Political Methodology II
Section: Randomized Experiments

Apoorva Lal

January 22, 2020

Stanford University

Roadmap

Understanding Estimands

Identification under random assignment

Estimation under random assignment

Testing in small samples: Randomization inference

Roadmap

Understanding Estimands

Identification under random assignment

Estimation under random assignment

Testing in small samples: Randomization inference

Review of Potential Outcomes Notation

- Treatment status: $D_i = 1$ if observation i gets the treatment and $D_i = 0$ if i doesn't get the treatment.
- Potential outcome with treatment: Y_{1i}, or sometimes $Y_i(1)$
- Potential outcome without treatment: Y_{0i}, or sometimes $Y_i(0)$.
- Individual-level treatment effect: $\tau_i = Y_{1i} - Y_{0i}$.

Review of Potential Outcomes Notation

- Treatment status: $D_i = 1$ if observation i gets the treatment and $D_i = 0$ if i doesn't get the treatment.
- Potential outcome with treatment: Y_{1i}, or sometimes $Y_i(1)$
- Potential outcome without treatment: Y_{0i}, or sometimes $Y_i(0)$.
- Individual-level treatment effect: $\tau_i = Y_{1i} - Y_{0i}$.
- Observed outcome: $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$,

Review of Potential Outcomes Notation

- Treatment status: $D_i = 1$ if observation i gets the treatment and $D_i = 0$ if i doesn't get the treatment.
- Potential outcome with treatment: Y_{1i}, or sometimes $Y_i(1)$
- Potential outcome without treatment: Y_{0i}, or sometimes $Y_i(0)$.
- Individual-level treatment effect: $\tau_i = Y_{1i} - Y_{0i}$.
- Observed outcome: $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$, or equivalently:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

Review of Potential Outcomes Notation

- Treatment status: $D_i = 1$ if observation i gets the treatment and $D_i = 0$ if i doesn't get the treatment.
- Potential outcome with treatment: Y_{1i}, or sometimes $Y_i(1)$
- Potential outcome without treatment: Y_{0i}, or sometimes $Y_i(0)$.
- Individual-level treatment effect: $\tau_i = Y_{1i} - Y_{0i}$.
- Observed outcome: $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$, or equivalently:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

The *fundamental problem of causal inference* is that we can never observe Y_{1i} and Y_{0i} simultaneously.

Review of Potential Outcomes Notation

This model of potential outcomes already makes a big assumption. What is it?

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

Review of Potential Outcomes Notation

This model of potential outcomes already makes a big assumption. What is it?

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

SUTVA. Why?

Review of Potential Outcomes Notation

This model of potential outcomes already makes a big assumption. What is it?

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

SUTVA. Why?

SUTVA is fundamentally an assumption of how much missing data we have on potential outcomes. If there are N units in the population, how many possible ways are there to assign a binary treatment to each unit?

Review of Potential Outcomes Notation

This model of potential outcomes already makes a big assumption. What is it?

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

SUTVA. Why?

SUTVA is fundamentally an assumption of how much missing data we have on potential outcomes. If there are N units in the population, how many possible ways are there to assign a binary treatment to each unit? 2^N. If potential outcomes for unit i are a function of everyone's treatment assignment, how many potential outcomes do we have?

Review of Potential Outcomes Notation

This model of potential outcomes already makes a big assumption. What is it?

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$
$$Y_i = Y_{0i} + \tau_i D_i$$

SUTVA. Why?

SUTVA is fundamentally an assumption of how much missing data we have on potential outcomes. If there are N units in the population, how many possible ways are there to assign a binary treatment to each unit? 2^N. If potential outcomes for unit i are a function of everyone's treatment assignment, how many potential outcomes do we have? 2^N. If SUTVA holds, then we assume unit i's potential outcomes are only a function of its own treatment status, and there are only two possible ways to assign treatment to unit i.

The Assignment Mechanism

The Assignment mechanism is the procedure that determines which units are selected to receive the treatment.

The Assignment Mechanism

The Assignment mechanism is the procedure that determines which units are selected to receive the treatment.

Formally, it is a row-exchangeable function $\mathbf{Pr}\left(\mathbf{D}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right)$ taking on values values in $[0, 1]$ [Imbens and Rubin 2015, Chap 2].

The Assignment Mechanism

The Assignment mechanism is the procedure that determines which units are selected to receive the treatment.

Formally, it is a row-exchangeable function $\mathbf{Pr}\left(\mathbf{D}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right)$ taking on values values in $[0, 1]$ [Imbens and Rubin 2015, Chap 2]. Examples include

- random assignment
- selection on observables
- selection on unobservables

The Assignment Mechanism

The Assignment mechanism is the procedure that determines which units are selected to receive the treatment.

Formally, it is a row-exchangeable function $\mathbf{Pr}\left(\mathbf{D}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right)$ taking on values values in $[0, 1]$ [Imbens and Rubin 2015, Chap 2]. Examples include

- random assignment
- selection on observables
- selection on unobservables

Key goal of modern causal inference training is to rewire your brains to think about the assignment mechanism (instead of focussing on variation in the outcome Y).

Understanding the assignment mechanism is the key pre-requisite for moving from correlation to causation.

Understanding estimands: ATE

What is the average treatment effect?

Understanding estimands: ATE

What is the average treatment effect? Recall that each individual has an individual-level treatment effect $\tau_i = Y_{1i} - Y_{0i}$. The average treatment effect is simply the expected value of the individual τ_i's:

$$ATE = \mathbb{E}[\tau_i]$$
$$ATE = \mathbb{E}[Y_{1i} - Y_{0i}]$$
$$ATE = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$$

Understanding estimands: ATT and ATC

What are the average treatment effect on the treated (ATT) and the average treatment effect on the control (ATC)?

Understanding estimands: ATT and ATC

What are the average treatment effect on the treated (ATT) and the average treatment effect on the control (ATC)?

- ATT is the average treatment effect among those units that *actually received* the treatment:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$$

- Similarly, ATC is the average treatment effect among those units that *did not receive* the treatment:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 0]$$

Understanding estimands: ATT and ATC

What are the average treatment effect on the treated (ATT) and the average treatment effect on the control (ATC)?

- ATT is the average treatment effect among those units that *actually received* the treatment:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$$

- Similarly, ATC is the average treatment effect among those units that *did not receive* the treatment:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 0]$$

What's notable compared to the ATE is that the ATT and ATC depend on the treatment assignment in the sample.

Relationship between ATE, ATT, and ATC

We can decompose ATE into a weighted sum of the ATT and ATC. Use the law of iterated expectations to rewrite $\mathbb{E}[Y_{1i}]$ and $\mathbb{E}[Y_{0i}]$:

$$\begin{aligned}\mathbb{E}[Y_{1i}] &= \mathbb{E}\Big[\mathbb{E}[Y_{1i} \mid D_i]\Big] \\ &= \mathbb{E}[Y_{1i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]P(D_i = 0)\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Y_{0i}] &= \mathbb{E}\Big[\mathbb{E}[Y_{0i} \mid D_i]\Big] \\ &= \mathbb{E}[Y_{0i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{0i} \mid D_i = 0]P(D_i = 0)\end{aligned}$$

Relationship between ATE, ATT, and ATC

Using that notation, we can rewrite the ATE:

Relationship between ATE, ATT, and ATC

Using that notation, we can rewrite the ATE:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\
&= \Big(\mathbb{E}[Y_{1i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]P(D_i = 0)\Big) - \\
&\quad \Big(\mathbb{E}[Y_{0i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{0i} \mid D_i = 0]P(D_i = 0)\Big) \\
&= P(D_i = 1)\Big(\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]\Big) + \\
&\quad P(D_i = 0)\Big(\mathbb{E}[Y_{1i} \mid D_i = 0] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big)
\end{aligned}
$$

Relationship between ATE, ATT, and ATC

Using that notation, we can rewrite the ATE:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\
&= \Big(\mathbb{E}[Y_{1i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]P(D_i = 0)\Big) - \\
&\quad \Big(\mathbb{E}[Y_{0i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{0i} \mid D_i = 0]P(D_i = 0)\Big) \\
&= P(D_i = 1)\Big(\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]\Big) + \\
&\quad P(D_i = 0)\Big(\mathbb{E}[Y_{1i} \mid D_i = 0] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big) \\
&= P(D_i = 1)\big(\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]\big) + P(D_i = 0)\big(\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 0]\big)
\end{aligned}
$$

Relationship between ATE, ATT, and ATC

Using that notation, we can rewrite the ATE:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\
&= \Big(\mathbb{E}[Y_{1i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]P(D_i = 0)\Big) - \\
&\quad \Big(\mathbb{E}[Y_{0i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{0i} \mid D_i = 0]P(D_i = 0)\Big) \\
&= P(D_i = 1)\Big(\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]\Big) + \\
&\quad P(D_i = 0)\Big(\mathbb{E}[Y_{1i} \mid D_i = 0] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big) \\
&= P(D_i = 1)\big(\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]\big) + P(D_i = 0)\big(\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 0]\big) \\
ATE &= P(D_i = 1)ATT + P(D_i = 0)ATC
\end{aligned}
$$

Relationship between ATE, ATT, and ATC

Using that notation, we can rewrite the ATE:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\
&= \Big(\mathbb{E}[Y_{1i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]P(D_i = 0)\Big) - \\
&\quad \Big(\mathbb{E}[Y_{0i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{0i} \mid D_i = 0]P(D_i = 0)\Big) \\
&= P(D_i = 1)\Big(\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]\Big) + \\
&\quad P(D_i = 0)\Big(\mathbb{E}[Y_{1i} \mid D_i = 0] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big) \\
&= P(D_i = 1)\big(\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]\big) + P(D_i = 0)\big(\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 0]\big) \\
ATE &= P(D_i = 1)ATT + P(D_i = 0)ATC
\end{aligned}
$$

So the ATE is a weighted average of the ATT and ATC, with weights given by the proportion of treated units.

Naive Difference in Means

Why can't we simply use the difference in observed outcomes for treated and control units as an estimate of the ATE? Recall the bias decomposition:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] =$$

Naive Difference in Means

Why can't we simply use the difference in observed outcomes for treated and control units as an estimate of the ATE? Recall the bias decomposition:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] =$$
$$\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0] =$$

Naive Difference in Means

Why can't we simply use the difference in observed outcomes for treated and control units as an estimate of the ATE? Recall the bias decomposition:

$$
\begin{aligned}
&\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \\
&\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0] = \\
&\mathbb{E}[Y_{1i} \mid D_i = 1] + (\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]) - \mathbb{E}[Y_{0i} \mid D_i = 0] =
\end{aligned}
$$

Naive Difference in Means

Why can't we simply use the difference in observed outcomes for treated and control units as an estimate of the ATE? Recall the bias decomposition:

$$
\begin{aligned}
&\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \\
&\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0] = \\
&\mathbb{E}[Y_{1i} \mid D_i = 1] + (\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]) - \mathbb{E}[Y_{0i} \mid D_i = 0] = \\
&\underbrace{\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]}_{\text{ATT}} + \underbrace{\{\mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]\}}_{\text{baseline bias}}
\end{aligned}
$$

Naive Difference in Means

We can go even further to get the naive difference in means in terms of the ATE plus two forms of bias:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \text{ATT} + \text{baseline bias}$$

Naive Difference in Means

We can go even further to get the naive difference in means in terms of the ATE plus two forms of bias:

$$\begin{aligned}&\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \text{ATT} + \text{baseline bias} \\ &= P(D_i = 1)\text{ATT} + P(D_i = 0)\text{ATT} + \text{baseline bias}\end{aligned}$$

Naive Difference in Means

We can go even further to get the naive difference in means in terms of the ATE plus two forms of bias:

$$
\begin{aligned}
&\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \text{ATT} + \text{baseline bias} \\
&= P(D_i = 1)\text{ATT} + P(D_i = 0)\text{ATT} + \text{baseline bias} \\
&= P(D_i = 1)\text{ATT} + P(D_i = 0)\text{ATT} + (P(D_i = 0)\text{ATC} - P(D_i = 0)\text{ATC}) + \text{baseline bias}
\end{aligned}
$$

Naive Difference in Means

We can go even further to get the naive difference in means in terms of the ATE plus two forms of bias:

$$
\begin{aligned}
&\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \text{ATT} + \text{baseline bias} \\
&= P(D_i = 1)\text{ATT} + P(D_i = 0)\text{ATT} + \text{baseline bias} \\
&= P(D_i = 1)\text{ATT} + P(D_i = 0)\text{ATT} + (P(D_i = 0)\text{ATC} - P(D_i = 0)\text{ATC}) + \text{baseline bias} \\
&= \underbrace{P(D_i = 1)\text{ATT} + P(D_i = 0)\text{ATC}}_{\text{ATE}} + \underbrace{P(D_i = 0)(\text{ATT} - \text{ATC})}_{\text{differential treatment effect bias}} + \text{baseline bias}
\end{aligned}
$$

So the naive difference in means will be a biased estimator of the ATE if we have either differential treatment effect bias or baseline bias.

Example

Suppose we are interested in the effect of college on an individual's earnings. We estimate this effect with a naive difference in average wages between individuals who attended college and those who did not.

What source of baseline bias might we be worried about?

Example

Suppose we are interested in the effect of college on an individual's earnings. We estimate this effect with a naive difference in average wages between individuals who attended college and those who did not.

What source of baseline bias might we be worried about? Individuals who attended college might have had higher baseline levels of potential earnings to begin with in the absence of going to college.

What source of differential treatment effect bias might we be worried about?

Example

Suppose we are interested in the effect of college on an individual's earnings. We estimate this effect with a naive difference in average wages between individuals who attended college and those who did not.

What source of baseline bias might we be worried about? Individuals who attended college might have had higher baseline levels of potential earnings to begin with in the absence of going to college.

What source of differential treatment effect bias might we be worried about? Attending college may have a greater effect on potential earnings for individuals that selected into attending college than it would have had for those who did not attend college. This is a difference between the ATC and ATT.

Roadmap

Understanding Estimands

Identification under random assignment

Estimation under random assignment

Testing in small samples: Randomization inference

Identification: Learning the lingo

What do we mean when we talk about *identification*?

Identification: Learning the lingo

What do we mean when we talk about *identification*?

'Over two dozen different terms for identification appear in the econometrics literature'; Lewbel, "The Identification Zoo", JEL 2020

Identification: Learning the lingo

What do we mean when we talk about *identification*?

'Over two dozen different terms for identification appear in the econometrics literature'; Lewbel, "The Identification Zoo", JEL 2020

- Informally, it refers to what the data can (even in theory) tell us about a parameter.

Identification: Learning the lingo

What do we mean when we talk about *identification*?

'Over two dozen different terms for identification appear in the econometrics literature'; Lewbel, "The Identification Zoo", JEL 2020

- Informally, it refers to what the data can (even in theory) tell us about a parameter.
- If a parameter is (point) identified, it means that if we had infinite data, we could calculate the true parameter value.

Identification: Learning the lingo

What do we mean when we talk about *identification*?

'Over two dozen different terms for identification appear in the econometrics literature'; Lewbel, "The Identification Zoo", JEL 2020

- Informally, it refers to what the data can (even in theory) tell us about a parameter.
- If a parameter is (point) identified, it means that if we had infinite data, we could calculate the true parameter value.
- If we couldn't calculate the parameter even with infinite data, the parameter is *unidentified*. This is the problem with the difference in means estimator.

Identification: Learning the lingo

What do we mean when we talk about *identification*?

'Over two dozen different terms for identification appear in the econometrics literature'; Lewbel, "The Identification Zoo", JEL 2020

- Informally, it refers to what the data can (even in theory) tell us about a parameter.
- If a parameter is (point) identified, it means that if we had infinite data, we could calculate the true parameter value.
- If we couldn't calculate the parameter even with infinite data, the parameter is *unidentified*. This is the problem with the difference in means estimator.
- Often we're interested in a causal parameter like the ATE. Typically we say a parameter is "causally identified" if we have an unbiased (or consistent) estimator for the causal parameter of interest under a set of "identifying assumptions."

Identification under random assignment

Random assignment of the treatment breaks the dependence between the potential outcomes and treatment status that causes bias.

Identification under random assignment

Random assignment of the treatment breaks the dependence between the potential outcomes and treatment status that causes bias. Formally:

Identification Assumption

Treatment Independence Assume that treatment assignment is independent of the potential outcomes, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$.

Identification under random assignment

Random assignment of the treatment breaks the dependence between the potential outcomes and treatment status that causes bias. Formally:

Identification Assumption

Treatment Independence Assume that treatment assignment is independent of the potential outcomes, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$.

Note that this assumption is satisfied by design under randomization (though literal randomization is not strictly necessary for it to hold). When this assumption holds, the difference in means identifies the ATE.

Identification under random assignment

Random assignment of the treatment breaks the dependence between the potential outcomes and treatment status that causes bias. Formally:

Identification Assumption

Treatment Independence Assume that treatment assignment is independent of the potential outcomes, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$.

Note that this assumption is satisfied by design under randomization (though literal randomization is not strictly necessary for it to hold). When this assumption holds, the difference in means identifies the ATE.

Identification Result (Difference in Means)

Under treatment independence, the difference in mean outcomes between treated and control groups is an unbiased estimator for the average treatment effect.

ATE Identification

Let's look again at the difference in means:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]}_{\text{ATT}} + \underbrace{\{\mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]\}}_{\text{bias}}$$

Under random assignment, $\mathbb{E}[Y_{1i} \mid D_i] = \mathbb{E}[Y_{1i}]$ and $\mathbb{E}[Y_{0i} \mid D_i] = \mathbb{E}[Y_{0i}]$. The treatment status doesn't contain any information about the value of (Y_{0i}, Y_{1i}), so the Y_{0i}'s and Y_{1i}'s that we actually observe are a random sample of *all* the Y_{0i}'s and Y_{1i}'s.

ATE Identification

Let's look again at the difference in means:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]}_{\text{ATT}} + \underbrace{\{\mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]\}}_{\text{bias}}$$

Under random assignment, $\mathbb{E}[Y_{1i} \mid D_i] = \mathbb{E}[Y_{1i}]$ and $\mathbb{E}[Y_{0i} \mid D_i] = \mathbb{E}[Y_{0i}]$. The treatment status doesn't contain any information about the value of (Y_{0i}, Y_{1i}), so the Y_{0i}'s and Y_{1i}'s that we actually observe are a random sample of *all* the Y_{0i}'s and Y_{1i}'s.

Therefore:

$$\begin{aligned}
\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] &= \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1] + \mathbb{E}[Y_{0i}] - \mathbb{E}[Y_{0i}] \\
&= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\
&= ATE
\end{aligned}$$

Other Estimands under Random Assignment

Definition (Quantile Function)

$Q_\theta(Y)$ is the $\theta-$th quantile of the distribution of Y:

$$\mathbf{Pr}\left(Y \leq Q_\theta(Y)\right) = \theta$$

Other Estimands under Random Assignment

Definition (Quantile Function)

$Q_\theta(Y)$ is the $\theta-$th quantile of the distribution of Y:

$$\mathbf{Pr}\left(Y \leq Q_\theta(Y)\right) = \theta$$

Since $Y_0, Y_1 \perp\!\!\!\perp D$, we can write

$$Y_0 \sim Y_0|D = 0 \sim Y|D = 0$$
$$Y_1 \sim Y_1|D = 1 \sim Y|D = 1$$

where $\sim$ means *has the same distribution as*. So, treatment effect at any quantile $Q_\theta(Y_1) - Q_\theta(Y_0)$ is identified.

Other Estimands under Random Assignment

Definition (Quantile Function)

$Q_\theta(Y)$ is the $\theta-$th quantile of the distribution of Y:

$$\mathbf{Pr}\left(Y \leq Q_\theta(Y)\right) = \theta$$

Since $Y_0, Y_1 \perp\!\!\!\perp D$, we can write

$$Y_0 \sim Y_0|D = 0 \sim Y|D = 0$$
$$Y_1 \sim Y_1|D = 1 \sim Y|D = 1$$

where $\sim$ means *has the same distribution as*. So, treatment effect at any quantile $Q_\theta(Y_1) - Q_\theta(Y_0)$ is identified.

$Q_\theta(Y_1 - Y_0)$ is not identified, however. Unlike for expectations, the difference of quantiles is not the same as the quantiles of the difference.

Roadmap

Understanding Estimands

Identification under random assignment

Estimation under random assignment

Testing in small samples: Randomization inference

Estimating the ATE using difference in means

The identification result tells us that $\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = ATE$. We can estimate these quantities using the sample analogues:

$$\widehat{\mathbb{E}}[Y_i \mid D_i = 1] = \frac{1}{N_1} \sum_{i:D_i=1}^{N} Y_i$$

$$\widehat{\mathbb{E}}[Y_i \mid D_i = 0] = \frac{1}{N_0} \sum_{i:D_i=0}^{N} Y_i$$

These are unbiased and consistent estimators for the true population quantities, so the difference in sample means is an unbiased and consistent estimator for the ATE.

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$\begin{aligned} Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\ &= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \end{aligned}$$

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= Y_{0i} + D_i (Y_{1i} - Y_{0i}) + \left((Y_{0i} - \overline{Y}_0) D_i \cdot \left[(Y_{1i} - \overline{Y}_i) - (Y_{0i} - \overline{Y}_0) \right] \right)
\end{aligned}
$$

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= Y_{0i} + D_i (Y_{1i} - Y_{0i}) + \left((Y_{0i} - \overline{Y}_0) D_i \cdot \left[(Y_{1i} - \overline{Y}_i) - (Y_{0i} - \overline{Y}_0) \right] \right) \\
&= \overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0) D_i + \{ (Y_{0i} - \overline{Y}_0) + D_i [(Y_{1i} - Y_{0i}) - (\overline{Y}_1 - \overline{Y}_0)] \}
\end{aligned}
$$

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= Y_{0i} + D_i(Y_{1i} - Y_{0i}) + \left((Y_{0i} - \overline{Y}_0) D_i \cdot \left[(Y_{1i} - \overline{Y}_i) - (Y_{0i} - \overline{Y}_0)\right]\right) \\
&= \overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0) D_i + \{(Y_{0i} - \overline{Y}_0) + D_i[(Y_{1i} - Y_{0i}) - (\overline{Y}_1 - \overline{Y}_0)]\} \\
&= \underbrace{\overline{Y}_0}_{\alpha} + \underbrace{\tau_{ATE}}_{\beta} D_i + \underbrace{\{(Y_{0i} - \overline{Y}_0) + D_i(\tau_i - \tau_{ATE})\}}_{\epsilon_i}
\end{aligned}
$$

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= Y_{0i} + D_i(Y_{1i} - Y_{0i}) + \left((Y_{0i} - \overline{Y}_0) D_i \cdot \left[(Y_{1i} - \overline{Y}_i) - (Y_{0i} - \overline{Y}_0)\right]\right) \\
&= \overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0) D_i + \{(Y_{0i} - \overline{Y}_0) + D_i[(Y_{1i} - Y_{0i}) - (\overline{Y}_1 - \overline{Y}_0)]\} \\
&= \underbrace{\overline{Y}_0}_{\alpha} + \underbrace{\tau_{ATE}}_{\beta} D_i + \underbrace{\{(Y_{0i} - \overline{Y}_0) + D_i(\tau_i - \tau_{ATE})\}}_{\epsilon_i} \\
&= \alpha + \beta D_i + \epsilon_i
\end{aligned}
$$

When will the regression estimator $\hat{\beta}$ will be unbiased for the ATE?

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= Y_{0i} + D_i(Y_{1i} - Y_{0i}) + \left((Y_{0i} - \overline{Y}_0) D_i \cdot \left[(Y_{1i} - \overline{Y}_i) - (Y_{0i} - \overline{Y}_0)\right]\right) \\
&= \overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0) D_i + \{(Y_{0i} - \overline{Y}_0) + D_i[(Y_{1i} - Y_{0i}) - (\overline{Y}_1 - \overline{Y}_0)]\} \\
&= \underbrace{\overline{Y}_0}_{\alpha} + \underbrace{\tau_{ATE}}_{\beta} D_i + \underbrace{\{(Y_{0i} - \overline{Y}_0) + D_i(\tau_i - \tau_{ATE})\}}_{\epsilon_i} \\
&= \alpha + \beta D_i + \epsilon_i
\end{aligned}
$$

When will the regression estimator $\hat{\beta}$ will be unbiased for the ATE?

Estimating the ATE using regression

Recall that we can rewrite the potential outcomes model:

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= Y_{0i} + D_i(Y_{1i} - Y_{0i}) + \big((Y_{0i} - \overline{Y}_0) D_i \cdot \big[(Y_{1i} - \overline{Y}_i) - (Y_{0i} - \overline{Y}_0)\big]\big) \\
&= \overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0) D_i + \{(Y_{0i} - \overline{Y}_0) + D_i[(Y_{1i} - Y_{0i}) - (\overline{Y}_1 - \overline{Y}_0)]\} \\
&= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{\tau_{ATE}}_{\beta} D_i + \underbrace{\{(Y_{0i} - \bar{Y}_0) + D_i(\tau_i - \tau_{ATE})\}}_{\epsilon_i} \\
&= \alpha + \beta D_i + \epsilon_i
\end{aligned}
$$

When will the regression estimator $\hat{\beta}$ will be unbiased for the ATE? When $\mathbb{E}[\epsilon_i \mid D_i] = 0$. This means:

- $\mathbb{E}[Y_{0i} - \bar{Y}_0 \mid D_i = 0] = 0$
- $\mathbb{E}[Y_{0i} - \bar{Y}_0 \mid D_i = 1] + \mathbb{E}[\tau_i - \tau_{ATE} \mid D_i = 1] = 0$

Both are satisfied under random assignment, so regression gives us an unbiased estimate of the ATE.

Roadmap

Understanding Estimands

Identification under random assignment

Estimation under random assignment

Testing in small samples: Randomization inference

Lady Tasting Tea

Figure 1: R.A. Fisher (left), tea kettle (right)

Lady Tasting Tea

Dr. Bristol claims she can tell whether milk or tea has been poured first simply by tasting the cup of tea. Fisher devises a statistical test for whether her ability to do this is better than random guessing.

Lady Tasting Tea

Dr. Bristol claims she can tell whether milk or tea has been poured first simply by tasting the cup of tea. Fisher devises a statistical test for whether her ability to do this is better than random guessing.

- Randomly assign four out of eight cups of tea to have milk poured first.
- Ask Bristol to determine which ones had the milk poured first, and sum up the number of correct choices she made.

Lady Tasting Tea

Dr. Bristol claims she can tell whether milk or tea has been poured first simply by tasting the cup of tea. Fisher devises a statistical test for whether her ability to do this is better than random guessing.

- Randomly assign four out of eight cups of tea to have milk poured first.
- Ask Bristol to determine which ones had the milk poured first, and sum up the number of correct choices she made.

Sharp null hypothesis: Bristol's choice would be exactly the same under any ordering of the tea cups (e.g., she randomly guessed). Turns out she identified all 4 cups with milk poured first correctly. What's the probability of this happening by chance?

Lady Tasting Tea

There are $\binom{8}{4} = 70$ distinct possible orderings of the tea cups. The number of ways to correctly identify four out of four milk cups is $\binom{4}{4} = 1$ out of 70 total ways to choose 4 cups out of 8. So the probability she'd identify all four milk cups correctly by chance is only $1/70$. Another way to think about this is that Bristol's observed choice perfectly matches only one of the possible orderings of the cups (the one realized in the actual experiment).

Lady Tasting Tea

There are $\binom{8}{4} = 70$ distinct possible orderings of the tea cups. The number of ways to correctly identify four out of four milk cups is $\binom{4}{4} = 1$ out of 70 total ways to choose 4 cups out of 8. So the probability she'd identify all four milk cups correctly by chance is only $1/70$. Another way to think about this is that Bristol's observed choice perfectly matches only one of the possible orderings of the cups (the one realized in the actual experiment).

In contrast, there are 16 ways she could have gotten 3 out of 4 milk cups right ($\binom{4}{3}\binom{4}{1} = 16$), which would put the random guess probability at $16/70$ for just one wrong guess!

Randomization inference

Idea is to test the sharp null hypothesis $H_0 : Y_{1i} = Y_{0i}, \forall i$. Under the null hypothesis, we can impute the full schedule of potential outcomes.

D_i	Y_0	Y_1	τ_i
1	?	-4	?
1	?	5	?
0	1	?	?
0	-10	?	?

Randomization inference

Idea is to test the sharp null hypothesis $H_0 : Y_{1i} = Y_{0i}, \forall i$. Under the null hypothesis, we can impute the full schedule of potential outcomes.

D_i	Y_0	Y_1	τ_i
1	−4	−4	0
1	5	5	0
0	1	1	0
0	−10	−10	0

This allows us to characterize the sampling distribution of any estimator under the sharp null by re-estimating the statistic under every possible permutation of the treatment assignment vector.

Randomization inference

D_i	Y_0	Y_1	τ_i
1	−4	−4	0
1	5	5	0
0	1	1	0
0	−10	−10	0

Fix $N_1 = 2$. There are $\binom{4}{2} = 6$ possible treatment assignment vectors and corresponding ATE estimates:

Assignment vector	$\hat{\tau}_{ATE}$
$(1, 1, 0, 0)$	5
$(1, 0, 1, 0)$	1
$(1, 0, 0, 1)$	−10
$(0, 1, 1, 0)$	10
$(0, 1, 0, 1)$	−1
$(0, 0, 1, 1)$	−5

Randomization inference

D_i	Y_0	Y_1	τ_i
1	−4	−4	0
1	5	5	0
0	1	1	0
0	−10	−10	0

Fix $N_1 = 2$. There are $\binom{4}{2} = 6$ possible treatment assignment vectors and corresponding ATE estimates:

Assignment vector	$\hat{\tau}_{ATE}$
$(1, 1, 0, 0)$	5
$(1, 0, 1, 0)$	1
$(1, 0, 0, 1)$	−10
$(0, 1, 1, 0)$	10
$(0, 1, 0, 1)$	−1
$(0, 0, 1, 1)$	−5

The actual $\hat{\tau}_{ATE}$ we observe is 5. Using the exact distribution of the test statistic under the null, we can compute that in 2 out of 6 possible randomizations we observe a statistic at least this large. So we have a p-value of 0.33.

Randomization inference

- When we generate all possible treatment assignment vectors, randomization inference gives us the exact p-value for our hypothesis test.
- In practice, often we just take a random sample of the treatment assignment vectors since there are too many to compute all of them.
- Simulate different assignment vectors using exactly the same assignment mechanism.
- Randomization inference doesn't rely on any asymptotics or assumptions about distributions.
- Can be useful especially in small samples.
- Note that it tests a different null hypothesis!

Implementation

R code