Journal of Educational and Behavioral Statistics
Fall 2004, Vol. 29, No. 3, pp. 343–367

Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies

Donald B. Rubin
Harvard University

Inference for causal effects is a critical activity in many branches of science and public policy. The field of statistics is the one field most suited to address such problems, whether from designed experiments or observational studies. Consequently, it is arguably essential that departments of statistics teach courses in causal inference to both graduate and undergraduate students. This article discusses an outline of such courses based on repeated experience over more than a decade.

Keywords: *Bayes, causal inference, Fisher, Neyman, noncompliance, Rubin's causal model*

Framework

The courses that this article summarizes have been taught in three basic versions: an advanced course taught repeatedly at Harvard University for more than a decade, sometimes jointly with the Department of Economics (with Professor Guido Imbens, now at the University of California, Berkeley); a current introductory course for undergraduates at Harvard University with no background in statistics or mathematics; and a compact short course taught for various Chapters of the American Statistical Association, at the Joint Statistical Meetings, as well as at several European universities. The versions obviously differ in the depth and range of material that can be covered, and in the order of presentation of some of the ideas. The discussion here most closely follows the advanced versions, although indications are given of how the course is modified for the other versions. The full content of the advanced course is the basis for a textbook by Imbens and Rubin, due to appear in 2005. I also have a very introductory text under way, based on the initial material in the current introductory course.

Causal Effects Through Potential Outcomes

All versions of the course begin with the definition of *causal effects* through *potential outcomes.* Causal effects are comparisons of the potential outcomes that

Thanks are due Elizabeth Stuart and Samantha Cook for careful reading of this manuscript and for exceptional work as teaching assistants in various versions of this course. Also, I thank Milo Shield for facilitating a grant from the Keck Foundation that helped make the introductory versions of this course possible. In addition, I thank Howard Wainer for encouraging this submission to *JEBS*. Finally, I must thank Guido Imbens for his sharing the teaching of the advanced version of the course in its early years.

would have been observed under different exposures of *units* to *treatments*. Thus, the primitives (basic concepts) for causal inference are units, treatments, and potential outcomes. For example, the causal effect of your taking aspirin two hours ago involves the comparison of the state of your headache now with what it would have been had you not taken aspirin: $Y(0)$ is your outcome (a measure of your current headache pain) without aspirin, and $Y(1)$ is your outcome with aspirin, and the difference, $Y(1) - Y(0)$, is an obvious definition of the causal effect of the aspirin on your headache. More generally, $Y(W)$ is your outcome under treatment W, $W = 0, 1$.

This starting point is commonly referred to as "Rubin's causal model" (RCM) (Holland, 1986), but the formal notation in the context of randomization-based inference in randomized experiments dates back to Neyman (1923), discussed by Rubin (1990a), and the intuitive idea goes back centuries in various literatures; for an early reference in statistics, see Fisher (1918), and for ones in economics see Tinbergen (1930) and Haavelmo (1944). The label "RCM" arises because of extensions (e.g., Rubin, 1974, 1977, 1978) that generated a formal framework beyond randomized experiments and beyond randomization-based inference; this extended framework also allowed the formal consideration of complications, such as unintended missing data and noncompliance. The critical inferential idea of this starting point is to view the problem of causal inference as one of drawing an inference about the missing value, $Y_{mis} = (1 - W)Y(1) + WY(0)$, from the observed value, $Y_{obs} = WY(1) + (W - 1)Y(0)$. That is, if $W = 1$, $Y_{mis} = \mathrm{Y}(0)$ and $Y_{obs} = Y(1)$, whereas if $W = 0$, $Y_{mis} = Y(1)$ and $Y_{obs} = Y(0)$. That is, we can never actually observe both $Y(0)$ and $Y(1)$, for any unit; one or the other is always missing.

Because of the importance of making this very intuitive and transparent definition of causal effect using potential outcomes central to students' thinking about causality, the course devotes some brief time to the history of this idea, including the history of standard assumptions, such as "no-interference-between-units" (Cox, 1958) and the more encompassing "stable-unit-treatment-value" assumption (SUTVA) (Rubin, 1978, 1980) and a history of randomization itself, including the fundamental contributions of Fisher (1925) and Neyman (1923). Under stability (SUTVA), each unit has a potential outcome under treatment 1 and another potential outcome under treatment 0, and the full set of all potential outcomes for N units can be represented by an N row by two column array (see Table 1). This simplifi-

TABLE 1
Potential Outcomes for N *Subjects When SUTVA Is Assumed*

Subject	$Y(1)$	$Y(2)$
1	a = b	c = d
2	e = g	f = h
. . .	. . .	. . .
N	y	z

Note. Entries for first two subjects correspond to Table 2.

cation is important because we learn about casual effects through replication: having more than one unit exposed to each of the treatments.

For an example of complications that can exist without this stability assumption, suppose you and I are in the same room with headaches, and *your* taking aspirin for your headache affects the state of *my* headache whether or not I take aspirin (if you do not take it and eliminate your headache, your complaining will be unbearable!) (see Table 2). Without some such "exclusion restrictions" (to use language common in economics), which limit the range of potential outcomes, causal inference is impossible. Understanding this limitation is essential: all causal inference relies on assumptions that restrict the possible potential outcomes so that we can learn something about causal effects from observable data. Nothing is wrong with making assumptions; on the contrary, such assumptions are the strands that join the field of statistics to scientific disciplines. The quality of these assumptions and their precise explication, not their existence, is the issue.

The use of potential outcomes, $Y(1)$ and $Y(0)$, rather than the observed outcome, Y_{obs}, to define the causal effect for each unit is especially critical in nonrandomized studies, as is made clear by the consideration of "Lord's Paradox" (Lord, 1967), as discussed by Holland and Rubin (1983). This example is a wonderful pedagogical tool to convince students of the value of the potential outcomes framework for thinking about causal inference.

This framework using potential outcomes to define causal effects in general is now relatively well accepted in many fields. For example, in psychology, see Wilkinson et al. (1999); in economics, see the transition to adopt it reflected by comparing Heckman (1979) to Heckman (1989), and Pratt and Schlaifer (1984) to Pratt and Schlaifer (1988), after discussion by Holland (1989) and Rosenbaum and Rubin (1984a), respectively. Also see Baker (1998), Dempster (1990), Efron and Feldman (1991), Gelman and King (1991), Greenland and Poole (1988), Greenland, Robins, and Pearl (1999), Holland (1988a, 1988b), Kadane and Seidenfeld (1990), Robins (1987, 1989), Rosenbaum (1987), Smith and Sugden (1988), Sobel (1990, 1995, 1996), Sugden (1988), and their references, and so forth. A recent article exploring whether the full potential outcomes framework can be avoided when conducting causal inference is Dawid (2000) with discussion. Also see Cox (1992).

TABLE 2
Potential Outcomes for Two Subjects When SUTVA Is Not Assumed

Subject	$Y(11)$	$Y(10)$	$Y(01)$	$Y(00)$
1	a	b	c	d
2	e	f	g	h

Note. $Y(jk)$ = outcome given Subject 1 receives treatment j and Subject 2 receives treatment k: j, $k = 1, 0$.

The Need for Posited Assignment Mechanisms

The essential role of the assignment mechanism is then introduced: Without a model for how treatments get assigned to units, formal causal inference, as least using probabilistic statements, is impossible. This does *not* mean that we cannot do anything unless we know the assignment mechanism. But it does mean that probabilistic statements about causal effects cannot be made without positing such mechanisms. It is critical that students appreciate this, especially because most published articles purporting to conduct causal inference, at least in many areas of the application of statistics, never even explicitly consider what the assignment mechanism might be unless the study was randomized.

To illustrate the need to posit an assignment mechanism, consider the trivial situation depicted in Table 3, where there are four patients, each assigned to one of two medical operations by a physician: Y is the number of years lived after the operation. Column 1 refers to the potential outcomes under operation 1, the vector $Y(1)$, and column 2 to the potential outcomes under operation 0, the vector $Y(0)$; this representation makes the stability assumption. The individual causal effects, defined as the years lived after operation 1 minus the years lived after operation 0, are given in the third column. We can, of course, never directly observe any of these causal effects. The fourth column, labeled W, indicates the operation each patient received, and the corresponding asterisked potential outcomes are the observed values of Y, the vector Y_{obs}; the unasterisked values comprise the vector Y_{mis}. Now, the doctor in this example is one each of us would want to use because each patient gets assigned the better operation—not better for the average patient—but better for the individual patient. But what general conclusions do the data suggest? The average observed length of survival for those given operation 1 is one year more than for those given operation 0, so the obvious, but incorrect, conclusion is that operation 1 is superior for the pool of patients for which these four patients are representative. This conclusion is wrong because the typical causal effects (e.g., the average or median of the individual causal effects in column 3) favor operation 0 for these patients, giving an average benefit of three years.

TABLE 3
Artificial Example of a Confounded Assignment Mechanism

Subject	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	W
1	1	6*	–5	0
2	3	12*	–9	0
3	9*	8	1	1
4	11*	10	1	1
Population mean	6	9	–3	
Observed mean	10**	9**		

* Observed values.

** Observed means—the difference of these (–1) has the wrong sign as an estimate of the true average causal effect (3).

The point of this little example is that the assignment mechanism matters to valid inference: simply comparing observed values under the treatments only works in general if patients are randomly assigned treatments. If, as in this little example, the assignment mechanism is not randomized, any analysis of the observed data must take this into account: because at least half the potential outcomes (which define causal effects) are missing, the process that makes them missing must be part of the inferential model (Rubin, 1976a, p. 581).

A simple classification of assignment mechanisms then follows, where W is the vector of treatment assignments for the units, and $Y(1)$ and $Y(0)$ are the vectors (or matrices) of the potential outcomes; X is the matrix of covariates, unaffected by treatment assignment, such as the units' ages or sexes.

Unconfounded (with the potential outcomes) assignment mechanisms have

$$\Pr[W|X, Y(1), Y(0)] = \Pr(W|X).$$

Ignorable assignment mechanisms have

$$\Pr[W|X, Y(1), Y(0)] = \Pr(W|X, Y_{obs}).$$

Nonignorable assignment mechanisms do not satisfy this expression, and therefore the assignment W is entangled with the missing potential outcomes, Y_{mis}, as in the "perfect doctor" example.

At this point, the advanced and the introductory courses take somewhat different paths. The introductory versions of the course introduce a whole sequence of examples, starting with simple randomized experiments and progressing through to more general ignorable designs, both unconfounded (e.g., Efron, 1971) and confounded (e.g., Ware, 1989), and ending with some nonignorable examples. These examples help to introduce basic probability calculations to students in the introductory courses who cannot be expected to have been exposed to them before.

The advanced versions of the course take a far more formal approach, relying on the superior technical background of these students. All versions point out that classical randomized experiments are unconfounded with each unit having a positive probability of receiving each treatment:

$$0 < p(W_i = 1|X_i) < 1.$$

All versions of the course then turn to methods for causal inference in the simplest setting: the completely randomized experiment.

Formal Statistical Modes of Causal Inference and Their Application in Classical Randomized Experiments

Fundamentally, there are three formal statistical modes of causal inference; one is Bayesian, which treats the potential outcomes as random variables, and two are based only on the assignment mechanism, which treat the potential outcomes as fixed but unknown quantities. Rubin (1990b) describes these three as well as a combination, which is fundamentally not as conceptually tight, and so is not discussed here as a distinct mode. Of the two distinct forms of randomization-based inference, one is that of Neyman (1923) and the other is that of Fisher (1925). All are first introduced in the absence of covariates, X. All versions of the course begin with the assignment-based models because they extend classical randomization-based methods in randomized experiments.

Fisherian Randomization-Based Inference

Fisher's approach is the more direct conceptually and is introduced before the others. It is closely related to the mathematical idea of proof by contradiction. It basically is a "stochastic proof by contradiction" giving the significance level (or p value)—really, the plausibility—of the "null hypothesis," which often is that there is absolutely no treatment effect whatsoever.

The first element in Fisher's mode of inference is the null hypothesis, which is nearly always $Y(1) \equiv Y(0)$ for all units: the treatments have absolutely no effect on the outcomes. Under this null hypothesis, all potential outcomes are known from the observed outcome Y_{obs} because $Y(1) \equiv Y(0) \equiv Y_{obs}$. It follows that, under this null hypothesis, the value of any statistic, S, such as the difference of the observed averages for units exposed to treatment 1 and units exposed to treatment 0, $\bar{y}_1 - \bar{y}_0$, is known not only for the observed assignment, but for all possible assignments W.

Suppose we choose a statistic, S such as $\bar{y}_1 - \bar{y}_0$, and calculate its value under each possible assignment (assuming the null hypothesis) and also calculate the probability of each assignment under the randomized assignment mechanism. In most classical experiments, these probabilities are either zero or a common value for all possible assignments. For example, in a completely randomized experiment with $N = 2n$ units, n are randomly chosen to receive treatment 1 and n to receive treatment 0. Then any assignment W that has n 1's and n 0's has probability $1/C^N_n$, and all other W's have zero probability. Knowing for each W the value of S and its probability, we can then calculate the probability (under the assignment mechanism and the null hypothesis) that we would observe a value of S as "unusual" as, or more unusual than, the observed value of S, S_{obs}. "Unusual" is defined a priori, typically by how discrepant S_{obs} is from the typical values of S. This probability is the plausibility (p value or significance level) of the observed value of the statistic S under the null hypothesis: the probability of a result (represented by the value S_{obs} of the statistic, S) as rare or more rare than the actual observed value if the null hypothesis were true, where the probability is over the distribution induced by the assignment mechanism.

This form of inference is elegant: Unless the data suggest that the null hypothesis of no treatment effect is false (for an appropriate choice of statistic, S), it is not easy to claim evidence for differing efficacies of the treatments.

Fisher's approach is then extended to other "sharp" null hypotheses, that is, a null hypothesis such that from knowledge of Y_{obs}, $Y(1)$ and $Y(0)$ are known; for example, an additive null, which asserts that for each unit, $Y(1) - Y(0)$ is a specified constant (e.g., 3). The collection of such null hypotheses that do not lead to an extreme p value can be used to create interval estimates of the causal effect assuming additivity, which bridges to Neyman's approach. Extensions to other statistics and other designs are only mentioned, and are considered later in all versions of the course.

Neymanian Randomization-Based Inference

Neyman's form of randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism in order to calculate a confidence interval for the typical causal effect. The essential idea is the same as in Neyman's (1934) classic article on randomization-based (now often called "designed-based") inference in surveys. Typically, an unbiased estimator of the causal estimand (the typical causal effect, e.g., the average) is created, and an unbiased, or upwardly biased, estimator of the variance of that unbiased estimator is found (bias and variance both defined with respect to the randomization distribution). Then, an appeal is made to the central limit theorem for the normality of the estimator over its randomization distribution, whence a confidence interval for the causal estimand is obtained.

To be more explicit, the causal estimand is typically the average causal effect $\overline{Y(1)} - \overline{Y(0)}$, where the averages are over all units in the population being studied, and the traditional statistic for estimating this effect is the difference in sample averages for the two groups, $\bar{y}_1 - \bar{y}_0$, which can be shown to be unbiased for $\overline{Y(1)} - \overline{Y(0)}$ in a completely randomized design. A common choice for estimating the variance of $\bar{y}_1 - \bar{y}_0$ over its randomization distribution, in completely randomized experiments with $N = n_1 + n_2$ units, is $se^2 = s_1^2 / n_1 + s_0^2 / n_0$, where s_1^2, s_0^2, n_1, and n_0 are the observed sample variances and sample sizes in the two treatment groups. Neyman (1923) showed that se^2 overestimates the actual variance of $\bar{y}_1 - \bar{y}_0$, unless additivity holds (i.e., unless all individual causal effects are constant), in which case se^2 is unbiased for the variance of $\bar{y}_1 - \bar{y}_0$. The standard 95% confidence interval for $\overline{Y(1)} - \overline{Y(0)}$ is $\bar{y}_1 - \bar{y}_0 \pm 1.96se$, which, in large enough samples, includes $\overline{Y(1)} - \overline{Y(0)}$ in at least 95% of the possible random assignments.

Neyman's form of inference is less direct than Fisher's. It is really aimed at evaluations of procedures: in repeated applications, how often does the interval $\bar{y}_1 - \bar{y}_0 \pm 1.96se$ include $\overline{Y(1)} - \overline{Y(0)}$? Nevertheless, it forms the basis for much of what is done in important areas of application (e.g., the world of pharmaceutical development and approval, the world of randomized clinical trials in medicine), and therefore it is important that students understand the framework and how it differs from Fisher's. Neyman's approach is not prescriptive in telling us what to do, but rather it tells us how to evaluate a proposed procedure for drawing causal inferences. The third approach is more direct.

Bayesian Inference

The third form of statistical inference for causal effects is Bayesian, as developed in Rubin (1978), where the model for the assignment mechanism, $\Pr[W|X, Y(1), Y(0)]$, is supplemented with a model for the data, $\Pr[Y(1), Y(0)|X]$. A causal inference is obtained by conditioning on what is observed to calculate the conditional distribution of the causal estimand given observed values, which follows by Bayes theorem from the observed data and the models for the assignment mechanism and the data.

The advanced versions of the course continue more explicitly. Suppose that the causal estimand is, as before, $\overline{Y(1)} - \overline{Y(0)}$; then its "posterior" distribution; that is, its conditional distribution given the model specifications and the observed values of W, X, and Y_{obs}, is written as $\Pr(\overline{Y(1)} - \overline{Y(0)} \mid W, X, Y_{obs})$, which follows from the "posterior predictive" distribution of the missing values,

$$\Pr(Y_{mis} \mid W, X, Y_{obs}),$$

where this expression is evaluated at the observed values of W, X, and Y_{obs}. The analytic analysis in a completely randomized experiment with no covariates is then presented using a simple normal model (Rubin, 1990a). The answer is shown to agree very closely with Neyman's confidence approach:

$$\Pr[\overline{Y(1)} - \overline{Y(0)} \mid W, X, Y_{obs}] \doteq N\left(\bar{y}_1 - \bar{y}_0, \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0} - c\right),$$

where c is directly proportional to the variance of the individual causal effects, and thus, $c = 0$ under additivity, thereby agreeing with the Neymanian answer pragmatically.

In all versions of the course, intuitive methods of multiply imputing (Rubin, 1987) the values in Y_{mis} are considered. Multiple imputation is used because all Bayesian calculations in the course (nearly all in the advanced course) are described in terms of simulating the posterior predictive distribution of the missing values—basically by multiply imputing the missing potential outcomes, Y_{mis}. For example, in a completely randomized experiment with no covariates, the missing values of $Y(1)$ for control units are drawn, essentially, at random from the observed values of $Y(1)$ for the treated units, and the missing values of $Y(0)$ for the treated units are drawn from the observed values of $Y(0)$ for the control units. In each imputed data set, the causal estimand is then just calculated, and the causal estimand's posterior distribution is simply the set of its values over the full set of multiple imputations. This simulation-based approach opens the door for the much more demanding computational methods used in the advanced versions of the course, for example, MCMC (Markov Chain Monte Carlo) methods, and shows that the predictive approach does not require the usual formal models.

It is emphasized that, when formal models are used for predicting the missing potential outcomes, the parameters of the models are generally not causal effects. For example, it may be that log Y (e.g., Y = income) is considered normal in the treatment and control groups, but the causal effect of interest is the comparison of total incomes of the units under the two treatment conditions. The investigator defines the causal estimand, but nature chooses the distributional forms for the variables, and inference via simulation is ideally suited for addressing such questions. After the missing values are imputed, the causal effect is calculated; for example, the average causal effect is found by simple arithmetic once the missing causal effects have been imputed.

Because of its use of models for the data as implemented by simulation, the Bayesian approach is by far the most direct and flexible of the modes of inference for causal effects. However, it achieves these benefits by postulating a distribution for the data, $\Pr[Y(1), Y(0) \mid X]$, which the assignment-based (randomization-based) approaches avoid. Such a distribution can be very useful, but it can be like a loaded gun in the hands of a child, fraught with danger for the naive data analyst. This inherent flexibility and danger is a critical message to convey to students, especially in the context of observational studies, discussed later in the course.

Roles for Covariates in the Modes of Inference

As stated earlier, covariates are variables whose values are not affected by the treatment assignment, for example, variables that are recorded before randomization into treatment groups (e.g., year of birth, pretest scores in an educational evaluation). In classical randomized experiments, covariates can be used to increase efficiency of estimation. If a covariate is used in the assignment mechanism, as with a blocking variable in a randomized block design, that covariate must be used in analysis, from all three perspectives, for valid inference. The point about efficiency gains is made in the context of a completely randomized experiment.

The efficiency role for covariates is most transparent with Bayesian inference: for example, in an educational experiment, observed pretest scores are typically predictive of missing posttest scores (e.g., posttest scores under the treatment not received). Therefore, using pretest scores improves prediction of Y_{mis} and reduces the variance of the imputed Y_{mis} potential outcomes. For example, when using a linear regression model, the residual variance of posttest scores after adjusting for pretest scores is less than the variance of raw posttest scores. This use of regression estimation leads, under the usual normal model for the potential outcomes, to the "covariance-adjusted" estimator of causal effects. In the advanced versions of the course, ordinary least squares is formally introduced at this point as a Bayesian method for predicting the missing potential outcomes.

From the Fisherian and Neymanian perspectives, we can use covariates to define a new statistic to estimate causal estimands. For example, one can use the difference in average observed gain scores, $(\bar{y}_1 - \bar{x}_1) - (\bar{y}_0 - \bar{x}_0)$—where $\bar{x}_1$ and $\bar{x}_0$ are average observed pretest scores for those exposed to $W = 1$ and $W = 0$, respectively—rather than the difference in average posttest scores, to estimate $\overline{Y(1)} - \overline{Y(0)}$. If X and Y are correlated, from the Neymanian perspective, the variance of the difference in average gain scores will be less than the variance of the difference in average posttest scores, which translates into smaller estimated variances and shorter confidence intervals. From the Fisherian perspective, this reduced variance translates into more sensitive tests of the null hypothesis: under an alternative hypothesis, smaller real deviations from the null hypothesis are more likely to lead to more significant p values.

This point is easily made clear by examples. Suppose as an extreme case, that the new treatment adds essentially 10 points to everyone's pretest score, whereas the old treatment does nothing. The observed gain scores have essentially zero variance in each treatment group, whereas the posttest scores have the same variances as the pretest scores. This means that the Neymanian confidence interval for the treated minus control difference in gain scores is much shorter than the corresponding interval for posttest scores. Also, the observed value of the difference of gain scores is the most extreme value that can be observed under the Fisherian null hypothesis, and so the observed result with gain scores is as significant as possible, which is not true for the difference in posttest scores.

In the advanced versions of the course, it is emphasized that the Fisherian and Neyman approaches can use any statistic. For example, they can be applied to the covariance adjusted estimate found by the Bayesian approach.

Novel Features of the RCM Approach

It is now noted that the approach to causal inference being taught in these courses is fundamentally different from the approach in traditional "regression-based" courses or "design of experiments" courses, and incorporates three novel features. First, potential outcomes are used to define causal effects in all cases, no matter what the actual design of the study, observational or experimental. Although Neyman's potential outcome notation was standard in the context of randomized experiments, prior to Rubin (1974, 1977, 1978), the observed outcome notation was standard in the context of observational studies. From the RCM perspective, the potential outcomes (and covariates) are defined as scientific entities, unaffected by whether we try to learn about them via experiments or observational studies.

Second, a formal model for the assignment mechanism, the process that creates missing and observed potential outcomes, is defined explicitly. Displaying the dependence on the potential outcomes is a novel (if not uncontroversial) feature of this approach, and it allows the formal benefits of randomized and sequentially randomized designs to be stated; this cannot be done relying on the observed outcome notation. Assignment-based inference relies solely on this model for its inferences for causal effects.

Third, the framework allows, but does not require, a distribution for the potential outcomes (and covariates), thereby completing the model specification for all variables. Thus, the framework accommodates both assignment-based and Bayesian modes of inference in one coherent approach rather than using distinct approaches for experiments and observational studies. Previously, this was thought impossible by some (see, e.g., Kempthorne, 1976, p. 497).

Regular Designs

The course then turns to inference in "regular designs," which are like completely randomized experiments except that the probabilities of treatment assignment are allowed to depend on covariates, and so can vary from unit to unit. In the introductory course, these designs are simply called "more complex randomized experiments," and the formality for them is omitted. The randomized paired comparison and randomized block designs are particularly important, especially in the introductory versions of the course.

Notation

Regular designs have two features. First, they are unconfounded,

$$\Pr[W|X, Y(1), Y(0)] = \Pr(W|X),$$

(e.g., older males have probability .8 of being assigned the new treatment; younger males, .6; older females, .5; and younger females, .2). Second, they have as key elements the individual assignment possibilities as a function of unit i's value of the covariates, $p_i \equiv \Pr(W_i \mid X_i)$, which are strictly between zero and one,

$$0 < p_i < 1,$$

and

$$\Pr(W|X) = g(W)\prod_{1}^{N} p_i$$

for some exchangeable function $g(\cdot)$. These assignment probabilities, p_i, are called propensity scores (Rosenbaum & Rubin, 1983a). Regular designs are the major template for the analysis of observational, nonrandomized studies. That is, with an observational data set, we try to structure it so that we can conceptualize the data as having arisen from an underlying regular assignment mechanism.

Two situations need to be distinguished: when the propensity scores are known and when they are not, although the techniques of analysis are similar in both settings.

Known Propensity Scores

When the propensity scores are known, the assignment mechanism is essentially known. As a result, simple generalizations of Fisherian and Neymanian modes of inference can be applied. In particular, Horvitz-Thompson (1952) estimation (also see Cochran, 1963), where observations are weighted by the inverse probabilities of their being observed, plays a key role for both assignment-based modes of inference because these estimates are unbiased for average treatment effects over the randomization distribution with no modeling assumptions. As the overlap in propensity scores in the treatment and control groups becomes less, the Neymanian variance of the estimators for the average causal effect increases, and the Fisherian randomization distribution has more of its probability mass on the observed randomization, with the result that it becomes very difficult to get a "significant" p value. If there is no or little overlap in the propensity scores in the treatment groups, no causal inference is possible without strong external assumptions. This is a critical issue that all students in all versions of the course must be made to appreciate.

In general, with the assignment-based modes of inference and known propensity scores that take many values, it is acceptable to create several (e.g., 5–10) subclasses of propensity scores to recreate a randomized block experiment (i.e., a series of completely randomized experiments with different propensities). Alternatively, pairs of treatment-control units can be created that are matched on the propensity scores, thereby recreating a paired comparison experiment.

With Bayesian inference, a regular assignment mechanism is ignorable (Rubin, 1978), so that after including the covariates that determine the assignment probabilities (or an "adequate summary" [Rubin, 1985], such as, possibly, the vector of propensity scores in the model), analysis in principle proceeds as in a classical randomized experiment. In this situation, however, there can be much greater sensitivity to the model specification for $\Pr[Y(1), Y(0) | X]$ than in a classical randomized experiment because of the extrapolation involved when there is little overlap in the propensity scores' distributions in treated and control groups. This sensitivity to model extrapolation is the Bayesian's analogue of the Neymanian increased variance of the Horvitz-Thompson estimator and the Fisherian decreasing lack of ability to achieve significant differences in such designs as they become more unbalanced.

The sensitivity of formal models can be reduced by fitting separate models within propensity score subclasses, where the distributions of treatment and control covariates are similar. As mentioned earlier, depending on the level of the course, the way such models are described varies tremendously. In the introductory course, intuitive matching methods are the focus (e.g., to impute the missing $Y_i(0)$ for treated unit i, find and sample from the "donor pool" consisting of the five controls with the "closest" values of X to X_i). In the advanced course, the modeling is typically fully Bayesian and can involve MCMC methods to impute the missing potential outcomes.

Unknown Propensity Scores

When the propensity scores are unknown but the assignment mechanism is regular, an important first step from the assignment-based perspective is to estimate them. Differing methods to estimate propensity scores are discussed in all courses (e.g., discriminant analysis, logistic regression), but the students are not expected to be able to use these methods in the introductory course. Again, if there is little or no overlap in the distributions of the estimated propensity scores in the treatment groups, there is no hope of drawing valid causal inferences from these data without making strong external assumptions involving model-based extrapolation, because the estimated propensities will all be essentially either 0 or 1. The message that sometimes a data set cannot support a decent causal inference is very important to convey to students at all levels.

Generally, with a design that is known to be regular, the issues that arise with estimated propensity scores are the same as with known ones, and the reduction to a paired-comparison or randomized-block design is acceptable when there is enough overlap in the estimated propensity scores. The main difference that arises with sufficient overlap is that the use of estimated rather than true propensity scores typically yields more precise, not less precise estimates (Rubin & Thomas, 1992b). This point is simply mentioned in the introductory courses but developed in the more advanced ones.

Matched Sampling and Subclassification

When the design is regular, many issues are common to cases with known and unknown propensity scores, for example the use of matched sampling to create treated-control pairs whose values of X are "close." Various definitions of close are discussed: techniques for scalar X (Rubin, 1973a), ones for multivariate X summarized by the scalar (estimated) propensity score or best linear discriminant (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1983a), Mahalanobis metric matching (Rubin, 1976b, 1976c, 1979), Mahalanobis metric matching within propensity score calipers (Rosenbaum & Rubin, 1985; Rubin & Thomas, 1996). The level and amount of technical detail that is presented varies dramatically from the introductory to the advanced versions of the course.

The closely related technique of subclassification is also presented, starting with Cochran (1968). The example in Rubin (1997) on treatments for breast cancer is a particularly simple one that can be used to display basic ideas. The possible need to discard irrelevant units from the control group is emphasized using real examples (e.g., from the tobacco litigation [Rubin, 2002]). Sometimes it may even be necessary to discard some treated units at "unmatchable."

In the introductory versions of the course, the focus is on conveying the intuition behind these approaches: the ability of matching and subclassification to balance simultaneously the distribution of many covariates in the treated and control groups. In the advanced versions of the course, some theoretical results, for instance, concerning "Equal Percent Bias Reducing" (EPBR) (Rubin, 1976b, 1976c) matching methods are mentioned, as are extensions involving affinely invariant matching methods with ellipsoidal distributions (Rubin & Thomas, 1992a), and students are encouraged to explore topics in this relatively unstudied but important area.

The Combination of Propensity-Based and Predictive Methods

Once matched samples and/or subclassified samples have been obtained, the use of predictive models can be very helpful for obtaining improved estimates. This message is old, having appeared in Rubin (1973b), Cochran and Rubin (1973), and Rubin (1979), with analytic and simulation-based justification. The combination is especially important for obtaining tight interval estimates. Many illustrative examples are given, using matching, subclassification, and these combined with model building (e.g., Reinisch et al., 1995; Rosenbaum & Rubin, 1984b, 1985; Rubin, 1997; Smith, 1997) assuming the assignment mechanism is regular. The propensity-based methods tend to remove bias whatever the relationships between potential outcomes and covariates if the model for the assignment mechanism is appropriate, whereas the model-based methods tend to remove bias and reduce variance when the assumed model is appropriate. This result is especially important when the propensity scores are not known and is developed more fully in that context.

Using these real examples, including ones where there exist nearly parallel randomized and nonrandomized studies (e.g., as in Dehejia & Wahba, 1999; LaLonde, 1986), it is reinforced that in general the combination of both propensity score methods and Bayesian (predictive) model building is superior to either alone for the objective of achieving essentially the same answer from an observational study as from the parallel randomized experiment. In the advanced courses, some attention is paid to theoretical results, such as in Rubin and Thomas (2000). Because the most important ideas are easily conveyed by real examples, this theory is only mentioned in the introductory courses.

Observational Studies—Basic Approach

The course is now ready to tackle observational studies for causal effects. The critical issue here is that, in order to draw statistical inferences, a model for the assignment mechanism is needed, and thus, we need a template into which we can map the data from an observational study. That is, we need to posit a particular form for the assignment mechanism, and the major template is the class of regular designs.

Template for Observational Studies

Although regular designs with unknown propensity scores are not that common in practice (because of the need to know all covariates used in the assignment mechanism), they are the most critical template for inference for causal effects from observational data. That is, we assemble data with enough covariates that it becomes plausible (or initially arguable) to claim that the unknown assignment mechanism is unconfounded, and moreover regular, given these covariates. This is also known as the assumption of "strongly ignorable treatment assignment" (Rosenbaum & Rubin, 1983a).

Given an observational study with the assumption of strongly ignorable treatment assignment, we can apply the techniques for a regular design with unknown propensity scores to draw valid causal inferences (valid under that assumption). In fact, the real examples of regular designs with unknown propensity scores that were presented as examples of regular designs were, in fact, mostly observational studies. They are now reviewed again where the assumption of strongly ignorable treatment assignment, instead of being simply accepted, is evaluated.

Designing an Observational Study

One of the key ideas in this part of the course is that like good experiments, good observational studies are designed, not simply "found." When designing an experiment, we do not have any outcome data, but we plan the collection, organization, and analysis of the data to improve our chances of obtaining valid, reliable, and precise causal answers. The same exercise should be done in an observational study. Even if outcome data are available at the design stage, they should be put aside. The example of designing an observational study to assess the effects of smoking is used to illustrate this important point (Rubin, 2002).

Matching, subclassification, and propensity score methods are especially appropriate at this stage of design because they do not use outcome data. The students are reminded that with matching, treated and control matched pairs are formed on the basis of covariates, thereby trying to reconstruct a paired comparison randomized experiment, whereas with subclassification, groups of treated and control units are assembled that are similar with respect to covariates, thereby trying to reconstruct a series of completely randomized experiments with changing probabilities of treatment assignment. Propensity score methods are especially important here when there are many covariates to control.

Model-based methods cannot be used at this design stage, because they involve outcome data. However, plans can be made for how to use them, for example, regression adjustments on matched-pair differences or within subclasses, and such adjustments can be important for removing conditional biases remaining after the matching or subclassification. Important references include William Cochran's work on this topic, reviewed in Rubin (1984), and real examples (e.g., Imbens, Rubin, & Sacerdote, 1999; Reinisch et al., 1995; Rosenbaum & Rubin, 1984b).

Analyzing an Observational Study

Because the template for the basic analysis of an observational study is a complex randomized experiment (i.e., a regular design), the basic analysis parallels that of a regular design. Here is where the bridge to standard regression-based analyses of observational data is made, but in all versions of the course it is emphasized that the regression models fundamentally are being used to predict the missing potential outcomes. That is, predict the missing $Y(1)$ values for those assigned $W = 0$ from models for $Y(1)$ given X fitted using those assigned $W = 1$, and predict the missing $Y(0)$ values for those assigned treatment ($W = 1$) using models for $Y(0)$ given X fitted using those assigned control ($W = 0$). Of course, issues of model extrapolation are of major concern in observational studies, especially when the propensity scores do not overlap much. The advanced courses also consider examples of the analysis of observational data from other perspectives, especially from economics (e.g., Ashenfelter & Card, 1985; Ashenfelter & Krueger, 1993).

Sensitivity Analyses and Bounds for Nonignorable Designs

Because the conceptualization of an observational data set as arising from a regular (and thus, ignorable) assignment mechanism is an assumption, it is often important to consider deviations from that assumption. Two related techniques are described in all versions of the course: sensitivity analysis and the creation of bounds.

Sensitivity analyses allow the assignment mechanism to be nonignorable by positing some unobserved covariates U so that the assignment mechanism is ignorable only when given U (and moreover, regular in most examples), and U is correlated with the potential outcomes, even given observed covariates X. A classic example is Cornfield et al.'s (1959) analysis of the causal effects of smoking on lung cancer. The advanced course considers an especially transparent situation described by Rosenbaum and Rubin (1983b), where we can analytically see how conclusions change a U's relationships with assignment and potential outcomes change. Extensions in Rosenbaum (2002) are briefly considered in the advanced course but are beyond the scope of the introductory versions.

Bounds on point estimates can be viewed as the extreme estimates that might be obtained over extreme distributions of the unobserved covariates U. Although often very broad, such bounds can play an important role in informing us about the sources of sharpness in inferences. Some relevant references for this approach include Manski et al. (1992), Manski and Nagin (1998), and Horowitz and Manski (2000). The idea of bounds is easily conveyed to students at all levels.

Related techniques for assessing nonignorable designs receive some attention in the advanced level course but are beyond the scope of the introductory course. These include the formal role of a second control group (Rosenbaum, 1987) and tests of unconfoundedness (Rosenbaum, 1984). Generally, the flavor of all versions of the course is to try to use evidence to produce better estimates rather than to test assumptions. That is, if there is evidence in the data available to test an assumption, then there is evidence for how to generalize the questionable assumption, and thereby improve the estimation of causal effects. Consequently, only limited time, even in the advanced course, is spent on such tests of the unconfoundedness assumption.

Complications

The final part of all versions of the course deals with complications, such as dropout and missing data, or noncompliance and surrogate outcomes. Many of these are on the cutting edge of research, and so the specific topics and treatments not only vary across versions of the course but also change from year to year. The discussion here is only intended to give the reader a flavor of the topics that might be covered when this article was drafted.

Principal Stratification

Principal stratification (Frangakis & Rubin, 2002) refers to the situation where some sort of adjustment is to be made for an outcome variable that is "intermediate to" or "on the causal pathway" to the final outcome Y. This intermediate outcome is called D with corresponding potential outcomes $[D(0), D(1)]$ and observed value D_{obs}, where $D_{obs,\,i} = W_i\, D_i(1) + (1 - W_i)D_i(0)$.

A common mistake is to treat D_{obs} as if it is a covariate, which it is not, and do an analysis stratified by D_{obs}. This mistake was even made by Fisher, in his classic textbook *The Design of Experiments,* from the first edition in 1935 to the last. The correct procedure is to stratify on the joint values $[D(1), D(0)]$, which are unaffected by W and so can be treated as a vector covariate. Thus, stratifying by $[D(1), D(0)]$ is legitimate and is called "principal stratification."

There are many examples of this general principle, the most common being noncompliance with assigned treatment. This is an important topic to bring to students' attention because it is the bridge to the economists' tool of instrumental variables, as well as an important introduction to more complex examples of principal stratification.

Noncompliance, CACE, and IV Estimation

At all levels, this topic is introduced using the example of Sommer and Zeger (1991) and the perspective of Angirst, Imbens, and Rubin (1996). This is a very large randomized experiment assessing the effect of vitamin A supplements on infant mortality in Indonesia, where vitamin A is only available to those assigned to take it. There are two principal strata here, defined by whether or not the units would take vitamin A if assigned it: compliers and noncompliers. The strata are observed for the units assigned to take vitamin A, but the principal strata are not observed for the units assigned to control. The objective is to estimate the causal effect of taking vitamin A within each principal stratum, that is, for compliers (the Complier Average Causal Effect [CACE]) and noncompliers (the Noncomplier Average Causal Effect [NACE]). We can estimate the overall average causal effect of assignment on $\overline{Y(1)} - \overline{Y(0)}$ using the usual estimate, $\bar{y}_1 - \bar{y}_0$. In addition, we can estimate the proportion of compliers by looking in the random half assigned to take vitamin A, say p_c. Thus, for these simple estimates we have

$$\bar{y}_1 - \bar{y}_0 = p_c \cdot \text{cace} + (1 - p_c) \cdot \text{nace},$$

where "cace" is the estimated causal effect of assignment on Y for compliers (the estimated CACE), and "nace" is the estimated causal effect of assignment on Y for noncompliers. Suppose we assume that there is no effect of being assigned to take vitamin A on mortality for those who would not take it whether or not they are assigned to take it: the exclusion restriction. Then nace = 0, and

$$\text{cace} = (\bar{\text{y}}_1 - \bar{\text{y}}_0)/p_c,$$

which is the instrumental variables estimate (IVE) from economics (see Haavelmo, 1944; Tinbergen, 1930 for background).

The IVE is the estimated intention-to-treat effect on Y divided by the difference in the proportion who receive the new treatment in the new treatment and control groups. The IVE has been reinvented several times despite its old history (e.g., Bloom, 1984; Zelen, 1979). More recent highly relevant work includes Baker (1998) and Baker and Lindeman (1994).

Bayesian and Likelihood Analysis With Noncompliance

Although the IV (instrumental variables) approach to noncompliance is very effective at revealing how simple assumptions can be used to address noncompliance, the associated IVE is a "method of moments" estimate, which is generally relatively inefficient. The statistically more generally principled approach is based on likelihood and Bayesian principles where the unknown compliance status is explicitly treated as missing data, as in Imbens and Rubin (1997). This treatment requires the application of iterative maximization or iterative simulation techniques, and thus is well beyond the scope of an introductory course. Such techniques are, however, important components of an advanced course.

More explicitly, when treating true compliance status as missing data within the likelihood framework, the EM algorithm (Dempster, Laird, & Rubin, 1977) or its extensions, such as ECM (Meng & Rubin, 1993), ECME (Liu & Rubin, 1994), AECM (Meng & van Dyk, 1997), and PXEM (Liu, Rubin, & Wu, 1998) can be used to find maximum likelihood estimates of the CACE; the associated confidence intervals can be based on large sample likelihood theory. An even more revealing approach is to use the fully Bayesian paradigm and the stochastic version of EM, the Data Augmentation algorithm, DA (Tanner & Wong, 1987), or its MCMC extensions (e.g., Gelman et al., 2003). A particular advantage of the Bayesian approach is the freedom to relax the exclusion restriction. The EM and DA approaches are developed in Imbens and Rubin (1997) and applied to the example used by Sommer and Zeger (1991). Another nice example is Little and Yau (1998). Also see Ettner (1996) for a more traditional approach from the econometric perspective and Goetghebeur and Molenberghs (1996) for related work.

An even more developed example of the fully Bayesian approach is that of Hirano et al. (2000), where extra encouragement to receive flu shots is randomly assigned, but many patients do not comply, thereby creating both never-takers and always-takers in addition to compliers. This example shows that the Bayesian approach can reveal evidence against the exclusion restriction in one group. This whole area is in rapid development and is a great source of research topics for graduate students (see, e.g., Barnard et al., 2002).

Combining Propensity Scores, Covariate Modeling, and Principal Stratification Modeling

A very important message to convey is that the propensity scores, covariate modeling, and principal stratification modeling approaches are not in competition, but are addressing different aspects of the problem of causal inference and can, and often should be, used in concert. This message is important at all levels of the course.

That is, generally in observational studies, it is most appropriate to use propensity score methods, matching, and subclassification to help design an observational study so that there is reasonable overlap in the distribution of observed covariates in the treatment and control groups. Then regression modeling of the covariates can be applied to the matched or subclassified groups to adjust for any remaining bias and increase efficiency of estimation. At this point, observed covariates should be well balanced or we cannot proceed with statistically defensible causal inferences.

The issue of unobserved covariates remains, and in some cases, before turning to sensitivity analysis or bounds to investigate how answers might change, principal stratification analysis can be appropriate. For example, suppose that in a particular propensity subclass, the treated and control groups are well balanced on the observed covariates, so that we can use the template for a randomized experiment there. Suppose in addition, however, that the treatment we really care about is not the one that has been used to define the well balanced groups but rather an exposure (or not) that can be viewed as a result of compliance or not with the assigned treatment. Then principal stratification techniques can be applied within the propensity controlled groups. That is, propensity matching and subclassification should be used, within the larger template of a randomized experiment with noncompliance, to create "assigned treatment" and "assigned control" groups that are well balanced on all covariates. The covariate modeling and principal stratification modeling should be applied to estimate the effects of "received treatment" for the subgroup of true compliers. This topic is at the cutting edge of current research, and thus can only be developed in the advanced versions of the courses.

Other Complications

In many ways, the material presented thus far is more than can be covered in one course, even in an advanced graduate course. Arguably, the material through the Observational Studies—Basic Approach section is enough for an introductory course, even when most theory is omitted, and enough for an advanced course if the theory and mathematical development related to propensity methods, matched sampling, and covariate modeling are included.

Then a second graduate level course could begin with the larger template of noncompliance and fully develop it and further extensions and complications. For example, the problem of missing data, both in covariates and outcomes, is very common. Standard methods (e.g., as in Little & Rubin, 2002; Rubin, 1987) need to be taught, and special methods, for instance, for dealing with missing covariates in propensity score analyses (D'Agostino & Rubin, 1999), also should be taught. Outcomes that are censored (e.g., survival data) can be viewed as a special but very important case of missing or coarsened data (Heitjan & Rubin, 1991). Moreover, dealing with combined complications, such as missing outcomes with noncompliance (Frangakis & Rubin, 1999), should receive attention, as should clustering in design issues (Frangakis, Rubin, & Zhou, 2002).

Extensions to multilevel treatments open a whole new collection of issues, even in classical randomized experiments, where the entire area of classical experimental design is essentially devoted to multilevel treatments (e.g., latin squares, split plots, repeated measures, and incomplete block designs). This also opens the door to longitudinal data and sequentially randomized experiments and other specialized techniques (Robins, 1997).

These extensions are important, especially for active researchers in areas of applications. But all such techniques are based on the foundations developed in the "first course," outlined in the first four sections of this article. Moreover, these foundations must be firmly grasped before complications and extensions are considered.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 434, [as Applications Invited Discussion Article with discussion and rejoinder, 444–472].

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics, 67,* 648–660.

Ashenfelter, O., & Krueger, A. (1993). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review, 84,* 1157–1173.

Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association, 93,* 929–934.

Baker, S. G., & Lindeman, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine, 13,* 2269–2278.

Barnard, J., Frangakis, C., Hill, J., & Rubin. D. B. (2002). School choice in NY City: A Bayesian analysis of an imperfect randomized experiment. In C. Gatsonis, B. Carlin, & A. Carriquiry (Eds.), *Case studies in Bayesian statistics: Vol. 5* [With discussion and rejoinder] (pp. 3–97). New York: Springer.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8,* 225–246.

Cochran, W. G. (1963). *Sampling techniques* (2nd ed.). New York: John Wiley.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 295–313.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya—A, 35,* 417–446.

Cornfield, J., Haenszel, E., Hammond, C., Lilienfeld, A., Shimkin, M. D., & Wynder, E. L. (1959). Smoking and lung cancer; recent evidence and a discussion of some questions. *Journal of the National Cancer Institute, 22,* 173–203.

Cox, D. R. (1958). *Planning of experiments.* New York: John Wiley.

Cox, D. R. (1992). Causality: Some statistical aspects. *Journal of the Royal Statistical Society, Series A, 155,* 291–301.

D'Agostino, R., Jr., & Rubin, D. B. (1999). Estimation and use of propensity scores with incomplete data. *Journal of the American Statistical Association, 95,* 749–759.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 407–424 [With discussion].

Dehejia, R. H., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Dempster, A. P. (1990). Causality and statistics. *Journal of Statistical Planning and Inference, 25,* 261–278.

Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38. [With discussion and reply].

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

Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association, 86,* 9–17.

Ettner, S. L. (1996). The timing of preventive services for women and children: The effect of having a usual source of care. *American Journal of Public Health, 86,* 1748–1754.

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

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Boyd.

Frangakis, C., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika, 86,* 366–379.

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

Frangakis, C., Rubin, D. B., & Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. [With discussion and rejoinder]. *Biostatistics, 3,* 147–177.

Gelman, A., & King, G. (1991). Estimating incumbency advantage without bias. *American Journal of Political Science, 34,* 1142–1164.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Goetghebeur, E., & Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association, 91,* 928–934.

Greenland, S., & Poole, C. (1988). Invariants and noninvariants in the concept of interdependent effects. *Scandinavian Journal of Work and Environmental Health, 14,* 125–129.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science, 14,* 29–46.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica, 15,* 413–419.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153–161.

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

Heitjan, D., & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics, 19,* 2244–2253.

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

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–970.

Holland, P. W. (1988a). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology,* 449–484.

Holland, P. W. (1988b). Comment on "Employment discrimination and statistical science" by A. P. Dempster. *Statistical Science, 3,* 186–188.

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 and V. Hotz. *Journal of the American Statistical Association, 84,* 875–877.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In Wainer & Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 3–25). Hillsdale, NJ: Erlbaum.

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariates and outcome data. *Journal of the American Statistical Association, 95,* 77–84.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association, 47,* 663–685.

Imbens, G., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics, 25,* 305–327.

Imbens, G. B., Rubin, D. B., & Sacerdote, B. (1999). Estimating the effect of unearned income on labor supply, earnings, savings and consumption: Evidence from a survey of lottery players. Tilberg University, Center for Economic Research, Discussion Paper 9934. Also, National Bureau of Economic Research Working Paper 7001.

Kadane, J. B., & Seidenfeld, T. (1990). Randomization in a Bayesian perspective. *Journal of Statistical Planning and Inference, 25,* 329–346.

Kempthorne, O. (1976). Comments on "On rereading R. A. Fisher." *Annals of Statistics, 4,* 495–497.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review, 76,* 604–620.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.

Little, R. J., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 147–159.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

Liu, C. H., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika, 81,* 633–648.

Liu, C. H., Rubin, D. B., & Wu, Y. N. (1998). Parameter expansion for EM acceleration: The PX-EM algorithm. *Biometrika, 85,* 755–770.

Manski, C. F., Sandefur, G. D., McLanahan, S., & Powers, D. (1992). Alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association, 87,* 25–37.

Manski, C. F., & Nagin, D. S. (1998). Bounding disagreements about treatment effects: A study of sentencing and recidivism. *Sociological Methodology, 28,* 99–137.

Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm. A general framework. *Biometrika, 80,* 267–278.

Meng, X. L., & van Dyk, D. A. (1997). The EM algorithm—An old folk song sung to a fast new tune [with discussion]. *Journal of the Royal Statistical Society, Series B, 59,* 511–567.

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

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, Series A, 97,* 558–606.

Pratt, J. W., & Schlaifer, R. (1984). On the nature and discovery of structure. *Journal of the American Statistical Association, 79,* 9–33.

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

Reinisch, J., Sanders, S., & Mortensen, E. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *The Journal of the American Medical Association, 274,* 1518–1525.

Robins, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases 40*(Suppl. 2), 139S–161S.

Robins, J. M. (1989). The control of confounding by intermediate variables. *Statistics in Medicine, 8,* 679–701.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality: Lecture notes in statistics, Vol. 120* (pp. 69–117). New York: Springer.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study [with Discussion]. *Statistical Science, 2,* 292–316.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York: Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *The Journal of the Royal Statistical Society, Series B, 45,* 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1984a). Estimating the effects caused by treatments. Discussion of "On the nature and discovery of structure" by Pratt and Schlaifer. *Journal of the American Statistical Association, 79,* 26–28.

Rosenbaum, P. R., & Rubin, D. B. (1984b). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score. *The American Statistician, 39,* 33–38.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics, 29,* 159–183. [Printer's correction note 30, 728.]

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29,* 184–203.

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

Rubin, D. B. (1976a). Inference and missing data. *Biometrika, 63,* 581–592. [With discussion and reply.]

Rubin, D. B. (1976b). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics, 32,* 109–120. [Printer's correction note p. 955.]

Rubin, D. B. (1976c). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics, 32,* 121–132. [Printer's correction note p. 955.]

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26. [Printer's correction note 3, p. 384.]

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

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *The Journal of the American Statistical Association, 74,* 318–328.

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *The Journal of the American Statistical Association, 75,* 591–593.

Rubin, D. B. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In Rao & Sedransk (Eds.), *W. G. Cochran's impact on statistics* (pp. 37–69). New York: John Wiley.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. In Bernardo, DeGroot, Lindley, & Smith (Eds.), *Bayesian statistics, Vol. 2* (pp. 463–472). North Holland.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley.

Rubin, D. B. (1990a). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5,* 472–480.

Rubin, D. B. (1990b). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference, 25,* 279–292.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127,* 757–763.

Rubin, D. B. (2002). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2,* 169–188.

Rubin, D. B., & Thomas, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics, 20,* 1079–1093.

Rubin, D. B., & Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal covariates. *Biometrika, 79,* 797–809.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology, 27,* 325–353.

Smith, T. M. F., & Sugden, R. A. (1988). Sampling and assignment mechanisms in experiments, surveys and observational studies. *International Statistical Review, 56,* 165–180.

Sobel, M. E. (1990). Effect analysis and causation in linear structural equation models. *Psychometrika, 55,* 495–515.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In Arminger, Clogg, & Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York: Plenum.

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods & Research, 24,* 353–379.

Sommer, A., & Zeger, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine, 10,* 45–52.

Sugden, R. A. (1988). The 2 × 2 table in observational studies. In Bernardo, DeGroot, Lindley, & Smith (Eds.), *Bayesian statistics 3* (pp. 785–790). New York: Oxford University Press.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82,* 543–546.

Tinbergen, J. (1930). Determination and interpretation of supply curves: An example. *Zeitschrift fur Nationalokonomie.* Reprinted in: *The foundations of economics,* Henry & Morgan (Eds.).

Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statistical Science, 4,* 298–340. [With discussion and rejoinder.]

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist, 54,* 594–604.

Zelen, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine, 300,* 1242–1245.

Author

DONALD B. RUBIN is John L. Loeb Professor of Statistics, Department of Statistics, Harvard University, 1 Oxford Street, Science Center, Cambridge, MA 02138; rubin@stat.harvard.edu. His areas of specialization are causal inference in experiments and observational studies; inference in sample surveys with nonresponse and in more general missing data problems; application of Bayesian and empirical Bayesian techniques; and developing and applying statistical models to data in a variety of scientific disciplines.

Manuscript received February 19, 2003
Revision received June 2, 2003
Accepted June 2, 2003