Randomized experiments

Amanda Coston[1,2], Mateo Dulce Rubio[1,3], and Edward H. Kennedy[3]

[1] Heinz College of Information Systems & Public Policy
[2] Machine Learning Department
[3] Department of Statistics & Data Science
Carnegie Mellon University

{acoston,mdulceru}@andrew.cmu.edu, edward@stat.cmu.edu

1 Introduction

In this chapter we give an introduction and review of randomized experiments. More details and exposition can be found in Hernán and Robins [2010], Imbens and Rubin [2015], Rosenbaum [2002], Tsiatis [2006], van der Laan and Robins [2003], for example, among many others. Notation will be defined as it is introduced, but we give a reference table in the Appendix.

1.1 Why Do We Need Randomization?

Suppose we observe outcomes $(Y_1, ..., Y_n)$ for n subjects, each of whom are either treated $(A = 1)$ or not $(A = 0)$, and we want to learn the causal effect of the treatment A on the outcome Y, say on average. An initial idea might be to compare the average outcome for those n_1 who receive treatment versus that of the n_0 who receive control:

$$\frac{1}{n_1} \sum_{i:A_i=1} Y_i \quad \text{versus} \quad \frac{1}{n_0} \sum_{i:A_i=0} Y_i.$$

However, any differences we see could be spurious, i.e., explained by something else.

Example 1. As an example from criminal justice, consider the pre-trial release setting where a judge has to determine whether to release a defendant pre-trial. This decision might be based on their reasoning of how likely the defendant is to fail to appear for trial, and how likely the defendant is to be arrested for a new crime before the trial. Jurisdictions across the country are increasingly using data-driven prediction instruments, such as the Public Safety Assessment (PSA), to help judges make these decisions. While these risk assessments often contain separate predictions for failure to appear and new arrests, for simplicity consider the prediction of new arrest. Suppose we are interested in evaluating an algorithmic risk assessment instrument that predicts the likelihood of re-arrest based on the defendant's prior criminal record, current charges, and age [Imai et al., 2020]. We would then like to investigate questions such as: *Do judges make better decisions when they have access to these risk assessment scores? Are observed re-arrests rates lower for decisions made based on the risk assessments?*

In the above example, if judges could decide when to see the re-arrest risk assessment instrument, we may expect judges to be more likely to do so for difficult cases where they are uncertain, compared to cases where it is clear that the defendant should be released. In such a setting, because the control cases are less risky than the treated cases, we cannot attribute differences in outcomes to the risk assessment. More generally, differences in outcomes might be due to the units receiving treatment being inherently different from those receiving control. This is popularly recognized as "correlation does not imply causation".

What if we think we can identify cases that are similarly difficult (or easy) and would like to compare outcomes under treatment against outcomes under control for cases that are similar in level of difficulty? This would require us to measure any and all variables X that might explain differences in outcomes between treated and control units. In the pre-trial release decision setting, one might try to measure every possible criminal risk factor, such as economic opportunity, behavioral characteristics, mental health, community support, social ties, and so on. There are at least three difficulties with this approach:

1. We often simply do not know every single $X = (X_1, X_2, ..., X_{1000}, ...)$ that could explain *any* differences in outcomes between subjects with different treatment levels.
2. Even if we did know every single possible X with certainty, it might be impossible or too expensive to measure every single one of them.
3. And even if we could measure every single X, there may be so many that few if any subjects would have the same or similar X's in every dimension, meaning it is impossible to find any untreated units with the same X values as a treated unit. This so-called “curse of dimensionality" would make estimation impossibly difficult.

We can avoid these difficulties when we can control who gets treatment. A simple yet beautiful solution is to assign treatment randomly. For example, if one could flip a coin to decide whether each subject gets treatment versus control (e.g., in our example, flip a coin to decide in which cases the judge is presented with the AI risk score), then we could properly study the causal effect of treatment A on outcome Y. Surprisingly, the benefits of randomization were largely unknown until relatively recently in the long history of science: according to the Oxford English Dictionary, its first recorded use was due to R.A. Fisher in 1926.

1.2 Why Does Randomization Work?

Why does assigning treatments at random allow for valid estimation of causal effects? Randomization ensures that treatment is completely independent of *all* subject characteristics, whether measured or not. In other words, the treated look *exactly the same* as the untreated, in expectation, and not only for all measured variables X but also for *any* unmeasured variables U (we will see later that finite-sample differences are handled with appropriate variance estimators). Thus any observed differences in the outcomes for the treated versus untreated must be due to the treatment: it is the only systematic way in which the groups differ.

We can formalize this argument using the *potential* or *counterfactual outcomes* Y_i^a, i.e., the hypothetical outcome we would have observed if subject i had received treatment $A_i = a$. We briefly discuss the concept and notation of potential outcomes, and then we revisit the question of why randomization works. Note that Y_i represents what was actually observed, whereas Y_i^a represents what would have been observed under a treatment that might not have been received in the real world. This is a very important distinction. One can imagine Y_i^a for different values of a representing different outcomes that would have existed in parallel universes where everything else is the same except for the treatment assignment being $A_i = a$.

For instance, under a binary treatment each subject has two potential outcomes: Y_i^1 if treated ($A_i = 1$) and Y_i^0 if not ($A_i = 0$). The former is observed only for the units under treatment and the latter only for the control (untreated) units, so that the observed Y_i can be written as $Y_i = A_i Y_i^1 + (1 - A_i) Y_i^0$. In our criminal justice example, Y_i^1 could be whether defendant i is re-arrested had the judge used the AI risk score to make her pre-trial decision, and Y_i^0 could be whether re-arrest occurred had the risk score not been used.

Remark 1. Often we will drop the i subscript, and if it is clear what is being intervened upon we will just write Y^1 or Y^0. We use superscripts as in Y^a to denote potential outcomes, but other references may use subscripts Y_a or parentheses $Y(a)$.

Now let's consider why randomization works in terms of the potential outcomes. Specifically, by randomly assigning treatment A, we are taking two random samples: one of the Y^1 values and another of the Y^0 values. Since random samples yield unbiased estimators of population means, the average observed outcomes in the two groups will be unbiased estimates of the corresponding average potential outcomes.

We can prove randomization works mathematically, as illustrated in the following result. In addition to randomization, this proposition makes the assumption that an unit's observed outcome depends only on its own treatment assignment (the assumption of consistency). This assumption is violated for instance when there is interference between units, an issue we revisit in the advanced topics at the end of the chapter.

Proposition 1. *Let $(A, Y) \sim \mathbb{P}$ and assume:*

1. *Consistency: $Y = Y^a$ whenever $A = a$.*
2. *Randomization: $A \perp\!\!\!\perp Y^a$ for each a.*

Then

$$\mathbb{E}(Y \mid A = a) = \mathbb{E}(Y^a).$$

Proof. We have

$$\mathbb{E}(Y \mid A = a) = \mathbb{E}(Y^a \mid A = a) = \mathbb{E}(Y^a),$$

using consistency in the first equality and randomization in the second. □

Proposition 1 shows that we can *identify* the expected value of the potential outcome Y^a as the expected value of the observed outcome Y for those under treatment level $A = a$. Identification refers to the process of expressing causal parameters (e.g., $\mathbb{E}(Y^a)$) in terms of the distribution we actually sample from (e.g., (Y, A)).

For example, suppose we are interested in comparing the re-arrest rate if a risk assessment instrument is always used $\mathbb{E}(Y^1)$ versus the re-arrest rate if it is never used $\mathbb{E}(Y^0)$. If we can randomly assign when the risk assessment scores are presented to the judges, then Proposition 1 guarantees that we would be able to properly estimate $\mathbb{E}(Y^0)$ and $\mathbb{E}(Y^1)$ using the observed outcomes Y, and therefore we can estimate the causal effect of the AI system on the re-arrest rates.

Remark 2. It is important not to confuse $A \perp\!\!\!\perp Y^a$ with $A \perp\!\!\!\perp Y$: these are very different. $A \perp\!\!\!\perp Y^a$ means treatment is independent of potential outcomes (which can be viewed as "pre-treatment" variables that exist just prior to the treatment assignment), and reflects that treatment is not confounded. $A \perp\!\!\!\perp Y$ means treatment is independent of the *observed* outcome, and would for example be a consequence of treatment not only being unconfounded but also ineffective (e.g., $Y^1 = Y^0$). Always remember to distinguish potential from observed outcomes.

Remark 3. Although Proposition 1 gives an identification result for the mean potential outcome, its assumptions are sufficient for identifying the entire distribution of potential outcomes as $\mathbb{P}(Y^a \leq t) = \mathbb{P}(Y \leq t \mid A = a)$.

Proposition 1 also shows that treatment assignment need not necessarily be a subject-specific coin flip – for the purposes of achieving identification of the potential outcome distribution, treatment just needs to be independent of potential outcomes. This leads to the following definition of a randomized experiment:

Definition 1 (Randomized experiment). *A study is a randomized experiment if the treatment assignment is both probabilistic and known.*

There are many types of experimental designs. Letting $A^n = (A_1, ..., A_n)$ denote the vector of treatment assignments for the n study subjects, we have the following designs:

- Completely randomized: n_1 of n subjects are randomly assigned to treatment, i.e., $\mathbb{P}(A^n = a^n) = 1/\binom{n}{n_1}$ for $\sum_i a_i = n_1$.
- Bernoulli: Treatments are assigned via independent coin flips, i.e., $\mathbb{P}(A^n = a^n) = (1/2)^n$ for every $a^n = (a_1, ..., a_n) \in \{0, 1\}^n$.
- Stratified Bernoulli: Treatments are assigned via independent *biased* coin flips depending on covariates, i.e., $\mathbb{P}(A^n = a^n \mid X^n) = \prod_i \mathbb{P}(A_i = a_i \mid X_i)$.
- Matched pairs: Matched pairs are constructed with exacty one treated in each pair, i.e., $\mathbb{P}(A^n = a^n \mid X^n) = 1/2^{n/2}$.

Depending on the particulars of the experimental setting, one design may be favored over another for reasons such as efficiency or feasibility.

1.3 Connections to AI for Social Impact

Once deployed in the real world, even the most carefully designed artificial intelligence (AI) systems may fail to achieve their intended goals or may have adverse unintended consequences. How should researchers assess whether the AI actually improved outcomes? Randomized experiments are the gold standard for evaluation. They enable one to isolate the effect of the AI from other potentially confounding factors. Examples of AI systems for social impact that have been deployed and evaluated in the real world abound: Wang et al. [2019] studies the effect on the adenoma detection rate of using real-time AI-assisted colonoscopies, Mohler et al. [2015] evaluates the effect of using a predictive policing model on crime rates in Los Angeles, Mate et al. [2021] assesses the effect on dropout of using a model to prioritize the follow-up of participants in a maternal and child care information program, etc.

2 Simple Randomized Experiments

2.1 Testing: Fisher's Sharp Null

Jerzy Neyman seems to have been the first to introduce potential outcomes [Neyman, 1990], but R.A. Fisher was perhaps the first to really advocate for randomization [Fisher, 1925].

Fisher was interested in testing the *sharp null hypothesis*

$$H_0 : Y_i^1 = Y_i^0 \text{ for all } i,$$

which says that treatment has no effect whatsoever – not only is the mean of Y^1 exactly equal to that of Y^0, but the entire distributions are equal, and further each individual potential outcome is exactly the same under both treatment and control. This is a strong null with lots of structure, in line with Fisher's perspective that one should "make your theories elaborate" [Rosenbaum, 2002].

To test a generic null hypothesis H_0 we need (1) a statistic T, and (2) its distribution under the null. Then one can obtain a a p-value, i.e., $\mathbb{P}_{H_0}(T \geq t_{obs})$, the chance under the null of seeing data as extreme as that which was actually observed.

To test Fisher's sharp null, we can use as a statistic any summary measure of how treatment changes outcomes; for example, a simple yet common choice is the absolute difference-in-means

$$T(A^n, Y^n) = \left| \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i \right| = \left| \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} \right|,$$

where $\mathbb{P}_n(Z)$ denotes the empirical mean of Z, $\frac{1}{n}\sum_{i=1}^n Z_i$. Note this test statistic will be large if the treated versus untreated means differ, but not if the treatment only changes non-central aspects of the distribution, e.g., the variance.

Armed with a test statistic, we now need to know its distribution under the null. This is actually easy, and typically part of the motivation for using the sharp null: it yields tractable null distributions, which can be computed in a non-asymptotic and distribution-free manner. To illustrate, consider a completely randomized experiment where the observed value of the difference-in-means test statistic for the observed data $(A^n = (A_1, \ldots, A_n), Y^n = (Y_1, \ldots, Y_n))$ is $T(A^n, Y^n) \approx 0.9$. We can also compute the value of this statistic under the null, for any randomization, since under the null the potential outcomes are exactly the same, i.e., $Y^0 = Y^1 = Y$. Therefore we can obtain the null distribution of T by permuting the A^n vector (according to the known treatment assignment mechanism), while keeping the Y^n vector fixed, computing the corresponding value of the test statistic T, which yields the corresponding null distribution $\mathbb{P}_{H_0}(T \leq t)$. A p-value can be computed by simply counting the proportion of permutations with test statistics larger than that which was observed.

Example R code to test Fisher's sharp null can be found in the Appendix. The data and results are shown in Figure 1. In this simulation, the p-value is 0.084 and so there is sufficient evidence to reject the sharp null hypothesis of no individual treatment effect at level $\alpha = 0.10$.

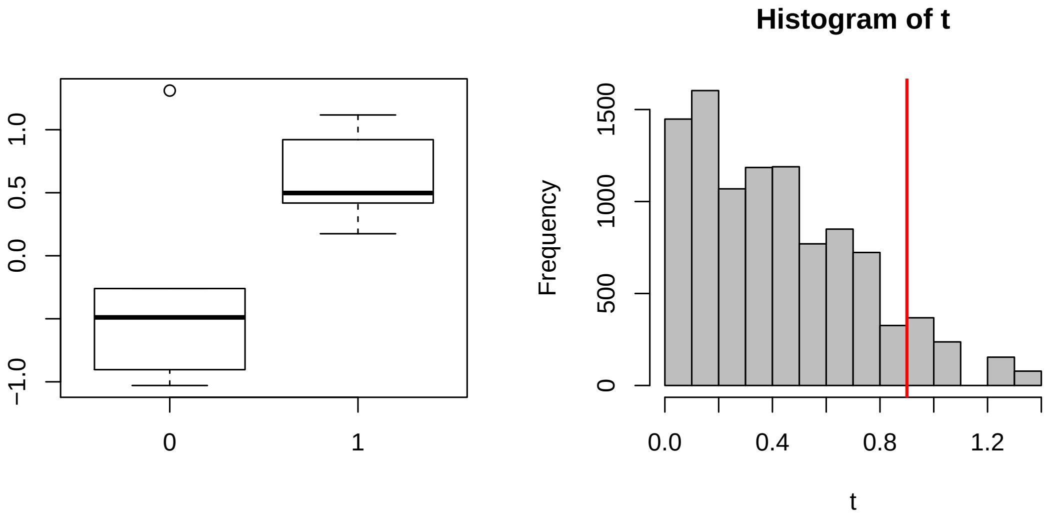

Fig. 1: Boxplot of simulated outcomes under control 0 and treatment 1 (left), and histogram of permutation-based null distribution with the red vertical line denoting the observed test statistic value (right).

Mathematically, for a completely randomized experiment where a fixed number n_1 are treated, the null distribution can be written as

$$\begin{aligned}\mathbb{P}_{H_0}(T \geq t) = \mathbb{P}_{H_0}\{T(A^n, y^n) \geq t\} &= \sum_{a^n \in \mathcal{A}} \mathbb{1}\{T(a^n, y^n) \geq t\}\mathbb{P}(A^n = a^n) \\ &= \sum_{a^n : \sum_i a_i = n_1} \frac{\mathbb{1}\{T(a^n, y^n) \geq t\}}{\binom{n}{n_1}}.\end{aligned}$$

In theory we can compute this distribution exactly; in practice if n is large we may need to resort to simulation (e.g., sample K of the $\binom{n}{n_1}$ randomizations). However the distribution can be simulated with arbitrarily high accuracy by taking K large enough.

Remark 4. The null distribution calculation above treats the (potential) outcomes y^n as fixed; this can be viewed as an assumption that Y^a is not a random variable, or the probability can just be defined conditionally, given the random potential outcomes.

Fisher's permutation-style test is simple but impressive: it gives an exact distribution-free p-value for testing H_0, which is valid for any n. Nonetheless here are some caveats:

- The power of the test depends heavily on the choice of statistic, e.g., the difference-in-means test statistic will have no power against a treatment that makes outcomes bimodal or otherwise more variable.
- Fisher's test is of the *sharp null* of no individual effect, not of no average effect – in fact rejecting Fisher's null could still mean there is no effect on average.

2.2 Estimation: Sample Average Effects

In the 1920s and 1930s, Fisher and Neyman had some heated debates about whether testing Fisher's sharp null should be the primary goal or not [Lehmann, 2011]; in contrast to Fisher, Neyman advocated more for estimation rather than testing, and focused on average effects. Average effects might be considered more relevant for policy decisions, since they indicate how a population would fare on average if all versus none were treated, e.g., how would public safety change if all vs. none of the judges use an AI risk score in their pre-trial decisions? In contrast, rejecting the sharp null only indicates that treatment has *some* effect, without saying much about what kind.

The *sample average treatment effect* is given by

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0).$$

This parameter is different from those we will study later in that it is a functional of the particular sample, rather than of a population distribution (i.e., strictly speaking it is a data-dependent parameter, which is why we index it with n).

Remark 5. In this section we again treat potential outcomes as fixed, not random; or equivalently we treat probability statements as conditional on the potential outcomes. Note however that even if the potential outcomes are fixed, the observed outcome is random since it is a function of the random treatment: $Y = Ay^1 + (1 - A)y^0$.

A natural estimator for ψ_n in completely randomized experiments is the difference-in-means

$$\widehat{\psi} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)}. \tag{1}$$

We will now characterize the bias and variance of this estimator and discuss inference.

Proposition 2. *The difference-in-means estimator (1) is unbiased for ψ_n in a completely randomized experiment, assuming only consistency (i.e., $Y = Ay^1 + (1-A)y^0$).*

Proof. By definition, in a completely randomized experiment, we have

$$\begin{aligned}\mathbb{P}(A_1 = 1) &= \sum_{\sum_{i>1} a_i = n_1 - 1} \mathbb{P}(A_1 = 1, A_2 = a_2, ..., A_n = a_n) \\ &= \sum_{\sum_{i>1} a_i = n_1 - 1} \binom{n}{n_1}^{-1} = \frac{\binom{n-1}{n_1-1}}{\binom{n}{n_1}} = \frac{n_1}{n},\end{aligned}$$

and similarly for all other $i > 1$. Therefore

$$\mathbb{E}(\widehat{\psi}) = \frac{1}{n_1} \sum_{i=1}^{n} \mathbb{E}(A_i) y_i^1 - \frac{1}{n_0} \sum_{i=1}^{n} \{1 - \mathbb{E}(A_i)\} y_i^0 = \frac{1}{n} \sum_{i=1}^{n} (y_i^1 - y_i^0),$$

where the first equality follows by consistency, and the second since $\mathbb{E}(A_i) = \mathbb{P}(A_i = 1) = n_1/n$. □

As mentioned previously, the intuition behind unbiasedness in this setup is that the treatments pick out random samples of Y^1 and Y^0 potential outcomes, and random sampling allows for unbiased estimation of means.

Now we will explore the variance of $\widehat{\psi}$, which is critical for constructing confidence intervals and hypothesis tests; its calculation requires some care since the A_i's are not independent (e.g., in the $n = 2$ case, if $A_1 = 1$ then it must be the case that $A_2 = 0$).

Proposition 3. *For a completely randomized experiment, and assuming consistency, the variance of the difference-in-means estimator is given by*

$$var(\widehat{\psi}) = \frac{\sigma_n^2(y^1)}{n_1} + \frac{\sigma_n^2(y^0)}{n_0} - \frac{\sigma_n^2(y^1 - y^0)}{n}, \tag{2}$$

where

$$\sigma_n^2(v) = \frac{1}{n-1} \sum_{i=1}^{n} \left(v_i - \frac{1}{n} \sum_{j=1}^{n} v_j \right)^2$$

denotes the finite sample variance of $(v_1, ..., v_n)$.

Proof. See the appendix of Chapter 6 in Imbens and Rubin [2015]. □

Regarding inference, a finite-sample central limit theorem implies under some regularity conditions that

$$\frac{\widehat{\psi} - \psi_n}{\sqrt{\text{var}(\widehat{\psi})}} \rightsquigarrow N(0, 1).$$

Therefore to construct large-sample confidence intervals, one needs to estimate the variance $\text{var}(\widehat{\psi})$ in (2). The first two terms in this variance can be estimated with

$$\widehat{\sigma}_n^2(y^a) = \frac{1}{n_a - 1} \sum_{i:A_i=a} \left(Y_i - \frac{1}{n_a} \sum_{j:A_j=a} Y_i \right)^2 ,$$

but the third term is the finite-sample variance of the *individual* treatment effects $(y_i^1 - y_i^0)$, and involves product terms like $y_i^1 y_i^0$ which can never be observed together. Thus the third term cannot be consistently estimated. However we can upper bound the variance as, for example

$$\text{var}(\widehat{\psi}) \leq \frac{\sigma_n^2(y^1)}{n_1} + \frac{\sigma_n^2(y^0)}{n_0}, \tag{3}$$

which will yield conservative inference (at worst), when used to construct confidence intervals. Tighter bounds can be achieved with the Cauchy-Schwarz inequality or Frechet-Hoeffding bounds [Aronow et al., 2014].

2.3 Population Average Effects

In this section we move to population rather than finite-sample effects. These effects can be useful for at least three reasons:

1. Population effects are often of particular substantive interest: typically we might view our sample as haphazard and not particularly special, except insofar as they tell us something about some larger population from which they were drawn.
2. Often population effect estimators can also be used for estimating sample effects, without modification, while the converse is not necessarily true; thus by studying population effects we can kill two birds with one stone. This is discussed in more detail shortly.
3. Population effects can be simpler to study, easing theoretical analyses without losing much in terms of main ideas.

Therefore here we suppose we observe an iid sample $(Z_1, ..., Z_n)$ from population distribution $\mathbb{P}$ with $Z = (A, Y)$. Our goal is to estimate the population average effect

$$\psi = \mathbb{E}(Y^1 - Y^0),$$

rather than the sample average effect ψ_n from before. Recall that average effects ask how a population outcome would change on expectation if all versus none were treated.

Now we will study three basic properties of the difference-in-means estimator $\widehat{\psi}$ given in (1): its bias, variance, and limiting distribution. We will see that, using this estimator, precise estimation and inference are possible for the causal effect ψ in Bernoulli trials, under the following three (quite weak) assumptions:

1. Consistency: $Y = Y^a$ if $A = a$.
2. Bernoulli randomization: $A \perp\!\!\!\perp Y^a$ with $\mathbb{P}(A = 1) = \pi$.
3. Finite variance: Y has finite conditional variance given $A = a$.

Remark 6. Note that, in the above Bernoulli trial where each A_i is assigned via an independent coin flip, the observed number of treated subjects $N_1 = \sum_i A_i \sim \text{Bin}(n, \pi)$ is random, not fixed. In this section we also view the potential outcomes as random variables, and not fixed.

Properties of the Difference-in-Means Estimator

Theorem 1. *Assume consistency. In a Bernoulli trial, the difference-in-means estimator (1) is unbiased for ψ and has variance no greater than*

$$\frac{2}{(n+1)}\left(\frac{\sigma_1^2}{\pi}+\frac{\sigma_0^2}{1-\pi}\right),$$

where $\sigma_a^2 = var(Y \mid A=a)$.

Proof. Since the assumptions from Proposition 1 hold we have

$$\psi = \mathbb{E}(Y^1 - Y^0) = \mathbb{E}(Y \mid A=1) - \mathbb{E}(Y \mid A=0).$$

Let $\widehat{\pi} = \mathbb{P}_n(A)$ and just consider the first term $\widehat{\mu}_1 = \mathbb{P}_n(AY)/\widehat{\pi}$ as an estimator of $\mu_1 = \mathbb{E}(Y \mid A=1)$. We have

$$\begin{aligned}\mathbb{E}(\widehat{\mu}_1 \mid A^n) &= \frac{1}{\widehat{\pi}}\mathbb{E}\Big\{\mathbb{P}_n(AY) \mid A^n\Big\} = \frac{1}{\widehat{\pi}}\mathbb{P}_n\Big\{A\mathbb{E}(Y \mid A^n)\Big\} \\ &= \frac{1}{\widehat{\pi}}\mathbb{P}_n\Big\{A\mathbb{E}(Y \mid A=1)\Big\} = (\widehat{\pi}\mu_1)/\widehat{\pi} = \mu_1,\end{aligned}$$

by iterated expectation and the iid assumption. Unbiasedness now follows by iterated expectation, and consistency follows from the weak law of large numbers and continuous mapping theorem. The logic is the same for $\widehat{\mu}_0 = \mathbb{P}_n\{(1-A)Y\}/(1-\widehat{\pi})$. By the law of total variance we have

$$\text{var}(\widehat{\mu}_1) = \text{var}\Big\{\mathbb{E}(\widehat{\mu}_1 \mid A^n)\Big\} + \mathbb{E}\Big\{\text{var}(\widehat{\mu}_1 \mid A^n)\Big\}.$$

Note $\text{var}\{\mathbb{E}(\widehat{\mu}_1 \mid A^n)\} = \text{var}(\mu_1) = 0$ from above, and

$$\begin{aligned}\text{var}(\widehat{\mu}_1 \mid A^n) &= \left(\frac{1}{n\widehat{\pi}}\right)^2 \sum_{i=1}^n A_i \text{var}\left(Y_i \mid A^n\right) \\ &= \left(\frac{1}{n\widehat{\pi}}\right)^2 \sum_{i=1}^n A_i \sigma_1^2 = \frac{\sigma_1^2}{N_1}\mathbb{1}(N_1 > 0),\end{aligned}$$

where we used independence and defined $\sigma_1^2 = \text{var}(Y \mid A=1)$ and $N_1 = n\widehat{\pi} \sim \text{Bin}(n,\pi)$. Now

$$\text{var}(\widehat{\mu}_1) = \mathbb{E}\Big\{\text{var}(\widehat{\mu}_1 \mid A^n)\Big\} \leq \frac{2\sigma_1^2}{(n+1)\pi}$$

by the expected binomial reciprocal result (Lemma A.2) of Devroye et al. [1996]. The same logic applies to $\widehat{\mu}_0$, and iterated expectation shows that the covariance term $\text{cov}(\widehat{\mu}_1, \widehat{\mu}_0)$ is exactly zero, which gives the result. □

Theorem 2. *Assume consistency. For a Bernoulli trial, the difference-in-means estimator is root-n consistent and asymptotically normal with*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\left(0, \ \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}\right),$$

where $\sigma_a^2 = var(Y \mid A=a)$.

Proof. We again focus on μ_1 and its estimator. Note we have

$$\begin{aligned}
\widehat{\mu}_1 - \mu_1 &= \frac{\mathbb{P}_n(AY)}{\widehat{\pi}} - \mu_1 = \mathbb{P}_n\left\{ \frac{A}{\widehat{\pi}}(Y - \mu_1) \right\} \\
&= \mathbb{P}_n\left\{ \frac{A}{\pi}(Y - \mu_1) \right\} + \left(\frac{1}{\widehat{\pi}} - \frac{1}{\pi} \right) \mathbb{P}_n\left\{ A(Y - \mu_1) \right\} \\
&= \mathbb{P}_n\left\{ \frac{A}{\pi}(Y - \mu_1) \right\} + O_{\mathbb{P}}(1/\sqrt{n}) O_{\mathbb{P}}(1/\sqrt{n}),
\end{aligned}$$

where the last equality follows by the central limit theorem, which implies $\sqrt{n}\{\mathbb{P}_n(V) - \mathbb{E}(V)\} = O_{\mathbb{P}}(1)$ for any iid V with finite mean and variance, together with the fact that $(\widehat{\pi}, \pi)$ are bounded away from zero. Therefore

$$\widehat{\mu}_1 - \mu_1 = \mathbb{P}_n\left\{ \frac{A}{\pi}(Y - \mu_1) \right\} + o_{\mathbb{P}}(1/\sqrt{n}),$$

since $O_{\mathbb{P}}(1/\sqrt{n}) O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})$. Therefore by the central limit theorem (together with Slutsky's theorem) we have

$$\sqrt{n}\Big(\widehat{\mu}_1 - \mu_1 \Big) \rightsquigarrow N\left(0, \text{var}\left\{ \frac{A}{\pi}(Y - \mu_1) \right\} \right).$$

The logic for the $\widehat{\mu}_0$ part is analogous. □

Theorem 1 is quite powerful in showing that, in Bernoulli trials, mean counterfactuals can be estimated very precisely (i.e., with zero bias and variance that scales like $1/n$) using no assumptions other than consistency and finite variance. In other words: randomization allows accurate and essentially assumption-free causal inference.

Similarly, Theorems 1 and 2 also pave the way for inference, in the form of confidence intervals and hypothesis tests. Namely, finite sample confidence intervals could be constructed based on Theorem 1 using bounds on the conditional variances σ_a^2, and Theorem 2 implies for example that an asymptotic 95% CI is given by

$$\widehat{\psi} \pm \left(\frac{1.96}{\sqrt{n}} \right) \widehat{\text{sd}} \left\{ \frac{A(Y - \widehat{\mu}_1)}{\pi} - \frac{(1 - A)(Y - \widehat{\mu}_1)}{1 - \pi} \right\}.$$

Remark 7. We saw above that the asymptotic variance of the difference-in-means estimator in a Bernoulli experiment is given by

$$\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1 - \pi}.$$

One interesting thing to note about this variance comes from the perspective of experimental design: what is the best choice of π for optimizing efficiency? In fact, it is straightforward to show that

$$\arg\min_{\pi} \left(\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1 - \pi} \right) = \frac{\sigma_1}{\sigma_0 + \sigma_1},$$

therefore for optimal efficiency the proportion treated should match the standard deviation of treated outcomes, as a fraction of the total standard deviation for treated and untreated outcomes. This matches intuition: if outcomes are more variable among treated patients than among those receiving control (i.e., if $\sigma_1 > \sigma_0$) then more patients should be assigned to treatment, in order to counterbalance this extra noise among the treated.

Sample versus Population Effects Here we point out an interesting connection between sample effect estimation in completely randomized experiments and population effect estimation in Bernoulli experiments.

Based on Theorem 2, an asymptotic 95% CI for ψ in a Bernoulli experiment is given by

$$\widehat{\psi} \pm 1.96\sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}},$$

where $\widehat{\sigma}_a^2 \equiv \sigma_n^2(y^a)$ is the usual sample variance among the treated ($a = 1$) and controls ($a = 0$), which we used in our analysis of the difference-in-means as an estimator of the *sample* average effect in completely randomized experiments (e.g., Proposition 3).

In fact, $\widehat{\psi}$ is the exact same point estimate of the sample effect that we analyzed in completely randomized experiments, and similarly the exact same confidence interval

$$\widehat{\psi} \pm 1.96\sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

is also valid (possibly conservative) in completely randomized experiments, guaranteeing at least 95% coverage of the sample effect. (This results from using the naive bound of $\sigma_n^2(y^1 - y^0) \geq 0$ as in (3)). Thus, not only is the estimator for the population effect exactly the same as that for the sample effect, but confidence intervals for the population effect are also valid for the sample effect, being at worst conservative. This is an archetypal example of how finite-sample and population-based frameworks can coincide.

Note that, although population-based confidence intervals are valid for sample effects, the converse is not necessarily true: it is easier to estimate sample effects, in the sense that the same estimators have smaller variances relative to sample versus population effects. Thus a confidence interval for a sample effect may not be valid for a population effect. For example, Imbens [2004] shows that

$$\mathbb{E}\{(\widehat{\psi} - \psi_n)^2\} = \mathbb{E}\{(\widehat{\psi} - \psi)^2\} - \frac{\text{var}(Y^1 - Y^0)}{n} + o(1/n),$$

so that the difference-in-means has smaller variance when estimating the sample effect ψ_n. For some intuition, imagine both potential outcomes were observed for each subject: then the sample effect would be estimated without error, but not the population effect.

Difference-in-Means versus Horvitz-Thompson Estimators Note that the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} = \mathbb{P}_n\left\{\left(\frac{A}{\widehat{\pi}}\right)Y - \left(\frac{1-A}{1-\widehat{\pi}}\right)Y\right\},$$

which suggests a different estimator, where we replace the estimated proportion treated $\widehat{\pi}$ with its known population value π:

$$\widehat{\psi}_{ht} = \mathbb{P}_n\left\{\left(\frac{A}{\pi}\right)Y - \left(\frac{1-A}{1-\pi}\right)Y\right\}.$$

This estimator is known as the Horvitz-Thompson estimator, hence the *ht* subscript.

Since we are replacing an estimated quantity $\widehat{\pi}$ with its known value π, it may appear as if we should gain efficiency. Here we study whether this is actually the case. It is straightforward to check that the Horvitz-Thompson estimator is unbiased and consistent, like the difference-in-means; and since it is exactly equal to a sample average, we can apply the central limit theorem to immediately obtain

$$\sqrt{n}(\widehat{\psi}_{ht} - \psi) \rightsquigarrow N\left(0, \text{var}\left\{\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right) Y\right\}\right).$$

Thus both the difference-in-means and Horvitz-Thompson estimators are unbiased, root-n consistent, and asymptotically normal. To determine whether it is beneficial or not to replace the estimate $\widehat{\pi}$ with its known value π, we will compare asymptotic variances.

Let $\phi = \frac{A}{\pi}(Y - \mu_1) - \frac{1-A}{1-\pi}(Y - \mu_0)$ and $\phi_{ht} = \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right) Y$ denote the functions whose variances correspond to the asymptotic variances of $\widehat{\psi}$ and $\widehat{\psi}_{ht}$, respectively. Then we have

$$\begin{aligned} \text{var}(\phi_{ht}) &= \text{var}\left(\phi + \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) \\ &= \text{var}(\phi) + \text{var}\left(\frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right), \end{aligned}$$

where the last line follows since $\mathbb{E}(\phi \mid A) = 0$ implies that

$$\text{cov}\left(\phi, \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) = 0$$

by iterated expectation.

Therefore

$$\text{var}(\phi_{ht}) \geq \text{var}(\phi),$$

and thus the Horvitz-Thompson estimator is actually *less efficient* than the difference-in-means. This is somewhat counterintuitive: replacing an estimated quantity with its known population counterpart actually reduces efficiency! Often when we estimate things we get something *less precise* than if we just used the true quantity.

One way to think about this paradox is as follows. Rather than viewing $\widehat{\psi}_{ht}$ as replacing an estimated quantity with a known quantity, one can instead view it as moving away from the sample average $\widehat{\psi} = \widehat{\mu}_1 - \widehat{\mu}_0$ with a noisier version

$$\widehat{\psi}_{ht} = \left(\frac{\widehat{\pi}}{\pi}\right)\widehat{\mu}_1 - \left(\frac{1-\widehat{\pi}}{1-\pi}\right)\widehat{\mu}_0,$$

which should degrade performance, merely since sample averages are efficient estimators of means. In other words, the Horvitz-Thompson estimator is using the expected number of treated $n\pi$ rather than the actual number $n\widehat{\pi}$, so that when the actual number differs from its expectation, the averages are not correctly weighted.

3 Randomized Experiments with Covariates

3.1 Identification with Covariates

So far we have considered settings where we have access to an iid sample

$$(A_1, Y_1), ..., (A_n, Y_n) \sim \mathbb{P},$$

but it is very common to also observe auxiliary covariate information (e.g., demographics like age or gender, socio-economic status, prior criminal records or baseline outcome measures, etc.). Thus in practice we often have an iid sample

$$(X_1, A_1, Y_1), ..., (X_n, A_n, Y_n) \sim \mathbb{P}$$

for covariates or features $X \in \mathbb{R}^d$. If we believe these covariates might be useful in predicting our outcome of interest (e.g., re-arrest), we might wonder whether incorporating knowledge of these covariates could improve our estimates of the causal effect $\psi = \mathbb{E}(Y^1 - Y^0)$, i.e., the mean outcome in the population if all versus none were treated.

In detail, the questions we consider here are: Does our previous identification result $\psi = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0)$ in the setting without covariates still hold? Are there any new identification results that the covariates buy us? We will see that the answer to both questions is yes for the experimental setting, in which we can assume

1. Consistency: $Y = AY^1 + (1 - A)Y^0$.
2. Randomization: $A \perp\!\!\!\perp (X, Y^a)$ for $a \in \{0, 1\}$ with $\mathbb{P}(A = 1 \mid X) = \pi$.

Proposition 4. *Assume consistency and randomization as given above. Then*

$$\begin{aligned}\mathbb{E}(Y^1 - Y^0) &= \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) \\ &= \mathbb{E}\Big\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\Big\} \\ &\equiv \int \Big\{\mathbb{E}(Y \mid X = x, A = 1) - \mathbb{E}(Y \mid X = x, A = 0)\Big\}\ d\mathbb{P}(x).\end{aligned}$$

Proof. Since $A \perp\!\!\!\perp (X, Y^a)$, standard independence calculations show that this implies $A \perp\!\!\!\perp Y^a$ and $A \perp\!\!\!\perp Y^a \mid X$. We know from the previous section that $A \perp\!\!\!\perp Y^a$ implies

$$\mathbb{E}(Y^a) = \mathbb{E}(Y^a \mid A = a) = \mathbb{E}(Y \mid A = a)$$

by randomization and consistency. For the second identification result note that

$$\mathbb{E}(Y^a) = \mathbb{E}\{\mathbb{E}(Y^a \mid X)\} = \mathbb{E}\{\mathbb{E}(Y^a \mid X, A = a)\} = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\},$$

where the first equality follows by iterated expectation, the second by $A \perp\!\!\!\perp Y^a \mid X$, and the third by consistency. □

The two identification results above suggest (at least) two different estimators for $\psi_1 = \mathbb{E}(Y^1)$, for example, namely:

$$\widehat{\psi}_1 = \mathbb{P}_n(Y \mid A = 1) \ \text{ versus } \ \widehat{\psi}_1 = \mathbb{P}_n\{\widehat{\mathbb{E}}(Y \mid X, A = 1)\}.$$

In what follows we will consider which estimator is “better”, and whether and how covariate information should be incorporated.

3.2 Logistic Regression & Collapsibility

In this section suppose $Y \in \{0, 1\}$ is a binary outcome. With binary outcomes, perhaps the most common approach in practice is to assume the logistic regression model

$$\text{logit}\ \mathbb{P}(Y = 1 \mid X, A) = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X,$$

and call β_1 “the effect” of treatment. What “effect” does this actually represent?

First, this may not be an effect at all, because the logistic model is probably not exactly correct in practice. In reality such a model probably leaves out important covariate interactions, higher-order terms, covariate-treatment interactions, non-logit links, etc. We often fit logistic regression models because they are fast and easy, not because they are particularly realistic.

Nevertheless, for the sake of argument, assume that the logistic model is correct. Then

$$\exp(\beta_1) = \frac{\text{odds}(Y = 1 \mid X, A = 1)}{\text{odds}(Y = 1 \mid X, A = 0)} = \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)},$$

where in the second equality we used consistency and randomization. This is a *conditional odds ratio* (OR). Importantly

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1) \neq \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)},$$

hence it is not an average treatment effect (on the risk difference scale). In fact, even if the model is correct, in general

$$\frac{\text{odds}(Y^1 = 1)}{\text{odds}(Y^0 = 1)} \neq \frac{\text{odds}(Y^1 = 1 \mid X)}{\text{odds}(Y^0 = 1 \mid X)},$$

and it is neither a population odds ratio effect (even if the conditional OR is constant). This follows since

$$\begin{aligned}
\text{odds}(Y^1 = 1) &= \frac{\mathbb{P}(Y^1 = 1)}{\mathbb{P}(Y^1 = 0)} = \frac{\mathbb{P}(Y = 1 \mid A = 1)}{\mathbb{P}(Y = 0 \mid A = 1)} \\
&= \frac{\mathbb{E}\{\mathbb{P}(Y = 1 \mid X, A = 1)\}}{\mathbb{E}\{\mathbb{P}(Y = 0 \mid X, A = 1)\}} = \frac{\mathbb{E}\{\text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} X)\}}{1 - \mathbb{E}\{\text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} X)\}} \\
&\neq \frac{\text{expit}\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}}{1 - \text{expit}\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\}} = \exp\{\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} \mathbb{E}(X)\},
\end{aligned}$$

since $\mathbb{E}\{f(X)\} \neq f\{\mathbb{E}(X)\}$ for nonlinear f. This is called the problem of *non-collapsibility* [Freedman, 2008, Greenland et al., 1999]. Thus we say the odds ratio is not collapsible since the average of the conditional ORs is not generally equal to the marginal OR. In fact the marginal OR can be bigger or smaller than all of the conditional ORs; this is counterintuitive.

The main take-away is that coefficients in general non-linear models are conditional and do not correspond to marginal (i.e., population averaged) effects – even if the model is correct. This subtlety is often missed.

However, this problem does not arise in a (correctly specified) linear model, e.g., of the form

$$\mathbb{E}(Y \mid X, A) = \beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X.$$

If the above model is correct, then

$$\beta_1 = \mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0),$$

so the coefficient is a conditional effect. Moreover, under the linear model assumption and $A \perp\!\!\!\perp Y^a \mid X$, we have

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\} = \beta_1,$$

therefore the parameter is also a marginal effect.

We have seen that going after coefficients in nonlinear regression models can be sub-optimal in experiments. Namely, we typically have to assume the model is correct (a sometimes heroic assumption, which is not guaranteed by randomization) and, even if the model is correct, the coefficient in that case will be a conditional effect which does not correspond to a well-defined effect in the whole population. In what follows we will discuss how to deal with the second issue, and then after that the first issue.

3.3 Recovering Population Effects via Regression

In the previous section we saw that, under parametric model assumptions (with randomization), the coefficient from a logistic regression model recovers a conditional odds ratio. Here we consider the question of how we might use this fit to estimate a marginal average treatment effect.

First, when we fit a logistic (or any other) regression model we are estimating the conditional expectation function

$$\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a).$$

For example with logistic regression we estimate this function with

$$\widehat{\mu}_a(x) = \text{expit}(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^{\mathrm{T}} x)$$

for $\widehat{\beta}$ the maximum likelihood estimates. Now recall that under the randomization assumption $A \perp\!\!\!\perp (Y^a, X)$ (with consistency) the average treatment effect is given by (Proposition 4)

$$\psi = \mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\},$$

which suggests the estimator

$$\widehat{\psi} = \mathbb{P}_n\Big\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\Big\} = \frac{1}{n}\sum_{i=1}^{n}\Big\{\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big\}.$$

This estimator is sometimes called the *plug-in*, *g-computation*, or *standardization* estimator. Conceptually, it is taking the estimated conditional effect $\widehat{\mu}_1(x) - \widehat{\mu}_0(x)$ and standardizing it to the (empirical) population distribution of covariates. You can also think of it as "imputing" an estimate of the effect $\mu_1(x) - \mu_0(x)$ for each person and averaging.

In practice one can fit a logistic regression and obtain predicted values under $A = 1$ and $A = 0$ separately, for everyone, regardless of actual observed treatment, then take the difference for each person, and average across people. The Appendix provides code in R to recover population effects via regression. Note there is no particular reason to favor logistic regression for constructing the regression estimates $\widehat{\mu}_a(x)$; one might instead consider linear regression, probit regression, regression trees, kernel estimators, splines, generalized additive models, the lasso, boosting, random forests, neural networks, deep learning, etc. However, we will see in Section 3.4 that this estimator can in general be improved.

Properties of the Plug-in Estimator In this section we analyze the simple plug-in estimator $\widehat{\psi} = \mathbb{P}_n\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\}$ used above. Note that here $\widehat{\mu}_a(X)$ need not be a logistic regression model, but any regression model that estimates the conditional expectation $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$. We analyze the plug-in estimator by finding answers to three standard questions: Is the estimator consistent? What is its convergence rate? What is its asymptotic limiting distribution?

Remark 8. Notice that if we let $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$, then $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ and $\psi = \mathbb{E}(f)$. Thus our estimator is a sample average of an estimated function, and our target estimand is an expectation of the true function f. Thus its performance will be very closely tied to the errors in estimating the function f with $\widehat{f}$. In the Appendix we give a short review and discussion of properties of estimated functions.

Consistency tells us whether our estimator is at least converging to the correct target as sample size increases (the lowest bar we would hope an estimator would clear), convergence rates tell us how quickly this convergence occurs (i.e., how much information in the sample does the estimator make use of), and asymptotic distributions tell us whether our estimator is well-behaved enough to give us hope for constructing confidence intervals and doing inference.

More specifically, our goal will often be to express $\widehat{\psi} - \psi$ as a (centered) sample average, plus some noise. We know how to analyze sample averages, since for any fixed function g of the iid observations Z, we have $(\mathbb{P}_n - \mathbb{P})g(Z) = (\mathbb{P}_n - \mathbb{E})g(Z) = O_{\mathbb{P}}(1/\sqrt{n})$ and in particular

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})g(Z) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[g(Z_i) - \mathbb{E}\{g(Z)\} \Big] \rightsquigarrow N\Big(0, \text{var}\{g(Z)\}\Big)$$

by the central limit theorem. Therefore the problem will be reduced to analyzing whatever the extra noise is.

First we will introduce a foundational decomposition for $\widehat{\psi}$ (in fact, for any estimator that takes a similar form); this will be crucial for many estimators we analyze throughout the chapter. More details and review can be found in Kennedy [2022].

Lemma 1. *Let $\widehat{\psi} = \mathbb{P}_n(\widehat{f}) = \frac{1}{n}\sum_i \widehat{f}(Z_i)$ be an estimator of the generic expectation $\psi = \mathbb{P}(f) = \mathbb{E}\{f(Z)\}$ based on n samples $(Z_1, ..., Z_n)$, where $\widehat{f}$ can be any estimator and $f : \mathcal{Z} \mapsto \mathbb{R}$ any function. Then we have the decomposition*

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2, \tag{4}$$

where

$$\begin{aligned} Z^* &= (\mathbb{P}_n - \mathbb{P})f, \\ T_1 &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f), \\ T_2 &= \mathbb{P}(\widehat{f} - f). \end{aligned}$$

Proof. We have

$$\begin{aligned} \widehat{\psi} - \psi &= \mathbb{P}_n(\widehat{f}) - \mathbb{P}(f) \\ &= (\mathbb{P}_n - \mathbb{P})\widehat{f} + \mathbb{P}(\widehat{f} - f) \\ &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) + (\mathbb{P}_n - \mathbb{P})f + \mathbb{P}(\widehat{f} - f) \\ &\equiv T_1 + Z^* + T_2, \end{aligned}$$

where the first line follows by definition, the second by adding and subtracting $\mathbb{P}(\widehat{f})$ (which we recall is not the same as $\mathbb{E}(\widehat{f})$, see the Appendix), and the third by adding and subtracting the quantity $(\mathbb{P}_n - \mathbb{P})f = (\mathbb{P}_n - \mathbb{E})f = \frac{1}{n}\sum_i [f(X_i) - \mathbb{E}\{f(X_i)\}]$. □

Lemma 1 applies to the plug-in estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ if as before we let $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$, and the target parameter $\psi = \mathbb{E}(f)$ is the population expectation of the true function f. Consequently, Lemma 1 allows us to express our estimator as a centered sample average plus noise. The first term Z^* in (4) is a nice centered sample average, and therefore by the central limit theorem it behaves as a normally distributed variable with variance $\text{var}(f)/n$, up to error $o_{\mathbb{P}}(1/\sqrt{n})$. Thus our problem is reduced to analyzing the two noise terms, denoted T_1 and T_2.

Remark 9. Note that the decomposition in (4) only relied on the estimator being a sample average of an estimated function $\widehat{f}$, and on the estimand being an expectation of a true function f. There was nothing special about ψ being the average treatment effect, or f being a regression function. The decomposition (4) thus arises often, since many estimands are expected values of (sometimes complicated) generic functions, which can be estimated by corresponding sample averages of estimates of these functions.

First, it turns out that the term T_1 is typically of smaller order than even the Z^* term. In fact, $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ under some regularity conditions, as long as $\widehat{f}$ is consistent for f in L_2 norm, i.e., as long as

$$\|\widehat{f} - f\|_2^2 = \int \{\widehat{f}(x) - f(x)\}^2 \, d\mathbb{P}(x) = o_{\mathbb{P}}(1),$$

(see Kennedy [2022]). Intuitively, however, this should not be too surprising, since T_1 is a centered sample average (just like Z^*), but in fact the quantity it is averaging is shrinking to zero with n (as long as $\widehat{f}$ is tending to f). This is like taking larger and larger centered sample averages of a random variable whose variance shrinks with n.

Now we turn to the last noise term T_2, which is the really interesting one. For many estimators we discuss in the chapter, the T_2 term will be particularly crucial, driving the rate of convergence and limiting distribution.

The Parametric Plug-in Estimator First we consider analyzing $T_2 = \mathbb{P}(\widehat{f} - f)$ in the case where $\widehat{f}$ is estimated with a (correct) parametric model, i.e., where

$$\widehat{f}(x) = f(x; \widehat{\beta}),$$

for some finite-dimensional parameter $\beta \in \mathbb{R}^p$. For example, when using the logistic regression model as before, we would have $f(x;\beta) = \text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} x) - \text{expit}(\beta_0 + \beta_2^{\mathrm{T}} x)$. Note in the parametric case we can view

$$T_2 = \mathbb{P}(\widehat{f} - f) = \int \{f(x; \widehat{\beta}) - f(x; \beta)\} \, d\mathbb{P}(x) \equiv g(\widehat{\beta}) - g(\beta)$$

as a simple difference in functions $\widehat{\beta}$ and β, where the function g will be smooth if f is. Therefore we will first understand the error between $\widehat{\beta}$ and β, and then use the delta method.

For most smooth parametric models, the estimator $\widehat{\beta}$ will solve an estimating equation based on some mean-zero estimating function m that is smooth in β. For example, the logistic regression estimator solves an estimating equation based on the estimating function (or score function)

$$m(Z; \beta) = \begin{pmatrix} 1 \\ A \\ X \end{pmatrix} \Big\{ Y - \text{expit}(\beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X) \Big\},$$

so that

$$\mathbb{P}_n\{m(Z; \widehat{\beta})\} = 0$$

by definition. The next result shows that such solutions to finite-dimensional estimating equations behave like sample averages (for a proof, see Theorem 5.23 of van der Vaart [2000]).

Lemma 2. *Suppose the estimator $\widehat{\beta} \in \mathbb{R}^p$ solves an estimating equation so that $\mathbb{P}_n\{m(Z; \widehat{\beta})\} = 0$. Assume $m(z; \beta) \in \mathbb{R}^p$ is Lipschitz in β, and that $\mathbb{E}\{m(z; \beta)\}$ is differentiable at the population β satisfying $\mathbb{E}\{m(Z; \beta)\} = 0$ with nonsingular derivative matrix. Then*

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[\left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z; \beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}), \tag{5}$$

Corollary 1. *Under the conditions of Lemma 2, the estimating equation estimator $\widehat{\beta}$ is root-n consistent and asymptotically normal.*

Now we have all the tools we need to analyze the quantity $\mathbb{P}(\widehat{f} - f)$ and thus the estimator $\widehat{\psi}$ in the parametric case.

Theorem 3. *Let $f(x) = \mu_1(x) - \mu_0(x)$ and $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$, so that $\psi = \mathbb{E}\{f(X)\}$ is the average treatment effect. Assume the parametric model $\mu_a(x) = \mu_a(x; \beta)$ for some $\beta \in \mathbb{R}^p$, and that the estimator $\widehat{\beta}$ satisfies the conditions of Lemma 2. Then*

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z; \beta) + o_{\mathbb{P}}(1/\sqrt{n}),$$

where

$$g(z; \beta) = f(x; \beta) + \frac{\partial \mathbb{E}\{f(X; \beta)\}}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(z; \beta),$$

and so is root-n consistent and asymptotically normal.

Proof. By Lemma 1 we have

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2$$

where $Z^* = (\mathbb{P}_n - \mathbb{P})f$ and T_1 and T_2 defined accordingly. By Lemma 2 we have

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[\left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z; \beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}),$$

which also is enough to imply $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Further by the delta method we have

$$T_2 = g(\widehat{\beta}) - g(\beta) = (\mathbb{P}_n - \mathbb{P}) \left[\frac{\partial g(\beta)}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z; \beta) \right] + o_{\mathbb{P}}(1/\sqrt{n})$$

for $g(\beta) = \mathbb{E}\{f(Z; \beta)\}$. Combining the terms gives the result. □

To summarize, when $\widehat{\mu}$ is estimated with a correct parametric model, the resulting plug-in estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ for $f = \mu_1 - \mu_0$ is root-n consistent for the causal effect ψ and asymptotically normal.

When the parametric model for μ is correct, this plug-in estimator is most efficient (this follows from classical low-dimensional parametric maximum likelihood theory); the intuition is that with a correct model we can "predict" the treatment effect much more precisely than say with the difference-in-means estimator. Further confidence intervals can be constructed using estimates of the closed-form asymptotic variance given above, or via bootstrap (which is typically easier).

Of course, we may have serious doubts that our parametric models are actually correct, especially when X contains some continuous covariates or is high-dimensional. At best, such models may be a modestly biased approximation, but at worst – when very misspecified – they may irreparably bias our estimation procedure, yielding estimates that are not only far away from the truth, but to an unknown extent.

The Nonparametric Plug-in The discussion at the end of the previous section raises the question of how the plug-in estimator would behave if we used a more flexible estimator to construct $\widehat{\mu}$, say random forests or the lasso or deep learning. In this case, the central limit theorem term Z^* in the decomposition (4) still behaves as a mean-zero normally distributed random variable with variance $\mathrm{var}(f)/n$, since it does not depend on the estimated $\widehat{f}$. Further, even when μ is treated as a potentially infinite-dimensional function and estimated flexibly and data-adaptively, the term T_1 can still be of smaller order (though one may need to use sample splitting, as discussed in more detail in Kennedy [2022] and later).

Unfortunately the picture is not as rosy for the important T_2 term in (4). If all we know about the flexible estimator $\widehat{f}$ is a high-level rates of convergence, say in L_2 norm, then typically all we can say about T_2 is

$$T_2 = \mathbb{P}(\widehat{f} - f) \leq \sqrt{\mathbb{P}\{(\widehat{f} - f)^2\}} = \|\widehat{f} - f\|_2,$$

where the second inequality uses Cauchy-Schwarz. This means in general we would expect the plug-in estimator $\widehat{\psi}$ to inherit the (typically slow) rate of convergence of the nonparametric estimator $\widehat{f}$. This is a problem since for most realistic infinite-dimensional function classes the L_2 error will be far away from $1/\sqrt{n}$. For example when f lies in a Hölder class with smoothness s (i.e., all partial derivatives up to order $s-1$ are bounded and s^{th} derivatives continuous) then for *any* estimator $\widehat{f}$ the rate cannot be any faster than

$$\|\widehat{f} - f\|_2 \ \gtrsim \ n^{-s/(2s+d)}$$

uniformly over the Hölder class [Tsybakov, 2009]; note this rate is always slower than $\sqrt{n}$. Neural network classes are known for yielding dimension-independent rates [Györfi et al., 2002], but even these are roughly of order $n^{-1/4}$, somewhat of a far cry from $1/\sqrt{n}$.

Further, when $\widehat{\mu}$ is estimated flexibly with modern nonparametric tools, we do not only pay a price in the rate of convergence – it will generally also be true that, even if we can derive a tractable limiting distribution, there will be some smoothing bias, so confidence intervals will not be correctly centered (even using the bootstrap) and thus will not cover at the nominal level. However, often complex nonparametric estimators do not even yield tractable limiting distributions, even uncentered.

3.4 Efficient Model-Free Estimation

At this point it appears the analyst is in a bit of a quandary. One could use the simple difference-in-means estimator, which is root-n consistent and asymptotically normal *under no modeling assumptions*; however it completely ignores covariate information and so may be quite inefficient. Alternatively one could model the regression function and use the plug-in estimator. However if parametric models are used to achieve root-n rates and small confidence intervals, one risks bias due to model misspecification; on the other hand, if one models the regression functions nonparametrically, then the curse of dimensionality subjects us to slow rates of convergence, and at a loss for confidence intervals and inference.

What should be done? Is there any way to get the best of both worlds, using the covariates to gain efficiency over the difference-in-means estimator, but retaining its model-free benefits and not risking bias?

The Doubly Robust Estimator It turns out there exists a bias-corrected estimator, whose validity is based on randomization, yet which can incorporate regression predictions to increase efficiency:

$$\widehat{\psi} = \mathbb{P}_n \left[\Big\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \Big\} + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi} \right) \Big\{ Y - \widehat{\mu}_A(X) \Big\} \right], \tag{6}$$

where $\widehat{\mu}_a(x)$ is an estimate of the regression function $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$ and $\pi = \mathbb{P}(A = 1)$ is the (known) randomization probability.

The estimator (6) can be viewed as the plug-in estimator $\mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ plus a "correction" term that incorporates the randomization probabilities π. It goes by various names, including: model-assisted Horvitz-Thompson, bias-corrected plug-in, semiparametric or semiparametric efficient, augmented inverse-probability-weighted (AIPW), doubly robust, double machine learning, etc. We mostly refer to it as doubly robust.

The estimator (6) has an interesting and somewhat difficult-to-trace history across fields. Here is an abbreviated and limited portion of its path across the literature: In survey sampling problems, Cochran [1977] and others used simple regression models in an agnostic way to improve the efficiency of the unbiased Horvitz-Thompson estimator from 1952. Robins and Rotnitzky [1995], Robins et al. [1994, 1995] studied efficient semiparametric estimation in general missing data problems (extending work by Bickel et al. [1993] and Pfanzagl [1982] and others), and presented a version of this estimator (6) where nuisance quantities were estimated with parametric models. Robins and Wang [2000] started referring to the estimator (6) as "doubly protected", and Robins and Rotnitzky [2001] and Bang and Robins [2005] as "doubly robust". In a series of papers, Tsiatis and colleagues [Davidian et al., 2005, Leon et al., 2003, Yang and Tsiatis, 2001, Zhang et al., 2008] applied the theory from Robins and others to randomized experiments, focusing on efficiency concerns. These papers are a nice introduction to the estimator (6) in the experimental setup. In the early to mid 2000s, van der Laan and Robins [2003] and others started developing theory for the case where nuisance estimators such as $\widehat{\mu}_a$ are estimated nonparametrically. The estimator and related methods have been recently studied in econometrics [Chernozhukov et al., 2018], with more of a focus on high-dimensional sparse models.

In fact it can be shown that any (regular) $\sqrt{n}$-consistent and asymptotically normal estimator can be written in the form (6), for some choice of $\widehat{\mu}_a$ [Tsiatis, 2006, van der Laan and Robins, 2003]. For example, the difference-in-means estimator is recovered if $\widehat{\mu}_a = \mathbb{P}_n(Y \mid A = a)$, and the Horvitz-Thompson or inverse-probability-weighted estimator if $\widehat{\mu}_a = 0$. Shortly we will study some cases where, surprisingly, the parametric plug-in takes this form with for example $\widehat{\mu}_a = g(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^{\mathrm{T}} x)$. This is one of the reasons it is a bit unclear where the estimator originated, since it includes many variants as a special case.

Remark 10. As we did above, at several points in this chapter we will refer to *regular* estimators. For the time being, a regular estimator can be taken to mean an estimator whose limiting distribution is insensitive to local perturbations of the data-generating process. Imposing regularity rules out *super-efficient* estimators, for example, which trade very good performance at a particular $\mathbb{P}$ for very bad performance "near" $\mathbb{P}$. More discussion can be found in Tsiatis [2006] and van der Vaart [2000].

As mentioned earlier, the estimator (6) can be interpreted as a corrected version of the plug-in estimator $\widehat{\psi}_{pi} = \mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ since

$$\widehat{\psi} = \widehat{\psi}_{pi} + \mathbb{P}_n\left[\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \widehat{\mu}_A(X)\right\}\right],$$

where we will see the correction term removes any bias afflicting the regression estimator $\widehat{\mu}_a$.

Analogously, the doubly robust estimator can also be viewed as a corrected (or "augmented") version of the inverse-probability weighted (Horvitz-Thompson) estimator $\widehat{\psi}_{ipw} = \mathbb{P}_n\{\left(\frac{AY}{\pi} - \frac{1-A}{1-\pi}\right)Y\}$ since

$$\widehat{\psi} = \widehat{\psi}_{ipw} + \mathbb{P}_n\left[\left(1 - \frac{A}{\pi}\right)\widehat{\mu}_1(X) - \left(1 - \frac{1-A}{1-\pi}\right)\widehat{\mu}_0(X)\right].$$

We know from the previous chapter that $\widehat{\psi}_{ipw}$ is already unbiased; thus the above augmentation term is reducing variance rather than bias.

Remark 11. Note that the doubly robust estimator requires no extra model fitting beyond that already required to construct the plug-in estimator. The appendix presents example code showing how to correct the plug-in estimator in R.

A natural question about the doubly robust estimator is: where does the correction come from, and why does it take that specific form? A complete answer to this is non-trivial; however some short discussion is still useful. The form of the correction comes from nonparametric efficiency theory for functional estimation [Bickel et al., 1993, Kennedy, 2022, Tsiatis, 2006, van der Laan and Robins, 2003], and there are at least two high-level heuristics available. The first is that the average treatment effect parameter $\psi = \psi(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}(\mu_1 - \mu_0)$ is a "smooth" functional, when viewed as a map from probability distributions $\mathbb{P}$ to the real line; and this smoothness allows for convenient and effective bias correction. The second is that a randomized experiment with known treatment mechanism leads to a semiparametric model for the distribution $\mathbb{P}$ from which we sample: part of the distribution $\mathbb{P}$ is known (the conditional distribution of treatment given any covariates) while the rest is left unrestricted (the covariate distribution and the conditional distribution of the outcome given covariates and treatment). Under this semiparametric model, one can use tools from efficiency theory to derive the form of *all* possible (regular) asymptotically normal estimators of the parameter ψ, and subsequently find the one with the smallest variance.

Answering the question of *why* the correction works is easier than answering where it comes from. This is the focus of the next section.

Remark 12. In what follows we consider the case where the regression estimator $\widehat{\mu}_a$ is constructed from a separate training sample D^n independent of the experimental sample $Z^n = \{(X_1, A_1, Y_1), ..., (X_n, A_n, Y_n)\}$. In practice one can simply randomly split the sample, using half as D^n for training and the other half as Z^n for estimation. Note that in this case, variance results should really be framed in terms of $n/2$ instead of n. However one can combat this loss of efficiency with an easy fix: after constructing the sample-split estimator, swap the samples, using Z^n for training and D^n for estimation, and then average the resulting estimators (similarly to k-fold sample-splitting). This approach recovers full sample size efficiency [Chernozhukov et al., 2018, Kennedy, 2022, Robins et al., 2008, Zheng and van der Laan, 2010].

There are two reasons for doing sample splitting: the first is that the analysis is more straightforward, and the second is that it prevents overfitting and allows for the use of arbitrarily complex estimators $\widehat{\mu}_a$ (e.g., random forests, boosting, neural nets). Without sample splitting, one would have to restrict the complexity of the estimator $\widehat{\mu}_a$ via empirical process conditions. Intuitively, this is because the estimator $\widehat{\psi}$ is using the data twice: once to estimate the unknown function μ_a and once to estimate the bias correction. Sample splitting ensures that these tasks are accomplished independently.

Properties of the Doubly Robust Estimator In this section we study the bias, variance, and limiting distribution of the estimator (6).

As in our analysis of the plug-in estimator in the previous section, we note that our estimator can be written as a sample average of an estimated function. Namely $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ where now $\widehat{f} = f(\widehat{\mu}) \equiv f_1(\widehat{\mu}) - f_0(\widehat{\mu})$ for

$$f_a(\overline{\mu}) \equiv \overline{\mu}_a(X) + \frac{\mathbb{1}(A = a)}{\mathbb{P}(A = a)}\Big\{Y - \overline{\mu}_A(X)\Big\}. \tag{7}$$

First we tackle the bias of $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$, under *no modeling assumptions whatsoever.*

Theorem 4. *Consider an iid Bernoulli experiment with known $\mathbb{P}(A = 1) = \pi$. Then the doubly robust estimator $\widehat{\psi}$ in (6) is unbiased for the average treatment effect when the regression estimates $\widehat{\mu}_a$ are constructed from a separate independent sample.*

Proof. We derive the bias for $\psi_1 = \mathbb{E}(Y^1)$ with $\widehat{\psi}_1 = \mathbb{P}_n(\widehat{f}_1)$ since the logic is exactly the same for $\mathbb{E}(Y^0)$ and the difference $\psi = \psi_1 - \psi_0$. First note that for *any* $\overline{\mu}_1$

$$\begin{aligned} \mathbb{P}\{f_1(\overline{\mu})\} &= \mathbb{P}\left[\overline{\mu}_1(X) + \frac{A}{\pi}\Big\{Y - \overline{\mu}_1(X)\Big\}\right] \\ &= \mathbb{P}\left[\overline{\mu}_1(X) + \frac{\pi}{\pi}\Big\{\mu_1(X) - \overline{\mu}_1(X)\Big\}\right] \\ &= \mathbb{E}\{\mu_1(X)\} = \psi_1, \end{aligned} \tag{8}$$

where the second equality used iterated expectation and the Bernoulli randomization. Therefore we have $\mathbb{E}(\widehat{\psi}_1 \mid D^n) = \mathbb{P}\{f(\widehat{\mu}_1)\} = \psi_1$, where the first equality uses the fact that $\widehat{\mu}_a(x)$ is fixed given independent D^n and the iid assumption, and the second (8). □

Theorem 4 is a simple but powerful result. It shows the doubly robust estimator is exactly unbiased, *for any choice of* regression estimator $\widehat{\mu}_a$. Hence, although the estimator $\widehat{\psi}$ exploits covariate information, its bias is not at all affected by accidentally misspecified models or biased regression estimators with slow convergence rates.

Remark 13. Theorem 4 also has an important implication for understanding the variance and limiting distribution of $\widehat{\psi}$. Namely, the logic in the proof shows that

$$\mathbb{P}\{f(\overline{\mu})\} = \psi$$

for *any* (fixed) $\overline{\mu}$. This means that, since $\widehat{\psi}$ is a sample average of an estimated function and thus the decomposition from Lemma 1 holds, we can write

$$\begin{aligned} \widehat{\psi} - \psi &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - \overline{f}) + \mathbb{P}(\widehat{f} - \overline{f}) + (\mathbb{P}_n - \mathbb{P})\overline{f} \\ &\equiv T_1 + T_2 + Z^* \end{aligned} \tag{9}$$

for *any* $\overline{f} = f(\overline{\mu})$. Since it will be useful in our analysis for $\widehat{f}$ to be consistent for $\overline{f}$, we will simply define $\overline{f} = f(\overline{\mu})$ to be the corresponding probability limit, i.e., by taking $\overline{\mu}_a$ to be a fixed function such that $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$. We will see that this will allow us to completely sidestep whether the estimator $\widehat{\mu}_a$ is consistent for the *true* regression function μ_a, and instead just require that it be consistent for *something.*

Now we tackle the limiting distribution of $\widehat{\psi}$. Recall we know Z^* in the decomposition (9) is asymptotically normal, so we only need to understand the T_1 and T_2 terms. First we provide a general analysis of the first term $T_1 = (\mathbb{P}_n - \mathbb{P})(\widehat{f} - \overline{f})$ in that decomposition.

Lemma 3. *Let $\mathbb{P}_n$ denote the empirical measure over $Z^n = (Z_1, ..., Z_n)$, and let $\widehat{f}(z)$ be any function estimated from a sample $D^N = (Z_{n+1}, ..., Z_{n+N})$, which is independent of Z^n. Then*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right).$$

Proof. See Kennedy et al. [2020]. □

Lemma 3 shows that T_1 terms are asymptotically negligible, i.e., that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$, as long as $\widehat{f}$ is consistent for f (or $\overline{f}$ in our case, which will hold by definition). The next result gives the limiting distribution of the doubly robust estimator, under no assumptions beyond the experiment design (and iid sampling) and that the regression estimators $\widehat{\mu}_a$ converge to anything at any rate.

Theorem 5. *Consider an iid Bernoulli experiment with known $\mathbb{P}(A = 1) = \pi$. Suppose the regression estimators $\widehat{\mu}_a$ are:*

1. *constructed from a separate independent sample, and*
2. *consistent (at any rate) for some functions $\overline{\mu}_a$ (not necessarily the true regression functions μ_a) in the sense that $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$.*

Then the doubly robust estimator $\widehat{\psi}$ is root-n consistent and asymptotically normal with

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\Big(0, var(\overline{f})\Big),$$

where $\overline{f} = f(\overline{\mu})$ is defined as in (7).

Proof. By Lemma 1 we can write the decomposition (9) with $\overline{f} = f(\overline{\mu})$ for any $\overline{\mu}$. We will define $\overline{\mu}$ as the probability limit of $\widehat{\mu}$, as in the statement of the theorem.

By Lemma 3, we have $T_1 = O_{\mathbb{P}}(\|\widehat{f} - \overline{f}\|/\sqrt{n})$. Now note

$$\begin{aligned}\|\widehat{f}_1 - \overline{f}_1\|^2 &= \left\| \Big\{ \widehat{\mu}_1 - \overline{\mu}_1 \Big\} \left\{ 1 - \frac{A}{\pi} \right\} \right\|^2 \\ &= \left(\frac{\text{var}(A)}{\pi^2} \right) \int \Big\{ \widehat{\mu}_1(x) - \overline{\mu}_1(x) \Big\}^2 \, d\mathbb{P}(x) \\ &= \left(\frac{1 - \pi}{\pi} \right) \|\widehat{\mu}_1 - \overline{\mu}_1\|' 2\end{aligned}$$

where the second equality used the Bernoulli randomization. The same logic applies to $\|\widehat{f}_0 - \overline{f}_0\|$, and so by the triangle inequality

$$T_1 = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - \overline{f}\|}{\sqrt{n}} \right) = O_{\mathbb{P}}\left(\frac{\|\widehat{\mu}_1 - \overline{\mu}_1\| + \|\widehat{\mu}_0 - \overline{\mu}_0\|}{\sqrt{n}} \right),$$

which is $o_{\mathbb{P}}(1/\sqrt{n})$ since $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$ by definition. For the T_2 term, we have $\mathbb{P}(\widehat{f} - \overline{f}) = 0$ by (8). This gives the result. □

Theorem 5 shows that not only is the doubly robust estimator $\widehat{\psi}$ unbiased for any choice of regression estimator, it is also root-n consistent and asymptotically normal – even if the estimators $\widehat{\mu}_a$ are completely misspecified, and or converging at arbitrarily slow rates. This is a surprisingly impressive result, given that essentially no model assumptions were used.

This immediately implies that distribution-free confidence intervals can be constructed as in the following corollary.

Corollary 2. *Under the assumptions of Theorem 5, a distribution-free asymptotic 95% confidence interval for the average treatment effect ψ is given by*

$$\widehat{\psi} \pm 1.96 \sqrt{\frac{\widehat{var}\{f(\widehat{\mu})\}}{n}}.$$

Further, finite-sample variance bounds can be constructed using the same logic as in the proof of Theorem 5.

Efficiency In the previous section we learned the surprising result that the sample-split doubly robust estimator is exactly unbiased for any choice of regression estimator $\widehat{\mu}_a$, and root-n consistent and asymptotically normal as long as $\widehat{\mu}_a$ converges to some fixed function at any rate. As would be expected, the efficiency of the doubly robust estimator depends on the probability limits $\overline{\mu}_a$ that the regression estimators $\widehat{\mu}_a$ converge to. This raises some important questions: What is the best possible (i.e., most efficient) probability limit $\overline{\mu}_a$? Is the doubly robust estimator necessarily more efficient than the difference-in-means or Horvitz-Thompson estimator? Since the difference-in-means and Horvitz-Thompson estimators can be written as variants of the doubly robust estimator, for particular choices of $\widehat{\mu}_a$, the best choice of $\overline{\mu}_a$ will dominate others in this class.

The next result shows what one might expect: the best limit $\overline{\mu}_a$ in terms of efficiency is the *true* regression function μ_a (recall this limit is irrelevant for bias since $\widehat{\psi}$ is unbiased for any $\widehat{\mu}_a$).

Theorem 6. *Define $f(\overline{\mu})$ as in (7). Then for any $\overline{\mu} = (\overline{\mu}_1, \overline{\mu}_0)$ with $\overline{\mu}_a : \mathcal{X} \mapsto \mathbb{R}$*

$$var\{f(\overline{\mu})\} \geq var\{f(\mu)\},$$

where $\mu = (\mu_1, \mu_0)$ denotes the true regression functions.

Proof. We have

$$\begin{aligned}
\text{var}\{f(\overline{\mu})\} &= \text{var}\left[\overline{\mu}_1(X) - \overline{\mu}_0(X) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \overline{\mu}_A(X)\right\}\right] \\
&= \text{var}\{f(\mu)\} + \text{var}\left\{\left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\} \\
&\quad + 2\text{cov}\left\{f(\mu), \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\}.
\end{aligned}$$

But the latter covariance is zero since

$$\begin{aligned}
&\text{cov}\left\{f(\mu), \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\} \\
&\quad = \mathbb{E}\left[(\mu_1 - \mu_0 - \psi)\left\{\left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\}\right] = 0,
\end{aligned}$$

where the second equality follows from iterated expectation since $\mathbb{E}\{f(\mu) \mid X, A\} = \mu_1 - \mu_0$, and since $A \perp\!\!\!\perp X$ so that $\mathbb{E}\{Ag(X)\} = \pi\mathbb{E}\{g(X)\}$ for any g. This gives the result □

Theorem 6 is critically informative about how to construct the doubly robust estimator $\widehat{\psi}$ in practice. Namely, it indicates that we should estimate the regression functions as flexibly as possible: bias is zero regardless, and efficiency is optimized when the regression functions are estimated consistently. This is a special case not often seen in statistics where there is essentially no penalty (at least asymptotically) for slow rates of convergence, and important benefits for consistency.

However the second question still remains: when based on a misspecified model for μ_a, does the doubly robust estimator necessarily still improve efficiency (say relative to the Horvitz-Thompson estimator)? In fact, this is not necessarily so, for particularly misspecified choices of $\widehat{\mu}_a$. However, there are multiple approaches that can be used to guarantee efficiency gains. One simple option proposed by Rubin and van der Laan [2008] is to posit a working parametric model $\mu_a(x) = \mu_a(x;\beta)$, but rather than estimating the parameters via maximum likelihood, instead estimate parameters by picking those that minimize an estimator of the variance, i.e., use

$$\widetilde{\beta} = \arg\min_{\beta} \ \widehat{\text{var}}\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \mu_A(X;\beta)\right\}\right].$$

Other similar approaches are also possible [Tan, 2010].

Back to the Plug-In In the previous section we saw strong evidence that, if one wants to remain agnostic about the data-generating process beyond the known randomization probabilities, retaining robustness while exploiting covariate information to gain efficiency, then the doubly robust estimator (6) using a flexible regression estimator is a good choice. In particular, it will be root-n consistent and asymptotically normal as long as the regression estimator $\widehat{\mu}_a$ converges to *anything at any rate*, and if the regression estimator $\widehat{\mu}_a$ is consistent for the true regression function (again *at any rate*) then it will be asymptotically efficient.

However, in practice, applied researchers often use ordinary least squares or plug-in estimators based on parametric models. Is there any basis for trusting such results? In fact, the following result shows that some if not many parametric plug-in estimators can be represented in the doubly robust form: they are doubly robust estimators disguised as plug-ins. (Though it is important to note that this is not true of all plug-in estimators).

Proposition 5. *Suppose regression predictions $\widehat{\mu}_a$ satisfy*

$$\mathbb{P}_n\left[(1,A)^{\mathrm{T}}\Big\{Y-\widehat{\mu}_A(X)\Big\}\right]=0 \tag{10}$$

where $A \perp\!\!\!\perp X$ is randomized according to a Bernoulli experiment. Then the parametric plug-in estimator $\mathbb{P}_n\{\widehat{\mu}_1(X)-\widehat{\mu}_0(X)\}$ is numerically equivalent to the doubly robust estimator

$$\mathbb{P}_n\left[\Big\{\widehat{\mu}_1(X)-\widehat{\mu}_0(X)\Big\}+\left(\frac{A}{\pi}-\frac{1-A}{1-\pi}\right)\Big\{Y-\widehat{\mu}_A(X)\Big\}\right].$$

Proof. Since $\mathbb{P}_n[(1,A)^{\mathrm{T}}\{Y-\widehat{\mu}_A(X)\}]=0$ it follows that

$$\frac{1}{\pi}\mathbb{P}_n\left[A\Big\{Y-\widehat{\mu}_A(X)\Big\}\right]=\frac{1}{1-\pi}\mathbb{P}_n\left[A\Big\{Y-\widehat{\mu}_A(X)\Big\}\right]=\frac{1}{1-\pi}\mathbb{P}_n\left[\Big\{Y-\widehat{\mu}_A(X)\Big\}\right]=0.$$

Therefore

$$\mathbb{P}_n\left[\left(\frac{A}{\pi}-\frac{1-A}{1-\pi}\right)\Big\{Y-\widehat{\mu}_A(X)\Big\}\right]=0$$

so that the correction term in the doubly robust estimator is zero, and the plug-in and doubly robust estimator are equal. □

The sufficient condition (10) in Proposition 5 says that the $\widehat{\mu}_a$ residuals must average to zero both in the whole sample and among the treated. This will hold for example in generalized linear models with an intercept and main effect term for treatment. Thus Proposition 5 shows that, although our earlier analysis of the parametric plug-in appeared to hinge on restrictive parametric model assumptions, this is not necessarily so – at least in Bernoulli experiments, and for plug-ins based on models with an intercept and main effect for treatment. Such parametric plug-in estimators will be root-n consistent for the average treatment effect (and asymptotically normal), even under misspecification of the regression estimator $\widehat{\mu}_a$, as long as it converges in probability to anything at any rate (a very weak condition).

Remark 14. Although a plug-in whose regression estimates satisfy (10) will take on all the advantageous robustness and efficiency properties of the doubly robust estimator, note that variance estimates must be based on the doubly robust variance as in Corollary 2. Otherwise corresponding confidence intervals and hypothesis tests may not be valid.

Remark 15. The condition (10) is generally not enough to ensure that a plug-in will be doubly robust outside of Bernoulli experiments, e.g., if treatment is not completely independent of covariates. For example, in conditionally randomized experiments, plug-ins would in general require correctly specified outcome regression models for fast root-n rates.

4 Selected Advanced Topics

4.1 Conditional Randomization

In conditionally randomized experiments, the randomization probabilities can differ by covariate values, e.g., in a stratified Bernoulli experiment one sets

$$\mathbb{P}(A = 1 \mid X = x, Y^a) = \pi(x)$$

where the function $\pi(x)$ can vary with x (recall $\pi(x) = \pi$ in a Bernoulli experiment).

Experiments may use conditional or stratified randomization to improve efficiency (e.g., by treating more units at covariate values where treated outcomes are more variable than control outcomes), or to improve subject outcomes (e.g., by treating more units at covariate values where treated outcomes are likely to be higher than control outcomes, and treating fewer when treatment is ineffective or even harmful).

Doubly robust estimators take the same form as (6), but replace π with $\pi(x)$, i.e.,

$$\mathbb{P}_n \left[\left\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \right\} + \left\{ \frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)} \right\} \left\{ Y - \widehat{\mu}_A(X) \right\} \right]$$

and the logic of the theoretical analysis is the same as well.

There are some important differences between simple Bernoulli experiments and conditionally randomized designs, however. First, the difference-in-means estimator is no longer a valid estimator, since it is no longer the case that $A \perp\!\!\!\perp Y^a$ or $A \perp\!\!\!\perp (X, Y^a)$; instead, in a conditionally randomized experiment it only holds that $A \perp\!\!\!\perp Y^a \mid X$. Second, as mentioned above, plug-in estimators are not in general doubly robust in conditionally randomized designs, even when they satisfy the condition (10); this is because the randomization probabilities cannot be brought outside the average as in the proof of Proposition 5.

4.2 Cluster Randomized Trials

Another important and common experimental design is the cluster randomized trial, in which treatments are randomized to groups of individuals (e.g., states, hospitals, schools, etc.). Importantly, in this setting there is no variation in treatment within a cluster (only across clusters); thus all individuals in a particular cluster receive the same treatment. Hence the data structure is n observations of

$$Z_i = \{X_i = (V_i, X_{i1}, ..., X_{iN_i}), A_i, Y_i = (Y_{i1}, ..., Y_{iN_i})\},$$

where V_i denotes cluster-level covariates for the ith cluster, X_{ij} denotes covariates measured on the jth individual in cluster i, N_i is the number of individuals in cluster i, A_i is an indicator of whether cluster i received treatment, and Y_{ij} is the outcome for the jth individual in cluster i. Here randomization ensures that for each $i = 1, ..., n$

$$A_i \perp\!\!\!\perp (Y_{i1}^a, ..., Y_{iN_i}^a) \mid V_i, (X_{i1}, ..., X_{iN_i}).$$

It is commonly assumed that the cluster observations Z_i are independent and identically drawn from some target population, while units within clusters can be arbitrarily dependent. (However it is also possible to conduct design-based inference only using the treatment distribution).

One can consider average effects defined via expected values of quantities like

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}^{a},$$

i.e., the average outcome if all individuals in all clusters received treatment $A = a$. For example letting $\overline{Y}_i^a = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}^a$ denote the average outcome in cluster j, the expected value of the above can be written $\mathbb{E}(\overline{Y}^a)$ if clusters are independent and identically distributed. Letting $\overline{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$ denote the average outcome in cluster i, a doubly robust estimator analogous to (6) is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\widehat{\mathbb{E}}(\overline{Y}_i \mid X_i, A_i = 1) - \widehat{\mathbb{E}}(\overline{Y}_i \mid X_i, A_i = 0)\right\} + \left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right)\left\{\overline{Y}_i - \widehat{\mathbb{E}}(\overline{Y}_i \mid X_i, A_i)\right\}\right],$$

with analogous variance estimators and limiting distributions defined as in previous sections. Note that the regression predictions here are $\mathbb{E}(\overline{Y}_i \mid X_i, A_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{E}(Y_{ij} \mid X_i, A_i)$, which can be estimated by regressing the average cluster outcomes on all cluster-level and individual-level covariates (for everyone in the cluster, as well as treatment), since $X_i = (V_i, X_{i1}, ..., X_{iN_i})$. This would in general be a high-dimensional regression problem, and therefore sometimes dimension-reducing independence assumptions are employed (e.g., assuming individuals' outcomes are conditionally independent of others' covariate values).

Finally we note that there are important challenges unique to cluster randomized trials, including for example alternative definitions of causal effects and multiple possible asymptotic regimes and independence assumptions; we refer to Balzer et al. [2019, 2021], Benitez et al. [2021] for more details.

4.3 Interference

So far we have relied on the assumption that the potential outcomes of unit i do not depend on treatments assigned to units $j \neq i$, via the consistency assumption that $Y = Y^a$ whenever $A = a$. However, this may be unrealistic in settings where subjects are connected in a network, so that there is *interference* between units. For example in an experiment studying the effectiveness of a vaccine, whether an individual becomes sick depends not only on their own vaccination status, but also on the vaccination status of their community. For an example related to AI for social impact, consider an evaluation of whether intelligent tutoring systems improve student outcomes [Feng et al., 2014]. There could be spillover effects from the treatment into control groups if, for example, control students studied with the treated, e.g., if the students from the treatment group shared their knowledge indirectly with their control peers.

With intereference, the potential outcome for unit i can depend on the treatment assignment of other units (potentially all other units), and so potential outcomes need to be indexed by these other treatment assignments. For example the potential outcome for subject i if all n subjects were treated at levels $(a_1, a_2, ..., a_n) \equiv a^n$ can be written as

$$Y_i(a^n)$$

(we move from subscripts to parentheses to ease notation). Letting $A^n = (A_1, ..., A_n)$ denote the vector of observed treatment assignments for all n units, the observed outcome is then given by $Y_i = Y_i(A^n)$. If every subject's outcome can depend on all other subjects' treatments, there would be no way to estimate treatment effects without very strong assumptions. Thus one can also introduce an $n \times n$

adjacency matrix M whose elements $M_{ij} = 1$ if subjects i and j are connected, and are zero otherwise.

The earliest work on interference (e.g., Hudgens and Halloran [2008]) typically considered partial interference, where there are natural groups (e.g., households, classrooms), and interference can occur within but not across groups. In this setting we can let $\overline{a}_i = \{a_j : M_{ij} = 1\}$ denote the treatments of those units in the same group as subject i, and then

$$Y_i(a^n) = Y_i(a_i, \overline{a}_i),$$

i.e., unit i's outcome is only affected by treatment assignments of those units in the same group. One can then consider for example individual direct effects of treatment

$$Y_i(a_i = 1, \overline{a}_i) - Y_i(a_i = 0, \overline{a}_i)$$

which represents how unit i's outcome would change if they received treatment versus control, but all other units' treatments in their group remained unchanged. Similarly an indirect effect is

$$Y_i(a_i = 0, \overline{a}_i) - Y_i(a_i = 0, \overline{a}_i')$$

which represents how unit i's outcome would change if they received control but others' assignments in their group changed from $\overline{a}_i$ to $\overline{a}_i'$. Hudgens and Halloran [2008] provide unbiased estimators and inferential tools for averages of the above direct and indirect effects in two-stage randomized experiments, where for example groups are randomized to receive a higher or lower proportion of treatment, with individuals within groups randomized based on that proportion.

For more recent work on interference in more complex network settings we refer to [Aronow and Samii, 2017, Ogburn et al., 2017], for example.

5 Discussion

In this chapter we have reviewed concepts behind and methods for the statistical analysis of randomized experiments. After starting with simple randomized experiments consisting of only treatment and outcome data collected, we paid special attention to the common setting where auxiliary covariate information is also available. In these cases one needs to take special care to exploit the covariate information for efficiency gains, while relying solely on randomization for inference and not risking bias from model misspecification. We showed in particular how a semiparametric doubly robust estimator allows for such balance of efficiency and robustness.

There are many crucial topics beyond the scope of this chapter, including noncompliance, missing data, longitudinal treatments, heterogeneous effect estimation, and more. For these and other topics, as well as more details on concepts included here, we refer to Hernán and Robins [2010], Imbens and Rubin [2015], Rosenbaum [2002], Tsiatis [2006], van der Laan and Robins [2003] and other similar resources, such as those mentioned in the text.

Acknowledgements

EK gratefully acknowledges support from NSF DMS Grant 1810979, NSF CAREER Award 2047444, and NIH R01 Grant LM013361-01A1.

Bibliography

P. M. Aronow and C. Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

L. B. Balzer, W. Zheng, M. J. van der Laan, and M. L. Petersen. A new approach to hierarchical data analysis: targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Statistical methods in medical research*, 28(6):1761–1780, 2019.

L. B. Balzer, M. van der Laan, J. Ayieko, M. Kamya, G. Chamie, J. Schwab, D. V. Havlir, and M. L. Petersen. Two-stage tmle to reduce bias and improve efficiency in cluster randomized trials. *arXiv preprint arXiv:2106.15737*, 2021.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

A. Benitez, M. L. Petersen, M. J. van der Laan, N. Santos, E. Butrick, D. Walker, R. Ghosh, P. Otieno, P. Waiswa, and L. B. Balzer. Comparative methods for the analysis of cluster randomized trials. *arXiv preprint arXiv:2110.09633*, 2021.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press, 1993.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

W. G. Cochran. *Sampling Techniques.* John Wiley & Sons, 1977.

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

M. Feng, J. Roschelle, N. Heffernan, J. Fairman, and R. Murphy. Implementation of an intelligent tutoring system for online homework support in an efficacy trial. In *International Conference on Intelligent Tutoring Systems*, pages 561–566. Springer, 2014.

R. A. Fisher. *Statistical methods for research workers.* Oliver and Boyd, 1925.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

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

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2002.

M. A. Hernán and J. M. Robins. *Causal inference.* CRC Boca Raton, FL, 2010.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

K. Imai, Z. Jiang, J. Greiner, R. Halen, and S. Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845*, 2020.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arxiv*, 2022.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008–2030, 2020.

E. L. Lehmann. *Fisher, Neyman, and the creation of classical statistics.* Springer Science & Business Media, 2011.

S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

A. Mate, L. Madaan, A. Taneja, N. Madhiwalla, S. Verma, G. Singh, A. Hegde, P. Varakantham, and M. Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. *arXiv preprint arXiv:2109.08075*, 2021.

G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American statistical association*, 110(512):1399–1411, 2015.

J. Neyman. On the application of probability theory to agricultural experiments, essay on principles: Section 9 (1923), translated. *Statistical Science*, 5(4):465–472, 1990.

E. L. Ogburn, O. Sofrygin, I. Diaz, and M. J. Van Der Laan. Causal inference for social network data. *arXiv preprint arXiv:1705.08527*, 2017.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429): 106–121, 1995.

J. M. Robins, L. Li, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421, 2008.

P. R. Rosenbaum. *Observational studies.* Springer, 2002.

D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3): 661–682, 2010.

A. A. Tsiatis. *Semiparametric Theory and Missing Data.* New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* New York: Springer, 2009.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge: Cambridge University Press, 2000.

P. Wang, T. M. Berzin, J. R. G. Brown, S. Bharadwaj, A. Becq, X. Xiao, P. Liu, L. Li, Y. Song, D. Zhang, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, 68(10):1813–1819, 2019.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.

W. Zheng and M. J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 273:1–58, 2010.

6 Appendix

6.1 Notation

Y^a	the counterfactual/potential outcome of Y under $A = a$
$X \perp\!\!\!\perp Y \mid Z$	X is independent of Y given Z
iid	independent and identically distributed
$\mathbb{P}_n$	the empirical measure, so $\frac{1}{n}\sum_i f(Z_i) = \mathbb{P}_n\{f(Z)\} = \mathbb{P}_n(f)$
$X_n \rightsquigarrow Y$	X_n converges in distribution to Y
$X_n = o_{\mathbb{P}}(r_n)$	X_n/r_n converges to zero in probability, i.e., $X_n/r_n \xrightarrow{p} 0$
$X_n = O_{\mathbb{P}}(r_n)$	X_n/r_n is bounded in probability
$\Vert\widehat{f} - f\Vert_1$	L_1 distance: $\int \vert\widehat{f}(x) - f(x)\vert \ d\mathbb{P}(x)$
$\Vert\widehat{f} - f\Vert_2$	L_2 distance: $\sqrt{\int \{\widehat{f}(x) - f(x)\}^2 \ d\mathbb{P}(x)}$
$\Vert\widehat{f} - f\Vert_\infty$	L_∞ distance: $\sup_{x \in \mathcal{X}} \vert\widehat{f}(x) - f(x)\vert$
$\mathbb{P}(\widehat{f})$	$\int \widehat{f}(z) \ d\mathbb{P}(z)$, also written $\mathbb{P}\{\widehat{f}(Z)\}$

6.2 Estimated functions

Since many analyses involve differences between an estimated function $\widehat{f}$ and some true function f, i.e., errors in estimating f with $\widehat{f}$, we will need a notion of consistency for random functions $\widehat{f}$. Recall that a scalar (or Euclidean) estimator $\widehat{\psi}$ is consistent if $\widehat{\psi} - \psi = o_{\mathbb{P}}(1)$, i.e., if $\widehat{\psi}$ converges to ψ in probability.

For functions we can define an appropriate (scalar) distance measure, and then consistency will be defined as in the scalar case. Some popular distance measures for functions are the L_1, L_2, or L_∞ distances defined in the notation table above. Note that all of these distances are themselves random variables, since they depend on the estimated $\widehat{f}$. Now we are ready to define consistency of an estimated function.

Definition 2. *An estimated function $\widehat{f}(x)$ is consistent for a fixed target $f(x)$ in distance measure $d(\cdot, \cdot)$ if*

$$d(\widehat{f}, f) = o_{\mathbb{P}}(1).$$

Similarly, $\widehat{f}$ converges at rate $r_n \to \infty$ to f in distance d if

$$d(\widehat{f}, f) = o_{\mathbb{P}}(1/r_n).$$

In addition to having a notion of consistency or convergence for estimated functions $\widehat{f}$, it will also be useful for us to have some special notation for the expected value over a random function's argument, conditioning on the randomness in the function.

Definition 3. *For an estimated function $\widehat{f}(x)$ built from a sample $Z^n = (Z_1, ..., Z_n)$ we use the notation*

$$\mathbb{P}(\widehat{f}) = \mathbb{P}\{\widehat{f}(Z)\} \equiv \int \widehat{f}(z) \ d\mathbb{P}(z) = \mathbb{E}\left\{\widehat{f}(Z) \ \Big| \ Z^n\right\}$$

to denote expectations over a new independent observation Z, conditioning on the sample Z^n.

Remark 16. The heuristic interpretation of $\mathbb{P}(\widehat{f})$ is as follows: you construct the function $\widehat{f}(z)$ from a sample Z^n, and then take its average over new repeated independent draws of the argument Z. It is important to note that, for a *fixed* function $f(z)$ we have

$$\mathbb{E}\{f(Z)\} = \mathbb{P}(f),$$

whereas for a random estimated function $\widehat{f}(z)$ depending on a sample Z^n, we have

$$\mathbb{E}\{\widehat{f}(Z)\} = \mathbb{E}\left[\mathbb{E}\left\{\widehat{f}(Z) \;\middle|\; Z^n\right\}\right] \neq \mathbb{P}(\widehat{f}).$$

In particular, the quantity $\mathbb{P}(\widehat{f})$ on the right-hand-side is random (through its dependence on $\widehat{f}$ and Z^n), whereas the quantities on the left-hand-side are fixed.

6.3 R Code

Testing Fisher's Sharp Null Consider the following completely randomized experiment:

```
> set.seed(100)
> ## simulate fake data
> n <- 10; a <- rep(c(1,0),5); y <- a*rnorm(n,1)+(1-a)*rnorm(n,-1)
> cbind(a,y)
      a          y
 [1,] 1  0.4978076
 [2,] 0 -0.9037255
 [3,] 1  0.9210829
 [4,] 0 -0.2601595
 [5,] 1  1.1169713
 [6,] 0 -1.0293167
 [7,] 1  0.4182093
 [8,] 0 -0.4891437
 [9,] 1  0.1747406
[10,] 0  1.3102968
>
> ## compute test statistic
> (tobs <- abs(mean(y[a==1])-mean(y[a==0])))
[1] 0.9001721
```

Then we can test Fisher's Sharp Null in R with:

```
> ## permute treatments to simulate null
> t <- NULL; for (j in 1:10000){
+     asim <- sample(a)
+     t <- c(t, abs(mean(y[asim==1])-mean(y[asim==0]))) }
>
> ## compute p-value
> mean(t>=tobs)
[1] 0.0837
```

Recovering Population Effects via Regression In R we can fit a logistic regression model and evaluate it at the observed $(X_1, A_1), ..., (X_n, A_n)$ values with the `predict` command, as in:

```
lrmod <- glm(y ~ x+a, family=binomial)
muhat <- predict(lrmod, type="response")
```

Note that the `predict` function outputs predicted values under the *observed* treatment; in contrast here one needs predicted values under $A = 1$ and $A = 0$ separately, so you need the `newdata` argument. Here is some example code for a simulated dataset:

```
> cbind(x,a,y)
                x a y
 [1,] -0.44577826 0 0
 [2,] -1.20585657 0 0
 [3,]  0.04112631 1 1
 [4,]  0.63938841 0 0
 [5,] -0.78655436 0 1
 [6,] -0.38548930 0 1
 [7,] -0.47586788 1 0
 [8,]  0.71975069 1 1
 [9,] -0.01850562 1 1
...
[100,]  2.01893816 0 1
>
> mumod <- glm(y~x+a, family=binomial)
>
> mu1hat <- predict(mumod, newdata=data.frame(x,a=1) ,type="response")
> mu0hat <- predict(mumod, newdata=data.frame(x,a=0), type="response")
>
> cbind(x,a,y, mu1hat, mu0hat, mu1hat-mu0hat)
              x a y    mu1hat    mu0hat
1   -0.44577826 0 0 0.8178912 0.5709608 0.2469305
2   -1.20585657 0 0 0.7793358 0.5113598 0.2679760
3    0.04112631 1 1 0.8397107 0.6081932 0.2315174
4    0.63938841 0 0 0.8635670 0.6522366 0.2113304
5   -0.78655436 0 1 0.8012901 0.5443888 0.2569013
6   -0.38548930 0 1 0.8207133 0.5756239 0.2450893
7   -0.47586788 1 0 0.8164699 0.5686286 0.2478412
8    0.71975069 1 1 0.8665332 0.6579775 0.2085556
...
100  2.01893816 0 1 0.9073274 0.7436597 0.1636677
>
> mean(mu1hat-mu0hat)
[1] 0.2303284
```

So for the above simulated dataset, the estimated average treatment effect using logistic regression is $\widehat{\psi} = 0.23$.

Doubly Robust Estimator Here is example code showing how to correct the plug-in estimator we constructed using the doubly robust estimator:

```
> cbind(x,a,y)[1:5,]
                x a y
 [1,] -0.44577826 0 0
 [2,] -1.20585657 0 0
 [3,]  0.04112631 1 1
 [4,]  0.63938841 0 0
 [5,] -0.78655436 0 1
>
```

```
> mumod <- glm(y~x+a, family=binomial)
> mu1hat <- predict(mumod, newdata=data.frame(x,a=1) ,type="response")
> mu0hat <- predict(mumod, newdata=data.frame(x,a=0), type="response")
>
> pi <- 0.5; muahat <- a*mu1hat + (1-a)*mu0hat
>
> mean( (mu1hat-mu0hat) + (a/pi - (1-a)/(1-pi)) * (y-muahat) )
[1] 0.2303284
```