The Statistics of Causal Inference: The View from Political Methodology

Luke Keele*

December 9, 2014

Abstract

Many areas of political science focus on causal questions. Evidence from statistical analyses are often used to make the case for causal relationships. While statistical evidence can help establish causal relationships, it can also provide strong evidence of causality where none exists. In this essay, I provide an overview of the statistics of causal inference. Instead of focusing on statistical methods, such as matching, I focus more on the assumptions needed to give statistics estimates a causal interpretation. Such assumptions are often referred to as identification assumptions, and these assumptions are critical to any statistical analysis about causal effects. I outline a wide range of identification assumptions and highlight the design based approach to causal inference. I conclude with an overview of statistical methods that are frequently used for causal inference.

1 Introduction

One central task of the scientific enterprise is establishing causal relationships. Take one example from the international relations literature on the democratic peace. One well known finding is that democracies tend to engage in lower levels of interstate conflict. We can just treat this as a descriptive finding: democratic governance is correlated with lower levels of interstate conflict. This descriptive finding, however, begs a causal question: if a country becomes more democratic will it then engage in less conflict? Rarely are we content with statistical associations. Instead, we often seek to understand causal relationships.

Causality is something we all understand, since we use it in our daily life. It refers to the relational concept where one set of events cause another set of events. Causal inference is the

---

*Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 19130 Email: ljk20@psu.edu.
process by which we make claims about causal relationships. While causality seems a simple concept in everyday life, the establishment of causal relationships in many contexts is a difficult enterprise. Early models of causality focused on unique causes such as gravity. Gravity always causes things to fall to the earth and is the unique cause of that action. In biological and social applications outcomes rarely have unique causes, as causes tend to be contingent. In such contexts, the counterfactual model of causality is useful. Under the counterfactual model, rather than define causality purely in terms of observable events, causation is defined in terms of observable and unobservable events. Thus I say, if Iraq had been democratic, war would not have broken out. This is a counterfactual statement about the world that asserts that if a cause had occurred an effect would have followed. This counterfactual approach is based on the idea that some of the information needed to make a causal inference is unobserved and thus some assumptions must be made before I can make a causal inference.\(^1\)

In the social sciences, data and statistical evidence are often used to test causal claims. Over the last 15 years, the potential outcomes framework, a manifestation of the counterfactual model of causality, has come to dominate statistical thinking about causality. What is behind the popularity of this new approach? Why do some, myself included, view this framework as an improvement over the past and not simply a “re-labeling” of existing statistical concepts? First, I think the counterfactual approach has provided new insights into when data can reveal causal inferences. Specifically, there has been a renewed interest in the assumptions needed for causal inference and unpacking the exact meaning of those assumptions. The other key change is the emphasis on design and the design-based approach. As I discuss below, the phrase design-based approach does not have universal definition, but there is widespread agreement that causal inferences are stronger when steps are taken in the design stage to bolster assumptions before estimation.

In this essay I provide a roadmap to the statistics of causal inference. I divide the statistics of causal inference into three parts: identification strategies, the design-based approach, and statis-

\(^1\)See Hidalgo and Sekhon (2011) for a nice overview of different models of causality and the rise of the counterfactual model.
tical tools. I begin with an introduction to the concept of causal identification and identification analyses. An identification analysis identifies the assumptions needed for statistical estimates to be given a causal interpretation. Next, a researcher must select an identification strategy or research design. Here, I provide a brief overview of several common identification strategies.

Once an identification strategy has been selected, the analyst often bolsters the design through a design based approach. The design-based approach, as I define it, is a broad set of techniques that can make identification more credible generally without the use of statistical models and without using outcomes. Finally, I provide a brief overview of statistical tools like matching and inverse probability weighting, which are commonly applied statistical tools in causal inference. I also review how the mode of statistical inference changes when the focus is on causal effects. Through this structure I hope to clarify the distinctions between identification, design, and statistical analysis.

2 Identification

I begin with review of identification. Identification is an extremely important concept in the statistics of causal inference. One way to describe whether a statistical estimate can be given a causal interpretation is to discuss whether it is identified or not. Identification concepts are invoked (often implicitly) in any analysis that purports to present a causal effect. Identification has both a precise formal meaning and a number of more informal meanings. I do not present a formal definition. Instead I present an informal definition to convey how the word identification is used in the literature.

2.1 Basics of Identification

Informally, a parameter is said to be "identified' if changing the value of the true parameter that generated the data implies a different distribution of the observed data (Matzkin 2007, sec 3.1). As I outline below, identification is a central focus in causal inference. Identification problems are not confined to causal inference, however, as many statistical analyses face identification problems. For example, studies of ecological inference are based on the identification of mixtures
of probability distributions using only knowledge of the marginal distributions. More generally, Manski (1995) separates the problem of inference into two components: statistical and identification. Under the identification part of inference, we seek to describe the conclusions that can be drawn with an infinite sample. Studies of statistical inference focus on what can be learned in finite samples and as the sample size increases. If identification fails, nothing can be learned even if the sample is infinite.

A causal effect is (nonparametrically) identifiable when the distribution of the observed data is compatible with a single value of the treatment effect parameter. Conversely, we say that a causal effect is unidentifiable when the distribution of the observed data is compatible with several values of the true treatment effect measure. In particular, we are interested in nonparametric identification. In a nonparametric identification analysis, we formally prove which non-model based (functional form) assumptions are needed for identification. Nonparametric identification highlights the weakest set of assumptions needed for identification. In an nonparametric identification analysis, the analysts provides both a formal statement of the assumptions needed to identify a particular causal effect and a proof that those assumptions lead to an identified causal effect. For example, one could state and prove what assumptions must hold for a randomized experiment to identify a causal effect. More commonly one invokes an existing identification analysis and stipulates identification under that set of assumptions. An identification analysis (nonparametric or otherwise) precedes questions of statistical inference and estimation. Nonparametric identification refers to the possibility of correctly estimating a causal effect with some hypothetical set of data that is infinite in size and without reference to any specific statistical model. The ability to actually estimate a causal effect is another matter that depends on a specific data set. As I will outline later, nonparametric identification often leads to a preference for nonparametric estimation methods.

The potential outcomes framework (see, e.g. Rubin 1974), which often referred to as the Rubin Causal Model (RCM) (Holland 1986) is one formal way to demonstrate that causal questions in

\footnote{Linearity and additivity, for example, are model-based functional form assumptions.}
statistics are questions of identifiability. The RCM is the dominant model of causality at the moment. Like all models it is wrong, but it is also quite useful. In fact, the RCM is not the only model of causality that is embedded within a statistical framework. Dawid (2000) develops a decision theoretic approach to causality that rejects counterfactuals. Pearl (1995, 2009a) advocates for a model of causality based on nonparametric structural equations and path diagrams. Below, I use the RCM to explain the identification problem in a causal analysis.

In the potential outcomes model, each unit has multiple potential outcomes but only one actual outcome. Potential outcomes represent unit level behavior in the presence or absence of an intervention or treatment, and the actual outcome depends on actual treatment received. I denote a binary treatment with $D_i \in \{0,1\}$, though the treatment need not be binary. The potential outcomes are $Y_{iD} \in \{0,1\}$. The actual outcome is a function of treatment assignment and potential outcomes such that $Y_i = D_iY_{i1}+(1-D_i)Y_{i0}$. Under this framework, we can define various forms of the unit level causal effect of $D_i$, which are comparisons of unit level potential outcomes. One possible comparison is a difference in potential outcomes, $Y_{i1} - Y_{i0}$, but in general the comparisons can take different forms such as a ratio: $Y_{i1}/Y_{i0}$.

We cannot estimate this unit level causal effect since we do not observe the potential outcomes. The potential outcomes model formalizes the idea that the individual-level causal effect of $D_i$ is unobservable, which is sometimes called the fundamental problem of causal inference (Holland 1986). Instead, we focus on the average treatment effect (ATE):

$$ATE = \mathbb{E}[Y_{i1} - Y_{i0}]$$

This is known as a causal estimand. It is separate from the statistical estimator or a specific point estimate. In an identification analysis, we identify specific estimands. The ATE is the average difference in the pair of potential outcomes average over the entire population. Often causal estimands are defined as averages over specific subpopulations. For example, we might average over subpopulations defined by pretreatment covariates such as sex and estimate the ATE for
females only. When the estimand is defined for a specific subpopulation, it is said to be more local. Frequently, the average treatment effect is defined for the subpopulation exposed to the treatment or the average treatment effect on the treated (ATT).

\[ ATT = \mathbb{E} [Y_{i1} - Y_{i0} \mid D_i = 1] \]

Finally, we often define the relevant subpopulation partially in terms of potential outcomes. As I discuss later the most well-known estimand to be defined in terms of a subpopulation based on potential outcomes comes from instrumental variables.\(^3\)

I should note that I have followed common practice and written the estimands as averages. Causal identification rarely implies that only the middle (as represented by an average) of the treated and control distributions will differ. Analysts should always consider that causal effects might only be apparent at particular quantiles. I revisit this topic later when I consider methods of inference for causal effects.

For all these estimands, we still face an identification problem, since there are terms in the estimand that are unobservable. Even if we had samples of infinite size, we still could not estimate the average causal effect without observing both potential outcomes. Resolution of the problem requires a set of assumptions that warrant inferences based on observable quantities. Identification assumptions thus bridge theoretical and observable quantities. When identification assumptions holds it implies that one may then proceed to questions of statistical inference. Next, I use potential outcomes to clearly elucidate the unobservable quantities in the ATE estimand. I define \( \pi \) as the proportion of the sample assigned to the treatment condition. Using \( \pi \) I can

\(^3\)Here, I implicitly invoke the stable unit treatment value assumption (SUTVA), which permits the assumption that we are actually observing the potential outcomes associated with each treatment condition. I discuss SUTVA in more detail in the next section.
decompose the ATE as follows:

$$E[Y_i | D_i = 1] - E[Y_i | D_i = 0]$$

$$= \pi \{E[Y_{i1} | D_i = 1] - E[Y_{i0} | D_i = 1]\} + (1 - \pi) \{E[Y_{i1} | D_i = 0] - E[Y_{i0} | D_i = 0]\} \quad (1)$$

In this equation, the ATE is a function of five quantities. Without additional assumptions we can estimate three of those quantities directly from observed data. We can estimate $\pi$ using $E[D_i]$. We can also readily estimate $E[Y_{i1} | D_i = 1]$ and $E[Y_{i0} | D_i = 0]$ using $E[Y_i | D_i = 1]$ and $E[Y_i | D_i = 0]$. However, we cannot estimate $E[Y_{i1} | D_i = 0]$ and $E[Y_{i0} | D_i = 1]$ from the data without assumptions. That is, we face an identification problem, since these are counterfactual quantities. One is the average outcome under treatment for those units in the control condition, and the other is the average outcome under control for those in the treatment condition. No additional amount of data will allow us to estimate these quantities, we must find assumptions that allows for identification.

In causal inference, identification generally rests on the assumption that treatment status is independent of potential outcomes. Formally this assumption is:

$$Y_{i1}, Y_{i0} \perp \perp D_i.$$ 

Why does this assumption work? The expectation of the observed outcome conditional on $D_i = 1$ can be written as:

$$E[Y_i | D_i = 1] = E[Y_{i0} + D_i (Y_{i1} - Y_{i0}) | D_i = 1]$$

$$= E[Y_{i1} | D_i = 1]$$

$$= E[Y_{i1}]$$

where the last step follows from the assumption of independence. That is, taking the expectation of the observed outcome, provides the expectation of the potential outcome when independence
holds. Independence between treatment status and potential outcomes allows us to connect the observed outcomes to the potential outcomes. The same logic impiles that $\mathbb{E}[Y_i | D_i = 0] = \mathbb{E}[Y_{i0}]$. It follows from the above statements that

$$ATE = \mathbb{E}[Y_{i1} - Y_{i0}] = \mathbb{E}[Y_{i1}] - \mathbb{E}[Y(0)] = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0].$$

That is under the assumption of independence, the expectation of the unobserved potential outcomes is equal to the conditional expectations of the observed outcomes conditional on treatment assignment. The independence assumption allows us to connect unobservable potential outcomes to observed quantities in the data. When are we justified in assuming independence holds between the treatment and the potential outcomes? I take up that question in the next section.

### 3 Identification Strategies

Now I turn to identification strategies. An identification strategy is simply a research design (Angrist and Pischke 2010)\(^4\) Part of an identification strategy is an assumption or set of assumptions that will identify the causal effect of interest. To ask what is your identification strategy is to ask what research design (and assumptions) one intends to use for identification. The key distinction is that for one identification strategy, we can mechanistically ensure that independence between treatment status and potential outcomes holds. The following review of identification strategies is necessarily brief. Readers interested in an a more in-depth review from different perspectives should consult Angrist and Pischke (2009); Winship and Morgan (1999) and Rosenbaum (2010). Before reviewing identification strategies, I take up two additional topics: stable unit treatment

\(^4\)I should note that there are various inconsistencies in how identification strategy is defined since Angrist and Pischke (2009) define an identification strategy as “the manner in which a researcher uses observational data to approximate a real experiment.” This definition would seem to preclude experiments as identification strategy.
value assumption and threats to identification.

While the identification strategies below make a wide variety of assumptions about independence between treatment and potential outcomes, most of the strategies below also require the stable unit treatment value assumption (SUTVA) (Rubin 1986). SUTVA is made up of the two following components: 1) there are no hidden forms of treatment, which implies that for unit \( i \) under \( D_i = d \), we assume that \( Y_{id} = Y_i \) and 2) a subject’s potential outcome is not affected by other subjects’ exposure to the treatment.

The first component of SUTVA is often referred to as the consistency assumption in the epidemiological literature. The consistency assumption is somewhat controversial. Hernán and VanderWeele (2011) argue the consistency assumption must be evaluated by analysts since it links observed data to the counterfactual outcomes. They argue that in the absence of consistency, one would not know which counterfactual contrast is being estimated by the data. For example, if the treatment were “fifteen minutes of exercise," there are many different forms of exercise. They contend that it will be difficult to justify any decision making based on effect estimates since we may not know which form of exercise actually made the treatment effective. In contrast, van der Laan, Haight and Tager (2005) says consistency is an axiom which can be taken for granted, while Pearl (2010) maintains that consistency immediately follows so long as the causal model is correct. In some sense, there are elements of truth to both sides. Under an appropriate identification strategy, it is possible to estimate the effect of 15 minutes of exercise. However, generating policy recommendations may be difficult given the fact that the treatment may contain a large number of elements.

The second part of the SUTVA assumption tends to be a more serious problem in many social science settings. The problem is that if we treat a unit and that unit can then spread some of that treatment to a control unit or units, the comparison is no longer between treated and control, but between treated and partially treated. If one specifies a model of contagion for how the treatment spreads, one can make some progress toward identification, but if we no knowledge of treatment spillovers, causal parameters will not be identified. Taking interference into account is currently
Next, I briefly review the two most common threats to identification. The first is confounding due to a common cause. Figure 1(a) contains a causal diagram of confounding. Confounding in many statistical texts is referred to as a spurious relationship. In this diagram, we might think that $D_i$ is a cause of $Y_i$, but in fact $L_i$ is cause of both while $D_i$ is actually independent of $Y_i$. However, if we estimate the statistical association between $D_i$ and $Y_i$, we will find them spuriously correlated as represented by the dashed line. Concerns about confounding then are concerns about identifiability.

The next threat to identification stems from when $D_i$ and $Y_i$ both condition on a common effect. Rosenbaum (1984) identified this threat to identification as conditioning on a post-treatment covariate. Figure 1(b) represents this situation. The box around $L_i$ represents conditioning. Often this is defined as selection since it arises from selection on the dependent variable. Selection tends to be a more subtle problem than confounding, since it can cause a failure of identification even when the treatment is independent of potential outcomes and confounding is ruled out. An example is useful. Let’s say that I administer a treatment designed to enhance governance in local councils. Here, let’s assume the treatment is independent of potential outcomes. The outcome is measured using voting records in council elections. I return a year later and find that for many councils the voting records were lost. Unless missing data status is independent of potential outcomes, identification will fail. The analysis will now implicitly condition on whether
voting records are observable.

### 3.1 Randomized Experiments

The randomized experiment is often considered the “gold standard” among identification strategies. Here subjects are assigned to $D_i$ via some random mechanism like the toss of a fair coin. The typical estimand in a basic randomized design is the average treatment effect, which under this identification strategy is equivalent to the average treatment effect on the treated. Of course, randomization does not necessarily imply that only averages will differ across the treated and control groups. As I discuss later, other features of the treated and control distributions may also be of interest. I do not spend anytime on the particulars of randomized experiments. That is covered in much greater detail elsewhere (Rosenbaum 2010; Gerber and Green 2012, ch. 2). Instead, I want to convey what it is special about this identification strategy. The key strength of experiments is the researcher has the ability to impose independence between treatment status and potential outcomes on a set of units because he or she has control over the assignment process. As I outlined above, if the treatment is independent of the potential outcomes, then the treatment effect parameter is identified. Randomization makes such independence hold. Short of incorrectly generating random treatment assignments, under this identification strategy the analyst knows that independence holds which allows the researcher to assert that the treated and control groups will be identical in all respects, observable and unobservable, save receipt of the treatment with arbitrarily high probability as the sample size grows large. This implies that randomization allows us to rule out confounding due to a common cause.

Experiments are not without their difficulties. One key threat to identification in an experiment is attrition, which is a form of selection. That is, subject outcomes are not available after randomization, and this missingness is a function of randomization. Another complication in experiments is noncompliance. It is often the case that subjects do not comply with their assigned treatment status. A full discussion of noncompliance and attrition are beyond the scope of this article, below I discuss noncompliance, since it forms a separate identification strategy.
Finally, a randomized experiment identifies the treatment effect within the population used in the study. This treatment effect may or may not extrapolate to other populations. To ensure valid extrapolation, one either needs random sampling in addition to randomization of treatment or additional assumptions. Given this fact, experiments are often said to be internally valid, but they may lack external validity. Whether this is a feature or a bug is a matter of substantial disagreement. Many of those who label themselves as interested in causal inference tend to value internal validity over external validity. If our concern is observing a causal effect, we might place more value on a well executed lab experiment than an observed association from a very large representative sample of data. As we will see, some identification strategies explicitly work based on comparisons of comparable but unrepresentative subpopulations. I explain the logic behind the value placed on internal validity in the next section.

3.2 Natural Experiments

The next identification strategy is based on natural experiments. I define a natural experiment as a real-world situation that produces haphazard assignment to a treatment. The hope is that a natural intervention will create as-if randomized treatment assignment and produce independence between treatment assignment and unit level potential outcomes. Of course, randomization in an experiment is a fact, while haphazard treatment assignment often requires considerable judgment to justify it as as-if random. The circumstances of the natural experiment speak to whether the claim of as-if random assignment is credible, but there is no way to know whether assignment is as good as randomized. An example is helpful.

Lyall (2009) seeks to understand whether indiscriminate violence increased insurgent attacks. To that end, he exploits shelling patterns by Russian troops in Chechnya that appear to be at worst indiscriminate and at best as-if random. He does find that the treatment, being shelled, appears to be uncorrelated with pre-treatment covariates as would be the case in a randomized experiment. The difficulty is that unlike with randomization we don’t know whether the patterns are truly random since they are beyond the control of the analyst. As such, natural experiments
often require careful justification for the as-if random nature of assignment. The basic template, however, is present in the study by Lyall (2009). He finds a real-world situation that appears to mimic a randomized experiment. Exploiting such circumstances is often a very credible identification strategy. Like randomized experiments, the focus is on internal validity. We have no way of knowing whether the causal effect in Chechnya would hold in another circumstance, but what we hope to observe is a causal effect operating in relative isolation from the very real threats of confounding.

3.3 Instrumental Variables

Informally, an instrument is a random push to accept a treatment, but the push can only affect the outcome if it induces units to take the treatment. Holland (1988) outlined the randomized encouragement design as the prototype of an instrument. He described this design as an experiment where some participants are encouraged to exercise. While subjects are randomly encouraged to exercise, subjects then select their exposure to the exercise treatment in that they select whether to exercise or not. Moreover, some of those assigned to the non-exercise arm will decide to exercise. Later all participants are measured on the outcome. Here, we might use a measure of lung function as an outcome. There are two effects of interest in designs of this type. In this design, the effect of being assigned to encouragement is identified since this has been randomly assigned. This estimand is often called the intention-to-treat (ITT) effect. This estimand tells us whether encouragement changes the outcome. Under additional assumptions, the method of instrumental variables (IV) identifies the effect of the treatment, exercise, as opposed to the effect of being assigned to exercise encouragement (Angrist, Imbens and Rubin 1996). Specifically, IV identifies the average effect among those induced to take the treatment by a randomized encouragement. The IV estimand is often referred to as either the complier average causal effect (CACE) or the local average treatment effect (LATE). The IV estimand is local since it is defined for a subpopulation: the compliers. However, this subpopulation is defined in terms of potential outcomes (Angrist, Imbens and Rubin 1996).
For IV to provide valid causal inferences, the five assumptions outlined by Angrist, Imbens and Rubin (1996) must hold. The assumptions needed for the IV estimand to be identified are (1) ignorable (as-if random) assignment of the encouragement; (2) the stable unit treatment value assumption (SUTVA); (3) no direct effect of the instrument (here encouragement) on the outcome also known as the exclusion restriction; (4) monotonicity; and (5) the instrument must have a nonzero effect on the treatment. The first two assumptions are identical to those need to identify the ITT effect. The other three are additional assumptions needed to identify the CACE.

Real life circumstances can create circumstances that mimic the randomized encouragement design. More broadly, we can define an instrument as a haphazard nudge to accept a treatment. Here, IV becomes identification strategy based on a type of natural experiment. Hansford and Gomez (2010) is one example of using IV as a natural experiment identification strategy. They seek to understand whether lower turnout reduces the vote share for the Democratic party. Hansford and Gomez (2010) exploit the fact that rainfall appears to decrease turnout on election day. They use rainfall as an as-if random discouragement for turnout. If rainfall is a valid instrument, this allows them to identify the local effect of turnout on vote share among the counties discouraged to vote by rain on election day. While the IV identification strategy can be credible, when used as a natural experiment it requires great care. See Bound, Jaeger and Baker (1995) for one example of a fairly spectacular failure of IV. See Sovey and Green (2011) for a more detailed overview of the IV identification strategy.

3.4 Regression Discontinuity Designs

The regression discontinuity (RD) design is another identification strategy that is typically classified as a type of natural experiment. In a regression discontinuity design, assignment of the binary treatment, $D_i$, is a function of a known continuous covariate, $S_i$, usually referred to as the forcing variable or the score. In the sharp RD design, treatment assignment is a deterministic function of the score, where all units with score less than a known cutoff in the score, $c$, are assigned to the control condition ($D = 0$) and all units above the cutoff are assigned to the treatment condition.

---

5See Sovey and Green (2011) for a more detailed introduction to IV analysis assumptions.
$D_i = 1$). In the fuzzy RD design, assignment to the treatment is a random variable given the score, but the probability of receiving treatment conditional on the score, $P(D_i = 1|S_i)$, must jump discontinuously at $c$. This implies that it is possible for some units with scores below $c$ to receive the treatment. The fuzzy RD design results in an equivalence between RD and IV (Hahn, Todd and van der Klaauw 2001). See Lee and Lemieux (2010) for a much lengthier description of RD designs.

Hahn, Todd and van der Klaauw (2001) demonstrate that for the sharp RD design to be identified the potential outcomes must be a continuous function of the score in the neighborhood around the discontinuity. Under this continuity assumption, the potential outcomes can be arbitrarily correlated with the score, so that, for example, people with higher scores might have higher potential gains from treatment. The continuity assumption is a formal statement of the idea that individuals very close to the cutoff but on opposite sides of it are comparable or good counterfactuals for each other. Thus, continuity of the conditional regression function is enough to identify the causal effect at the cutoff.

The RD design is another example of where the estimand changes as a function of the design. The RD design identifies a local average treatment effect for the subpopulation of individuals whose value of the score is at or near $c$. Estimation of this treatment effect proceeds by selecting a subset of units just above and below the discontinuity and calculating the difference across these two groups. So the estimand is restricted to a subset of units on either side of the threshold that are thought to be good counterfactuals. In this design, it is only possible to identify the treatment effect among a small subpopulation around the cutoff.

Lee and Lemieux (2010) note one strength of the RDD is that it is a design. Like randomization, some decision-maker must implement a treatment assignment mechanism based on a continuous score and a cutoff for a population of subjects. Lee and Lemieux (2010) emphasize this aspect of an RDD which distinguishes it from many natural experiments that rely on an instrument such as rainfall, which is certainly stochastic in some sense, but is not a controlled treatment assignment mechanism. Moreover, RD designs have gained further credibility
by recovering experimental benchmarks (Cook, Shadish and Wong 2008; Green et al. 2009).

### 3.5 Selection on Observables

Under the “selection on observables” identification strategy, the analyst asserts that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow, Cain and Goldberger 1980). Under this assumption, there are no unobservable differences between the treated and control groups. This assumption has a number of different names, which include “conditional ignorability” and “no omitted variables.” All of these are statements of the same idea: we seek to make the treatment independent of the potential outcomes conditional on observed covariates. Critically, the selection on observables assumption is nonrefutable, insofar as it cannot be verified with observed data (Manski 2007).

Given this set of “correct” covariates, we can use an statistical adjustment methods such as regression, matching or weighting to make conditional independence hold. In regression terms, this implies that we tend to prefer longer specifications to shorter specifications. Of course, there are dangers in pursuing overly long specifications. While we need to include all covariates that predict the outcome and treatment, we cannot condition on any covariates that are affected by treatment (Rosenbaum 1984). Even in a randomized experiment, conditioning on covariates that are affected by the treatment will bias our estimate of the treatment effect. This is sometimes known as over or bad control. See Angrist and Pischke (2009, pg. 69) for an accessible review of the formal statement of the bias that arises from controlling for post-treatment covariates. In an experiment, we can clearly delineate between the pre-treatment and post-treatment time periods. In observational data, that is often more difficult. In survey data, for example, it can be difficult to delineate any covariates as either pre or post-treatment.

In a further complication, Pearl (2009a,b) warns that adjustment for certain types of pre-treatment covariates can cause bias. This is known as “M-Bias” and arises from conditioning when there is a particular structure of unobservable covariates that create what is known as a “collider.” Ding and Miratrix (Forthcoming) show that while M-Bias is generally small, there are
rare cases where blind inclusion of pretreatment covariates can induce bias. As such, one must choose specifications with some care. I can’t emphasize enough that selection on observables is a very strong assumption. It is often difficult to imagine selection on observables is plausible in many contexts. Generally, selection on observables needs to be combined with a number of different design elements before it becomes plausible. I outline design elements in the next section.

3.6 Selection on Observables with Temporal Data

As I noted above, the selection on observables identification strategy requires that all differences between treated and control are observable. We can weaken this assumption when we observe units across multiple time periods. When there are data at multiple time periods, three different identification strategies are possible: fixed effects, differences-in-differences, and identification based on lags. See (Angrist and Pischke 2009, ch. 5) for a more in-depth overview of these related identification strategies.

Under the fixed effects identification strategy, if we use repeated observations on individuals, we can control for unobserved confounders that are time invariant. Therefore if confounders are time invariant it doesn’t matter if they are unobserved. However, we must also assume that the treatment effect is linear and additive, which is a strong constraint on how units respond to treatment. Differences-in-differences (DID) is a second identification strategy based on panel data. Angrist and Pischke (2009, pg. 228) describe DID as a fixed effects identification strategy using aggregate data. Here the key identifying assumption is that trends in the outcome would be the same across treated and control groups in absence of the treatment. That is we must assume that no other events beside the treatment alters the temporal path of either the treated or control groups.

The next identification strategy conditions on unobservables in a indirect fashion using past outcomes. Under this identification strategy, we assume selection on observables, but we condition on some number of lags of the outcome. Why is this an improvement over simply conditioning on
observables? The key insight is that lagged outcomes are a function of both observable covariates and unobservables. As such, if we condition on lagged outcomes we can indirectly condition on unobservables. The method of synthetic case control relies on this identification strategy (Abadie and Gardeazabal 2003; Abadie, Diamond and Hainmueller 2010).

While all these methods do allow for conditioning on unobservables, they all require the unobservables to have a very specific configuration. For all three strategies, the key assumption remains untestable. See Arceneaux, Gerber and Green (2006) for one example of where identification based on lags fails. This should serve a useful reminder that identification under any version of the selection on observables assumption is fraught with uncertainty.

3.7 Partial Identification

The goal under most identification strategies is point identification—identification of a single parameter that describes the causal effect of $D_i$. An alternative approach is to instead place bounds on the treatment effect, which can typically be done with weaker assumptions. The method of partial identification is most closely linked to the work of Manski (1990, 1995). See Mebane and Poast (2013) and Keele and Minozzi (2012) for examples in political science. Under partial identification, the analyst acknowledges that there is a fundamental tension between the credibility of assumptions and the strength of conclusions. As such the analysis proceeds by starting with the no-assumption bounds and adding assumptions about the nature of treatment response or assignment. By adding the assumptions individually, it allows one to observe exactly which assumption provides an informative inference. Assumptions can also be combined for sharper inferences. Under the strict Manski approach, typically one cannot rule out a zero order treatment effect.

The partial identification strategy can be very useful. First, the use of assumptions is completely transparent. In many statistical analyses, there is no sense of which assumption is essential for identification. The discipline of adding assumptions in a specific order and debating the credibility of those assumptions is an important exercise. Second, it can be applied to any
identification strategy. Lee (2009) uses a partial identification approach for randomized experiments with missing outcome data. Balke and Pearl (1997) use partial identification to relax the monotonicity assumption and exclusion restriction under the instrumental variables identification strategy. Finally, partial identification also underpins many forms of sensitivity analysis.

### 3.8 Mediation Analysis

I conclude this section with one final identification strategy that is rather different from those above. In an analysis of causal effects, we can broadly define three types of effects: total, direct, and indirect effects. The total effect is equivalent to the average treatment effect. In a mediation analysis, we seek to decompose the total effect into indirect and direct effects. One criticism of the total effect is that it cannot tell the analyst why the treatment works only that it does or not. In a mediation analysis, the analysts posits a causal mechanism which depends on $M_i$ known as a mediating variable, which occurs post-treatment and is assumed to be affected by the treatment. The causal mediation effect represents the indirect effect of the treatment on the outcome through the mediating variable (Pearl 2001; Robins 2003). While the indirect effect represents the effect of the treatment through $M_i$, the direct effect represents the effect of the treatment through all other possible mediators. The goal in a mediation analysis is to decompose the total effect into its indirect and direct components.

Identification in a mediation analysis proceeds in two parts. First, one makes the case for identifiability of the total effect. Identification of the direct and indirect effects require an additional assumption. Typically analysts use an assumption known as sequential ignorability, which rules out confounding between $M_i$ and $Y_i$ (Imai et al. 2011). That is the analysts must assume that all pre-treatment covariates that might confound the relationship between the mediator and outcome are observed. Thus the focus here is on the identification assumptions for the indirect and direct effects, while identification of the total effect depends on one of the identification strategies listed above. As such, this identification strategy is generally secondary since one must first make a case for the identifiability of the total effect. If identifiability of the total effect is
doubtful, there is little use in pursuing a mediation analysis.

3.9 Reasoning About Assumptions

I conclude this section by highlighting one of the more important skills needed to successfully test for causal effects. My argument for this skill is based on the fact that the plausibility of an identification strategy depends on the empirical context. For every identification strategy outlined above, one can find contexts where it is plausible and other contexts where that same strategy is indefensible.

Take the selection on observables identification strategy which is generally viewed as the weakest identification strategy. Sekhon and Titiunik (2012) present an example of estimating incumbency effects based on the redistricting process where selection on observables is credible. In their example, voters are moved into an incumbency treatment in the redistricting process. Under redistricting, we are able to identify the decision makers in the treatment selection process as state legislators. Moreover, we know state legislators use census data on race and election data on registration, turnout rates, and vote returns to decide how to draw districts. As such, we have good reason to believe that the specification of our statistical models will be nearly correct since we can simply model the selection process that occurs when states redistrict. Moreover, Sekhon and Titiunik (2012) note that treatments assigned in the redistricting process share an important aspect of experiments: the individuals in charge of assigning treatment are separate from the population that receives the treatment. As such, while state legislators rely on observable measures such as as vote share, they do not consider any individual-level characteristics of voters. Importantly, this implies that unobservables should not play an important role in the treatment selection process. Thus redistricting makes selection on observables a plausible identification strategy.

Take DID as a second example. Gordon (2011) is one example where a DID identification strategy is highly plausible. Alternatively, Keele and Minozzi (2012) outline an example where a DID identification strategy generally fails. As such, reasoning about identification assumptions is
a critical part of any statistical analysis that purports to be causal. Since untestable assumptions are unavoidable in causal inference, it only through careful understanding of those assumptions that one can make a case for their plausibility in a specific empirical context. When an analysis is meant to be causal, the researcher must think deeply about the assumptions and part of the analysis should be a well reasoned defense of the identification strategy. Qualitative information is often critical for defending the identification strategy. Reasoning about assumptions is often not part of a statistical analysis, but it must be when the goal is to identify causal effects.

A number of important advances in causal inference stem from a re-articulation of identification assumptions in a way that allows for a better understanding of those assumptions. For example, Lee (2008) developed a useful way to interpret the continuity assumption in the RD design. He defines the score as $S_i = W_i + e_i$, where $W_i$ represents efforts by agents to sort above and below $c$ and $e_i$ is a stochastic component. When $e$ is small, this implies that agents are able to precisely sort around the threshold, and treatment is mostly determined by self-selection and identification is less plausible. However, when $e$ is larger agents will have difficulty self-selecting into treatment, and whether an agent is above or below the threshold is essentially random. This behavioral interpretation of the continuity assumption allows aids in the assessment of the RD design.

The writing of IV in the potential outcomes framework is another example of how restating assumptions can be incredibly important. Angrist, Imbens and Rubin (1996) take the traditional statement of IV assumptions based on covariance restrictions and restate them into a form that allows for reasoning about their plausibility. Many of the mistakes that are made with IV as a natural experiment identification strategy could be avoided if researchers used the potential outcomes framework to reason about the IV assumptions. One way to do this is to use the randomized encouragement design as a template for any IV based natural experiment. I find that this generally helps understand whether IV assumptions are plausible in a given setting. In sum, reasoning about identification assumption is a critical skill.
4 The Design-Based Approach

I now turn to the topic of the design-based approach. Throughout the causal inference literature one will invariably notice many references to the importance of design and a general emphasis on the design-based approach. Unfortunately there isn’t a widely agreed upon definition of what it means to use a design-based approach. Dunning (2012) maintains that only natural experiments can be classified design-based. Imbens (2010, pg. 403) uses a much broader definition saying that under the design-based approach the analyst places an explicit emphasis on reducing heterogeneity, clarity about identifying assumptions, a concern about endogeneity, and the role of research design.

We might define the design-based approach by saying it is a mode of statistical analysis that emphasizes design rather than statistical modeling. This begs the question of what is design. Rubin (2008, p.810) defines design as all contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data. The reason design is understood to be important is that randomized experiments rule out unobserved confounding through the use of randomization a design based element applied before any collection of outcome data. As such, analysts can use the design to develop analyses that are less sensitive to confounding. For example, selection of an identification strategy is itself one part of the design which can clearly contribute to reduced sensitivity to confounding.

Here I outline a non-exhaustive list of important insights and techniques that have become part of the the design-based approach. Many of the topics below are quasi-experimental techniques that allow the analyst to argue that he or she is more likely to distinguish treatment effects from plausible alternatives or biases. As such, these methods can generally be combined with any identification strategy to rule out bias.

---

6Dunning generally uses the phrase design-based inference instead of design-based approach. I exclusively use the term design-based approach to avoid confusion with an older use of the term design-based inference used in the literature on survey sampling.
4.1 Reducing Heterogeneity

As I noted earlier, within the statistics of causal inference, interval validity is often valued over external validity. Successful natural experiments often focus on unrepresentative portions of the population where heterogeneity is lower. In general, there tends to be a greater concern about bias rather than efficiency in the analysis of observational studies. Why is efficiency a secondary concern in observational studies? The basic insight is from Cochran and Chambers (1965) who demonstrates that if there is a fixed bias that does not decrease as the sample size grows, then as the sample size increases this bias will dominate the mean squared error for the estimate of the treatment effect. In other words, increasing the sample size can shrink the confidence intervals to a point that excludes the true treatment effect point estimate. In a randomized experiment, where the estimate is known to be unbiased adding additional observations simply increases power. In an observational study any additional data that contributes to the heterogeneity may increase bias.

In general the call to reduce heterogeneity arises from differential concerns about sampling uncertainty and uncertainty from unobserved confounding. In observational data, the bias from unobserved confounders amounts to a far greater source of uncertainty than does a limited sample size. Increasing the sample size, moreover, does nothing to reduce the bias from unobserved confounders. Rosenbaum (2004, 2005a) has analytically demonstrated that reducing unit heterogeneity in observational data reduces sensitivity to bias from unobserved confounders. Reducing unit heterogeneity amounts to restricting the analysis to a more homogeneous subset of the entire data set. One might argue that the concomitant reduction in sample size will reduce the power to detect treatment effects, but this is not the case. Rosenbaum (2004, 2005a) proves that when treatments are nonrandomly assigned reducing unit heterogeneity reduces both sampling variability and sensitivity to bias from unobserved covariates. The RD design embodies this approach to heterogeneity. In an RD design, casual inferences are possible since one focuses on the much smaller homogenous population just above and below the cutoff. In short, there are strong reasons for focusing on small samples where differences across treated and control are reduced.
not by statistical means but by the design.

This move to reduce heterogeneity has lead to a specific practice in observational studies. Sometimes it is quite difficult to find a control group that we judge to be similar enough to the treated group. In short the analyst judges that there is too much heterogeneity across the two groups. Often this occurs because there are treated observations that are very different from any of the control units. One solution is to drop the treated units from the study and restrict the analysis to a subset of the treated that are comparable. Crump et al. (2009) and Rosenbaum (2012) both developed algorithms for dropping incomparable treated observations. See Zubizarreta et al. (2013) and Keele, Titiunik and Zubizarreta (2014) for examples of analyses of this type. Importantly these methods change the estimand. As soon as a single treated unit is dropped, the estimand is some more local version of the ATT. The difficultly is that we now longer have a well defined estimand. As such a tension develops between having a well-defined causal estimand and making a credible claim that treated and control are comparable in all observable respects.

Is this defensible? I would argue that generally yes it is. Identification under the RD design presents a similar dilemma. Strictly speaking the causal effect is identified exactly at the cutoff, but in practice, we use some subset of of observations above and below the cutoff. While there are a number of principled methods for selecting this neighborhood, we are selecting a somewhat arbitrary set of the treated units that are deemed comparable to the controls. As Rosenbaum (2012) notes, “...often the available data do not represent a natural population, and so there is no compelling reason to estimate the effect of the treatment on all people recorded in this source of data...” In general, I think it is not worth holding the estimand inviolate in the face of observable bias. So researchers have two choices when subjects lack comparability. Give up and declare the identification strategy implausible, or alter the estimand and focus on a subset of the sample where heterogeneity is not a threat. If the analysts adopts the later strategy, they should be quite clear that the estimand had to change in order to make the identification strategy credible.

See Imbens and Kalyanaraman (2012) and Calonico, Cattaneo and Titiunik (2013) for recent methods on selecting the neighborhood.
4.2 Falsification Tests

Falsification tests come in various forms, but generally focus on testing for treatment effects in places where the analysts know they should not exist. Causal theories may do more than predict the presence of an association; causal theories may also predict an absence of causal effects. When we find causal effects where they should not be, this is often a sign of hidden confounders and a failure of the identification strategy.

Rosenbaum (2002b) relates a useful example of using falsification test. In a study of treated and control groups, researchers were interested in whether eating fish contaminated with methylmercury caused chromosomal damage. In this study, the researchers used a selection on observables identification strategy in forming the treated and control groups, where the treated group was known to have consumed contaminated fish. One way we might understand whether selection on observables is reasonable is to use a falsification test. We cannot prove selection on observables holds, but we may find clear evidence that it does not hold. In the study, researchers collected data on a number of health related outcomes including whether subjects had asthma. There is currently no evidence that methylmercury causes asthma in any form. Researchers could then test for a treatment effect on asthma since it is an outcome known to be unaffected by the treatment. The presence of an effect on asthma would serve as evidence against the selection on observables assumption. That is, a treatment effect on asthma indicates that there is some unobservable difference across the treated and control groups that creates a treatment effect where none should exist. Falsification tests are often used with RD designs. In an RD design, we shouldn’t find that the discontinuity has an effect on any pretreatment covariates. Falsification tests of this type are often referred to as placebo tests.

4.3 Sensitivity Analysis

Sensitivity analyses are another element of a design-based approach. Many sensitivity analyses are based on a partial identification strategy, where bounds are placed on quantities of interest while

---

8The original study was conducted by (Skerfving et al. 1974).
a key assumption is relaxed. The phrase “sensitivity analysis” is often used informally. Formally, a sensitivity analysis is designed to quantify the degree to which a key identification assumption must be violated in order for a researcher’s original conclusion to be reversed. A sensitivity analysis provides a quantifiable statement about the plausibility of an identification strategy. If a causal inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions. The first sensitivity analysis explored whether it was possible for an unobserved confounder to explain the leftover variation in lung cancer rates after accounting for the association with smoking (Cornfield et al. 1959). While a sensitivity analysis can be conducted for any identification strategy, most sensitivity analyses focus on the selection on observables assumption (Imbens 2003; Rosenbaum 1987). For many identification strategies, specific forms of sensitivity analysis have not yet been developed.

Briefly, I outline the logic behind one form of sensitivity analysis. Rosenbaum (2002b) has developed a method of bounds to understand whether the selection on observables identification assumption is sensitive to the presence of a hidden confounder. Under this method, one places bounds on quantities such as the treatment effect point estimate or p-value based on a conjectured level of confounding. That is, the analyst states that he or she thinks the level of the confounding is a given magnitude. For that level of confounding, one can calculate bounds on the treatment effect point estimate. If zero is included in those bounds, a failure of the identification strategy would reverse the study conclusions for that level of confounding. One can vary the level of confounding to observe whether a small or large amount of confounding would reverse the study conclusions.

4.4 Pattern Specificity

I conclude this section with one final observation. Outside of identification strategies based on randomized experiments, statistical results from a single analysis are rarely considered to definite proof of a causal relationship. Instead analysts demonstrate causal relationships by building a multifaceted pattern of evidence. Rosenbaum (2005b) uses the phrase “pattern specificity” to
describe the evidence building process needed in a causal analysis. The concept behind pattern specificity is simple: one should test as many relevant implications of a causal hypothesis as possible. Confirmation of each additional implication strengthens the evidence for a causal effect. Thus a pattern of specific confirmatory tests provides better evidence than a single test. As Cook and Shadish (1994, pg. 95) write: “Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations.” Under pattern specificity, part of the design is the generation and testing of a large number of hypotheses based on the causal theory. If all hypotheses are confirmed, it lends greater credibility to the causal theory. Many of the techniques described above are often key elements in pattern specificity as one might use falsification tests and sensitivity analysis as part of a single research design.

In this section, I have highlighted the importance of the design-based approach. In general, causal analysis under a design-based approach seeks a plausible identification strategy and then often employs the techniques above to bolster the credibility of that strategy. While none of these techniques in isolation can rule out the presence of hidden bias, they can often increase the credibility of many identification strategies.

5 Tools for Causal Inference

In this final section, I provide an overview of a number of methods that are often used in the analysis of treatment effects. I provide little detail on these various methods as they are covered in much greater depth elsewhere. The appendix contains links and references to the software tools available for the methods discussed below.

5.1 DAGs

One tool that is sometimes applied in the literature on causal inference is that of causal graphs or directed acyclic graphs (DAGs) (Pearl 1995). Unlike the other methods outlined in this section, DAGs are a tool for identification as opposed to statistical analysis. DAGs are often useful for reasoning about causal structure, since they allow us to formalize identification concepts in a graphical manner. From a given graph, we can derive nonparametric identification results and
identify which variable or sets of variables are necessary for identification. Pearl (2009a) maintains that DAGs are essential to any causal analysis. A more limited view of DAGs would say that a DAG is meant to represent the analysts reasoned view of the causal structure between a set of variables. Once the DAG is written down, it can be defended as a causal representation of a theory. Based on that structure one can then derive whether a causal effect is nonparametrically identified or not. However, in cases where identification conditions are well-understood a DAG adds little to the analysis. That is, in a well conducted randomized experiment or a good natural experiment, the design creates such a simple DAG that they are of little use. However, under selection on observables, DAGs can be a useful way to clarify the necessary conditioning set for identification to hold.

5.2 Estimation Methods

The number and variety of statistical methods used in the estimation of causal effects is well beyond the scope of this article. Below I provide a high-level summary of the methods used. While identification is strictly speaking separate from estimation, an emphasis on nonparametric identification tends to influence estimation. When nonparametric identification holds, it implies a valid nonparametric estimator. Thus if a convincing case can be made for nonparametric identification, in theory nonparametric estimation provides a straightforward way to estimate the identified treatment effect.

What is the problem with straying too far from the implied nonparametric estimator? The danger is that if the analysts selects an overly restrictive method of statistical estimation, estimates of identified causal effects will be biased due to overly restrictive modeling assumptions. For example, assume that selection on observables holds but unit response to treatment is nonlinear. If the analysts applies an estimation method that assume a linear response to treatment, functional form mis-specification may bias the effect such that one might think the treatment is without effect when in fact the effect is simply nonlinear. It would be unfortunate to waste identification due to functional form misspecification. The possibility of bias from functional form misspecification
leads to a strong preference for nonparametric or semiparametric estimation methods. While data or other practical limitations may make nonparametric estimation infeasible, as I highlight below many of the methods used in causal analyses tend to be either nonparametric or semiparametric.

### 5.2.1 Regression

Here, I use the phrase regression broadly to include not only via least squares but also models with nonlinear links such as logistic regression models. The primary use of regression models is to adjust for confounders under selection on observables. However, regression models may be used in conjunction with most of the identification strategies described in this essay. For example, regression based methods are often used under both the instrumental variables and RD design identification strategies. This illustrates why statistical techniques are secondary to identification strategies. The credibility of the estimator is often a function of the identification strategy, and many methods of estimation have some applicability across different identification strategies.

Many researchers view regression models as estimators of causal effects with suspicion given the strong functional form assumptions needed. Regression models need not be wedded to restrictive functional forms. They can me made more flexible through the use of splines or kernel methods (Keele 2008; Hainmueller and Hazlett 2013). Hill, Weiss and Zhai (2011) and Hill (2011) shows how very flexible nonparametric methods that are loosely regression based can be used to estimate causal effects.

Many critiques of regression, however, extend beyond the restrictive functional form. Regression models have been strongly critiqued as a method of the estimation of causal effects (Berk 2006; Freedman 2005). For example, regression models often produce treatment effect estimates based on extrapolation that is not readily observable to the analyst. The basic interpretation of the regression coefficient as a marginal effect can lead to causal interpretations of regression models where identifiably is questionable. That is the statement that the $\beta$ coefficient in a regression model is the amount $Y$ changes for a unit change in $X$ is an implicitly causal statement that is unjustified without careful consideration of the identification strategy.

---

9There always exceptions. See Angrist and Pischke (2009, ch. 3) for a dissenting view
Regression models, however, also serve auxiliary purposes in a causal analysis. For example, the propensity score is the probability of being exposed to a specific treatment, and they are often used in matching or weighting analyses. In both cases, a logistic regression model is typically used to estimate the propensity score and thus is not the estimator of the causal effect, but the regression model serves a key role in the analysis.

5.2.2 Matching

Matching is one statistical method frequently linked to causal inference. Often it is used in conjunction under selection on observables to make treated and control groups identical in terms of observed covariates. Matching is equivalent to a specific form of nonparametric regression. See Angrist and Pischke (2009, pg. 69) for a discussion of the equivalencies. Matching, like regression, has a wide variety of uses across different identification strategies. Often natural experiments based on instruments require statistical adjustment; this form of adjustment can also be done via matching (Rosenbaum 2002a). Recently, matching has been adapted to RD designs (Keele, Titiunik and Zubizarreta 2014). I credit the more recent popularity of matching to work in economics where matching recovered the estimate from a randomized experiment based on observed covariates (Dehejia and Wahba 1999). This has also lead to some confusion, where matching has been mistaken for an identification strategy. See Sekhon (2009) and Arce- neaux, Gerber and Green (2006) for overviews of this confusion. However, it is worth repeating that matching is a statistical technique that is devoid of any identification assumptions. When matching is applied to an IV application, the identification assumptions are completely different from when matching is applied to an application where identifiability is based on selection on observables.

The main attraction of matching is that it is a completely nonparametric form of adjustment. I also think it has advantages in that one can completely customize the form of statistical adjustment. For example, one might dictate very close or exact matches on key variables and looser constraints on covariates that are less important. Balance testing also makes it readily apparent whether matching has succeeded in creating a comparable control group for the treated.
Matching, however, is simply a tool and cannot compensate for a poor identification strategy. Matching can also be part of the design. For example, matching can be used as a form of blocking in randomized experiments (Greevy et al. 2004). Here, units are made more comparable before treatments are assigned.

5.2.3 Weighting

Besides regression methods and matching, inverse probability (IP) weighting is the other major statistical method that has been developed specifically for the estimation of treatment effects (Robins, Rotnitzky and Zhao 1994; Robins 1999). IP weighting methods can be used to estimate treatment effects in a variety of situations but have seen widespread use in contexts with repeated and time varying treatments. Glynn and Quinn (2010b) provide a useful overview of these methods in a social science context.

Under this method of estimation, the analyst re-weights observations to create a pseudo population where treated and control units are conditionally independent of treatment status. This pseudo population is created by weighting each unit in the study by the inverse of the propensity score. The treatment effect estimate is simply the difference in means across treatment status within the pseudo population. A number of alternative methods for estimating weights are available, and the estimation of these weights forms an area of active research. IP weighting techniques are also closely identified with what are known as “doubly-robust” methods, though double robustness can also be achieved using matching methods (Ho et al. 2007). Double robust methods model both the treatment assignment mechanism and the outcome. If at least one of these models is correctly specified, the estimate of the average treatment effect will be consistent (Scharfstein, Rotnitzky and Robins 1999). The double-robust property is no magic bullet since poor estimation of the weights or misspecification of both models may cause bias (Kang, Schafer et al. 2007). One advantage of IP weighting is that it can also be use to model missingness in the outcomes. Moreover, variance calculations that take into account uncertainty in both the model of treatment and outcome are also straightforward.
5.3 Inferential Methods

In the analysis of causal effects, one could easily assume that little changes in terms of statistical testing. For example, in the analysis of an experiment, the usual $t$-test is typically applied to test whether the average treatment effect is zero. In reality, a subtle change has occurred. The standard justification for statistical inference is to characterize uncertainty about a random sample from a population. Of course, many experiments are not conducted with representative samples, and yet they can still lead to valid inferences about causal effects for the units under study. Generally in studies of causal effects, the mode of statistical inference is different. Our main source uncertainty is about whether a causal effect is real or instead a chance outcome due to the stochastic nature of the treatment assignment mechanism. The difference prompted Rubin (1991) to advise analysts to ask: what is your mode of inference?

This question is important since in the study of causal effects, statistical measures of uncertainty are dependent on how the treatment is assigned. The simplest example arises in randomized experiments. In many randomized experiments, treatments are assigned at the unit level. For example, a GOTV treatment could be assigned to individual level voters. However, we might instead conduct a group randomized trial, where groups of units are assigned to treatment or control. Under a group RCT, the GOTV treatment might be assigned to households or entire precincts. The difference in assignment mechanisms has implications for measures of statistical uncertainty. If we analyze the group trial as if it were an individual level trial, the analyst will underestimate statistical uncertainty, since the number of groups is more relevant than the number of individuals. Thus it is important to have clarity about how treatments are assigned, since statistical inference directly depends on the treatment assignment mechanism. The mode of inference question becomes more complex outside of experiments since we often do not directly observe how treatments were assigned. In observational data, it is often unclear whether the treatment assignment mechanism operates at a unit or group level, so analysts must carefully consider how to characterize statistical uncertainty. As such, it is important that analysts understand how statistical inference differs when causal effects are the goal. Identification may be
wasted if the mode of inference is wrong and measures of statistical uncertainty are incorrect.

Statistical inference for treatment effects is typically defined using one of two different frameworks. The first framework is associated with Jerzy Neyman, and the second framework was developed by Ronald Fisher. The Neyman framework is the most widely used. Here, I briefly point out differences between the two frameworks and discuss why I think it is important to be familiar with both frameworks. Under the Neyman framework, we ask what would be the average outcome if all units were exposed to treatment and how would that compare to the average outcome if all units were exposed to control? The statistical test under the Neyman framework is whether the average causal effect is zero. In the Fisherian framework, we test what is known as the sharp null hypothesis. Under the sharp null hypothesis, the analyst tests whether the treatment effect is zero for every unit. In potential outcomes notation, if the sharp null hypothesis holds then $Y_{1i} = Y_{0i}$ for every $i$. In the Fisherian approach, there is no way to test the null hypothesis that the average effect is zero (Imbens and Rubin 2015). This might strike some readers as a major drawback, since this would seem to be a very restrictive null hypothesis. One advantage of testing average effects is we can accommodate heterogenous responses to treatment. That is, under a test of the average effect, the units can have some mix of positive and negative responses to treatment.

However, only testing for average causal effects has pathologies of its own. Take an example from Imbens and Rubin (2015). Let’s say that $Y_{0i} = 2$. For 1/3 of the units in the study, the treatment effect is 2, but for 2/3 of the units, the effect is -1. Here, the average effect is zero, but the sharp null is not. Again the mode of inference matters, in that we might detect an effect with one mode of inference but miss it with another. One particular strength of the Fisherian framework is that it can accommodate a wide variety of tests about quantities other than averages. Thus far, I have described estimands only as averages. However, there is nothing specifically that implies that a treatment will only change the middle of the treated and

\[10\] There are a number of other features that are unique to the Fisherian framework, including that it can be used as a method of estimation. Keele, McConnaughy and White (2012) provide a basic overview of Fisher’s approach.
control distributions as summarized by the average. In the most extreme example, the treatment might only change the variance of the treated distribution. Under the Fisherian framework, we can apply the Kolmogorov-Smirnov (KS) statistic, which tests the maximum discrepancy in the empirical CDFs and can detect difference in any of the moments of the distribution. The Fisherian framework has also been extended in other fruitful ways. It serves as the basis for one common method of sensitivity analysis (Rosenbaum 2002a). Bowers, Fredrickson and Panagopoulos (2013) use it to analyze empirical applications with treatment spillovers.

I would argue that analysts need to be familiar both frameworks. A clear understanding of both is useful in two ways. One it clarifies how the mode of statistical inference matters. Under the Fisherian framework, it is obvious how the mode of inference changes depending on the assignment mechanism. Moreover, it allows for testing quantities other than average effects. The Neyman framework, however, allows for tests of average effects which are at a minimum a useful starting point. This framework also accommodates sampling from populations as well, which arises when randomized experiments are conducted with random samples from populations. Generally, after an examination of average effects, analysts should consider whether other features of the treated and control distributions differ and test for such differences.

6 Discussion

The reader may notice that this essay is heavily tilted toward identification rather than on intricacies of matching methods or relative merits of doubly robust estimators. It is not because estimation and inference aren’t important, but it is due to the fact that no statistical method can save a poor identification strategy. Many of the pathologies in the statistical analysis of causal effects stem from confusion over the separate roles of identification and estimation. Understanding this distinction, provides an important check on what it is that analysts think they can learn from data. Much of the language in statistics has long obscured the importance of assumptions. To say that a model is unbiased when correctly specified is a true statement, and yet seriously understates how difficult it can be to achieve the correct specification when the goal is estimation
of causal effects. An understanding of what it means for something to be correctly specified (i.e., identified) reveals the limits of what can generally be learned from data about causal effects, especially with observational data. Moreover, it reveals that complex statistical estimators may do nothing to aid an identification strategy. Causal inference, particularly in relation to topics where randomized experiments are impossible, will probably remain a difficult task that requires a series of different identification strategies across a number of different contexts before conclusions can be reached.

To that end, one goal in this essay was to illuminate an important paradox that lies within the statistics of causal inference. This paradox occurs in the fact that the most credible causal inferences require the least amount of statistical analysis. In fact, when a causal inference is credible, most of the work will have been done before the outcome data is collected. If the analyst has taken the time to develop a randomized treatment assignment mechanism and reduce or eliminate noncompliance and attrition, often the analysis of the causal effect is reduced to a simple contrast in measures of distributional location. The analyst is successful at identifying the causal effect not because of the complex statistical methods that are applied to the data, but due to the effort in developing a design before data is collected. A quote by Fisher (1938) is instructive on this point: “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.”
References


Calonico, Sebastian, Matias Cattaneo and Rocío Titiunik. 2013. “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs.”


**URL:** http://CRAN.R-project.org/package=CausalGAM


Appendices

A.1 Software Details

Below I list software tools and links for many of the types of analyses discussed in the essay.

A.1.1 Matching

- Genetic Matching (Sekhon 2011) http://sekhon.berkeley.edu/matching/
- MatchIt (Ho et al. 2007) http://gking.harvard.edu/matchit

A.1.2 Sensitivity Analysis

- The rbounds package (Keele 2013) http://www.personal.psu.edu/ljk20/rbounds/20vignette.pdf
- The causalsens package (Blackwell 2014) http://www.mattblackwell.org/software/causalsens/

A.1.3 Weighting

- The ipw package (van der Wal and Geskus 2011) http://www.jstatsoft.org/v43/i13/paper
- The causalGAM package (Glynn and Quinn 2010b,a) http://cran.r-project.org/web/packages/CausalGAM/CausalGAM.pdf

A.1.4 RD Designs

- A suite of software tools. (Calonico, Cattaneo and Titiunik 2014) https://sites.google.com/a/umich.edu/rdrobust/

A.1.5 Mediation

- The mediation package (Tingley et al. 2014) http://cran.r-project.org/web/packages/mediation/vignettes/mediation.pdf
- Some software macros (Valeri and VanderWeele 2013) http://www.hsph.harvard.edu/tyler-vanderweele/tools-and-tutorials/