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 $\vec{W} = (W_1, \ldots, W_n)$.

• 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, $Y_i(w_1, \ldots, w_n)$ is the *potential outcome* for unit i. It is a function of all possible *treatments* (i.e., the potential outcome if $W_1 = w_1, \ldots, W_n = w_n$).

• 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 $\mathbb{E}[Y \mid W = 1]$ 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. $\mathbb{E}[Y_i \mid W_i = 0]$
4. $\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$

Solution

Notice all quantities except $\mathbb{E}[Y \mid W = 0]$ involve potential outcomes, which are causal and theoretical (in contrast, $\mathbb{E}[Y \mid W = 0]$ is a statistical quantity that can be estimated from the data). Thus, all quantities except $\mathbb{E}[Y \mid W = 0]$ are causal quantities.

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*, $Y_i(W_1, \ldots, W_n) = Y_i(W_i)$ (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.[1]

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

For unit i, there are 2^n potential outcomes since each w_j in $Y_i(w_1, \ldots, w_n)$ can be either 0 or 1. In contrast, if we assume SUTVA, for unit i, there are only 2 potential outcomes since w_i in $Y_i(w_i)$ can be either 0 or 1.

3.3 Treatment Assignment

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

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).

[1] 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*, $P(W_i = 1) = \frac{n_1}{n}\ \forall\ i \in \{1, \ldots, n\}$, 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 $\{Y_i(0), Y_i(1)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^*$, 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 $n_0 = \sum_{i=1}^n (1 - W_i)$. □

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, $\theta_1 = \mathbb{E}[Y_i \mid W_i = 1]$ and $\theta_0 = \mathbb{E}[Y_i \mid W_i = 0]$. By fundamental bridge, $\theta_1 = P(Y_i = 1 \mid W_i = 1)$ and $\theta_0 = P(Y_i = 1 \mid W_i = 0)$. We can relate the conditional likelihood of θ_0, θ_1 to the MLE of Bernoulli data.

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

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

Recall for $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$, $\hat{\theta}_{\text{MLE}} = \bar{Y}$. The first term is the log-likelihood of θ_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, $\hat{\tau}_{\text{PATE,MLE}} = \bar{Y}_1 - \bar{Y}_0$. □

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).[2] 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

SATE is a causal quantity as it involves potential outcomes. Additionally, we will never have enough data to observe the SATE due to the fundamental problem of causal inference. However, we can still infer about it!

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 $\vec{Y}(w) = \{\vec{Y}(0), \vec{Y}(1)\}$ (i.e., the set of potential outcomes). First, notice the following:

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

[2] 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, $W_iY_i = W_iY_i(1)$. Similarly, $(1-W_i)Y_i = (1-W_i)Y_i(0)$.
Next, notice the following:

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

This implies $y_i(1) = \frac{\mathbb{E}[W_iY_i|\vec{Y}(w)=\vec{y}(w)]}{\mathbb{E}[W_i]}$ by algebra. Similarly, $y_i(0) = \frac{\mathbb{E}[(1-W_i)Y_i|\vec{Y}(w)=\vec{y}(w)]}{\mathbb{E}[1-W_i]}$.
Now, we want an estimator for SATE.

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

$$\hat{\tau}_{\text{SATE,MOM}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{\mathbb{E}[W_i]} - \frac{(1-W_i)Y_i}{\mathbb{E}[1-W_i]}\right) \quad \text{by definition of MOM}$$

Importantly, $\mathbb{E}[W_iY_i \mid \vec{Y}(w) = \vec{y}(w)]$ 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 $\mathbb{E}[W_iY_i \mid \vec{Y}(w) = \vec{y}(w)]$ is W_iY_i. □

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

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

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

Solution

We know the distribution of W_i and can thus compute $\mathbb{E}[W_i]$. E.g., in a Bernoulli randomization, $\mathbb{E}[W_i] = \mathbb{E}[1-W_i] = \frac{1}{2}$.

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 $\hat{\sigma}_w^2 = \frac{1}{n_w - 1}\sum_{i:W_i=w}\left(Y_i - \bar{Y}_w\right)^2$). Why would this be a conservative estimator in the finite-population approach? Intuitively, which approach should have less uncertainty?

[3] 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 $S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i(w) - \bar{y}(w)\right)^2$ and $S_{01} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i(1) - \bar{y}(1)\right)\left(y_i(0) - \bar{y}(0)\right)$.

Solution

The $(y_i(1) - y_i(0))^2$ is difficult to estimate due to the fundamental problem of causal inference, so it is usually dropped (e.g., in the estimator above). Thus, the true variance should be less since we're omitting the subtraction of non-negative terms, so the estimator is conservative. Intuitively, the finite-sample approach should have less uncertainty since we're generalizing to only our sample (even though, interestingly, both approaches usually use this estimator)!

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

Solution

Let $G_i = \frac{W_i Y_i}{\mathbb{E}[W_i]} - \frac{(1-W_i)Y_i}{\mathbb{E}[1-W_i]}$.

$$\begin{aligned}
\mathbb{E}[G_i \mid \vec{Y}(w) = \vec{y}(w)] &= \mathbb{E}\left[\frac{W_i Y_i}{\mathbb{E}[W_i]} - \frac{(1-W_i)Y_i}{\mathbb{E}[1-W_i]} \mid \vec{Y}(w) = \vec{y}(w)\right] \\
&= \frac{\mathbb{E}[W_i Y_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[W_i]} - \frac{\mathbb{E}[(1-W_i)Y_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[1-W_i]} \\
&= \frac{\mathbb{E}[W_i Y_i(1) \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[W_i]} - \frac{\mathbb{E}[(1-W_i)Y_i(0) \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[1-W_i]} \\
&= y_i(1)\frac{\mathbb{E}[W_i \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[W_i]} - y_i(0)\frac{\mathbb{E}[(1-W_i) \mid \vec{Y}(w) = \vec{y}(w)]}{\mathbb{E}[1-W_i]} \\
&= y_i(1)\frac{\mathbb{E}[W_i]}{\mathbb{E}[W_i]} - y_i(0)\frac{\mathbb{E}[(1-W_i)]}{\mathbb{E}[1-W_i]} \\
&= y_i(1) - y_i(0)
\end{aligned}$$

Now, $\hat{\tau}_{\text{SATE,MOM}} = \frac{1}{n}\sum_{i=1}^n G_i$, so $\mathbb{E}[\hat{\tau}_{\text{SATE,MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[G_i \mid \vec{Y}(w) = \vec{y}(w)] = \frac{1}{n}\sum_{i=1}^n (y_i(1) - y_i(0)) = \tau_{\text{SATE}}$.

Concept Checker 7. Suppose we're working in the super-population framework. Additionally, suppose we use complete randomization such that $\hat{\tau}_{\text{SATE,MOM}}$ is the difference in means estimator. Show the connection between τ_{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

$$
\begin{aligned}
\tau_{\text{PATE}} &= \mathbb{E}[\hat{\tau}_{\text{DIM}}] && \text{by unbiasedness} \\
&= \mathbb{E}_{\text{sampling}}[\mathbb{E}_{\text{treatment}}[\hat{\tau}_{\text{DIM}} \mid \vec{Y}(w)]] && \text{by Adam's Law} \\
&= \mathbb{E}_{\text{sampling}}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\right] && \text{by conditional unbiasedness} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\text{sampling}}[Y_i(1) - Y_i(0)] && \text{by linearity} \\
&= \frac{1}{n}\sum_{i=1}^{n}\tau_{\text{PATE}} && \text{by definition of PATE} \\
&= \tau_{\text{PATE}} && \text{by simplifying}
\end{aligned}
$$

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., $H_0 : \tau_i = 0 \ \forall \ i \in \{1, \ldots, n\}$.

• 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 $\hat{\tau}_{\text{SATE,MOM}}$ for *any* hypothetical assignment vector $\vec{W}$, not just the one we observed.

• **Randomization test:** Generate B i.i.d. draws $\vec{W}^{(1)}, \ldots, \vec{W}^{(B)}$ 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

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

Intuitively, if the null is true, the observed $\hat{\tau}_{\text{SATE,MOM}}$ 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., $\tau_1 = 1, \tau_2 = -1, \ldots$ 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:

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

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

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

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

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

$\hat{\tau}_{\text{DIM}} = \frac{1}{2}(6+7) - \frac{1}{2}(3+2) = 4$. Under Fisher's null, $Y_i(1) = Y_i(0) = Y_i$, so for $\vec{W}^{(1)}$, $\hat{\tau}_{\text{DIM}}^{(1)} = \frac{1}{2}(3+7) - \frac{1}{2}(6+2) = 1$. We repeat this for $\vec{W}^{(1)}, \ldots, \vec{W}^{(B)}$ (and their corresponding estimates $\hat{\tau}_{\text{DIM}}^{(1)}, \ldots, \hat{\tau}_{\text{DIM}}^{(B)}$) to calculate a p-value.

7 Recap

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

	Super-Population	**Finite-Sample**
Population	Units are drawn i.i.d. from a larger super-population	The n units in the study are fixed and finite
Inference	ATE for super-population	ATE for sample
Source(s) of randomness	Treatment assignment, sampling	Treatment assignment
Potential outcomes	$\{Y_i(0), Y_i(1)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^*$	A fixed set: $\vec{Y}(0) = \vec{y}(0), \vec{Y}(1) = \vec{y}(1)$
Random variables	$W_i, Y_i(0), Y_i(1)$	W_i
Data	$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$	$Y_i = W_i y_i(1) + (1 - W_i) y_i(0)$
Estimand	$\tau_{\text{PATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]$	$\tau_{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i(1) - y_i(0))$
Estimator	DIM/MOM/MLE: $\hat{\tau}_{\text{PATE,DIM}} = \bar{Y}_1 - \bar{Y}_0$	MOM/DIM: $\hat{\tau}_{\text{SATE,MOM}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i W_i}{\mathbb{E}[W_i]} - \frac{Y_i(1-W_i)}{\mathbb{E}[1-W_i]}\right)$

Concept Checker 9. Y_i is deterministic from W_i in the finite-sample approach, but this is not true in the super-population approach. Explain this mathematically and intuitively.

Solution

Mathematically, $Y_i(0)$ and $Y_i(1)$ are still random in the super-population approach, so from the switching equation, $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. It is not enough to only know W_i! Intuitively, suppose we sample $n = 1$ unit—either Emily or Ricky. If you only know their treatments (e.g., $W_{\text{Emily}} = 0, W_{\text{Ricky}} = 1$), it is unknown who will be included in the sample, so Y_1 is still random from the sampling! Notice this is not the case in the finite-sample approach.

8 Practice Problems

Problem 1. Suppose we are interested in the causal effect of hands-on tutoring on SAT score for all high school students. There are 3 treatment levels: $W = 2$ for hands-on tutoring, $W = 1$ for hands-off tutoring, and $W = 0$ for control. We randomly sample n students and randomize treatment assignment. Let Y_i be the SAT score for student i. Assume SUTVA and $P(W = 2) > 0$.

(a) Would the super-population approach or finite-sample approach be more appropriate for this problem?

(b) Suppose we are interested in $\mathbb{E}[Y(2)]$. What type of quantity is $\mathbb{E}[Y(2)]$? State what it is clearly in words.

(c) Identify $\mathbb{E}[Y(2)]$ as a statistical quantity.

(d) Construct a consistent estimator: $\hat{\mathbb{E}}[Y(2)]$.

Solution

(a) First, the super-population approach is more appropriate since we are interested in extrapolating to the larger population of all students.

(b) Related, $\mathbb{E}[Y(2)]$ is the expected potential outcome under treatment level 2 (i.e., hands-on tutoring) across all high school students. This is a causal quantity and thus fundamentally unobservable.
(c) However, we can identify it.

$$\begin{aligned}\mathbb{E}[Y(2)] &= \mathbb{E}[Y(2) \mid W = 2] && \text{by unconfoundedness (RCT)}\\ &= \mathbb{E}[Y \mid W = 2] && \text{by consistency (SUTVA)}\end{aligned}$$

Next, notice the following.

$$\begin{aligned}\mathbb{E}[Y\mathbb{1}\{W = 2\}] &= \mathbb{E}[Y\mathbb{1}\{W = 2\} \mid W = 2]P(W = 2)\\ &\quad + \mathbb{E}[Y\mathbb{1}\{W = 2\} \mid W = 1]P(W = 1)\\ &\quad + \mathbb{E}[Y\mathbb{1}\{W = 2\} \mid W = 0]P(W = 0) && \text{by LOTE}\\ &= \mathbb{E}[Y\mathbb{1}\{W = 2\} \mid W = 2]P(W = 2) && \text{by simplifying}\\ &= \mathbb{E}[Y \mid W = 2]P(W = 2) && \text{by simplifying}\end{aligned}$$

This implies $\mathbb{E}[Y \mid W = 2] = \frac{1}{P(W=2)}\mathbb{E}[Y\mathbb{1}\{W = 2\}]$ by algebra. To remember this important result, recall the analogous definition of conditional probability: $P(A \mid B) = \frac{1}{P(B)}P(A, B)$. With this, we continue.

$$\begin{aligned}\mathbb{E}[Y(2)] &= \mathbb{E}[Y \mid W = 2] && \text{from before}\\ &= \frac{1}{P(W = 2)}\mathbb{E}[Y\mathbb{1}\{W = 2\}] && \text{by substituting}\end{aligned}$$

(d) To construct an estimator, we can replace the theoretical quantities with their sample analogs.

$$\begin{aligned}\hat{\mathbb{E}}[Y(2)] &= \frac{1}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i = 2\}}\left(\frac{1}{n}\sum_{i=1}^n Y_i\mathbb{1}\{W_i = 2\}\right) && \text{by sample quantities}\\ &= \frac{1}{\sum_{i=1}^n \mathbb{1}\{W_i = 2\}}\sum_{i=1}^n Y_i\mathbb{1}\{W_i = 2\} && \text{by simplifying}\\ &= \bar{Y}_2 && \text{by rewriting}\end{aligned}$$

Finally, let's show this is consistent. First, $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i = 2\} \xrightarrow{p} \mathbb{E}[\mathbb{1}\{W_i = 2\}] = P(W = 2)$ by LLN and fundamental bridge, so $\frac{1}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i=2\}} \xrightarrow{p} \frac{1}{P(W=2)}$ by CMT with $g(x) = x^{-1}$. Next, $\frac{1}{n}\sum_{i=1}^n Y_i\mathbb{1}\{W_i = 2\} \xrightarrow{p} \mathbb{E}[Y\mathbb{1}\{W = 2\}]$ by LLN. Thus, $\frac{1}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i=2\}}\left(\frac{1}{n}\sum_{i=1}^n Y_i\mathbb{1}\{W_i = 2\}\right) \xrightarrow{p} \frac{1}{P(W=2)}\mathbb{E}[Y\mathbb{1}\{W = 2\}]$ by Theorem 3.5.7., so $\hat{\mathbb{E}}[Y(2)] \xrightarrow{p} \mathbb{E}[Y(2)]$ by substituting. We conclude $\hat{\mathbb{E}}[Y(2)]$ is consistent for $\mathbb{E}[Y(2)]$. For some intuition, notice this estimator is the sample mean of those under treatment level 2—of which there are $\sum_{i=1}^n \mathbb{1}\{W_i = 2\}$.

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

Solution

(a) First, let's find $\text{Cov}[W_i, W_j]$.

$$\begin{aligned}
\text{Cov}[W_i, W_j] &= \mathbb{E}[W_i W_j] - \mathbb{E}[W_i]\mathbb{E}[W_j] && \text{by definition of covariance} \\
&= P(W_i = 1, W_j = 1) - \left(\frac{n_1}{n}\right)^2 && \text{by fundamental bridge} \\
&= P(W_i = 1)P(W_j = 1 \mid W_i = 1) - \left(\frac{n_1}{n}\right)^2 && \text{by multiplication rule} \\
&= \left(\frac{n_1}{n}\right)\left(\frac{n_1 - 1}{n - 1}\right) - \left(\frac{n_1}{n}\right)^2 && \text{by given probabilities} \\
&= -\frac{n_1 n_0}{n^2(n-1)} && \text{by simplifying}
\end{aligned}$$

(b) Next, let's show $\hat{\tau}_{\text{SATE,MOM}}$ is the difference in means estimator.

$$\begin{aligned}
\hat{\tau}_{\text{SATE,MOM}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\mathbb{E}[W_i]} - \frac{(1-W_i)Y_i}{\mathbb{E}[1-W_i]}\right) && \text{from before} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{n_1/n} - \frac{(1-W_i)Y_i}{n_0/n}\right) && \text{by substituting} \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\frac{W_i Y_i}{n_1/n} - \sum_{i=1}^{n}\frac{(1-W_i)Y_i}{n_0/n}\right) && \text{by simplifying} \\
&= \frac{1}{n}\left(n\sum_{i=1}^{n}\frac{W_i Y_i}{n_1} - \sum_{i=1}^{n} n\frac{(1-W_i)Y_i}{n_0}\right) && \text{by factoring} \\
&= \sum_{i=1}^{n}\frac{W_i Y_i}{n_1} - \sum_{i=1}^{n}\frac{(1-W_i)Y_i}{n_0} && \text{by simplifying} \\
&= \frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-W_i)Y_i && \text{by factoring} \\
&= \frac{\sum_{i=1}^{n} Y_i W_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} Y_i(1-W_i)}{\sum_{i=1}^{n}(1-W_i)} && \text{by substituting} \\
&= \bar{Y}_1 - \bar{Y}_0 && \text{by rewriting}
\end{aligned}$$

Problem 3. Suppose in a study of $n = 6$ units under complete randomization with $n_1 = n_0 = 3$, the observed $\hat{\tau}_{\text{SATE,MOM}} = 4$. 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

(a) Count how many $|\hat{\tau}^{(b)}|$ are at least as extreme as the observed $|\hat{\tau}_{\text{SATE,MOM}}| = 4$. That includes $\hat{\tau}^{(b)} \in \{-4, 4\}$, so 2 out of 5 draws qualify. Thus $p = 2/5 = 0.4$. We do *not* reject at $\alpha = 0.05$.

(b) Fisher's null does *not* hold: $\tau_i \neq 0$ for every unit. However,

$$\tau_{\text{SATE}} = \frac{1}{6}(8 - 8 + 8 - 8 + 8 - 8) = 0,$$

so Neyman's null *does* hold. Individual effects can be large and of opposite sign, perfectly canceling in the average. This confirms Fisher's null $\implies$ Neyman's null, but not the converse.

(c) The randomization test requires Fisher's sharp null precisely because it fills in the missing potential outcomes. Under H_0^F, $Y_i(0) = Y_i(1) = Y_i$, so the full table of potential outcomes is observable and we can recompute the test statistic for every hypothetical re-randomization. Under Neyman's null alone, we only know $\frac{1}{n}\sum_i \tau_i = 0$. The individual $Y_i(0)$ and $Y_i(1)$ remain unknown, so the randomization distribution cannot be constructed. The asymptotic Normal pivot addresses this entirely by relying on the CLT rather than the assignment mechanism.