1

Causal Inference in Randomized and Non-Randomized Studies: The Definition, Identification, and Estimation of Causal Parameters

Michael E. Sobel

INTRODUCTION

The distinction between causation and association has figured prominently in science and philosophy for several hundred years at least, and, more recently, in statistical science as well, indeed, since Galton, Pearson and, Yule developed the theory of correlation.

Statisticians have pioneered two approaches to causal inference that have proven influential in the natural and behavioral sciences. The oldest dates back to Yule (1896), who wrote extensively about 'illusory' correlations, by which he meant correlations that should not be endowed with a causal interpretation. To distinguish between the illusory and non-illusory correlations, Yule invented partial correlation to 'control' for the influence of a common factor, arguing in context that because the relationship between pauperism and out relief did not vanish when 'controlling' for poverty, this relationship could be deemed causal. A half century later, philosophers, psychologists and social scientists (e.g., Reichenbach, 1956; Simon 1954; Suppes 1970) rediscovered Yule's approach to distinguishing between causal and non-causal relationships, and econometricians (e.g., Granger 1969) extended this idea to the time-series setting. Graphical models, path analysis and, more generally, structural

equation models, when these methods are used to make causal inferences, also rely on this type of reasoning. The theory of experimental design, which emerges in the 1920s and thereafter, and is associated especially with Neyman (1923) and Fisher (1925), forms the basis for a second approach to inferring causal relationships. Here, the use of good design, especially randomization, is emphasized, apparently obviating the need to worry about spurious relationships.

Despite these important contributions, during the majority of the twentieth century, most statisticians espoused the view that statistics had little to do with causation. But the situation has reversed dramatically in the last 30 years, since Rubin (1974, 1977, 1978, 1980) rediscovered Neyman's potential outcomes notation and extended the theory of experimental design to observational studies. Currently, there is a large and growing inference in statistics on the topic of causal inference and this second approach to inferring causal relationships is coming to dominate the first approach, even in disciplines such as economics, which rely on observational studies and where the first approach has traditionally dominated.

This chapter provides an introduction, tailored to the concerns of behavioral scientists, to this second approach to causal inference. Because causal inference is the act of making inferences about the causal relation and notions of the causal relation differ, it is important to understand what notion of causation is under consideration when such an inference is made. Thus, in the next section, I briefly review several notions of causation and also briefly examine the approach to causal inference that derives from Yule. In the section 'Unit and average causal effects' the second approach, which is built on the idea that a causal relation should sustain a counterfactual conditional statement, is introduced, and a number of estimands of interest are defined. The section 'Identification of causal parameters under ignorable treatment assignment' discusses the identification of causal effects and the next section, 'Estimation of causal parameters in randomized studies', discusses estimation. The section 'Mediation analyses' takes up the topic of mediation, which is of special interest to psychologists and prevention scientists. I show that the usual approach to mediation, which uses structural equation modeling, does not yield estimates of causal parameters, even in randomized studies. Several other approaches to mediation, including principal stratification and instrumental variables, are also considered.

CAUSATION AND PROBABILISTIC CAUSATION

Regularity theories of causation are concerned with the full (or philosophical) cause of an effect, by which is meant a set of conditions that is sufficient (or necessary or necessary and sufficient) for the effect to occur. This type of theory descends from Hume, who claimed that causation (as it exists in the real world) consists only of the following: (1) temporal priority, i.e., the cause must precede the effect in time; (2) spatiotemporal contiguity, i.e., the cause and effect are 'near' in time and space; and (3) constant conjunction, i.e., if the same circumstances are repeated, the same outcome will occur. Many subsequent writers argued that Hume's analysis cannot distinguish between regularities that are not causal, such as a relation between two events brought about by a common factor, and genuine causation. At the minimum, this suggests that Hume's account is incomplete. A number of philosophers [e.g., Bunge (1979) and Harré and Madden (1975)] argue that the causal relation is generative. While this idea is appealing, especially to modern scientists who speak of mechanisms, attempts to elaborate this idea have not been entirely successful. Another approach (examined later) is to require causal relationships to sustain counterfactual conditional statements.

Hume's analysis is also deterministic, and the literature on probabilistic causation that descends from Yule can be viewed as

an attempt to both relax this feature and distinguish between causal and non-causal regularities. The basic idea is as follows. First, there is a putative cause Z prior in some sense to an outcome Y. Further, Z and Y are associated (correlated). However, if the $Z - Y$ association vanishes when a (set of) variable(s) X prior to Z is conditioned on (or in some accounts, if such a set exists), this is taken to mean that Z 'does not cause' Y, that is, the relationship is 'spurious'. To complete the picture, various examples where this criterion would seem to work well have been constructed.

Granger causation and structural equation models use this type of reasoning to distinguish between empirical relationships that are regarded as causal and not causal. For example, consider a structural equation model, with X a vector of variables, Z an outcome occurring after X, and Y an outcome after Z. If the 'direct effect' of Y on Z is 0 (not 0), Z is not viewed (viewed) as a cause of Y. A sufficient condition for this direct effect to be 0 is that Y and Z are conditionally independent, given X. Of course, this same kind of reasoning can be extended to the consideration of other types of direct and indirect effects. For example, consider the case of a variable X associated with Z, with X and Y conditionally independent, given Z, implying the 'direct effect of X on Y is 0, and Z and Y not conditionally independent given X, implying the 'direct effect' of Z on Y is non-zero. Here there is an effect of X on Y, but the effect is indirect, through Z.

There are a number of problems with this approach. First and foremost, it confounds causation with the act of inferring causation, as evidenced by the fact that the criteria above for inferring causation are typically put forth independently of any explicit notion of the causal relation. As notions of the causal relation vary, this method of inferring causation may be appropriate for some notions of causation, for example, the case where causation is regarded as a predictive relationship among variables, but not for others. Nor (because the nature of the casual relation is not explicitly considered), is it clear whether causation is viewed as probabilistic in some inherent sense or if probability arises in some other way. The deficiencies of this approach are evident in the psychological literature on casual modeling, where a variety of extra-mathematical considerations (such as model specification) are used to suggest that model coefficients can be endowed with a causal interpretation (see, for example, Sobel 1995 on this point).

By way of contrast to regularity theories, manipulability theories view causes as variables that can be manipulated, with the outcome depending on the state of the manipulated variable. Here, as opposed to specifying all the variables and the functional relationship between these and the outcome (which would constitute a successful causal account of a phenomenon under a regularity theory), the goal is more modest, to examine the 'effect' of a particular variable. See Sobel (1995) for a reconciliation of these two approaches.

Manipulability theories require the causal relation to sustain a counterfactual conditional statement (e.g., eating the poison caused John to die means that John ate the poison and died, but had John not eaten the poison, he would not have died). This is closer to the way an experimentalist thinks of causation. However, many philosophers regard manipulability theories as anthropomorphic. Further, many questions that scientists ask are not amenable to experimentation, e.g, the effect of education on longevity or the effect of marriage on happiness. This would appear to seriously limit the value of this approach for addressing real scientific questions.

However, even without manipulating a person's level of education, one might imagine that had this person's level of education taken on a different value than that actually realized, this person might also have a different outcome. This suggests adopting the broader view that it is not the manipulation *per se*, but the idea that the causal variable could take on a different value than it actually takes on, which is key. This is the idea underlying counterfactual theories of

causation (e.g., Lewis 1973), where the closest possible world to the one we live in serves as the basis for the counterfactual.

Counterfactual theories also have their difficulties, both theoretical and practical. One criticism goes under the rubric of 'preemption'. Person A shoots person C in the head, and C dies. It seems natural to claim that person A caused person C to die. Yet, suppose that if A had not shot C, B would have done so and C would also have died. In that case, C dies whether or not A shoots him, so one cannot (under a simple counterfactual theory) say A caused C to die. This seems wrong.

In practice, the outcome may also depend on the way in which the cause is brought about. When an experiment is performed, this issue is not, *per se*, problematic, and the effect corresponding to that manipulation is well defined. Otherwise, as there may be different outcomes, 'the effect' is ill-defined, unless the closest world is specified; in some instances, this will be a very difficult task. This suggests that some questions, e.g., the effect of marriage on happiness, may be better left unasked (or at the minimum, one must specify the hypothetical intervention by which persons are exposed/not exposed to marriage and to marriage partners).

I now turn to the recent statistical literature on causal inference, which is also based on the idea that causal relations sustain counterfactual conditionals. This approach to casual inference (as suggested by the preceding material) is not concerned with elucidating the various causes of the outcome (effect) and the way in which these causes produce the effect, but with the more limited goal of inferring the effect (in a sense to be described) of a particular causal variable. Scientists who are interested in a fuller accounting of the causes of an effect and the pathways through which the effect is produced may find this approach less than entirely satisfying. However, as discussed subsequently, this approach can also be used to evaluate methods (such as structural equation models) that researchers sometimes use to provide a fuller account, and it is not hard to show that these methods rest on a number of implausible assumptions.

UNIT AND AVERAGE CAUSAL EFFECTS

The notion of causation congruous to recent statistical work on causal inference has two important properties. First, the causal relation is singular, i.e., it is meaningful to speak of effects at the individual level and these effects may vary over individuals (heterogeneity). Second, causal statements sustain counterfactual conditionals. Thus, we might state that attending health class caused Bill to drink less, by which we mean that Bill went to health class and later drank amount y, whereas had he not attended health class, later he would have drunk $y^* > y$. For Mary, perhaps the outcome is the same whether or not she attends, in which case we would say that attending health class did not cause Mary to drink less (or more). Note that only attending health class is considered as the cause. Other possible causes, e.g., sex, are regarded as part of the causal background (pretreatment covariates in statistical language).

The single most important contribution in this literature is the potential outcomes notation developed by Neyman (1923) and Rubin (1974) to express the ideas above. Using this notation allows causal effects to be defined independently of the association parameters that are actually estimated in studies; one can then ask whether and under what conditions these associations equal the causal effects. Consider the case of an experiment where unit i in a population $\mathcal{P}$ is assigned (or not) to receive a treatment. The data for this unit is typically written as: $(Z_i, Y_i, \underline{X}_i)$, where $Z_i = 1$ if i is assigned to receive the treatment, 0 otherwise, Y_i is the value of the outcome and $\underline{X}_i$ is a vector of covariates. Although this representation is adequate for descriptive modeling [e.g., the regression function $E(Y \mid Z, \underline{X})$], it does not adequately express the idea that Z might take on different values and that i's outcome might vary with this. One way to formalize this idea is to consider two outcomes for i,

$Y_{zi}(0)$, the outcome i would have if he is not assigned to receive treatment and $Y_{zi}(1)$, the outcome i would have if he is assigned to receive treatment. With this notation, it is then straightforward to define singular causal effects (unit effects) $h(Y_{zi}(0), Y_{zi}(1))$ [where there is no effect if $Y_{zi}(0) = Y_{zi}(1)$]; the unit effects then serve as the building blocks for various types of average effects.

Were it possible to take a sample from $(Y_z(0), Y_z(1))$, it would be a simple matter to obtain the unit effects $h(Y_{zi}(0), Y_{zi}(1))$ for the sampled units and to estimate various parameters that are functions of these. The literature on causal inference arises from the fact that it is only possible to observe one of the potential outcomes.

A limitation of the notation above is that only a scalar outcome and a binary treatment are considered. Because the generalization to a random object and to arbitrary types of treatments is trivial and does not generate substantially new issues, I continue to treat the case of a binary treatment and scalar outcome. A more serious limitation (though it is again not difficult to generalize this notation) that does generate new issues is that the notation above does not allow for interference (Cox, 1958), that is, i's potential outcomes are not allowed to depend on the treatment received by other units. Rubin (1980) calls this the stable unit treatment value assumption (SUTVA). Although this assumption is often reasonable and almost universally made, there are many instances in the social and behavioral sciences where it is untenable. For example, in schools, children in the same (or even different) classrooms may interfere with one another. This case has been studied by Gitelman (2005); for the more general case, see Halloran and Struchiner (1995) and Sobel (2006a, 2006b). Hereafter, I shall assume SUTVA holds.

Although the unit effects above cannot be determined (since only one of the potential outcomes is ever observed), it turns out remarkably that under suitable conditions (discussed later) various types of averages of these effects are nevertheless identifiable and can be consistently estimated.

As a simple example, consider the case above, with $h(Y_{zi}(0), Y_{zi}(1)) = Y_{zi}(1) - Y_{zi}(0)$. The 'intent to treat' estimand (hereafter ITT), which is commonly featured in connection with randomized clinical trials, is defined as:

$$E(Y_z(1) - Y_z(0)), \tag{1}$$

the average of the unobserved unit effects. Because the expected value is a linear operator, the ITT can also be expressed as $E(Y_z(1)) - E(Y_z(0))$. Thus, if it is possible to take a random sample of size n from $\mathcal{P}$ and then take random sub-samples from $Y_z(1)$ and $Y_z(0)$, the difference between the sample averages:

$$\frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n} (1 - Z_i) Y_i}{\sum_{i=1}^{n} (1 - Z_i)} \tag{2}$$

is an unbiased and consistent estimator of the ITT.

The ITT is one of a number of possible parameters of interest and may not always be of greatest scientific or policy relevance. It measures the effect of treatment assignment, and as subjects may not always take up the treatments to which they are assigned, the ITT does not measure the effect of treatment itself. A policy maker might nevertheless argue that the ITT is of primary interest because it measures the effect that would be actually be observed in the real world. As an example, consider the effect of a universally free school-breakfast program vs. the current Federal program (Crepinsek et al., 2006) on total food and nutrient intake. Some students will take up the free breakfast, others will not. From the policy maker's perspective, if the program is highly effective amongst those who take it up, but the takers are a small percentage of those who might benefit, the program may be judged a failure.

In observational studies where treatments are not assigned, one observes only whether or not a subject takes up a treatment ($D = 1$) or not ($D = 0$); defining the potential outcomes as $Y_{di}(0)$ and $Y_{di}(1)$, interest often centers on

the average treatment effect (hereafter ATE):

$$E(Y_d(1) - Y_d(0)) \tag{3}$$

or the effect of treatment on the treated (ATT):

$$E(Y_d(1) - Y_d(0) \mid D = 1). \tag{4}$$

Neyman (1923) first considered the ATE. The ATT was first considered by Belsen (1956) and discussed in detail by Rubin (1978). The ATE measures the average effect if all persons in the population are given the treatment, whereas the ATT measures the average effect of the treatment in the subpopulation that takes up the treatment. The ATE is a natural parameter of interest if the treatment corresponds to a policy that is under consideration for universal and mandatory adoption. In cases where adoption is voluntary, some economists have argued that only the ATT is relevant, because it reflects what would actually occur if the policy were to be implemented. However, one might also want to know if those who do not adopt the policy would benefit, because an affirmative answer might suggest to a policy maker that efforts focus on increasing the take up rate. Additionally, if the ATT is positive, and persons who have not take up the policy have access to this type of information, they might be more motivated to do so. This suggests that in general, for policy purposes, one might wish to know, in addition to the ATT, the average effect of treatment on the untreated (ATU) and the ATE, which is a weighted average of the ATT and ATU. The ATE will also be a more natural parameter of interest than the ATT in many contexts where the focus is on the basic science, where the causal variable may not be one that can be manipulated for policy purposes.

Various other parameters may also be of interest. First, for both scientific and policy reasons, one often wants to know whether the effects above vary in different sub-populations defined by characteristics of the units. Let $\underline{X}$ denote a vector of variables that are not affected by the treatment (or assignment variable), for example, take $\underline{X}$ to be a vector of pretreatment covariates. This leads to consideration of the parameters ITT($\underline{X}$), ATE($\underline{X}$), and ATT($\underline{X}$), where, for example, ATE($\underline{X}$) is defined as $\mathrm{E}(Y_d(1) - Y_d(0) \mid \underline{X})$, and the other parameters are defined analogously.

Although attention herein focuses on the parameters above, a number of other interesting and/or useful parameters have been defined and considered. Bjőrklund and Moffitt (1987) defined (and discussed the economic relevance of) the marginal treatment effect for subjects indifferent between participating or not in a program of interest. Quantile treatment effects, the difference between the marginal quantiles of $Y(1)$ and $Y(0)$, were defined by Doksum (1974) and Lehmann (1974); these effects have received some attention recently (e.g., Abadie, Angrist and Imbens, 2002). A parameter (discussed subsequently) that has received much attention lately is the local average treatment effect (LATE) considered by Angrist, Imbens and Rubin (1996).

Many other parameters might also be considered. A decision maker might wish to consider the utilities of the potential outcome values and ask whether a treatment increases average utility or some other measure of social welfare; this is a matter of considering $U(Y)$ as opposed to Y.

The average effects above take as building blocks the unit differences $Y_{di}(1) - Y_{di}(0)$ (or $Y_{zi}(1) - Y_{zi}(0)$). As will be evident later, because averages of these depend only upon the marginal distributions of $Y_d(0)$ and $Y_d(1)$, the effects in question are identified if the marginal distributions are identified. Parameters that depend on the joint distribution of the potential outcomes may also be defined (e.g., the proportion who would benefit from treatment), but the data, even from a randomized experiment, typically contain little or no information about this joint distribution, so these parameters will not be identifiable without introducing additional assumptions. While this may appear to be a serious limitation, several comments are in order. First, sometimes a transformation may produce an estimand of the desired form.

For example, for positive variables, with $h(Y_d(0), Y_d(1)) = Y_d(1)/Y_d(0)$, redefining the potential outcomes as $\log Y_d(0)$ and $\log Y_d(1)$ gives transformed effects in the desired form. Second, in decision making, the additional information contained in the joint distribution may be irrelevant (Imbens and Rubin, 1997) to the policy maker. Third, at least occasionally, plausible substantive assumptions could lead to identification of the joint distribution of the joint distribution. For example, let the outcome be death (1 if alive, 0 if dead) and suppose one wants to know the proportion benefiting under treatment; it is easy to see that if treatment is at least not harmful, the joint distribution is identified from the marginal distributions of the potential outcomes. However, in general, this is not the case, and as there is likely to be very little scientific knowledge about quantities like the joint distribution of potential outcomes in a study in which the marginal distribution is not even assumed to be known, the identification of parameters involving the joint distribution will typically require making assumptions that are substantively heroic (although perhaps mathematically convenient) and possibly quite sensitive to violations.

I now consider the identification of causal parameters.

IDENTIFICATION OF CAUSAL PARAMETERS UNDER IGNORABLE TREATMENT ASSIGNMENT

Random assignment is an assumption about the way units are assigned to treatments. At the heart of randomized experiments, this assumption enables identification of causal parameters. In the simplest case where each subject is assigned with probability $0 < \pi < 1$ to the control condition and probability $1-\pi$ to the treatment condition, random assignment implies treatment assignment is "ignorable", i.e., Z is independent of background covariates and potential outcomes:

$$Z \perp\!\!\!\perp \underline{X},\ Y_z(0),\ Y_z(1) \qquad (5)$$

To see how (5) is used for identification, note that whether or not randomization is assumed to assign subjects to treatments, what can actually be observed is a sample from the joint distribution of $(Y, Z, \underline{X})$. From this distribution, the conditional distributions $Y \mid Z = z$ for $z = 0, 1$ are identified, and as $Y = ZY_z(1)+(1-Z)Y_z(0)$, the distribution $Y \mid Z = z$ is the distribution $Y_z \mid Z = z$. Thus, the population means $E(Y \mid Z = 1) = E(Y_z(1) \mid Z = 1)$ and $E(Y \mid Z = 0) = E(Y_z(0) \mid Z = 0)$ are identifiable, so the difference:

$$E(Y \mid Z = 1) = E(Y_z(1) \mid Z = 1) - E(Y \mid Z = 0) = E(Y_z(0) \mid Z = 0) \qquad (6)$$

is also identified. In general, (6) does not equal (1) because the identified conditional distributions $Y_z \mid Z = z$, $z = 0, 1$, are not equal to the corresponding marginal distributions Y_z, $z = 0, 1$. But under the random assignment assumption, (5) holds, implying equality of the two sets of distributions; hence (6) = (1).

Often an investigator will also want to know if the value of the ITT depends on covariates of interest. The parameter of interest is then ITT $(\underline{X}) = \mathrm{E}(\mathrm{Y}_z(1) - \mathrm{Y}_z(0) \mid \underline{X})$. As (for the case above):

$$0 < \pi_z(\underline{X}) \equiv \Pr(Z = 1 \mid \underline{X}) < 1, \qquad (7)$$

and since assumption (5) implies:

$$Z \perp\!\!\!\perp Y_z(0),\ Y_z(1) \mid \underline{X}, \qquad (8)$$

that is, assignment is random within levels of $\underline{X}$, ITT$(\underline{X})$ is identifiable and equal to:

$$E(Y \mid Z = 1, \underline{X}) - E(Y_z(0) \mid Z = 0, \underline{X}). \qquad (9)$$

Just as the assumption of random assignment is the key to identifying causal parameters from the randomized experiment, the assumption of random assignment within blocks (sub-populations) is the key to identifying causal parameters from the randomized block experiment. It is also the key to making causal inferences from observational studies.

When assumptions (7) and (8) hold, treatment assignment is said to be strongly ignorable, given $\underline{X}$ (Rosenbaum and Rubin, 1983). In observational studies in the social and behavioral sciences, where subjects choose the treatment received (D), the assumption:

$$D \perp\!\!\!\perp \underline{X}, Y_d(0), Y_d(1) \tag{10}$$

[akin to (5)] is likely to be unreasonable, as typically evidenced by differences in the distribution of covariates in the treatment and control groups. However, if the investigator knows (as in a randomized block experiment) the covariates that account for the differential assignment of subjects into the treatment and control groups:

$$D \perp\!\!\!\perp Y_d(0), Y_d(1) \mid \underline{X}, \tag{11}$$

and if:

$$0 < \pi_d(\underline{X}) \equiv \Pr(D = 1 \mid \underline{X}) < 1, \tag{12}$$

for all $\underline{X}$, i.e., treatment received D is strongly ignorable given $\underline{X}$, the parameter ATE(X):

$$E(Y_d(1) - Y_d(0) \mid \underline{X}) \tag{13}$$

is identified and equals $E(Y \mid \underline{X}, D = 1) - E(Y \mid \underline{X}, D = 0)$. It also follows that ATE($\underline{X}$) = ATT($\underline{X}$):

$$E(Y_d(1) - Y_d(0) \mid \underline{X}, D = 1). \tag{14}$$

Typically, the investigator will be interested not only in ATE($\underline{X}$) and/or ATT($\underline{X}$), but also in ATE = $E_{\mathcal{P}}$(ATE($\underline{X}$)) and ATT = $E_{\mathcal{P}^*}$(ATT($\underline{X}$)), where $\mathcal{P}^*$ is the sub-population of units that receive treatment. Note that ATT $\neq$ ATE because these parameters are obtained by averaging over different units: the ATE is a weighted average of the ATT and the ATU.

More generally, as at the beginning of this section, under the types of ignorability assumptions above, it is possible to identify the marginal and conditional (given $\underline{X}$) distributions of the potential outcomes in $\mathcal{P}$ and to therefore consider any causal estimand that can be defined in terms of these distributions. I do not consider this matter further here, save to note that estimating these distributions will (when the outcome Y is metrical) typically be more difficult than estimating the average causal effects above.

Second, the average effects above can be identified under weaker ignorability assumptions than those given here. For example, ATE(X) and ATT(X) are identified under the marginal ignorability assumption:

$$D \perp\!\!\!\perp Y_d(d) \mid \underline{X} \tag{15}$$

for $d = 0, 1$, and occasionally using this weaker assumption is advantageous. It also obvious that for estimating means, ignorability assumptions can be replaced by the weaker condition of so-called 'mean independence', e.g., $E(Y_d \mid \underline{X}, D = d) = E(Y_d \mid \underline{X})$. However, it is difficult to think of situations where mean independence holds and ignorability does not. Additionally, mean independence does not hold for functions of Y, such as $U(Y)$, the utility of Y. Thus, I do not consider this further.

In observational studies, it will often be the case that an investigator is not sure if he/she has measured all the covariates $\underline{X}$ predictive of both the treatment and the outcome. Not surprisingly (as it is not possible to observe both potential outcomes), ignorability assumptions are not, *per se*, testable; attempts to assess such assumptions invariably rely on various types of auxiliary assumptions (Rosenbaum, 1987; Rosenbaum, 2002).

When an investigator believes there are variables he/she has not measured that predict both the treatment and the potential outcomes, it is nevertheless sometimes possible (using other types of assumptions) to estimate the parameters above (or parameters similar to these). The section 'Mediation analyses' examines the use of this approach in the context of mediation. Another approach that has been used involves the use of fixed effects models (and differences in differences) to remove the effects of unmeasured variables. When the investigator knows the treatment assignment rule, but the assignment probabilities are 0 and 1, as is the case in risk

based allocation (Thistlethwaite and Campbell, 1960), causal inferences necessarily rely on extrapolation. Nevertheless, in some cases, reasonable inferences can be made (Finkelstein, Levin and Robbins, 1996).

Other approaches in the absence of ignorability include bounding causal effects (Manski, 1995; Robins, 1989). Bounds that make few assumptions are often quite wide and not especially useful. Nevertheless, when assumptions leading to tighter bounds are credible, this approach may be quite helpful. Additionally, sensitivity analyses can also be very useful; if ignorability is violated due to an unmeasured covariate, but the results are robust to this violation, credible inferences can nevertheless be made (see Rosenbaum, 2002, for further material on this topic).

ESTIMATION OF CAUSAL PARAMETERS IN RANDOMIZED STUDIES

I consider ITT($\underline{X}$) and ITT in this section, using these cases to introduce the primary ideas underlying the estimation of causal effects in the simplest setting. The discussion is organized around two broad approaches: (1) using potential outcomes imputed by regression or some other method (e.g., matching) and using the observed and imputed outcomes to estimate ITT; and (2) reweighting the data in the treatment and control groups to reflect the composition of the population $\mathcal{P}$ (or an appropriate subpopulation thereof). The estimators considered under the first approach have been used in the experimental design literature for many years and will be familiar to most readers.

Estimation of ITT(X) and ITT in randomized studies

The simplest case, previously considered, estimates the ITT using (2) under the identification condition (5). It is also useful to note that (2) is also the coefficient $\hat{\tau}$ in the ordinary least squares regression of Y on Z:

$$Y_i = \alpha + \tau Z_i + \epsilon_i, \tag{16}$$

$i = 1, \ldots, n$, where the parameters are identified by the assumption $E(\epsilon) = 0$.

The estimator (2) also arises by predicting the missing outcomes $Y_{zi}(0)$ (if $Z_i = 1$) or $Y_{zi}(1)$ (if $Z_i = 0$) using the estimated minimum mean square error predictor $\widehat{E}(Y \mid Z)$. Let $\widehat{Y}_{zi}(0) = Y_{zi}(0)$ if $Z_i = 0$, and $\hat{\alpha} = \widehat{E}(Y \mid Z = 0)$ otherwise, $\widehat{Y}_{zi}(1) = Y_{zi}(1)$ if $Z_i = 1$, $\hat{\alpha}+\hat{\tau} = \widehat{E}(Y \mid Z = 1)$ otherwise; thus, (2) can also be written as:

$$n^{-1} \sum_{i=1}^{n} (\widehat{Y}_{zi}(1) - \widehat{Y}_{zi}(0)). \tag{17}$$

In the case of a randomized block experiment, where the probability of assignment to the treatment group depends on known covariates, assumption (5) will be violated, but if the covariates are unrelated to the potential outcomes:

$$Z \perp\!\!\!\perp Y_z(0), Y_z(1), \tag{18}$$

in which case (2) is still unbiased and consistent for ITT.

When the covariates are related to both treatment assignment and the potential outcomes, (8) provides the basis for extending the approach above. Let the covariates $\underline{X}$ take on L distinct values, corresponding to blocks $b = 1, \ldots, L$ and let $g(\underline{X}) \equiv B$ be the one to one onto function mapping $\underline{X}$ onto the blocking variable B. Within each block b, $n_{1b} = n_b \Pr(Z = 1 \mid B = b)$ of the n_b units are assigned to the treatment group, where $0 < \Pr(Z = 1 \mid B = b) < 1$ for all b. The matched pairs design is the special case where the sample size is $2n$, $L = n$, $n_{1b} = 1$, $n_{0b} = n_b - n_{1b} = 1$.

The regression corresponding to (16) is:

$$Y_i = \sum_{b=1}^{L} 1_{\{B_i\}}(b)(\alpha_b + \tau_b Z_i + \epsilon_i), \tag{19}$$

where $1_{\{B_i\}}(b) = 1$ if $\{B_i\} = b$, 0 otherwise, and $\mathrm{E}(\epsilon \mid \underline{X}, Z) = 0$. Thus, τ_b is the

value of ITT($\underline{X}$) in block b, with estimator $\hat{\tau}_b = \bar{Y}_{\{Z=1,B=b\}} - \bar{Y}_{\{Z=0,B=b\}}$, the difference between the treatment group and control group means in this block. The ITT can then be estimated using the estimated marginal distribution (or the marginal distribution if it is known) of the blocking variable:

$$\widehat{\text{ITT}} = \sum_{b=1}^{L} (\bar{Y}_{\{Z=1,B=b\}} - \bar{Y}_{\{Z=0,B=b\}}) \widehat{\Pr}(B=b). \tag{20}$$

Under random sampling from $\mathcal{P}$, $\widehat{\Pr}(B=b) = n_b/n$, and (20) = (17); For the matched pair design, in addition, (17) = (2).

As above, it is also easy to see that:

$$\bar{Y}_{\{Z=1,B=b\}} - \bar{Y}_{\{Z=0,B=b\}} = (n_b)^{-1} \sum_{i=1}^{n} 1_{\{B_i\}}(b)(\widehat{Y}_{zi}(1) - \widehat{Y}_{zi}(0)), \tag{21}$$

where the missing outcomes are imputed using the estimated 'best' predictor; thus the estimator (20) can also be obtained by imputing missing potential outcomes.

Another approach to estimating the ITT under (8) is to reweight the treatment group observations in such a way that the reweighted data from the treatment group (control group) would be a random sample from the distribution of $(Y_z(1), \underline{X})$ $(Y_z(0), \underline{X})$ and then apply the simple estimator (2) to the weighted data. This is the essence of 'inverse probability weighting' (Horvitz and Thompson, 1952).

To see how this works, suppose that $\pi_z(\underline{x})$ percent of the observations at level $\underline{x}$ of $\underline{X}$ are in the treatment group. Under random sampling from the distributions $(Y_z(1), \underline{X})$ and $(Y_z(0), \underline{X})$, the treatment and control groups should have the same distribution on $\underline{X}$. If the treated observations at level $\underline{x}$ are weighting by $\pi_z^{-1}(x)$ and the control group observations at $\underline{x}$ by $(1 - \pi_z(x))^{-1}$ the treated and controls will have the same distribution on $\underline{X}$ in the weighted data set and (2) can be applied to the weighted data, yielding the IPW estimator:

$$\frac{\sum_{i=1}^{n} \pi_z^{-1}(\underline{X}_i) Z_i Y_i}{\sum_{i=1}^{n} \pi_z^{-1}(\underline{X}_i) Z_i} - \frac{\sum_{i=1}^{n} (1 - \pi_z(\underline{X}_i))^{-1} (1 - Z_i) Y_i}{\sum_{i=1}^{n} (1 - \pi_z(\underline{X}_i))^{-1} (1 - Z_i)}. \tag{22}$$

Using $YZ = Y_z(1)Z$, elementary properties of conditional expectation, and assumption (8), leads to a more formal justification: $E(\pi_z^{-1}(\underline{X})ZY) = E(E(\pi_z^{-1}(\underline{X})ZY \mid \underline{X})) = E(\pi_z^{-1}(\underline{X})E(ZY_z(1) \mid \underline{X})) = E(\pi_z^{-1}(\underline{X})E(Z \mid \underline{X})E(Y_z(1) \mid \underline{X})) = E(Y_z(1))$. Finally, note that in the randomized block experiment $\pi_z(x) = \pi_z(g(x)) = \pi_z(b) = n_{1b}/n_b$, and the estimate (22) is identical to (20) under random sampling from $\mathcal{P}$.

Estimation of treatment effects in observational studies

In observational studies in the social and behavioral sciences, the assumption that treatment D is unrelated to the potential outcomes $Y_d(0)$ and $Y_d(1)$ is unlikely to hold. Estimation of the treatment effects ATE($\underline{X}$), ATT($\underline{X}$), ATE and ATT is therefore considered under the assumption (given by (11) and (7)) that treatment asignment is strongly ignorable, given the covariates $\underline{X}$ (Rosenbaum and Rubin, 1983). For a more extensive treatment of estimation under strongly ignorable treatment assignment, see the reviews by Imbens (2004) and Schafer and Kang (2007).

In principle, this case has already been considered. Nevertheless, new issues arise in attempting to use the estimators previously considered. There are serval reasons for this. First, in a randomized block experiment, the treatment and control group probabilities depend on the covariates in a known way, that is $\pi_z(\underline{X})$ is known. Thus, for example, if inverse probability weighting is used to estimate the ITT, the weights are known. This is not the case in an observational study. Second, in a randomized block experiment (8) is a byproduct of the study design, whereas

in observational studies the analogue (10) is an assumption; this issue was briefly discussed earlier. In practice, making a compelling argument that a particular set of covariates renders (8) true is the most difficult challenge facing empirical workers who want to use the methods below to make inferences about various types of treatment effects. Third, in a randomized block experiment, the covariates used in blocking take on (in principle) relatively few values, and thus the ITT can be estimated non-parametrically using the first appproach (as in (20)). In an observational study,where $\underline{X}$ is most likely high dimensional, this is no longer the case. And finally, it is necessary (though not difficult) to modify estimators of the ATE($\underline{X}$) and ATE to apply when it is desired to estimate ATT($\underline{X}$) and ATT.

Regression estimators use estimates ($\widehat{E}$) of the regressions $E(Y_d(1) \mid \underline{X})$ and $E(Y_d(0) \mid \underline{X})$ to impute missing potential outcomes: if $D_i = 1$, $\widehat{Y}_{di}(1) = Y_{di}(1)$, $\widehat{Y}_{di}(0) = \widehat{E}(Y_d(0) \mid \underline{X} = \underline{x}_i)$, and if $D_i = 0$, $\widehat{Y}_{di}(0) = Y_{di}(0)$, $\widehat{Y}_{di}(1) = \widehat{E}(Y_d(1) \mid \underline{X} = \underline{x}_i)$. These are then used to impute the unit effects and the ATE is then estimated by averaging over these:

$$n^{-1} \sum_{i=1}^{n} (\widehat{Y}_{di}(1) - \widehat{Y}_{di}(0)). \qquad (23)$$

As ATE($\underline{X}$) = ATT($\underline{X}$), the ATT can be obtained as above, averaging only over the n_1 treated observations.

As a starting point, consider the simplest (and still most widely used) regression estimator, where the covariates enter the response function linearly:

$$Y_i = \alpha + \tau D_i + \underline{\beta}' \underline{X}_i + \epsilon_i, \qquad (24)$$

where the parameters are identified by the condition $E(\epsilon \mid \underline{X}, D) = 0$. This leads to the well-known regression adjusted estimator of ATE(X):

$$\hat{\tau} = (\bar{Y}_{\{D=1\}} - \bar{Y}_{\{D=0\}}) - \underline{\hat{\beta}}' (\underline{\bar{X}}_{\{D=1\}} - \underline{\bar{X}}_{\{D=0\}}), \qquad (25)$$

where $\bar{Y}_{\{D=1\}}$ is the treatment group mean, $\bar{Y}_{\{D=0\}}$ is the control group mean, and $\underline{\bar{X}}_1$($\underline{\bar{X}}_0$) is the sample mean vector for the covariates in the treatment (control) group. It is easy to see that (25) = (23).

There are essentially two problems with the estimator $\hat{\tau}$. First, the investigator typically does not know the form of the response function and the linear form is chosen out of convenience. This form has very strong implications: τ = ATE(X) = ATE = ATT, that is, not only are the effects the same at all levels of X, but the ATE is also the ATT. When the regression functions are misspecified, using $\hat{\tau}$ can yield misleading inferences. Second, when there are 'regions' with little overlap between covariate values in the treatment and control groups, imputed values are then based on extrapolations outside the range of the data. For example, if the treatment group members have 'large' values of a covariate X_1 and the control group members have 'small' values, the imputations $\widehat{Y}_d(0)$ ($\widehat{Y}_d(1)$) for treatment (control) group members will involve extrapolating the control group (treatment group) regression to large (small) X_1 values. This may produce very misleading results.

To deal with the first of these difficulties, a natural alternative is to use nonlinear regresssion, or in the typical case where the form of the regression functions is not known, to estimate these non-parametrically (as in (19)). Imbens (2004) reviews this approach. Non-parametric regression can work well if $\underline{X}$ is not high dimensional, but when there are many covariates to control for, as is typical in observational studies, the precision of the estimator regression may be quite low. This problem then spills over to the imputations. [But see also Hill and McCulloch (2007), who propose using Bayesian Additive Regression Trees to fit the regression functions, finding that estimates based on this approach are superior to those obtained using many other typically employed methods.]

Sub-classification (also called blocking) is an older method used to estimate causal effects that is also non-parametric in spirit. Here, units with 'similar' values of $\underline{X}$ are grouped into blocks and the ATE is estimated as in the case of a randomized block

experiment considered above. To estimate the ATT, the distribution of the blocking variable B in the group receiving treatment (rather than the overall population) is used; equivalently, the imputed unit effects are averaged over the treatment group only. In a widely cited paper, Cochran (1968) shows in a concrete example with one covariate that subclassification with five blocks removes 90% of the bias.

Matching is another long-standing method that has been used which avoids parametrically modeling the regression functions. Although matching can be used to estimate the ATE, it has most commonly been used to estimate the ATT in the situation where the control group is substantially larger than the treatment group. In this case, each unit $i = 1, \ldots, n_1$ in the treatment group is matched to one or more units 'closest' in the control group and the outcome values from the matched control(s) are used to impute $\widehat{Y}_{di}(0)$. In the case of 'one to one' matching, unit i is matched to one control with value $Y* \equiv \widehat{Y}_{di}(0)$; if i is matched to more than one control, the average of the control group outcomes can be used.

There are many possible matching schemes. A unit can be matched with one or more others using various metrics to measure the distance between covariates X, and various criteria for when two units have covariate values close enough to constitute a 'match' can be used. In some schemes, matches are not reused, but in others are used again. In some schemes, not all units are necessarily matched. See Gu and Rosenbaum (1993) for a nice discussion of the issues involved in matching. Despite the intuitive appeal of matching, estimators that match on $\underline{X}$ typically have poor large sample properties (see Abadie and Imbens, 2006, for details).

The procedures above do not contend with the frequently encountered problem of insufficient overlap in the treatment and control groups. One alternative is to only consider regions where there is sufficient overlap, for example, to match only those treatment units with covariate values that are 'sufficiently' close to the values observed in the control group. When this is done, however, the quantity estimated is no longer the ATE or ATT because the average is only taken over the region of common support.

In an important paper, Rosenbaum and Rubin (1983) addressed the issue of overlap, proving that when (11) and (12) hold:

$$Y(0), Y(1) \perp\!\!\!\perp D \mid \pi_d(\underline{X}), \tag{26}$$

$$0 < \Pr(D = 1 \mid \pi_d(\underline{X})) < 1, \tag{27}$$

implying that any of the methods just discussed may be applied using the 'propensity score' $\pi_d(\underline{X})$ (which is a many to one function of $\underline{X}$), rather than $\underline{X}$; this cannot exacerbate the potential overlap problem and may help to lessen this problem. [For generalizations of the propensity score, applicable to the case where the treatment is categorical, ordinal or continuous, see Imai and van Dyk (2004), Imbens (2000) and Joffe and Rosenbaum (1999).]

Rosenbaum and Rubin (1983) also discuss (their corollary 4.3) using the propensity score to estimate the regression functions $E(Y_d(1) \mid \pi_d(\underline{X}))$ and $E(Y_d(0) \mid \pi_d(\underline{X}))$ when these are linear. Typically, the true form of the regression functions relating potential outcomes to the propensity score will be unknown. But these functions can be estimated non-parametrically more precisely using $\pi_d(\underline{X})$ than $\underline{X}$. However, the advantage of this approach is somewhat illusory, as in observational studies, $\pi_d(\underline{X})$ will be unknown and must be estimated. Logistic regression is often used, but this form is typically chosen for convenience. To the best of my knowledge, the impact of using a misspecified propensity score in this case has not been studied. If, however, a non-parametric estimator is used (for example, a sieve estimator as described in Imbens, 2004), the so-called 'curse of dimensionality' is simply transferred from estimation of the regression function to estimation of the propensity score.

A straightforward way to use subclassification on the propensity score is to divide the unit interval into L equal length intervals and group the observations by their

estimated propensity scores. Within interval I_ℓ, $\ell = 1, \ldots, L$, the ATE is estimated as:

$$\frac{\sum_{i\in I_\ell} D_i Y_i}{\sum_{i\in I_\ell} D_i} - \frac{\sum_{i\in I_\ell}(1-D_i)Y_i}{\sum_{i\in I_\ell}(1-D_i)}. \quad (28)$$

The ATE is then estimated by averaging the estimates (28), using weights $\frac{\sum_{i=1}^n 1_{A_i}(\ell)}{n}$, where $1_{A_i}(\ell) = 1$ if $i \in I_\ell$, 0 otherwise. To estimate the ATT, the weights should be modified to reflect the distribution of the observations receiving treatment: $\frac{\sum_{i=1}^n 1_{A_i}(\ell) D_i}{\sum_{i=1}^n D_i}$. Lunceford and Davidian (2004) study subclassification on the propensity score and compare this with IPW estimators (discussed below). Using simulations, Drake (1993) compares the bias in the case where the propensity score is known to the case where it is estimated, finding no additional bias is introduced in the latter case. She also finds that when the model for the propensity score is misspecified, the bias incurred is smaller than that incurred by misspecifying the regression function. This may be suggestive, but without knowing how to put the misspecification in the two different models onto a common ground, it is difficult to attribute too much meaning to this finding.

Matching on propensity scores is widely used in empirical work and has also been shown to perform well in some situations (Dehejia and Wahba, 1999). Corollary 4.1 in Rosenbaum and Rubin (1983) shows the ATE can be estimated by drawing a random sample $\pi_d(X_1), \ldots, \pi_d(X_n)$ from the distribution of $\pi_d(X)$, then randomly choosing a unit from the treatment group and the control group with this value $\pi_d(x)$, and taking the difference $Y(1) - Y(0)$, then averaging the n differences. In practice, of course, the propensity scores is usually unknown and must be estimated. [Rosenbaum (1987) explains the seemingly paradoxical finding that using the estimated propensity score tends to produce better balance than using the true propensity score.] Typically, the ATT is estimated by matching (using an estimate of the propensity score) each treated unit $i = 1, \ldots n_1$ to one or more control group units. The control group outcomes are then used to impute $\widehat{Y}_{di}(0)$ and the ATT is estimated as in the case of matching on $\underline{X}$.

Because the propensity score is a balancing score, after matching, the distribution of the covariates should be similar in the treatment and control groups. In practice, a researcher should check this balance and, if there is a problem, the model for the propensity score can be refitted (perhaps including interactions among covariates and/or other higher order terms) and the balance rechecked. In this sense, proper specification of the propensity score model is not really at issue here: the question is whether the matched sample is balanced.

Finally, in practice, it is often found that when the estimated propensity scores are near 0 or 1, the problem of insufficient overlap in the treatment and control groups may be lessened, but it is still present. In this case, the same kinds of issues previously discussed reappear.

As seen above, the propensity score also features prominently when methods that use inverse probability weighting are used to estimate treatment effects. There the covariates took on L distinct values, each with positive probability, and the IPW estimator is identical to the non-parametric regression estimator. This will no longer be the case. Parallelling the material above, the ATE may be estimated as:

$$\frac{\sum_{i=1}^n \hat{\pi}_d^{-1}(\underline{X}_i) D_i Y_i}{\sum_{i=1}^n \hat{\pi}_d^{-1}(\underline{X}_i) D_i} - \frac{\sum_{i=1}^n (1-\hat{\pi}_d(\underline{X}_i))^{-1}(1-D_i)Y_i}{\sum_{i=1}^n (1-\hat{\pi}_d(\underline{X}_i))^{-1}(1-D_i)}. \quad (29)$$

To estimate the ATT, it is necessary to weight the expression above by $\pi_d(\underline{X})$, giving

$$\bar{Y}_{\{D=1\}} - \frac{\sum_{i=1}^n \hat{\pi}_d(\underline{X}_i)(1-\hat{\pi}_d(\underline{X}_i))^{-1}(1-D_i)Y_i}{\sum_{i=1}^n \hat{\pi}_d(\underline{X}_i)(1-\hat{\pi}_d(X_i))^{-1}(1-D_i)}. \quad (30)$$

A problem with using these estimators is that probabilities near 0 and 1 assign large

weights to relatively few cases (Rosenbaum, 1987). IPW estimators do not require estimating the regression functions, but the weights must be estimated consistently in order that the estimator be consistent. When the model for the propensity score is misspecifed, the weights will be estimated incorrectly and the IPW estimator will not be consistent; if the estimated probabilities near 0 and 1 are not close to the true probabilities, the bias can be substantial. To contend with this, Hirano, Imbens and Ridder (2003) propose the use of a sieve estimator for the propensity score, while Shafer and Kang (2007) propose using a 'robit' model (a more robust model based on the cumulative distribution function of the t distribution, as opposed to the normal distribution (probit model) or logistic distribution (logistic regression).

Strategies for estimating treatment effects that combine one or more of the methods have also been proposed. For example, subclassification may still leave an imbalance between the covariates in the treatment and control groups. To reduce bias, linear regression of the outcome on D and $\underline{X}$ in each block may be used to adjust for the imbalance, e.g., as in (25) (Rosenbaum and Rubin, 1983). Matching estimators may be similarly modified.

Recently, a number of estimators that combine inverse probability weighting with regression have been proposed. These estimators have the property that so long as either the model for the propensity score is correct or the model for the regression function is correct, the estimator is consistent. Kang and Schafer (2007) do a nice job of explaining this idea, which originates in the sampling literature (Cassel, Sarndal and Wretman, 1976, 1977), and of summarizing the literature on this topic.

To give some intuition, consider estimation of the ATE. Suppose the population regression function is assumed to have the form:

$$Y_{di}(1) = g(\underline{X}_i) + \delta_i, \tag{31}$$

with $E(\delta_i \mid \underline{X}_i) = 0$, giving $E(Y_d(1) \mid \underline{X}) = g(\underline{X})$. As before, because (11) holds, the model can be estimated using the treated observations, and if the population model is correct and $\hat{g}$ is consistent, the estimator $n^{-1}\sum_{i=1}^{n}\hat{g}(\underline{X}_i)$ is consistent for $E(Y_d(1))$.

If the regression function is misspecified, the errors may not have 0 mean over $\mathcal{P}$. But if a good estimate of the δ_i can be obtained, $E(Y_d(1))$ can be estimated as $n^{-1}\sum_{i=1}^{n}\hat{g}(\underline{X}_i) + n^{-1}\sum_{i=1}^{n}\hat{\delta}_i$. To estimate the δ_i in $\mathcal{P}$, the propensity score can be used. If the model for the propensity score is correct, $E(D_i\pi_d(\underline{X}_i)\delta_i) = E(\delta_i)$, and thus the estimator

$$n^{-1}\sum_{i=1}^{n}\hat{g}(\underline{X}_i) + \sum_{i=1}^{n} D_i\hat{\pi}_d(\underline{X}_i)\hat{\delta}_i \tag{32}$$

will be consistent for $E(Y_d(1))$.

On the other hand, if the model for the regression function is correct, then whether or not the model for the propensity score is correct, $E(D_i\pi_d(\underline{X}_i)\delta_i) = EE(D_i\pi_d(\underline{X}_i)\delta_i \mid \underline{X}) = E(\pi_d(\underline{X}_i)E(D_i \mid \underline{X})E(\delta_i \mid \underline{X}) = 0$ as $E(\delta_i \mid \underline{X}) = 0$.

A consistent estimate for $E(Y_d(0))$ can be constructed in a similar manner. The weighted least-squares estimator, with appropriately chosen weights, is another example; more generally, the regression function can be estimated semiparametrically (Robins and Rotnitsky, 1995). Kang and Schafer (2007) discuss a number of other estimators that are consistent so long as either the regression function or the propensity score is specified correctly. Such estimators are often called 'doubly robust'; however, the reader should note that this terminology is a bit misleading. Statistical methods that operate well when the assumptions underlying their usage are violated are typically called robust. Here, the estimator is robust with misspecification to either the propensity score or the regression function, but not both. In that vein, Kang and Schafer's (2007) simulations suggest that when neither the propensity score nor the regression function is correctly specified, doubly robust estimators are often more biased than estimators without this attractive theoretical property.

MEDIATION ANALYSES

Mediation is a difficult topic and a thorough treatment would require an essay length treatment. The topic arises in several ways. First, even in randomized experiments, subjects do not always 'comply' with their treatment assignments. Thus, the treatment received D is an intermediate outcome intervening between Z and Y, and an investigator might want to know, in addition to the ITT (which measures the effect of Z on Y) the effect of D on Y. This might be of interest scientifically, and may also point, if the effect is substantial, but subjects don't take up the treatment, to the need to improve the delivery of the treatment package. Traditional methods of analysis that compare subjects by the treatment actually received or which compare only those subjects in the treatment and control groups that follow the experimental protocol are flawed because treatment received D is not ignorable with respect to Y. To handle this, Bloom (1984) first proposed using Z as an instrument for D. Subsequently, Angrist, Imbens and Rubin (1996) clarified the meaning of the IV estimand. Second, and more generally, researchers often have theories about the pathways (intervening variables) through which a particular cause (or set of causes) affects the response variable and the effects of both the particular cause(causes) and the intervening variables is of interest. To quantify these effects, psychologists and others often use structural equation models, following Baron and Kenny (1986), for example. However, the 'direct effects' of D on Y and Z on Y in structural equation models should not generally be interpreted as effects; conditions (which are unlikely to be met) when these parameters can be given a casual interpretation are also given below, as are conditions under which the IV estimated admits a causal interpretation.

Throughout, only the case of a randomized experiment with no covariates is considered; the results extend immediately to the case of an observational study where treatment assignment is ignorable only after conditioning on the covariates.

The local average treatment effect

As before, let $Z_i = 1$ if unit i is assigned to the treatment group, 0 otherwise. Let $D_{zi}(0)$ ($D_{zi}(1)$) denote the treatment i takes up when assigned to the control (treatment) group. Similarly, for $z = 0, 1$ and $d = 0, 1$, let $Y_{zdi}(0, 0)$ denote the response when i is assigned to treatment $z = 0$ and receives treatment $d = 0$; $Y_{zdi}(0, 1)$, $Y_{zdi}(1, 0)$, $Y_{zdi}(0, 0)$ are defined analogously. Let $Y_{zdi}(Z_i, D_{zi}(Z_i))$ denote i's observed response.

In a randomized experiment, the potential outcomes are assumed to be independent of treatment assignment:

$$D_z(0), D_z(1), Y_z(0, D_z(0)), Y_z(1, D_z(1)) \underline{\parallel} Z. \tag{33}$$

The two ITTs (hereafter ITT_D and ITT_Y) are identified as before by virtue of assumption (5); while these parameters are clearly of interest (and some would say these are the only parameters that should be of interest), neither parameter measures the effect of D on Y. That is because D is (in econometric parlance) 'endogenous'. To deal with such problems, economists have long used instrumental variables (including two stage least squares), in which 'exogenous' variables that are believed to affect Y only through D are used as an instrument for D. The IV estimand (in the simple case herein) is:

$$\begin{aligned} \frac{\text{cov}(Z, Y)}{\text{cov}(Z, D)} &= \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)} \\ &= \text{ITT}_Y / \text{ITT}_D. \end{aligned} \tag{34}$$

Recently, Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) clarified the meaning of the IV estimand (34) and the sense in which this estimand is a causal parameter.

ITT_Y is a weighted average over four compliance types: (1) compliers, with $D_{zi}(0) = 0, D_{zi}(1) = 1$; (2) never takers, with $D_{zi}(0) = 0, D_{zi}(1) = 0$; (3) always takers, with $D_{zi}(0) = 1, D_{zi}(1) = 1$; and (4) defiers, with $D_{zi}(0) = 1, D_{zi}(1) = 0$, who take up treatment if not assigned to treatment and who do not take up treatment if assigned

to treatment. Often it will be substantively reasonable to assume there are no defiers; this is the 'weak monotonicity assumption' $D_{zi}(1) \geq D_{zi}(0)$ for all i. Because the never takers and always takers receive the same treatment irrespective of their assignment, any effect of treatment assignment on Y for these types cannot be due to treatment D. If it is reasonable to assume the effect of treatment assignment operates only via the treatment, i.e., there is no 'direct effect' of Z on Y, then the unit effect of Z on Y for never takers and always takers is 0; this is called the exclusion restriction. Under weak monotonicity and exclusion, ITT_Y therefore reduces to:

$$E(Y_z(1, D_z(1)) - Y_z(0, D_z(0))) = E(Y(1, 1) - Y(0, 0)) \times \Pr(D_z(0) = 0, D_z(1) = 1). \quad (35)$$

As $\Pr(D_z(0) = 0, D_z(1) = 1) = E(D(1) - D(0))$ in the absence of defiers, provided this is greater than 0 (weak monotonicity and this assumption is sometimes called 'strong monotonicity'), (34) is the average effect of Z on Y for the compliers. If the direct effect of Z on Y for the compliers is also 0, (34) is also the effect of D on Y in this subpopulation; this is sometimes called the complier average causal effect (CACE) or the local average treatment effect (LATE). [For some further statistical work on compliance, see Imbens and Rubin, (1997), Little and Yau (1998), Jo (2002), Hirano, Imbens, Rubin and Zhou (2000)].

Because compliance is such an important issue, empirical researchers have been quick to apply the results above. But researchers who want to know the ATT or ATE might find the average effect of Z on Y for compliers or LATE to be of limited interest when the proportion of compliers is small (e.g., about 15% in the example presented by Angrist et al. (1996)]. Researchers who estimate the IV estimand (or who use instrumental variables or two stage least squares) should be careful not to forget that compliers may differ systematically from the never takers and always takers. However, LATE = ATT in the not uncommon case where the only way to obtain the treatment is by being in the treatment group. In addition, although the ATT is defined under the assumption that the exclusion restriction holds, the average effect of Z on Y for compliers equals the average effect of Z on Y for the treated in this case. Finally, in the (unlikely) case where the treatment effects are constant, LATE = ATT = ATE.

Empirical workers should also remember the exclusion restriction is very strong (even if applied only to the never takers and always takers), and in a 'natural' experiment or a randomized experiment that is not double blinded, this restriction may not hold. Researchers who are in the position of being able to design a double-blinded, randomized experiment should do so, and researchers who are relying on a natural experiment should think very seriously about whether or not this restriction is plausible. Finally, it is also important to remember that the compliers consitute an unobserved sub-population of $\mathcal{P}$, so that even if a policy maker were able to offer the treatment only to subjects in this subpopulation, he/she cannot identify these subjects with certainty.

The approach above also serves as the basis for the idea of principal stratification (Frangakis and Rubin, 2002). The essential idea is that for any intermediate outcome D (not necessarily binary), causal effects of D are defined within principal strata (subpopulations with identical values of $D_{zi}(0)$ and $D_{zi}(1)$).

Mediation and structural equation modeling

To facilitate comparison with the psychological literature, in which structural equation models are typically used to study mediation (Baron and Kenny, 1986; MacKinnon and Dwyer, 1993), I discuss the special case where D and Y are continuous, Z and D have additive effects on Y and the average effect of D on Y is linear (as described below); for a more general discussion, see Sobel (2008). As above, I make assumption (5) and examine the IV estimand;

the extension to the case where ignorability holds, conditional on covariates, is immediate.

Using potential outcomes, a linear causal model analogous to a linear structural equation model may be constructed:

$$D_{zi}(z) = \alpha_1^c + \gamma_1^c z + \varepsilon_{1zi}^c(z) \quad (36)$$

$$Y_{zdi}(z,d) = \alpha_2^c + \gamma_2^c z + \beta_2^c d + \varepsilon_{2zdi}^c(z,d), \quad (37)$$

where $E(\varepsilon_{1z}^c(z)) = E(\varepsilon_{2zd}^c(z,d)) = 0$; thus, $\gamma_1^c = \text{ITT}_D$, $\gamma_2^c = E(Y_{zd}(1,d) - Y_{zd}(0,d))$ for any d is the average unmediated effect of Z on Y, and $\beta_2^c = E(Y_{zd}(z,d+1) - Y_{zd}(z,d))$ for $z = 0, 1$ is the average effect of a one unit increase in D on Y.

A linear structural equation model for the relationship between Z, D and Y is given by:

$$D_i = \alpha_1^s + \gamma_1^s Z_i + \varepsilon_{1i}^s \quad (38)$$

$$Y_i = \alpha_2^s + \gamma_2^s Z_i + \beta_2^s D_i + \varepsilon_{2i}^s, \quad (39)$$

where the parameters are identified by the assumptions $E(\varepsilon_1^s \mid Z) = 0$ and $E(\varepsilon_2^s \mid Z, D) = 0$. Thus, the 'direct effects' of Z on D and Y, respectively, are: $\gamma_1^s = E(D \mid Z = 1) - E(D \mid Z = 0)$, $\gamma_2^s = E(Y \mid Z = 1, D = d) - E(Y \mid Z = 0, D = d)$. The 'direct effect' of D on Y is given by $\beta_2^s = E(Y \mid Z = z, D = d+1) - E(Y \mid Z = 0, D = d)$. The 'total effect' $\tau^s \equiv \gamma_2^s + \gamma_1^s \beta_2^s$.

By virtue of (33) $\gamma_1^s = \text{ITT}_D$ and $\tau^s = \text{ITT}_Y$ (Holland, 1988). However, neither γ_2^s nor β_2^s should generally be given a causal interpretation. To illustrate, consider $E(Y \mid Z = z, D = d) = E(Y_{zd}(z, D_z(z)) \mid Z = z, D_z(z) = d) = E(Y_{zd}(z, D_z(z)) \mid D_z(z) = d)$, where the last equality follows from (33). This gives $\gamma_2^s = E(Y_{zd}(1, D_z(1)) \mid D_z(1) = d) - E(Y_{zd}(0, D_z(0)) \mid D_z(0) = d)$. Because subjects with $D_z(0) = d$ are not the same subjects as those with $D_z(1) = d$, unless the unit effects of Z on D are 0, γ_2^s is a descriptive parameter comparing subjects across different subpopulations. Similar remarks apply to β_2^s.

It is also easy to see from the above that $\gamma_2^s = \gamma_2^c$ if $E(Y_{zd}(z, D_z(z)) \mid D_z(z) = d) = E(Y_{zd}(z, D_z(z)))$; a sufficient condition for this to hold is:

$$Y_z(z, D_z(z)) \underline{\perp\!\!\!\perp} D_z(z). \quad (40)$$

Similarly, $\beta_2^c = \beta_2^s$ under this condition. Results along these lines are reported in Eggleston, Scharfstein, Munoz and West (2006), Sobel (2008) and Ten Have, Joffe, Lynch, Brown and Maisto (2005). Unfortunately, this condition is unlikely to be met in applications, as it requires the intermediate outcome D to be ignorable with respect to Y, as if D had been randomized.

Holland (1988) showed that if (33) holds,the exclusion restriction $Y_{zdi}(1,d) - Y_{zdi}(0,d) = 0$ for all i holds, $ITT_D \neq 0$, and the other unit effects $D_{zi}(1) - D_{zi}(0)$ and $Y_{zdi}(z,d) - Y_{zdi}(z,d')$ are constant for all i, β_2^c is equal to the IV estimand (34). Unfortunately, the assumption that the effects are constant is even more implausible in the kinds of studies typically carried out in the behavioral and medical sciences than the assumptions needed to justify using structural equation models.

Sobel (2008) relaxes the assumption of constant effects, assuming instead:

$$E(\varepsilon_{2zd}^c(1, D_z(1)) - \varepsilon_{2zd}^c(0, D_z(0)) = 0. \quad (41)$$

Under (41), (33), the exclusion restriction $\gamma_2^c = 0$, and the assumption $\gamma_1^c \neq 0$, $\beta_2^c =$ the IV estimand (34). Further, assumption (41) is also weaker than the assumption (40) needed to justify using structural equation models.

The results above can be extended to the case where there are multiple instruments and multiple mediators. The results can also be extended to the case where compliance is an intermediate outcome prior to the mediating variable (Sobel, 2008) to obtain a complier average effect of the continuous mediator D on Y; this is the effect of the continuous mediator D on Y within the principal stratum (of the binary outcome denoting whether or not treatment is taken) composed of the compliers. Principal stratification itself can also be used to approach the problem of

estimating the effect of D on Y (Jo, 2008); here the idea would be to consider the effects of D on Y within strata defined by the pair of values of $(D_z(0), D_z(1))$.

DISCUSSION

In the last three decades, statisticians have generated a literature on causal inference that formally expresses the idea that causal relationships sustain counterfactual conditional statements. The potential outcomes notation allows causal estimands to be defined independently of the expected values of estimators. Thus, one can assess and give conditions (e.g., ignorability) under which estimators commonly employed actually estimate causal parameters. Prior to this, researchers estimated descriptive parameters and verbally argued these were causal based on other considerations, such as model specification, a practice that led workers in many disciplines, e.g., sociology and psychology, to interpret just about any parameter from a regression or structural equation model as a causal effect.

While this literature has led most researchers to a better understanding that a good study design (especially a randomized study) leads to more credible estimation of causal parameters than approaches using observational studies in conjunction with many unverifiable substantive assumptions, it is also important to remember the old lesson (Campbell and Stanley, 1963) that randomized studies do not always estimate parameters that are generalizable to the desired population. This is especially true for natural experiments, where the investigator has no control over the experiment, although the randomization assumption is plausible.

This literature has also led to clarification of existing procedures. In the process, new challenges have been generated. For example, while this literature reveals that the framework psychologists have been using for 25 years to study mediation is seriously flawed, as of yet, this literature cannot give adequate expression to and/or indicate how to assess the substantive theories that investigators have about the manner in which a treatment package may work through multiple mediators and the causal relationships among these mediators. These and many other issues are in need of much further work.

REFERENCES

Abadie, A. and Imbens. G. (2006) 'Large sample properties of matching estimators for average treatment effects', *Econometrica*, 74: 235–267.

Abadie, A., Angrist, J. and Imbens, G. (2002) 'Instrumental variables estimation of quantile treatment effects', *Econometrica*, 70: 91–117.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996) 'Identification of causal effects using instrumental variables', (with discussion) *Journal of the American Statistical Association*, 91: 444–472.

Baron A., Rubin M., and Kenny, D.A. (1986) 'The moderator–mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations', *Journal of Personality and Social Psychology*, 51: 1173–1182.

Belsen, W.A. (1956) A technique for studying the effects of a television broadcast', *Applied Statistics*, 5: 195–202.

Björklund, A. and Moffit, R. (1987) 'The estimation of wage gains and welfare gains in self-selection models', *The Review of Economics and Statistics*, 69: 42–49.

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

Bunge, M.A. (1979) *Causality and Modern Science* (3rd edn.). New York: Dover.

Campbell, D.T. and Stanley, J.C. (1963) *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976) 'Some results on generalized difference estimation and generalized regression estimation for finite populations', *Biometrika*, 63: 615–620.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977) *Foundations of Inference in Survey Sampling*. New York: Wiley.

Cochran, W.G. (1968) 'The effectiveness of adjustment by subclassication in removing bias in observational studies', *Biometrics*, 24: 205–213.

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

Crepinsek, M.K., Singh, A., Bernstein, L.S., and McLaughlin, J.E. (2006) 'Dietary effects of universal-free school breakfast: findings from the evaluation of the school breakfast program pilot project', *Journal of the American Dietetic Association*, 106: 1796–1803.

Dehejia, R.H. and Wahba, S. (1999) 'Causal effects in nonexperimental studies: reevaluating the evaluation of training programs', *Journal of the American Statistical Association*, 94: 1053–1062.

Doksum, K. (1974) 'Empirical probability plots and statistical inference for nonlinear models in the two-sample case', *Annals of Statistics*, 2: 267–277.

Drake, C. (1993) 'Effects of misspecication of the propensity score on estimators of treatment eect', *Biometrics*, 49: 1231–1236.

Eggleston, B., Scharfstein, D., Munoz, B. and West, S. (2006) 'Investigation mediation when counterfactuals are well-defined: does sunlight exposure mediate the effect of eye-glasses on cataracts?'. Unpublished manuscript, Johns Hopkins University.

Finkelstein, M.O., Levin, B. and Robbins, H. (1996) 'Clinical and prophylactic trials with assured new treatment for those at greater risk: II. examples', *American Journal of Public Health*, 86: 696–702.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*. London: Oliver and Boyd.

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

Gitelman, A.I. (2005) 'Estimating causal effects from multilevel group-allocation data', *Journal of Educational and Behavioral Statistics*, 30: 397–412.

Granger, C.W. (1969) 'Investigating causal relationships by econometric models and cross-spectral methods', *Econometrica*, 37: 424–438.

Gu, X.S, and Rosenbaum, P.R. (1993) 'Comparison of multivariate matching methods: structures, distances and algorithms', *Journal of Computational and Graphical Statistics*, 2: 405–420.

Halloran, M. E. and Struchiner, C.J. (1995) 'Causal inference in infectious diseases', *Epidemiology*, 6: 142–151.

Harre, R. and Madden, E.H. (1975) *Causal Powers: A Theory of Natural Necessity*. Oxford: Basil Blackwell.

Hill, J.L. and McCulloch, R.E. (2007) 'Bayesian nonparametric modeling for causal inference.' Unpublished manuscript, Columbia University.

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

Hirano, Keisuke, Imbens, Guido W., and Ridder, G. (2003) 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica*, 71: 1161–1189.

Holland, P.W. (1988) 'Causal inference, path analysis, and recursive structural equation models', (with discussion) in Clogg, C.C. (ed.), *Sociological Methodology*. Washington, D.C.: American Sociological Association. pp. 449–493.

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

Imai, K. and van Dyk, D.A. (2004) 'Causal inference with general treatment regimes: generalizing the propensity score', *Journal of the American Statistical Association*, 99: 854–866.

Imbens, G.W. (2000) 'The role of the propensity score in estimating dose-response functions', *Biometrika*, 87: 706–710.

Imbens, G.W. (2004) 'Nonparametric estimation of average treatment effects under exogeneity: a review', *Review of Economics and Statistics*, 86: 4–29.

Imbens, G.W., and J.D. Angrist (1994) 'Identification and estimation of local average treatment effects', *Econometrica*, 62: 467–475.

Imbens, G.W. and Rubin, D.B. (1997) 'Estimating outcome distributions for compliers in instrumental variables models', *Review of Economic Studies*, 64: 555–574.

Jo, B. (2002) 'Estimation of intervention effects with noncompliance: Alternative model specifications (with discussion),' *Journal of Educational and Behavioral Statistics*, 27: 385–415.

Jo, B. (2008) 'Causal inference in randomized experiments with mediational processes,' *Psychological Methods*, 13: 314–336.

Joffe, M.M. and Rosenbaum P.R. (1999) 'Propensity scores', *American Journal of Epidemiology*, 150: 327–333.

Kang, J.D.Y. and Schafer, J.L. (2007) 'Demystifying double robustness: a comparison of alternative strategies for estimating population means from incomplete data', *Statistical Science*, 22: 523–580.

Lehmann, E.L. (1974) 'Nonparametris: Statistical Methods Based on Ranks,' *Holden-Day*, Inc.: San Francisco, CA.

Lewis, D. (1973) 'Causation', *Journal of Philosophy*, 70: 556–567.

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

Lunceford, J.K., and M. Davidian. (2004) 'Stratication and weighting via the propensity score in estimation

of causal treatment effects: a comparative study', *Statistics in Medicine*, 23: 2937–2960.

MacKinnon, D.P. and Dwyer, J.H. (1993) 'Estimating mediating effects in prevention studies', *Evaluation Review*, 17: 144–158.

Manski, C.F. (1995) *Identication Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Neyman, J. (1923) 1990 'On the application of probability theory to agricultural experiments. essays on principles. Section 9', (with discussion) *Statistical Science*, 4: 465–480.

Reichenbach, H. (1956) *The Direction of Time*. Berkeley: University of California Press.

Robins, J.M. (1989) 'The analysis of randomized and nonrandomized aids treatment trials using a new approach to causal inference in longitudinal studies', in Sechrest, L., Freedman, H. and Mulley, A. (eds.), *Health Services Research Methodology: A Focus on AIDS*. Rockville, MD: US Department of Health and Human Services. pp. 113–159.

Robins, J.M. and Rotnitsky, A. (1995) 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association*, 90: 122–129.

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

Rosenbaum, P.R. (2002) *Observational Studies*. New York: Springer-Verlag.

Rosenbaum, P.R. and 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', *Journal of Educational Psychology*, 66: 688–701.

Rubin, D.B. (1977) 'Assignment to treatment groups on the basis of a covariate', *Journal of Educational Statistics*, 2: 1–26.

Rubin, D.B. (1978) 'Bayesian inference for causal effects: the role of randomization', *The Annals of Statistics*, 6: 34–58.

Rubin, D.B. (1980) 'Comment on "randomization analysis of experimental data: the Fisher randomization test" by D. Basu', *Journal of the American Statistical Association*, 75: 591–593.

Schafer, J.L. and Kang, J.D.Y. (2007) 'Average causal effects from observational studies: a practical guide and simulated example'. Unpublished manuscript, Pennsylvania State University.

Simon, H.A. (1954) 'Spurious correlation: a causal interpretation', *Journal of the American Statistical Association*, 49: 467–492.

Sobel, M.E. (1995) 'Causal inference in the social and behavioral sciences', in Arminger, G., Clogg, C.C. and Sobel, M.E. (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum. pp. 1–38.

Sobel, M.E. (2006a) 'Spatial concentration and social stratication: does the clustering of disadvantage "beget" bad outcomes?', in Bowles, S., Durlauf, S.N. and Hoff, K. (eds.), *Poverty Traps*, New York: Russell Sage Foundation. pp. 204–229.

Sobel, M.E. (2006b) 'What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference', *Journal of the American Statistical Association*, 101: 1398–1407.

Sobel, M.E. (2008) 'Identification of causal parameters in randomized studies with mediating variables,' *Journal of Educational and Behavioral Statistics*, 33: 230–251.

Suppes, P. (1970) *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

Tenhave, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S. A., and A.T. Beck (2007) 'Causal mediation analyses with rank preserving models,' *Biometrics*, 63: 926–934.

TenHave, T.R., Marshall, J., Kevin, L., Brown, G. and Maisto, S. (2005) *Causal Mediation Analysis with Structural Mean Models.* University of Pennsylvania Biostatistics, Working Paper.

Thistlethwaite, D.L. and Campbell, D.T. (1960) 'Regression-discontinuity analysis: an alternative to the *ex post facto* experiment', *Journal of Educational Psychology*, 51: 309–317.

Yule, G.U. (1896) 'On the correlation of total pauperism with proportion of out-relief ii: males over 65', *Economic Journal*, 6: 613–623.