1992). *Experimental Designs*, second ed. Wiley, New York. Reprinted as a "Wiley Classic".

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

de Finetti, B. (1963). Foresight: Its logical laws, its subjective sources. In: Kyburg, H.E., Smokler, H.E. (Eds.), *Studies in Subjective Probability*. Wiley, New York.

Dempster, A.P., Laird, N., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B* **39**, 1–38. (With discussion and reply.)

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

Fisher, R.A. (1918). The causes of human variability. *Eugenics Review* **10**, 213–220.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*, first ed. Oliver and Boyd, Edinburgh.

Frangakis, C.E., Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Freedman, D., Pisani, R., Purves, R. (1978). *Statistics*. Norton, New York.

Geman, S., Geman, D. (1984). Stochastic relaxation. Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** (November), 721–741.

Gelman, A., Carlin, J., Stern, H., Rubin, D. (2003). *Bayesian Data Analysis*, second ed. CRC Press, New York.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1995). *Markov Chain Monte Carlo in Practice*. CRC Press, New York.

Heckman, J.J. (1989). Causal inference and nonrandom samples. *J. Educational Statist.* **14**, 159–168.

Heckman, J.J., Hotz, J. (1989). Alternative methods for evaluating the impact of training programs. *J. Amer. Statist. Assoc.* **84**, 862–874. (With discussion.)

Hirano, K., Imbens, G., Rubin, D.B., Zhou, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.

Hodges, J.L., Lehmann, E. (1970). *Basic Concepts of Probability and Statistics*, second ed. Holden-Day, San Francisco.

Holland, P.W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81**, 945–970.

Holland, P.W. (1989). It's very clear. Comment on "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training" by J. Heckman, V. Hotz. *J. Amer. Statist. Assoc.* **84**, 875–877.

Imbens, G., Rubin, D.B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25**, 305–327.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York.

McCarthy, M.D. (1939). On the application of the z-test to randomized blocks. *Ann. Math. Statist.* **10**, 337.

Mealli, F., Rubin, D.B. (2003). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services Outcome Research Methodology*, 2–8.

Mill, J.S. (1973). A system of logic. In: *Collected Works of John Stuart Mill, vol. 7*. University of Toronto Press, Toronto.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Translated in *Statistical Science* **5** (1990), 465–480.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge.

Pitman, E.J.G. (1937). Significance tests which can be applied to samples from any population. III. The analysis of variance test. *Biometrika* **29**, 322–335.

Pratt, J.W. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc., Ser. B* **27**, 169–203. (With discussion.)

Pratt, J.W., Schlaifer, R. (1984). On the nature and discovery of structure. *J. Amer. Statist. Assoc.* **79**, 9–33. (With discussion.)

Pratt, J.W., Schlaifer, R. (1988). On the interpretation and observation of laws. *J. Econometrics* **39**, 23–52.

Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

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

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

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

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **7**, 34–58.

Rubin, D.B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75**, 591–593.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D.B. (2000). The utility of counterfactuals for causal inference. Comment on A.P. Dawid, 'Causal inference without counterfactuals'. *J. Amer. Statist. Assoc.* **95**, 435–438.

Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.* **5**, 472–480.

Rubin, D.B. (2004a). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. Reprinted with new appendices as a "Wiley Classic."

Rubin, D.B. (2004b). Direct and indirect causal effects via potential outcomes. *Scand. J. Statist.* **31**, 161–170; 195–198, with discussion and reply.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc., Ser. B* **13**, 238–241.

Tanner, M.A., Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528–550. (With discussion.)

Wallace, W.A. (1974). *Causality and Scientific Explanation: Classical and Contemporary Science, vol. 2*. University of Michigan Press, Ann Arbor.

Welch, B.L. (1937). On the z test in randomized blocks and Latin squares. *Biometrika* **29**, 21–52.

Zhang, J., Rubin, D.B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by 'death'. *J. Educational and Behavioral Statist.* **28**, 353–368.

Zhang, J., Rubin, D., Mealli, F. (2004). Evaluating the effects of training programs with experimental data. Submitted for publication.