

# Section 11: Causal Inference

Ricky Truong (rickytruong@college.harvard.edu),  
Emily Xing (exing@college.harvard.edu)

## 1 Introduction

### 1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

### 1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 - 11:30 AM in Cabot D-Hall (Emily).

## 2 Big Picture

“Correlation doesn’t mean causation”—but what does? Causal inference is the subfield of statistics concerned with drawing *causal* conclusions, which are more powerful than conclusions about association.

We begin with the *potential outcomes framework* to rigorously define the notation of causation. There are many *assumptions*, so often the challenge is justifying their credibility. For the sake of simplicity, we often assume *SUTVA* and *randomization*.

There are two main approaches to causal inference: *super-population* and *finite-sample*. Essentially, the choice depends on the group to which we want to generalize our inference. With the usual assumptions, we can arrive at the (arguably intuitive) *difference in means estimator* for both approaches, which can be extended with *confidence intervals* and *hypothesis tests*. Still, the formalization is important for extending to cases where the usual assumptions may not hold, such as with *observational data*.

## 3 Overview

**Idea.** *Statistical inference* is a whole lot of assumptions. *Causal inference* adds even more. Fundamentally, we cannot observe *causal* quantities, so we must use *assumptions* to *identify* them as *statistical* quantities (i.e., the estimands we’ve been working with throughout the course). From there, we can proceed with our usual inference! Informally, treat  $Y$  as our data (like before) and  $Y(0), Y(1)$  as quantities that only exist in theory.

### 3.1 Fundamentals

**Definition 1** (Outcome). For  $n$  units,  $Y_i$  is the *outcome* for unit  $i$ .

- E.g., suppose we are interested in the causal effect of tutoring on SAT score. Then  $Y_1$  is the SAT score for unit 1.

**Definition 2** (Treatment). For  $n$  units,  $W_i$  is the *treatment* for unit  $i$ . The *treatment vector* is  $\vec{W} = (W_1, \dots, W_n)$ .

- E.g., if  $W_i \in \{0, 1\}$  is binary, then  $W_1 = 1$  indicates unit 1 received tutoring while  $W_1 = 0$  indicates unit 1 did not.

**Definition 3** (Potential outcome). For  $n$  units,  $Y_i(w_1, \dots, w_n)$  is the *potential outcome* for unit  $i$ . It is a function of all possible *treatments* (i.e., the potential outcome if  $W_1 = w_1, \dots, W_n = w_n$ ).

- E.g., for  $n = 3$ ,  $Y_1(1, 0, 1)$  is the potential SAT score for unit 1 if, hypothetically, units 1 and 3 received tutoring while unit 2 did not.

**Definition 4** (Identification). The process of using *assumptions* to convert *causal quantities* to *statistical quantities*.

- E.g.,  $\mathbb{E}[Y(1)]$  is a causal quantity while  $\mathbb{E}[Y | W = 1]$  is a statistical quantity.

**Definition 5** (Individual treatment effect). For unit  $i$ ,  $\tau_i = Y_i(1) - Y_i(0)$ .

**Definition 6** (Fundamental problem of causal inference). The *individual treatment effect* can never be observed since we only see up to one *potential outcome* per unit.

- Thus, we rely on assumptions and focus on group-wide quantities.
- E.g., for us to observe unit 1's individual treatment effect, we'd have to observe their SAT score under tutoring and what their SAT score would've been that same day if everything else were the same except their tutoring status (i.e., we'd need an alternate universe)!

**Concept Checker 1.** Which of the following are causal quantities?

1.  $Y_i(1) - Y_i(0)$
2.  $\mathbb{E}[Y_i(0)]$
3.  $\mathbb{E}[Y_i | W_i = 0]$
4.  $\frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0))$

Solution

### 3.2 Assumptions

**Definition 7** (Consistency). Under *consistency*,  $Y_i = Y_i(W_i)$  (i.e., unit  $i$ 's outcome is their potential outcome under observed treatment  $W_i$ ).

- E.g., units 1 and 2 both receive tutoring, but unit 1 gets multiple hours of hands-on practice while unit 2 only gets a worksheet, so consistency is violated.

**Definition 8** (Non-interference). Under *non-interference*,  $Y_i(W_1, \dots, W_n) = Y_i(W_i)$  (i.e., unit  $i$ 's potential outcome is only a function of their own treatment).

- E.g., units 1 and 2 are best friends who study together and share exam advice. If unit 2 gets tutoring while unit 1 does not,  $Y_1(0, 1)$  is still different from  $Y_1(0, 0)$ , so non-interference is violated.

**Definition 9** (Stable unit treatment value assumption). Under *SUTVA*, both *consistency* and *non-interference* hold.

**Definition 10** (Unconfoundedness). Under *unconfoundedness*,  $Y_i(0), Y_i(1) \perp\!\!\!\perp W_i$  (i.e., unit  $i$ 's treatment is independent of their potential outcomes).

- E.g., unit 1 doesn't like tutoring while unit 2 desperately wants to improve their score. Magically, we know  $Y_1(0) = 1000$  and  $Y_1(1) = 1050$  while  $Y_2(0) = 1000$  and  $Y_2(1) = 1500$ . If students choose whether to get tutoring, there is self-selection since  $Y_2(1)$  being high affects  $W_2$ , so unconfoundedness is violated.

**Definition 11** (Binary treatment). Under *binary treatment*,  $W_i \in \{0, 1\}$ .

- Unless otherwise noted, we will often assume binary treatment and SUTVA since this dramatically simplifies notation.<sup>1</sup>

**Definition 12** (Switching equation). Assume *binary treatment* and *SUTVA*.  $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ .

**Concept Checker 2.** Assume binary treatment. Suppose there are  $n$  units. How many potential outcomes are there for each unit  $i$ ? What if we assume SUTVA?

Solution

### 3.3 Treatment Assignment

**Definition 13** (Set of potential outcomes). Assume *binary treatment* and *SUTVA*.  $\{\vec{Y}(0), \vec{Y}(1)\}$ , where  $\vec{Y}(0) = (Y_1(0), \dots, Y_n(0))$  and  $\vec{Y}(1) = (Y_1(1), \dots, Y_n(1))$ .

**Definition 14** (Assignment mechanism).  $P(\vec{W} = \vec{w} \mid \{\vec{Y}(0), \vec{Y}(1)\})$  (i.e., the joint probability mass function of the assignments given the potential outcomes).

- This is a function of the set of potential outcomes.

**Definition 15** (Randomization). An *assignment mechanism* where  $P(\vec{W} = \vec{w} \mid \{\vec{Y}(0), \vec{Y}(1)\}) = P(\vec{W} = \vec{w})$ .

- An experiment where treatment is randomized is called a *randomized control trial* (RCT).

<sup>1</sup>There has been much research done—and more to be done—on cases where these assumptions don't hold (e.g., with spatial, temporal, and network data).

- Notice the assumption of *unconfoundedness* holds under *randomization*.
- In *Bernoulli randomization*,  $P(W_i = 1) = p \in (0, 1) \forall i \implies Y_i(0), Y_i(1) \perp\!\!\!\perp W_i$ .
- In *complete randomization*,  $P(W_i = 1) = \frac{n_1}{n} \forall i \in \{1, \dots, n\}$ , where  $n_1$  is chosen in advance  $\implies Y_i(0), Y_i(1) \perp\!\!\!\perp W_i$ .

## 4 Super-Population Approach

**Idea.** To estimate causal quantities, we can use either the *super-population approach* (i.e., a model-based approach) or the *finite-sample approach* (i.e., a design-based approach). These are two complementary perspectives for our estimand of interest: *average treatment effect* (ATE).

In the super-population approach, we model the units as an i.i.d. sample from a larger (possibly infinite) population. Here, randomness comes from both *treatment assignment* and *sampling*. The causal estimand is the *population average treatment effect* (PATE).

Importantly, though potential outcomes are viewed as fixed for a given unit, they have a distribution across units. E.g., fix any person, and their height is not viewed as random. However, height is (approximately) normally distributed in the population, so when sampling, it is random which height will be observed.

**Definition 16** (Setup for super-population approach). Assume *binary treatment*, *SUTVA*, and *randomization*. Let  $\{Y_i(0), Y_i(1)\} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}^*$ , where  $\mathbb{P}^*$  is the super-population (i.e., we assume the potential outcomes are an i.i.d. sample from a statistical model).

**Definition 17** (Population average treatment effect).  $\tau_{\text{PATE}} = \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$ .

- I.e., the expected treatment effect across all units in the super-population.

**Definition 18** (DIM estimator for PATE). Assume *binary treatment*, *SUTVA*, and *randomization*. Let  $n_0 = \sum_{i=1}^n (1 - W_i)$  and  $n_1 = \sum_{i=1}^n W_i$  such that  $n = n_0 + n_1$ .  $\hat{\tau}_{\text{PATE, DIM}} = \bar{Y}_1 - \bar{Y}_0$ , where  $\bar{Y}_w = \frac{1}{n_w} \sum_{i: W_i=w} Y_i = \frac{1}{\sum_{i=1}^n \mathbb{1}\{W_i=w\}} \sum_{i=1}^n \mathbb{1}\{W_i=w\} Y_i$ .

- As we will show,  $\hat{\tau}_{\text{PATE, DIM}} = \hat{\tau}_{\text{PATE, MOM}}$ , and under *binary outcome*,  $\hat{\tau}_{\text{PATE, DIM}} = \hat{\tau}_{\text{PATE, MLE}}$ .

**Definition 19** (MOM estimator for PATE). Assume *binary treatment*, *SUTVA*, and *randomization*.  $\hat{\tau}_{\text{PATE, MOM}} = \bar{Y}_1 - \bar{Y}_0$ .

*Proof.*

$$\begin{aligned}
 \tau_{\text{PATE}} &= \mathbb{E}[Y_i(1) - Y_i(0)] && \text{by definition} \\
 &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] && \text{by linearity} \\
 &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] && \text{by unconfoundedness (RCT)} \\
 &= \mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0] && \text{by consistency (SUTVA)} \\
 &= \frac{\mathbb{E}[Y_i W_i]}{\mathbb{E}[W_i]} - \frac{\mathbb{E}[Y_i(1 - W_i)]}{\mathbb{E}[1 - W_i]} && \text{by conditional expectation}
 \end{aligned}$$

$$\begin{aligned}
 \hat{\tau}_{\text{PATE,MOM}} &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i W_i}{\frac{1}{n} \sum_{i=1}^n W_i} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - W_i)}{\frac{1}{n} \sum_{i=1}^n (1 - W_i)} && \text{by definition of MOM} \\
 &= \frac{\sum_{i=1}^n Y_i W_i}{\sum_{i=1}^n W_i} - \frac{\sum_{i=1}^n Y_i (1 - W_i)}{\sum_{i=1}^n (1 - W_i)} && \text{by simplifying} \\
 &= \bar{Y}_1 - \bar{Y}_0 && \text{by rewriting}
 \end{aligned}$$

Notice the first term is the sample mean of those treated—of which there are  $n_1 = \sum_{i=1}^n W_i$ —and the second term is the sample mean of those in control—of which there are  $n_0 = \sum_{i=1}^n (1 - W_i)$ .  $\square$

**Definition 20** (MLE for PATE). Assume *binary treatment*, *SUTVA*, and *randomization*. Additionally, assume *binary outcome*.  $\hat{\tau}_{\text{PATE,MLE}} = \bar{Y}_1 - \bar{Y}_0$ .

*Proof.* Let  $\theta_1 = \mathbb{E}[Y_i(1)]$  and  $\theta_0 = \mathbb{E}[Y_i(0)]$  such that  $\tau_{\text{PATE}} = \theta_1 - \theta_0$ . By the reasoning above,  $\theta_1 = \mathbb{E}[Y_i | W_i = 1]$  and  $\theta_0 = \mathbb{E}[Y_i | W_i = 0]$ . By fundamental bridge,  $\theta_1 = P(Y_i = 1 | W_i = 1)$  and  $\theta_0 = P(Y_i = 1 | W_i = 0)$ . We can relate the conditional likelihood of  $\theta_0, \theta_1$  to the MLE of Bernoulli data.

$$\begin{aligned}
 \mathcal{L}(\theta_0, \theta_1 | \vec{W} = \vec{w}; \vec{y}) &= P(\vec{Y} = \vec{y} | \vec{W} = \vec{w}; \theta_0, \theta_1) && \text{by likelihood} \\
 &= \prod_{i=1}^n P(Y_i = y_i | W_i = w_i; \theta_0, \theta_1) && \text{by i.i.d.}
 \end{aligned}$$

$$\begin{aligned}
 \ell(\theta_0, \theta_1 | \vec{W} = \vec{w}; \vec{y}) &= \log \left( \prod_{i=1}^n P(Y_i = y_i | W_i = w_i; \theta_0, \theta_1) \right) && \text{by log-likelihood} \\
 &= \sum_{i=1}^n \log(P(Y_i = y_i | W_i = w_i; \theta_0, \theta_1)) && \text{by properties of log} \\
 &= \sum_{i:W_i=0} \log(P(Y_i = y_i | W_i = w_i; \theta_0, \theta_1)) \\
 &\quad + \sum_{i:W_i=1} \log(P(Y_i = y_i | W_i = w_i; \theta_0, \theta_1)) && \text{by splitting} \\
 &= \sum_{i:W_i=0} \log(\theta_0^{y_i} (1 - \theta_0)^{1-y_i}) + \sum_{i:W_i=1} \log(\theta_1^{y_i} (1 - \theta_1)^{1-y_i}) && \text{by Bernoulli PMF} \\
 &= \sum_{i:W_i=0} y_i \log(\theta_0) + (1 - y_i) \log(1 - \theta_0) \\
 &\quad + \sum_{i:W_i=1} y_i \log(\theta_1) + (1 - y_i) \log(1 - \theta_1) && \text{by properties of log}
 \end{aligned}$$

Recall for  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$ ,  $\hat{\theta}_{\text{MLE}} = \bar{Y}$ . The first term is the log-likelihood of  $\theta_0$  for  $n_0$  observations. Thus,  $\hat{\theta}_{0,\text{MLE}} = \bar{Y}_0$ . By the same reasoning,  $\hat{\theta}_{1,\text{MLE}} = \bar{Y}_1$ . By invariance,  $\hat{\tau}_{\text{PATE,MLE}} = \bar{Y}_1 - \bar{Y}_0$ .  $\square$

**Definition 21** (Properties of DIM estimator for PATE). Assume *binary treatment*, *SUTVA*, and *randomization*. The DIM estimator is unbiased and achieves CRLB.

- $\mathbb{E}[\hat{\tau}_{\text{PATE,DIM}}] = \tau_{\text{PATE}}$ .
- $\text{Var}[\hat{\tau}_{\text{PATE,DIM}}] = \frac{\text{Var}[Y_i|W_i=1]}{n_1} + \frac{\text{Var}[Y_i|W_i=0]}{n_0}$ .

## 5 Finite-Sample Approach

**Idea.** In the finite-sample approach, we fix the  $n$  units in our sample, and we wish to generalize our findings to them. The causal estimand is the *sample average treatment effect* (SATE). Here, randomness comes only from *treatment assignment* as there is no sampling. By unit-level determinism, each unit  $i$  has fixed potential outcomes  $y_i(0)$  and  $y_i(1)$ , but it's random which one we'll observe (with the randomness coming from treatment).<sup>2</sup> Thus, inference is conditional on the set of potential outcomes, so we view them as fixed (even though half of them will never be observed).

**Definition 22** (Setup for finite-sample approach). Assume *binary treatment*, *SUTVA*, and *randomization*. We condition on  $\vec{Y}(0) = \vec{y}(0)$ ,  $\vec{Y}(1) = \vec{y}(1)$ . Importantly, by the *switching equation*,  $Y_i$  is deterministic from  $W_i$  given the potential outcomes.

**Definition 23** (Sample average treatment effect).  $\tau_{\text{SATE}} = \bar{\tau} = \frac{1}{n} \sum_{i=1}^n \tau_i = \frac{1}{n} \sum_{i=1}^n (y_i(1) - y_i(0))$ .

**Concept Checker 3.** SATE looks like a sample mean, but it's still an estimand. Why?

Solution

**Definition 24** (MOM estimator for SATE). Assume *binary treatment*, *SUTVA*, and *randomization*.  $\hat{\tau}_{\text{SATE,MOM}} = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i W_i}{\mathbb{E}[W_i]} - \frac{Y_i(1-W_i)}{\mathbb{E}[1-W_i]} \right)$ .

- It can be shown  $\hat{\tau}_{\text{SATE,MOM}}$  is the difference in means estimator under *complete randomization*.

*Proof.* To keep notation simple, let  $\vec{Y}(w) = \{Y(0), Y(1)\}$  (i.e., the set of potential outcomes). First, notice the following:

$$\begin{aligned}
 Y_i &= W_i Y_i(1) + (1 - W_i) Y_i(0) && \text{by switching equation} \\
 W_i Y_i &= W_i (W_i Y_i(1) + (1 - W_i) Y_i(0)) && \text{by multiplying} \\
 &= W_i^2 Y_i(1) + (W_i - W_i^2) Y_i(0) && \text{by distributing} \\
 &= W_i Y_i(1) + (W_i - W_i) Y_i(0) && \text{since } W_i \in \{0, 1\} \\
 &= W_i Y_i(1) && \text{by simplifying}
 \end{aligned}$$

<sup>2</sup>E.g., I have a weight  $y$  that is not random. It's just a number at the end of the day. We regard what my weight would be with and without medication— $y(1)$  and  $y(0)$ , respectively—as also not random. What is random is  $Y$ , which of the two will be observed. The randomness comes from treatment assignment:  $Y = y(0)$  if  $W = 0$  and  $Y = y(1)$  if  $W = 1$ .

Thus,  $W_i Y_i = W_i Y_i(1)$ . Similarly,  $(1 - W_i) Y_i = (1 - W_i) Y_i(0)$ .

Next, notice the following:

$$\begin{aligned} \mathbb{E}[W_i Y_i \mid \vec{Y}(w) = \vec{y}(w)] &= \mathbb{E}[W_i Y_i(1) \mid \vec{Y}(w) = \vec{y}(w)] && \text{by substituting} \\ &= y_i(1) \mathbb{E}[W_i \mid \vec{Y}(w) = \vec{y}(w)] && \text{by linearity} \\ &= y_i(1) \mathbb{E}[W_i] && \text{by unconfoundedness (RCT)} \end{aligned}$$

This implies  $y_i(1) = \frac{\mathbb{E}[W_i Y_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[W_i]}$  by algebra. Similarly,  $y_i(0) = \frac{\mathbb{E}[(1 - W_i) Y_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[1 - W_i]}$ .

Now, we want an estimator for SATE.

$$\begin{aligned} \tau_{\text{SATE}} &= \frac{1}{n} \sum_{i=1}^n (y_i(1) - y_i(0)) && \text{by definition} \\ &= \frac{1}{n} \sum_{i=1}^n \left( \frac{\mathbb{E}[W_i Y_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[W_i]} - \frac{\mathbb{E}[(1 - W_i) Y_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[1 - W_i]} \right) && \text{by substituting} \\ \hat{\tau}_{\text{SATE, MOM}} &= \frac{1}{n} \sum_{i=1}^n \left( \frac{W_i Y_i}{\mathbb{E}[W_i]} - \frac{(1 - W_i) Y_i}{\mathbb{E}[1 - W_i]} \right) && \text{by definition of MOM} \end{aligned}$$

Importantly,  $\mathbb{E}[W_i Y_i \mid \vec{Y}(w) = \vec{y}(w)]$  is conditional on the set of potential outcomes. Thus, the expectation is over the randomness in  $W_i$ . We observe one realization of  $W_i$ , so the sample analog of  $\mathbb{E}[W_i Y_i \mid \vec{Y}(w) = \vec{y}(w)]$  is  $W_i Y_i$ .  $\square$

**Definition 25** (Properties of MOM estimator for SATE). Assume *binary treatment*, *SUTVA*, and *randomization*. The MOM estimator is conditionally unbiased given  $\vec{Y}(w) = \vec{y}(w)$ .

- $\mathbb{E}[\hat{\tau}_{\text{SATE, MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \tau_{\text{SATE}}$ .
- $\text{Var}[\hat{\tau}_{\text{SATE, MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \frac{1}{n^2} \sum_{i=1}^n \left( \frac{(y_i(1))^2}{\mathbb{E}[W_i]} + \frac{(y_i(0))^2}{\mathbb{E}[1 - W_i]} - (y_i(1) - y_i(0))^2 \right)$ .<sup>3</sup>

**Concept Checker 4.** Why is  $\mathbb{E}[W_i]$  valid to include in the estimator?

Solution

**Concept Checker 5.** In practice, the variance must be estimated—usually with  $\widehat{\text{Var}}[\hat{\tau}_{\text{SATE, MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}$ , where  $\hat{\sigma}_w^2 = \frac{1}{n_w - 1} \sum_{i: W_i = w} (Y_i - \bar{Y}_w)^2$ . Why would this be a conservative estimator in the finite-population approach? Intuitively, which approach should have less uncertainty?

<sup>3</sup>It can be shown  $\text{Var}[\hat{\tau}_{\text{SATE, MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_1^2 + S_0^2 - 2S_{01}}{n}$  under complete randomization, where  $S_w^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i(w) - \bar{y}(w))^2$  and  $S_{01} = \frac{1}{n-1} \sum_{i=1}^n (y_i(1) - \bar{y}(1))(y_i(0) - \bar{y}(0))$ .

Solution

**Concept Checker 6.** Show  $\mathbb{E}[\hat{\tau}_{\text{SATE},\text{MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \tau_{\text{SATE}}$ .

Solution

**Concept Checker 7.** Suppose we're working in the super-population framework. Additionally, suppose we use complete randomization such that  $\hat{\tau}_{\text{SATE},\text{MOM}}$  is the difference in means estimator. Show the connection between  $\tau_{\text{PATE}}$  and  $\frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0))$  (i.e., the random/uncrystallized version of SATE) using the difference in means estimator.

## Solution

## 6 Confidence Intervals and Hypothesis Tests

**Idea.** Like before, we can do “better” than just the point estimator. We introduce two frameworks that give us two different flavors of hypothesis test: one uses the randomization itself as the source of inference (Fisher), and one uses the asymptotic Normal approximation (Neyman). Importantly, they test different things!

**Definition 26** (Fisher’s null hypothesis). The treatment effect is 0 for *each* unit. I.e.,  $H_0 : \tau_i = 0 \forall i \in \{1, \dots, n\}$ .

- Alternatively, there is a treatment effect for at least one unit. I.e.,  $H_A : \sum_{i=1}^n |\tau_i| > 0$ .
- Under  $H_0$ ,  $Y_i(0) = Y_i(1) = Y_i$  because  $\tau_i = Y_i(1) - Y_i(0) = 0$ .
- **Key insight:** Fisher’s null is called the *sharp* null because it fills in the missing potential outcomes. Under  $H_0$ , we know both  $Y_i(0)$  and  $Y_i(1)$  for every unit (they’re both just  $Y_i$ ). This means we can compute  $\hat{\tau}_{\text{SATE}, \text{MOM}}$  for *any* hypothetical assignment vector  $\vec{W}$ , not just the one we observed.
- **Randomization test:** Generate  $B$  i.i.d. draws  $\vec{W}^{(1)}, \dots, \vec{W}^{(B)}$  from the assignment mechanism. For each draw  $b$ , compute  $\hat{\tau}^{(b)}$  using the observed  $Y_i$ ’s and the hypothetical assignment  $\vec{W}^{(b)}$ . The randomization  $p$ -value is

$$p = \frac{1}{B} \sum_{b=1}^B \mathbb{1}\{|\hat{\tau}^{(b)}| \geq |\hat{\tau}_{\text{SATE}, \text{MOM}}|\}.$$

Intuitively, if the null is true, the observed  $\hat{\tau}_{\text{SATE}, \text{MOM}}$  should look like a typical draw from this randomization distribution. A small  $p$ -value means our observed test statistic is unusually large which is hard to explain by chance alone.

**Definition 27** (Neyman’s null hypothesis). The *average* treatment effect is 0. I.e.,  $H_0 : \tau_{\text{SATE}} = 0$ .

- Alternatively,  $H_A : \tau_{\text{SATE}} \neq 0$ .
- Fisher’s null implies Neyman’s null, but the converse is not true since individual effects can cancel. E.g.,  $\tau_1 = 1, \tau_2 = -1, \dots$  gives  $\tau_{\text{SATE}} = 0$ , but Fisher’s null fails.

• **Key insight:** Neyman’s null is weaker since it only cares about the average, not every individual. We test it via the asymptotic Normal pivot, using the conservative variance estimator from before:

$$T = \frac{\hat{\tau}_{\text{SATE},\text{MOM}}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}} \sim \mathcal{N}(0, 1).$$

Reject  $H_0$  when  $|T| > z_{\alpha/2}$ . A nominal  $100(1-\alpha)\%$  CI for  $\tau_{\text{SATE}}$  is  $\hat{\tau}_{\text{SATE},\text{MOM}} \pm z_{\alpha/2} \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}$ .

**Definition 28** (Super-population DIM estimator). Recall the conditional expectation and variance of  $\hat{\tau}_{\text{PATE},\text{DIM}}$ . We use the “plug-in” principle to substitute in estimators. We know MLE is asymptotically Normal.

- $\frac{\hat{\tau}_{\text{PATE},\text{DIM}} - \tau_{\text{PATE}}}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{\sum_{i=1}^n w_i} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\sum_{i=1}^n (1-w_i)}}} \mid \vec{W} = \vec{w} \sim \mathcal{N}(0, 1).$
- To test  $H_0 : \tau_{\text{PATE}} = 0$ , we use  $T = \frac{\hat{\tau}_{\text{PATE},\text{DIM}}}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{\sum_{i=1}^n w_i} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\sum_{i=1}^n (1-w_i)}}} \mid \vec{W} = \vec{w} \sim \mathcal{N}(0, 1).$

**Concept Checker 8.** 8 Suppose our observed data are  $\vec{Y} = (3, 6, 7, 2)$  and  $\vec{W} = (0, 1, 1, 0)$ . What is  $\hat{\tau}_{\text{DIM}}$ ? If we assume Fisher’s null, what would our data be for  $\vec{W}^{(1)} = (1, 0, 1, 0)$ ?

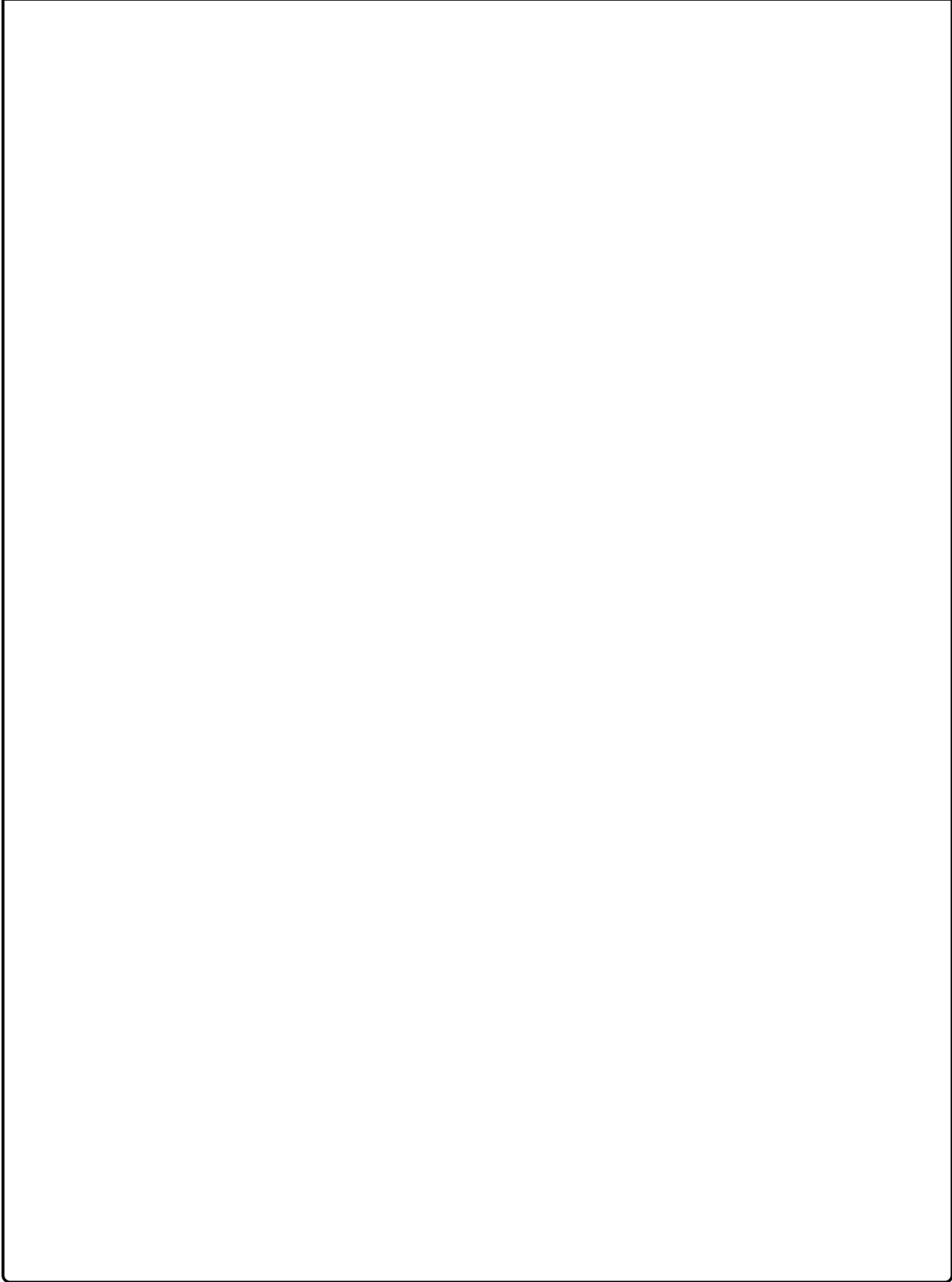
Solution

## 7 Recap

**Idea.** We conclude with a review of the two approaches and implications for future work.



Solution



**Problem 2.** Suppose we're working in the finite-population framework. Additionally, suppose we use complete randomization, where  $P(W_i = 1) = \frac{n_1}{n} \forall i$  with  $n$  and  $n_1$  known.

(a) Find  $\text{Cov}[W_i, W_j]$  for  $i \neq j$ .

(b) Show  $\hat{\tau}_{\text{SATE}, \text{MOM}}$  is the difference in means estimator.

Solution

**Problem 3.** Suppose in a study of  $n = 6$  units under complete randomization with  $n_1 = n_0 = 3$ , the observed  $\hat{\tau}_{\text{SATE}, \text{MOM}} = 4$ . Under Fisher's null, we re-randomize  $B = 5$  times and obtain  $\hat{\tau}^{(b)} \in \{-4, -2, 0, 2, 4\}$ .

- (a) Compute the randomization  $p$ -value. Do we reject at  $\alpha = 0.05$ ?
- (b) Now suppose the individual effects are  $\tau_1 = 8, \tau_2 = -8, \tau_3 = 8, \tau_4 = -8, \tau_5 = 8, \tau_6 = -8$ . Does Fisher's null hold? Does Neyman's null hold? What does this tell you about the relationship between the two nulls?
- (c) Why can't we use the randomization test to directly test Neyman's null?

Solution