Causal Inference

Kosuke Imai
Department of Politics, Princeton University

March 2, 2013

Throughout POL572 and 573, we will learn how to use various statistical methods in order to make *causal inference*, which is a main goal of social science research. We all know the mantra "correlation is not causation." The difficulty of inferring causality arises from the fact that we do not observe counterfactual outcomes, which are required to estimate causal effects. To formalize this intuition, we begin by describing the potential outcomes framework of causal inference.

1 Potential Outcomes Framework

First, we formally define the meaning of causality. What do we mean when we say "an event A causes another event B"? Here, we use the commonly accepted statistical framework of causality that is based on the notion of *potential outcomes*. This framework is often called the *Neyman-Rubin causal model* because the framework first appeared in Neyman (1923)'s analysis of randomized experiments and Rubin (1974) extended it to observational studies. The key idea is that any causal inference is based on both actual (or realized) and counterfactual outcomes.

For example, suppose that we are interested in the causal effect of a voter's exposure to a political TV advertisement on her/his voting behavior in an election. If this voter saw a particular advertisement during the campaign, then we will never know her voting behavior under the counterfactual scenario where she was not exposed to the advertisement. And yet, we would define the causal effect of the advertisement as the difference between the actual and counterfactual outcomes for voting behavior. Holland (1986) called this dilemma the *fundamental problem of causal inference*.

1.1 The Setup

We now formally define the potential outcomes, each of which corresponds to a particular value of the *treatment variable*, i.e., the causal variable of interest. Let T_i be the causal (or treatment) variable of interest for unit i. In our setting, T_i is a random variable which takes a value in a set of possible treatments $\mathcal{T}$. The potential outcome $Y_i(t)$ represents the outcome that would be observed for unit i if it receives the treatment value t, i.e., $T_i = t$ for $t \in \mathcal{T}$. The definition implies that the treatment variable determines which of the potential outcomes will be realized. If we let Y_i denote the actual outcome for unit i and assume that the treatment is binary, then, the observed outcome is a function of one's treatment variable, i.e., $Y_i = Y_i(T_i)$.

Going back to the previous example, the treatment variable is a binary indicator, which takes the value of one (zero), $T_i = 1$ ($T_i = 0$), if voter i is (not) exposed to the advertisement. Then, $Y_i(1)$ represents the turnout of the voter when exposed to the advertisement whereas $Y_i(0)$ is the outcome we would observe under no exposure. The fundamental problem of causal inference is that while the causal effect is defined as $Y_i(1) - Y_i(0)$, we only observe one of the two potential outcomes for any particular voter. For example, if voter i actually watched the advertisement, we observe $Y_i = Y_i(1)$ but $Y_i(0)$ is unobserved.

We assume that each potential outcome of a given unit is a fixed (but not necessarily observed) quantity whereas the treatment assignment is considered as a random variable In the context of randomized experiments, the observed outcome is a function of atual treatment status: Y_i is a random variable since it is a function of T_i. One possible objection to this assumption is that potential outcomes of a given unit should also be seen as random variables. For example, we should allow an experimental subject to flip a coin after he receives the treatment and before taking an action. Although such a scenario is plausible, it only adds an unnecessary complication. In fact, one can never empirically distinguish whether the potential outcomes are fixed or random. The exact same study, whether it is experimental or observational, cannot be repeated, and thus potential outcomes cannot be observed more than once.

1.2 Assumptions

The definition of potential outcomes given above makes an implicit but important assumption that the treatment status of one unit does not affect the potential outcomes of another unit. In the aforementioned example, it is assumed that the voter's (potential) voting behavior is not a function of whether or not someone else (e.g., her spouse) is exposed to the advertisement. Such an assumption may not hold in many social science research contexts where social interaction between units is unavoidable. Such a phenomenon is often called peer, diffusion, spill-over, or neighborhood effects. While it is possible to modify the current framework to accomodate these effects, we focus on the simplest setting as a starting point of our analysis.

Formally, the assumption of *no interference between units* can be written as follows. Let $\mathbf{T} = (T_1, T_2, \ldots, T_n)$ be a vector of treatment assignments for all n units. In general, we can write the potential outcome of unit i as a function of all n units, i.e., $Y_i(\mathbf{T})$. The assumption states that this potential outcome of unit i does not depend on other units' treatment status, i.e., $Y_i(\mathbf{T}) = Y_i(T_i)$.

In addition to the assumption of no interference, there are two assumptions that are implicitly made under the current framework. First, we assume no simultaneity. That is, the causal ordering between T_i and Y_i is pre-determined such that T_i affects Y_i but not vice versa. Second, it is assumed that there is only a single version of treatment across individuals. For example, suppose that a medical surgery is a treatment. If a different surgeon operates in a significantly different way, then we must consider it as a different treatment depending on who operates the surgery for each patient. In practice, any treatment is likely to differ in some aspects but the question is whether such differences are significant enough to be regarded as different treatments. All together, these assumptions constitute what Rubin (1990) called the *Stable Unit Treatment Value* (SUTVA) assumption.

1.3 Defining Causal Effects

Based on the framework described above, we define causal effects, which are also called treatment effects. We first consider the situation where the treatment variable T_i is binary (i.e., $\mathcal{T} = \{0, 1\}$). Thus, there are two potential outcomes for each unit, i.e., $Y_i(1)$ and $Y_i(0)$. One natural way to define a *unit treatment effect* (or unit causal effect) is to look at the difference between the two potential outcomes,

$$\tau_i \quad = \quad Y_i(1) - Y_i(0). \tag{1}$$

In the previously introduced example, if voter i only casts his ballot when exposed to the advertisement, then the causal effect is one for this voter. The difference is not the only quantity of interest. When the outcome variable is continuous, one may be interested in the ratio of two potential outcomes or the percentage increase due to treatment, which are defined as $Y_i(1)/Y_i(0)$ and $\{Y_i(1) - Y_i(0)\}/Y_i(0) \times 100$, respectively.

The key point here is that any causal quantity of interest is expressed as a function of two potential outcomes. Unfortunately, because the potential outcomes, $Y_i(1)$ and $Y_i(0)$, are never jointly observed, none of these quantities are known at an individual level. Indeed, the joint distribution of the potential outcomes, $P(Y_i(1), Y_i(0))$, cannot be directly inferred from the data. For example, the correlation between $Y_i(1)$ and $Y_i(0)$ is not identifiable because we never observe both quantities for any given unit. This implies that the distribution of unit causal effects, e.g., $P(Y_i(1) - Y_i(0))$, also cannot be estimated directly from the data without additional assumptions.

Fortunately, as we shall see later, the average treatment effect (ATE) can be identified in some situations. There are many ATEs of interest, depending on the procedure used to average unit causal effects. Here, we focus on the unit causal effect based on differences, i.e., $Y_i(1) - Y_i(0)$, and define the *sample average treatment effect* (SATE), which represents the sample average of unit treatment effects and is defined in the following way.

$$\text{SATE} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\} \tag{2}$$

Another quantity of interest is the sample average treatment effect for the treated (SATT), which is defined as,

$$\text{SATT} \quad = \quad \frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} T_i (Y_i(1) - Y_i(0)) \tag{3}$$

This quantity represents the average treatment effect among those who received the treatment. For program evaluation, this quantity may be of greater relevance because policy makers are often interested in the impact of program on those who actually enroll in the program rather than those who do not. If the treated units differ significantly from those in the control group, SATE and SATT are also expected to be quite different.

We can also define the average causal effects for a population rather than for a particular sample. This is the *population average treatment effect* (PATE) and the *population average treatment effect for the treated* (PATT), which are formally defined as,

$$\text{PATE} \quad = \quad \mathbb{E}\{Y_i(1) - Y_i(0)\} \tag{4}$$

$$\text{PATT} \quad = \quad \mathbb{E}\{Y_i(1) - Y_i(0) \mid T_i = 1\} \tag{5}$$

PATT represents the average treatment effect for a particular subpopulation, but more generally we are often interested in characterizing how the treatment effect varies as a function of pre-treatment covariates X_i. For example, the causal effects of exposure to political advertisement may differ between partisans and independents. This type of *treatment effect heterogeneity* is essential for testing the implications of specific social science theories. We define the conditional average treatment effect (CATE) as,

$$\text{CATE} \;=\; \mathbb{E}\{Y_i(1) - Y_i(0) \mid X_i\} \tag{6}$$

If the CATE differs depending on the value of X_i, we say that X_i moderates the treatment effect. Such a *causal moderation analysis* is an important part of causal inference in social sciences.

Finally, the average causal effects are not the only quantities of interest. One such quantity is the *α-quantile treatment effect* (QTE_α) for some $\alpha \in (0,1)$. For example, one may be interested in comparing the bottom 10 percentile wage of workers of a particular population under a job training program and the bottom 10 percentile wage of the same population of workers without such a program. This quantity is $\text{QTE}_{0.1}$. Formally, the α-quantile treatment effect is defined as,

$$\text{QTE}_\alpha \;=\; q_1(\alpha) - q_0(\alpha), \tag{7}$$

where $q_t(\alpha) = \inf\{y : \Pr(Y_i(t) \leq y) \geq \alpha\}$ represents α-quantile of potential outcome $Y_i(t)$ for $t = 0, 1$. Note that this causal effect is different from the α-quantile of the unit causal effects, i.e., $\text{median}\{Y_i(1) - Y_i(0)\}$, which cannot be identified from the data because we never observe unit treatment effects, i.e., $\tau_i = Y_i(1) - Y_i(0)$.

1.4 Causal Effects of Immutable Characteristics

Although the mathematical definition of causal effects are given above, causal effects may not always be well defined. Here, we discuss this question through some examples.

The first example concerns the causal effects of *immutable characteristics*. In the literature, it is often argued that the causal effects of immutable characteristics is not well defined because it is impossible to manipulate a defining feature such as gender or race. Yet, these characteristics are associated systematically with other attributes such as income, education, or beliefs. This led some to conclude that "no causation without manipulation" (Holland, 1986, p.959). In these situations, however, one can make inference about a redefined causal quantity. For example, Chattopadhyay and Duflo (2004) uses a randomized natural experiment to examine the causal effect of politicians' gender on policy outcomes where randomly selected local village councils were required to reserve certain policy making positions for women. Nevertheless, in this case, female politicians differ from their male counterparts in various characteristics other than their gender, and so the differences in observed policy outcomes cannot be solely attributed to policy makers' gender differences. Other factors such as education could be confounding factors for evaluating the effect of gender. This means that we cannot distinguish whether it is their "femaleness" or the kind of life experience of a woman who has chosen to become a politician. Nevertheless, the study *can* estimate the effect of having a female politician.

Another example is the field experiment designed and analyzed by Humphreys *et al.* (2006) where discussion leaders were randomly assigned to deliberation groups in order to examine whether discussion leaders can manipulate group decision outcomes. In particular, the authors investigate the extent to which leaders' policy preferences influence group discussion outcomes by computing their observed correlation (see Table 6). Because the leaders were randomly assigned to the groups, a positive correlation between the leaders preferences and group decision outcomes may imply that leaders were able to manipulate group discussion towards their own policy preferences. Unfortunately, this causal quantity does *not* measure the causal effect of leaders' preferences alone on group discussion outcomes. This is because policy preferences cannot be randomly assigned to leaders. As a consequence, leaders' preferences may be correlated with other observed and unobserved attributes of their own, making it difficult to isolate the causal effect that can be attributed to leaders' preferences alone. For example, those with higher education may favor different spending and also be more persuasive as a discussion leader. An alternative quantity of interest estimated in the Humphreys *et al.* (2006) study (see Section V) is the causal effect of leaders' *presence* on group discussion outcomes, rather than that of leaders' *preferences*. Here, there is no problem with immutable characteristics and causal inference could be straightforward. While we cannot directly randomize preferences, it is possible to randomize whether or not each group has a discussion leader and estimate the effect of the presence of a leader.

2 Estimation of Average Treatment Effects

We next describe the classical approach to the analysis of randomized experiments. This approach was first developed by Neyman (1923) and continues to be used by researchers even today.

2.1 Benefits of Randomization

Before randomized experiments were discovered, scientists used *controlled experiments* to make causal inferences. However, it became quickly clear that even in laboratory settings it is difficult (and perhaps impossible) to control for all the factors that could possibly influence the outcome of interest. Another equally significant limitation of controlled experiments was that when these differences are not eliminated, there is no way for researchers to quantify the error that result from those uncontrolled differences.

Of course, randomization does not completely solve this problem either. In fact, given any realized set of randomized treatment assignments in a particular experiment, one would never obtain treatment and control groups that are perfectly identical to each other in both their observed and unobserved characteristics. However, the randomization of treatment can achieve two things that controlled experiments cannot. It makes the distributions of these characteristics identical in a large sample (*asymptotic distributions*) or over repeated randomization in any finite sample (*randomization distributions*). Based on these two features, researchers can quantify the uncertainty about their resulting conclusions and conduct formal statistical tests of scientific hypotheses. In fact, when analyzing randomized experiments, we exploit the fact that experimenters have the complete knowledge of the randomized treatment assignment mechanism. We then base our inferences on either the asymptotic distributions or the randomization distributions that result from this randomization of the treatment.

2.2 Design-based Approach

Consider a completely randomized experiment where n units are selected into the treatment group of size n_1 and the control group of size n_0 where $n = n_1 + n_0$. The complete randomization of the treatment implies that the treatment assignment probability is the same for all units and equals n_1/n, i.e., $\Pr(T_i = 1) = n_1/n$ for all i. Formally, the randomization of treatment implies statistical independence between the treatment variable and the potential outcomes, $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$ for all i where $\perp\!\!\!\perp$ denotes statistical independence. In other words, the treatment assignment has nothing to do with potential outcomes.

The estimator we consider is the *difference-in-means* estimator, which represents the difference in the average outcome between the treatment and control groups. The estimator is formally defined as,

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i \tag{8}$$

Estimation of the SATE. We begin by showing that the difference-in-means estimator $\hat{\tau}$ is unbiased for the SATE defined in equation (2). To do this formally, let $\mathcal{O} = \{Y_i(1), Y_i(0)\}_{i=1}^{n}$ be a set of potential outcomes of all n units in the sample. To establish the unbiasedness, we consider the expectation of $\hat{\tau}$ over the (hypothetical) repeated randomization of treatment. What is the average value of our estimator when applied to all possible randomizations? Since n_1 units are assigned to the treatment group, there exist a total of $\binom{n}{n_1}$ ways to randomize, each of which is equally likely. In essence, we calculate the value of τ for each of these randomization schemes and take the average of all such values. Formally, we take the expectation of $\hat{\tau}$ by treating the potential outcomes of the sample as fixed quantities and the treatment assignment as random. This can be done by conditioning on this set $\mathcal{O}$ as follows,

$$\begin{aligned}
\mathbb{E}(\hat{\tau} \mid \mathcal{O}) &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(T_i Y_i(1) \mid \mathcal{O}) - \frac{1}{n_0}\sum_{i=1}^{n}\mathbb{E}\{(1-T_i)Y_i(0) \mid \mathcal{O}\} & (9)\\
&= \frac{1}{n_1}\sum_{i=1}^{n}\Pr(T_i = 1)Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\{1-\Pr(T_i=1)\}Y_i(0) & (10)\\
&= \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\} \;=\; \text{SATE} & (11)
\end{aligned}$$

where the second equality follows from the randomization of treatment and the fact that T_i is a binary random variable. This shows that the difference-in-means estimator $\hat{\tau}$ is unbiased (over repeated randomization of treatment) for the SATE.

In a similar manner, Neyman also derived the following exact expression for the variance that is based solely on the randomization of the treatment,

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) \;=\; \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right)$$

where S_1^2 and S_0^2 are the sample variance of the potential outcomes $Y_i(1)$ and $Y_i(0)$, respectively, and S_{01} is their sample covariance. Formally, these terms are defined as,

$$\begin{aligned}
S_t^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(t) - \overline{Y(t)})^2 & (12)\\
S_{01} &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(0) - \overline{Y(0)})(Y_i(1) - \overline{Y(1)}) & (13)
\end{aligned}$$

where $\overline{Y(t)} = \sum_{i=1}^n Y_i(t)/n$ is the sample mean of potential outcome for $t = 0, 1$.

It can be shown that the sample variances of potential outcomes, i.e., S_1^2 and S_0^2, can be estimated without bias using the sample variances of the observed outcomes for the treatment and control groups, which are defined as,

$$s_0^2 \;=\; \frac{1}{n_0 - 1}\sum_{i=1}^n (1 - T_i)(Y_i - \overline{Y}_0)^2 \tag{14}$$

$$s_1^2 \;=\; \frac{1}{n_1 - 1}\sum_{i=1}^n T_i(Y_i - \overline{Y}_1)^2 \tag{15}$$

where $\overline{Y}_0 = \sum_{i=1}^n (1 - T_i)Y_i/n_0$ and $\overline{Y}_1 = \sum_{i=1}^n T_iY_i/n_1$. That is, we have $\mathbb{E}(s_t^2 \mid \mathcal{O}) = S_t^2$ for $t = 0, 1$. In contrast, the sample covariance between the two potential outcomes, i.e., S_{01} cannot be estimated directly because we never observe the two potential outcomes, $Y_i(1)$ and $Y_i(0)$, jointly. We say that S_{01} is a *unidentifiable* parameter because no amount of data can help estimate it. That is, although we have derived the exact variance expression, we cannot estimate it however large one's sample size is.

One possible strategy here is to bound the variance. Specifically, we derive the bounds using the following *covariance inequality*,

$$-S_0S_1 \;\leq\; S_{01} \;\leq\; S_0S_1, \tag{16}$$

which is also equivalent to the fact that the correlation, i.e., $S_{01}/(S_0S_1)$, is bounded between -1 and 1. Substituting these lower and upper bounds into the expression given in equation (12), we obtain the following sharp bounds,

$$\frac{n_0n_1}{n}\left(\frac{S_1}{n_1} - \frac{S_0}{n_0}\right)^2 \;\leq\; \mathbb{V}(\hat{\tau} \mid \mathcal{O}) \;\leq\; \frac{n_0n_1}{n}\left(\frac{S_1}{n_1} + \frac{S_0}{n_0}\right)^2.$$

where the inequality for the upper (lower) bound becomes an equality if the sample correlation between $Y_i(1)$ and $Y_i(0)$ is 1 (-1). These bounds are said to be *sharp* (best possible) because they cannot be improved (shortened) without making an additional assumption.

Several remarks are in order here. First, the usual variance estimator for the difference-in-means estimator $s_1^2/n_1 + s_0^2/n_0$, is on average conservative. This can be seen from the following equality,

$$\frac{n_0n_1}{n}\left(\frac{S_1}{n_1} + \frac{S_0}{n_0}\right)^2 \;=\; \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} + \frac{1}{n}(S_1 - S_0)^2 \tag{17}$$

where $(S_1 - S_0)^2/n \geq 0$ means that the upper bound in equation (17) is never less than the expectation of the usual variance estimator. Second, assume that the treatment effect is identical across all units. This is the assumption of *constant additive unit causal effect*, which can be formally written as, $Y_i(1) - Y_i(0) = a$ for all units $i = 1, 2, \ldots, n$ and some constant a. This assumption is highly unlikely in social science research where unit level heterogeneity tends to be large, but under this assumption, we can write $\mathbb{V}(Y_i(1) - Y_i(0)) = S_1^2 + S_0^2 - 2\text{Cov}(Y_i(1), Y_i(0)) = 0$, which implies $S_1S_0 = (S_1^2 + S_0^2)/2$. Thus, we can identify the variance and it equals the expectation of the usual variance estimator,

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) \;=\; \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}. \tag{18}$$

Finally, while the exact variance is unidentifiable, one can use the expression given in equation (12) to obtain the optimal treatment assignment rule, which determines the relative sizes of the treatment and control groups so that the resulting variance is minimized. To do this, we calculate the first derivative of the variance and set it to zero. We find the following expression for the optimal sizes of the treatment and control groups,

$$n_1^{\text{opt}} \;=\; \frac{n}{1+S_0/S_1} \quad \text{and} \quad n_0^{\text{opt}} \;=\; \frac{n}{1+S_1/S_0} \tag{19}$$

This makes sense because a greater variation of the potential outcome under the treatment condition, for example, means that we need to allocate a greater sample size to the treatment group. By specifying the ratio S_1/S_0, experimenters can use this formula to choose the optimal treatment assignment rule.

Estimation of the PATE. Next, we discuss the estimation of the population average treatment effect, or PATE. Specifically, we assume that our sample is a simple random sample from the target population of interest. In this framework, we have two sources of randomness: random sampling from a population followed by the randomization of treatment assignment. The derivation of statistical properties of the difference-in-means estimator exactly follows this two-step procedure. To show the unbiasedness for the PATE, we use the law of iterated expectation,

$$\mathbb{E}(\hat{\tau}) \;=\; \mathbb{E}\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\} \;=\; \mathbb{E}(\text{SATE}) \;=\; \mathbb{E}(Y_i(1)-Y_i(0)) \;=\; \text{PATE} \tag{20}$$

where the third equality follows from the fact that the SATE is the sample average of differences in two potential outcomes, which under simple random sampling, is unbiased for the population average of those differences.

Similarly, the variance can be derived by applying the law of total variances and using the expressions derived in equations (11) and (12),

$$\begin{aligned}
\mathbb{V}(\hat{\tau}) \;&=\; \mathbb{E}\{\mathbb{V}(\hat{\tau} \mid \mathcal{O})\} + \mathbb{V}\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\} && (21)\\
&=\; \mathbb{E}\left\{\frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right)\right\} + \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)\right) && (22)\\
&=\; \frac{1}{n}\left(\frac{n_0}{n_1}\sigma_1^2 + \frac{n_1}{n_0}\sigma_0^2 + 2\sigma_{01}\right) + \frac{1}{n}\left(\sigma_1^2 + \sigma_0^2 - 2\sigma_{01}\right) && (23)\\
&=\; \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} && (24)
\end{aligned}$$

where $\sigma_t^2 = \mathbb{V}(Y_i(t))$ is the population variance of the potential outcome $Y_i(t)$ and $\sigma_{01} = \text{Cov}(Y_i(1), Y_i(0))$ is the population covariance of $Y_i(1)$ and $Y_i(0)$. Note that the first term in equation (21) represents the average within-sample variance while the second term is the across-sample variance. In equation (21), the conditional expectation and variance are computed with respect to the randomization of treatment given a particular sample whereas the outer expectation and variance are taken over repeated sampling of units.

We see that the unidentifiable covariance term goes away, implying that the standard variance estimator, $s_1^2/n_1 + s_0^2/n_0$, is unbiased (since $\mathbb{E}(s_t^2) = \sigma_t^2$ for $t = 0, 1$ by the law of iterated expectation). Thus, when estimating the SATE, the usual variance estimator is too conservative. However, when estimating the PATE, it estimates the true variance without bias.

2.3 Asymptotic Inference

Using the result we derived above, we can apply the law of large numbers and the central limit theorem in order to conduct asymptotic inference for the difference-in-means estimator. Recall that the law of large numbers and the central limit theorem are about the asymptotic behavior of sample averages. Thus, the first step is to rewrite the difference-in-means estimator as the sample mean of a random variable,

$$\hat{\tau} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_iY_i(1)}{k} - \frac{(1-T_i)Y_i(0)}{1-k}\right\} \tag{25}$$

where $k = n_1/n$ is the ratio of treatment group size relative to the total sample size. For example, if we assign half of the sample to the treatment group and the other half to the control group, we have $k = 0.5$. Our asymptotic inference assumes that this ratio k stays constant while the sample size n goes to infinity. In the above expression, we see that the difference-in-means estimator is now written as the sample mean of a random variable $T_iY_i(1)/k - (1-T_i)Y_i(0)/(1-k)$. Thus, we can now apply asymptotic theorems.

First, we apply the law of large numbers, which states that the sample mean converges in probability to the population mean as the sample size tends to infinity.

$$\hat{\tau} \;\xrightarrow{p}\; \mathbb{E}\left\{\frac{T_iY_i(1)}{k} - \frac{(1-T_i)Y_i(0)}{1-k}\right\} \;=\; \mathbb{E}(Y_i(1)-Y_i(0)) \;=\; \text{PATE} \tag{26}$$

Next, to apply the central limit theorem, we need the variance of this random variable as well as its expectation. The variance can be derived using the previous result as,

$$\mathbb{V}\left\{\frac{T_iY_i(1)}{k} - \frac{(1-T_i)Y_i(0)}{1-k}\right\} \;=\; n\mathbb{V}(\hat{\tau}) \;=\; \frac{\sigma_1^2}{k} + \frac{\sigma_0^2}{1-k} \tag{27}$$

Thus, the central limit theorem implies,

$$\sqrt{n}(\hat{\tau} - \text{PATE}) \;\xrightarrow{d}\; \mathcal{N}\left(0,\; \frac{\sigma_1^2}{k} + \frac{\sigma_0^2}{1-k}\right) \tag{28}$$

which in turn gives the following asymptotic approximation,

$$\hat{\tau} \;\overset{\text{approx.}}{\sim}\; \mathcal{N}\left(\text{PATE},\; \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right) \tag{29}$$

where we used the fact that $n_1 = kn$ and $n_0 = (1-k)n$. With this approximation, we can compute the following asymptotic standard error, which is the estimated standard deviation of the asymptotic sampling distribution of our estimator,

$$\text{s.e.} \;=\; \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}} \tag{30}$$

Finally, using this standard error, we can construct the $(1-\alpha) \times 100\%$ asymptotic confidence interval,

$$[\hat{\tau} - \text{s.e.} \times z_{1-\alpha/2},\; \hat{\tau} + \text{s.e.} \times z_{1-\alpha/2}] \tag{31}$$

where z_τ is a critical value with $\Phi(z_\tau) = \tau$ and $\Phi(\cdot)$ being the distribution function of the standard normal random variable (e.g., $\Phi(1.96) = 0.975$).

2.4 Model-based Approach

As we saw in the survey sampling lecture notes, the model-based approach is rather straightforward. In the context of randomized experiments, we assume that there are two populations, one representing the distribution of $Y_i(1)$ and the other representing the distribution of $Y_i(0)$. Under the model-based approach, we specify a probability distribution for each of these populations and assume that we have one random sample from each population. The inference can then be drawn solely based on the characteristics of these two probability distributions.

3 Identification of the Average Treatment Effects

We have shown that in randomized experiments the randomization of treatment assignment enables the unbiased estimation of average treatment effects. In observational studies, however, the treatment is not randomized and therefore the treated units may systematically differ from the control units in terms of both observed and unobserved characteristics. This selection problem does not go away even if the sample size is increased. In general, we say that a parameter is *unidentifiable* if there exists no consistent estimator for it. That is, we have an *identification problem* if even infinite amount of data does not allow one to nail down the true parameter value. Identification problem cannot be solved by statistical methods. Instead, one must look for alternative research designs or assumptions. We emphasize that identification is different from estimation. Estimation addresses the question of how much one can learn from a finite sample whereas identification analysis examines how much one can learn from infinite amount of data.

3.1 Point Identification

We consider a set of assumptions that are typically invoked in order to identify the average treatment effect in observational studies. These assumptions justify almost all regression-based inference. The first assumption is called *unconfoundedness* and is defined formally as,

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid X_i = x \quad \text{for all } x. \tag{32}$$

This assumption is also called *no omitted variable bias*, *ignorability*, or *selection on observables*. The variables that are associated with both treatment assignment and outcome are called *confounders*. The above assumption states that all confounders are measured and conditional on these observed confounders, the treatment assignment is independent of potential outcomes. In other words, we assume that the treatment assignment is essentially "randomized" among those units who share the exact same values of covariates X_i. This is not the assumption one can directly test from the data. The reason is that units whose observed characteristics are identical may differ on unobserved confounders.

The second assumption is called *overlap* and is defined as,

$$0 < \Pr(T_i = 1 \mid X_i = x) < 1 \quad \text{for all } x. \tag{33}$$

This assumption ensures that all units in the population has non-zero probability of being assigned to the treatment and control group. That is, the assumption guarantees that all units in the control group can be used to infer unobserved counterfactual outcomes of the treatment group units. In program evaluation, this assumption implies that those individuals who are ineligible for the program should be excluded from analysis. Similarly, those who are automatically enrolled in the program should also be removed from the sample. In practice, however, one does not know the exact probability of receiving the treatment and so verifying the validity of this assumption can be difficult.

Under these two assumptions, we can show that the average treatment effect is identified. We use the law of iterated expectation,

$$\begin{aligned} \mathbb{E}(Y_i(1) - Y_i(0)) &= \mathbb{E}\left\{\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i)\right\}, \qquad (34) \\ &= \mathbb{E}\left\{\mathbb{E}(Y_i(1) \mid T_i = 1, X_i) - \mathbb{E}(Y_i(0) \mid T_i = 0, X_i)\right\}, \qquad (35) \\ &= \mathbb{E}\left\{\mathbb{E}(Y_i \mid T_i = 1, X_i) - \mathbb{E}(Y_i \mid T_i = 0, X_i)\right\}, \qquad (36) \end{aligned}$$

where the second equality follows from the unconfoundedness assumption: once we condition on X_i, the treatment assignment T_i and potential outcomes are independent, which means that we can further condition on any value of T_i. We note that $\mathbb{E}(Y_i \mid T_i = t, X_i)$ can be estimated by regressing Y_i on T_i and X_i and setting $T_i = t$ for $t = 0, 1$. In general, this conditional expectation function, $\mathbb{E}(Y_i \mid T_i, X_i)$, is called *regression function.* We note that the outer expectation in the above expression is the distribution over covariates X_i. The intuition here is that $\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i)$ is the conditional ATE given a set of observed characteristics X_i, which basically characterizes the treatment effect heterogeneity as a function of X_i. By averaging this quantity over the distribution of X_i, we obtain the overall average treatment effect. In practice, when X_i is high dimensional (i.e., it contains a large number of variables), this averaging is done by first computing the estimated conditional ATE given each unit's observed characteristics X_i and then computing the sample average of these conditional ATEs,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i). \qquad (37)$$

The central difficulty of the exogeneity assumption is that one typically needs to collect a large number of pre-treatment covariates in order to make the unconfoundedness assumption credible. However, a large dimension of X_i means that regression modeling becomes a challenging task. Ideally, one would like to apply *nonparametric regression* techniques in order to model $\mathbb{E}(Y_i \mid T_i, X_i)$ without imposing a strong functional form assumption. However, the amount of data required by such nonparametric regressions increases exponentially as the number of covariates increases. This problem is known as the *curse of dimensionality.* We will further discuss these issues later in the course.

3.2 Partial Identification

In many situations, however, the exogeneity assumption may not be credible. It is always possible that unobserved confounders exist. What can we learn from the data if we do not impose the exogeneity assumption? Relaxing exogeneity means that even after conditioning on observed confounders X_i, the potential outcomes of the treatment and control groups differ systematically from each other. Manski (2007) and others have addressed this problem by deriving the sharp (i.e., best possible) bounds on the quantity of interest without making the assumptions that are not justifiable. Such bounds can serve as a starting point of *partial identification analysis* where researchers can examine the identification power of additional assumptions by investigating how the bounds change as one introduces these assumptions.

Here, we illustrate this partial identification analysis by considering a simple example where we relax the exogeneity assumption in observational studies. Suppose that the outcome variable is bounded. Without loss of generality, we consider the case where the outcome variable is bounded between 0 and 1, i.e., $0 \leq Y_i \leq 1$. To derive the no-assumption bound on the average treatment effect for the PATE, we use the following decomposition,

$$\begin{aligned}
\mathbb{E}(Y_i(1) - Y_i(0)) &= \mathbb{E}(Y_i(1) \mid T_i = 1) \Pr(T_i = 1) + \mathbb{E}(Y_i(1) \mid T_i = 0) \Pr(T_i = 0) && (38) \\
&\quad - \mathbb{E}(Y_i(0) \mid T_i = 1) \Pr(T_i = 1) - \mathbb{E}(Y_i(0) \mid T_i = 0) \Pr(T_i = 0) && (39) \\
&= \mathbb{E}(Y_i \mid T_i = 1) \Pr(T_i = 1) + \mathbb{E}(Y_i(1) \mid T_i = 0) \Pr(T_i = 0) && (40) \\
&\quad - \mathbb{E}(Y_i(0) \mid T_i = 1) \Pr(T_i = 1) - \mathbb{E}(Y_i \mid T_i = 0) \Pr(T_i = 0). && (41)
\end{aligned}$$

The final expression still contains the potential outcomes because we cannot infer the counterfactual outcome of the treatment group, for example, from the observed outcome of the control group. This means that we cannot identify the two terms that involve counterfactual outcomes, i.e., $\mathbb{E}(Y_i(0) \mid T_i = 1)$ and $\mathbb{E}(Y_i(1) \mid T_i = 0)$. To derive the sharp bound, we replace these unidentifiable quantities with their upper and lower bounds (for example, for the lower bound, we set $\mathbb{E}(Y_i(0) \mid T_i = 1) = 0$ and $\mathbb{E}(Y_i(1) \mid T_i = 0) = 1$). The resulting bound is,

$$\begin{aligned}
&[\{\mathbb{E}(Y_i \mid T_i = 1) - 1\} \Pr(T_i = 1) - \mathbb{E}(Y_i \mid T_i = 0) \Pr(T_i = 0), && (42) \\
&\qquad \mathbb{E}(Y_i \mid T_i = 1) \Pr(T_i = 1) + \{1 - \mathbb{E}(Y_i \mid T_i = 0)\} \Pr(T_i = 0)] && (43)
\end{aligned}$$

The width of this bound is exactly one, which is half of the length of the original bound, $[-1, 1]$, we had before looking at the data. Thus, although the resulting bound still contains zero, the data are to some extent informative about the PATE.

From this no-assumption bound, a variety of partial identification analysis is possible by imposing additional assumptions on unobserved counterfactual outcomes, $\mathbb{E}(Y_i(0) \mid T_i = 1)$ and $\mathbb{E}(Y_i(1) \mid T_i = 0)$. For example, consider the assumption of perfect self-selection where units self-select themselves into the treatment status that gives a better outcome. Formally, this can be written as, $Y_i(1) \geq Y_i(0) \iff T_i = 1$ and $Y_i(1) \leq Y_i(0) \iff T_i = 0$. Under this scenario, we have $0 \leq \mathbb{E}(Y_i(0) \mid T_i = 1) \leq \mathbb{E}(Y_i(1) \mid T_i = 1) = \mathbb{E}(Y_i \mid T_i = 1)$ and $0 \leq \mathbb{E}(Y_i(1) \mid T_i = 0) \leq \mathbb{E}(Y_i(0) \mid T_i = 0) = \mathbb{E}(Y_i \mid T_i = 0)$. Thus, the bound becomes,

$$[-\mathbb{E}(Y_i \mid T_i = 0) \Pr(T_i = 0), \quad \mathbb{E}(Y_i \mid T_i = 1) \Pr(T_i = 1)] \qquad (44)$$

For example, to derive the lower bound, you set $\mathbb{E}(Y_i(1) \mid T_i = 0) = 0$ and $\mathbb{E}(Y_i(0) \mid T_i = 1) = \mathbb{E}(Y_i \mid T_i = 1)$. This bound still contains zero but is narrower than the no-assumption bound given in equation (43). In this manner, partial identification analysis answers the question of how much information different assumptions can bring when inferring an unidentifiable parameter.

References

Chattopadhyay, R. and Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in india. *Econometrica* **72**, 5, 1409–1443.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–960.

Humphreys, M., Masters, W. A., and Sandbu, M. E. (2006). The role of leaders in democratic deliberations: Results from a field experiment in São tomé and Príncipe. *World Politics* **58**, 4, 583–622.

Manski, C. F. (2007). *Identification for Prediction and Decision.* Harvard University Press, Cambridge, MA.

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

Robins, J. M. (1989). *Health Research Methodology: A Focus on AIDS (eds. L. Sechrest, H. Freeman, and A. Mulley)*, chap. The Analysis of Randomized and Non-randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Logitudinal Studies. NCHSR, U.S. Public Health Service, Washington, D.C.

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

Rubin, D. B. (1990). Comments on “On the application of probability theory to agricultural experiments. Essay on principles. Section 9” by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.