Notes of Causal Inference

two points of view

Yue Wu

1 Introduction

In this note, we mainly discuss two points of view in casual inference.

They are the same for.. [Similarity of design-based analysis and super-population observed view: they want better estimator(in a sense of first order, second order); key difference of design-based analysis and super-population observed view: design-based analysis only focus on 'better' in a finite-n sense, but super-population analysis can focus on large-n sense] not talking about [4]

- Finite Population and Randomized Experiment View
- Super Population and Observational View

The causal task we discuss mainly focuses on the situation where we
Always be cautious:

- when you write a fraction, would there be a prob that the denominator is 0?
- when you do expectations, what is your population, what are you exactly calculating average of?
- the uncertainty comes from? (a way of testing whether understanding this point is to try if you can do simulation)
- how many distribution do we have, each has a distribution for i-th or iid samples from one distribution?
- random a scalar or a vector?

Sometimes, we need to think in a data-generating way, and think, do we get Y first, and assign Z, or do we get Y,Z pair together as a sample from an unknown population(and we must have some assumption of this distribution to make inference, just as the way we do in OLS regression analysis).

2 What we observed, and some canonical estimators(functions of observed data)

This chapter is just presenting/displaying some calculations/functions of the observed data we have, and we do not do any analysis. The notation $\hat{\tau}$ is introduced for convenience later, so its meaning can be ignored for now. We introduce these functions before defining the estimands we are interested in. (Here n is the size of sample we have, which is known and fixed.)

2.1 When we observed $\{Z_i, Y_i\}_{i=1}^n$

- Difference in means

  The Difference-in-means(DIM) estimator, as the name of it, is the difference of the means of observed Y_i of units between the treated group and controlled group.

$$\hat{\tau}^{DIM} = n_1^{-1} \sum_{i=1}^{n} Z_i Y_i - n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i \tag{1}$$

  A subtle point here is that by writing the estimator in this form, we are implicitly assuming that both n_0 and n_1 are nonzero, which is trivial if we assigned or observed units of both treated and controlled group. (Would either of them be 0 and the other equals to n? How possible would this happen?)

  DIM estimator can also be expressed in a regression view by running regression $Y \sim 1 + Z$, with the OLS estimate $\hat{\beta}_1$ numerically identical to DIM estimator.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 Z_i + \epsilon_i, \\ \hat{\tau}^{DIM} &= \hat{\beta}_1. \end{aligned} \tag{2}$$

- Inverse Propensity Weight(IPW) estimator

$$\hat{\tau}^{IPW} = n^{-1} \sum_{i=1}^{n} Y_i \left(\frac{\mathbf{1}_{\{Z_i=1\}}}{Pr(Z_i = 1)} - \frac{\mathbf{1}_{\{Z_i=0\}}}{Pr(Z_i = 0)} \right)$$

  if we know $Pr(Z_i = 1)$ and $Pr(Z_i = 0)$

2.2 When we observed $\{Z_i, Y_i, X_i\}_{i=1}^n$(discussed in later notes)

If we observed some covariate X_i, usually it means that we get additional information to make inference(why? in what way this information would help the estimation?). And various forms of estimators can be use, depending on how we would like to use this information.

- Difference in means(we can still calculate this value)
- Outcome modeling (or regression)/Linearly adjusted estimators/ANOVA estimator(s)

$$\hat{\tau}^{reg} = n^{-1} \sum_{i=1}^{n} (\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)) \tag{3}$$

- Inverse probability weighting (IPW)
- Matched-pairs estimator(s)

3 Finite Population and Randomized Experiment View

Before conducting any statistical analysis or inference, a question to ask is, **what population are we targeting**? This chapter focuses on cases where we care about the units we have observed, or to say, an interior way of thinking. In this sense, the sample is the population we are targeting and making inference on, and we reach internal validity by making inference on the specific units we have.

The Finite-Population Perspective: A Familiar Starting Point Recall what we do in standard regression analysis: (1) treat X_i(a scalar or a vector) as the fixed character and Y_i(a scalar) as a random variable for each observed units i; (2) make statistical assumption and inference on the general[in a finite sense] relation of Y(a random vector) and $X = (X_1, ..., X_n)$(a fixed vector or matrix) for all observed units; (3) by using Y^{obs}(the observed vector of size n) and X^{obs}(the observed vector or matrix of size n).

Now, in finite population view of causal inference, we: (1) treat $\{Y_i(1), Y_i(0), X_i\}$(a pair of scalars + a scalar or a vector) as the fixed character for each observed units i, where the pair is called potential outcomes and for each unit we only observe one of them; (3) by using Y^{obs}(the observed vector of size n) and Z^{obs}(the observed vector of size n) and possibly X^{obs}(the observed vector or matrix of size n); And now it is natural to think about: (2) **where the uncertainty comes from and what statistical assumption we plausibly make? What general thing we care about and want to make inference on for all observed units?**

3.1 Finite Population Estimands and Neymanian Repeated Sampling Inference

For Randomized Experiment, what do we mean by saying Randomized? In Neyman's view[site], the uncertainty comes solely from the assignment mechanism, which means that we treat $Z = (Z_1, ..., Z_n)$ as stochastic, i.e. a random vector of size n. We make assumption of the assignment mechanism, i.e. the statistical model of Z. The assignment mechanism indeed is an important thing, but it is not the interest of causal inference, so the assumption is always set without unknown parameters that need to be inferred(which is different from the regression analysis).

In causal inference, the general thing we really care about and want to make inference on for all units is some function/comparison of $\{Y_i(1), Y_i(0)\}$. A natural and basic choice of estimand is the Average Treatment Effect(ATE), following Neyman's questions: What would the average outcome be if all units were exposed to the treatment? How did that compare to the average outcome if all units were exposed to the control?

3.2 Set up

The units we care about are of size n, the potential outcomes of each units are $\{Z_i, Y_i\}_{i=1}^n$, what we can assign and observe are $\{Z_i, Y_i\}_{i=1}^n$.

Finite population quantities:

- means of potential outcomes
- variances of potential outcomes
- covariance of potential outcomes
- mean of individual effects(the average treatment effect(ATE)/average causal effect(ACE)/the difference in average potential outcomes/average of the differences in individual potential outcomes)
- variance of individual effects

And to summarize, we care about

- target population: finite population with fixed n and units of interest

- statistical assumption: fixed $Y_i(1), Y_i(0)$ and random assignment $Z = (Z_1, ..., Z_n)$
- target quantity/estimand: finite population average treatment effect $\tau = n^{-1} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$

3.3 Types of assumption of Treatment assignment in Randomized experiment

Treatment assignment is the process by which treatment was allocated to or chosen by(for) the participants in the study. And what do we mean by saying Randomized Experiment? How can the experiment conducted in what kind of Randomized way?

The **randomized** here denote that the assignment is independent of the potential outcomes of each participants. There are many ways to design a randomized experiment, and in this chapter we talk about two basic ways of randomized assignments.

Completely randomized experiment(CRE) For n_1 and n_0 with $n = n_1 + n_0$

$$\Pr(\mathbf{Z} = \mathbf{z}) = 1 \Big/ \binom{n}{n_1},$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ satisfies $\sum_{i=1}^{n} z_i = n_1$ and $\sum_{i=1}^{n} (1 - z_i) = n_0$.

The assignment vector $\mathbf{Z}$ has the following properties,

$$\mathbb{E}[\mathbf{Z}] = \frac{n_1}{n} \mathbf{1}_n, \quad \mathrm{Cov}(\mathbf{Z}) = \frac{n_1 n_0}{n(n-1)} \Big(\mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^\top \Big),$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{1}_n$ is the n-dimensional vector of ones.

This is a Multivariate Hypergeometric distribution and random permutation of n_1's and n_0's. It can be viewed as a random vector of n dependent random scalar:

$$\mathbb{E}[Z_i] = \frac{n_1}{n}, \quad \mathrm{Var}(Z_i) = \frac{n_1 n_0}{n(n-1)}, \quad \mathrm{Cov}(Z_i, Z_j) = -\frac{n_1 n_0}{n(n-1)} \quad (i \neq j),$$

$$Pr(Z_i = 1) = \mathbb{E}[Z_i] = \frac{n_1}{n} > 0$$

$$Z_i independent of Y$$

Bernoulli randomized experiment(BRE) (Note here the unit-level model is Bernoulli, and for n independent unit together is a ... vector random variable.)

It is a design-based way of thinking when we talk about randomized study. The assumption for randomized experiment is a statistical model, but we always know its exact distribution because we design the experiment. So is actually a **'fact assumption'**. We do not need to make 'plausible' assumption based on observed data since we already know the true statistical model.

If we think of the data-generating process and assume that we have not observed the data yet. Under CRE, we do know in advance that the realized data will satisfy $\sum_{i=1}^{n} Z_i = n_1$ but the real distribution of who get the treatment is random. Under BRE, we do not know the exact number of people to get the treatment, and we actually may end up with no treated sample

with a tiny probability. In that case, we need to re-randomize to get a sample that we can make inference with.

ex: With Large sample, for example the Randomized controlled trial in IT company, what can we assume, the sample means the population? Is the RCT more like CRE or BRE?

[write unconfoundedness and overlap]

3.4 Neymanian Repeated Sampling Inference under CRE

In this chapter, we derive some properties[summary/moments of distribution] of the ATE estimators, under the assumption of CRE. We focus on the first level(mean, unbiasedness) and second level(variance, as well as the estimator of sample variance) analysis.

3.4.1 DIM estimator

(1) DIM is an unbiased estimator for τ

$$E(\hat{\tau}^{DIM}) = \tau$$

(2) the variance of DIM estimator is

$$\text{Var}(\hat{\tau}^{DIM}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} + \frac{S^2(\tau)}{n}$$

where $S^2(1)$ and $S^2(0)$ is the variances of potential outcomes in the sample, defined as:

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(Y_i(1) - \bar{Y}(1)\big)^2, \quad S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(Y_i(0) - \bar{Y}(0)\big)^2,$$

and S_τ^2 is the sample variance of the individual treatment effects, defined as:

$$S_\tau^2 = (n-1)^{-1}\sum_{i=1}^{n}\big(Y_i(1) - Y_i(0) - \tau\big)^2$$

(3) the variance estimator

$$\hat{V} = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

is conservative for estimating $var(\hat{\tau}^{DIM})$ in the sense that

$$E(\hat{V}) - var(\hat{\tau}^{DIM}) = \frac{S_\tau^2}{n} \geq 0$$

3.4.2 OLS/regression adjusted estimator with X

Why is X useful here? a secondary level of use.

3.5 Does Randomized Experiment have anything to do with Super-population? Link internal and external validity. Can we think it in a super population view? (CRE)

[check words of 'finite population' or 'finite sample']

In previous chapters, we talk about Randomized Experiment, which assumes that ..., and .

By saying repeated Sampling, we are saying...

Two source of randomness.

reference: 3.5.1:[2]; 3.5.2:[3] Chapter 6.

For the previous part, the estimand τ is the finite sample population mean of individual effects, which is a function of the potential outcomes of the observed sample. It is a fixed but unknown value because we can not observe $Y_i(1), Y_i(0)$ at the same time, so we make assumptions and estimation for this quantity.

Understanding randomized sampling in this way is critizised because it only yields an outcome based solely on the sample units we observe. So a way to think a further step more, is to see the samples as a sample from a superpopulation.

[] So we need to answer: can we make inference on the quantity of the superpopulation distribution if we treat $Y(1), Y(0)$ as iid sample?

In this section, we use $\hat{\tau}$ for the DIM estimator for simplicity.

3.5.1 if observed units is IID from a infinite super population

Assume that random variables $(Y(1), Y(0))$ represent the pair of potential outcomes of an infinite super population. At the super population level, the average potential outcomes are $E\{Y(1)\}$ and $E\{Y(0)\}$, and the super-population quantities of average causal effect, variances of the potential outcomes and individual causal effect are

$$\tau_{sp} = E(Y(1) - Y(0)), \tag{4}$$

$$V_1 = \mathrm{Var}\{Y(1)\}, \quad V_0 = \mathrm{Var}\{Y(0)\}, \quad V_\tau = \mathrm{Var}\{Y(1) - Y(0)\}. \tag{5}$$

Take an independent and identically distributed (IID) finite population $\mathcal{S} = \{(Y_i(1), Y_i(0)) : i = 1, \ldots, n\}$ of size n from the infinite super population. At this finite sample/population level, we use τ_{fs} to denote the finite sample population mean of individual effects, which is a random variable itself. In this view, the τ we previously discussed is a realization of this value of the sample we observe. And from section 3.4 we know that

$$E(\hat{\tau} \mid \mathcal{S}) = \tau_{fs}, \quad \mathrm{Var}(\hat{\tau} \mid \mathcal{S}) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n}$$

Since $\mathcal{S}$ is IID sampled from the super population, we have

$$E_{sp}(\tau_{fs}) = \tau_{sp}, \quad Var_{sp}(\tau_{fs}) = Var_{sp}(\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) = \frac{V_\tau}{n},$$

$$E_{sp}(S_1^2) = V_1, \quad E_{sp}(S_0^2) = V_0, \quad E_{sp}(S_\tau^2) = V_\tau$$

Therefore, we can derive the result on $\hat{\tau}$ in a super population view, the DIM estimator is unbiased for the superpopulation mean of treatment effect:

$$\begin{aligned} E_{sp}(\hat{\tau}) &= E_{sp}\{E\{\hat{\tau} \mid \mathcal{S}\}\} \\ &= E_{sp}\{\tau_{fs}\} \\ &= \tau_{sp}, \end{aligned} \tag{6}$$

	means			variances		
	treatment	control	effect	treatment	control	effect
super population	$E\{Y(1)\}$	$E\{Y(0)\}$	$\tau_{sp} = E\{Y(1) - Y(0)\}$	V_1	V_0	V_τ
finite population	$\bar{Y}_1$	$\bar{Y}_0$	$\tau_{fs} = \bar{Y}_1 - \bar{Y}_0$	S_1^2	S_0^2	S_τ^2
sample	$\bar{Y}_1^{obs}$	$\bar{Y}_0^{obs}$	$\hat{\tau} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$	s_1^2	s_0^2	–

The sampling variance of $\hat{\tau}$ over sampling from the super-population equals

$$\begin{aligned}
\mathrm{Var}_{sp}(\hat{\tau}) &= E_{sp}\{\mathrm{Var}(\hat{\tau} \mid \mathcal{S})\} + \mathrm{Var}_{sp}\{E(\hat{\tau} \mid \mathcal{S})\} \\
&= E_{sp}\{\mathrm{Var}(\hat{\tau} \mid \mathcal{S})\} + \mathrm{Var}_{sp}(\tau_{fs}) \\
&= E_{sp}\{\mathrm{Var}(\hat{\tau} \mid \mathcal{S})\} + \frac{V_\tau}{n} \\
&= E_{sp}\{\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n}\} + \frac{V_\tau}{n} \\
&= \frac{V_1}{n_1} + \frac{V_0}{n_0} - \frac{V_\tau}{n} + \frac{V_\tau}{n} \\
&= \frac{V_1}{n_1} + \frac{V_0}{n_0}
\end{aligned} \tag{7}$$

which we can estimate without bias by substituting s_1^2 and s_0^2 for V_1 and V_0, $\hat{V} = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$. The estimator is identical to the previously conservative estimator of the sampling variance for the finite population average treatment effect estimator. Therefore, the conclusion is that: when considering the n observed units as a IID sample from an infinite super-population, the estimator $\hat{V}$ is not conservative but an unbiased estimate of the sampling variance of the estimator of the super-population average treatment effect.

3.5.2 if observed units is Simple Random Sample (SRS) from a super population

We can think of the super-population perspective and sampling in another way: with a large but countable number of units, and fixed value of $\{Y_i(1), Y_i(0)\}$ pairs, and our sample as a Simple Random Sample(SRS) from this population.

Let N_{sp} be the size of the super-population, with N_{sp} large, but countable. Let $\mathbf{Y}_{sp}(1)$ and $\mathbf{Y}_{sp}(0)$ denote the N_{sp}-component vectors with i-th element equal to $Y_i(1)$ and $Y_i(0)$ respectively. The super-population quantities of average causal effect, variances of the potential outcomes and individual causal effect are

$$\tau_{sp} = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} (Y_i(1) - Y_i(0)) = \bar{Y}_{sp,1} - \bar{Y}_{sp,0}, \tag{8}$$

$$\sigma_1^2 = -, \quad \sigma_0^2 = -, \quad \sigma^2 = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} (Y_i(1) - Y_i(0) - \tau_{sp})^2. \tag{9}$$

Suppose that the population of n subjects taking part in the completely randomized experiment is itself a simple random sample from a larger population. Simple Random Sample(SRS) is a sampling way often used in survey Sampling. [...]

	size	means			variances		
		treatment	control	effect	treatment	control	effect
super population	N_{sp}	$\bar{Y}_{sp,1}$	$\bar{Y}_{sp,0}$	$\tau_{sp} = \bar{Y}_{sp,1} - \bar{Y}_{sp,0}$	σ_1^2	σ_0^2	σ_τ^2
finite population	n	$\bar{Y}_1$	$\bar{Y}_0$	$\tau_{fs} = \bar{Y}_1 - \bar{Y}_0$	S_1^2	S_0^2	S_τ^2
sample	n	$\bar{Y}_1^{obs}$	$\bar{Y}_0^{obs}$	$\hat{\tau} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$	s_1^2	s_0^2	–

Here we have also two sources of randomization: (1) Let R_i denote the sampling indicator, so that $R_i = 1$ if unit i is sampled, and $R_i = 0$ if unit i is not sampled, with $\sum_{i=1}^{N_{sp}} R_i = n$. (2) For the units in the finite sample, we have $W_i = 1$ for units assigned to the treatment group, and $W_i = 0$ for units assigned to the control group. To simplify the exposition, let us assign $W_i = 0$ to all units not sampled (with $R_i = 0$). So we have $\sum_{i=1}^{N_{sp}} R_i W_i = n_1$. In this perspective[], we can rewrite the finite population effect mean as $\tau_{fs} = \frac{1}{N}\sum_{i=1}^{N_{sp}} R_i(Y_i(1) - Y_i(0))$. And from section 3.4 we know that

$$E(\hat{\tau} \mid \mathbf{R}) = \tau_{fs}, \quad \text{Var}(\hat{\tau} \mid \mathbf{R}) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n}.$$

Since the finite population is a SRS of the superpopulation, we have:

$$E_{sp}(\tau_{fs}) = \frac{1}{N}\sum_{i=1}^{N_{sp}} E(R_i)(Y_i(1) - Y_i(0)) = \frac{1}{N_{sp}}\sum_{i=1}^{N_{sp}}(Y_i(1) - Y_i(0)) = \tau_{sp},$$

$$Var_{sp}(\tau_{fs}) = E_{sp}\big[(\frac{1}{N}\sum_{i=1}^{N_{sp}} R_i(Y_i(1)-Y_i(0))-\tau_{sp})^2\big] = \frac{\sigma_\tau^2}{N}-\frac{\sigma_\tau^2}{N_{sp}}-\frac{1}{N_{sp}^2}\sum_{i=1}^{N_{sp}}\sum_{j\neq i}(Y_i(1)-Y_i(0)-\tau_{sp})(Y_j(1)-Y_j(0)-$$

Therefore, we can derive the result on $\hat{\tau}$ in a super population view, the DIM estimator is unbiased for the superpopulation mean of treatment effect:

$$\begin{aligned} E_{sp}(\hat{\tau}) &= E_{sp}\{E\{\hat{\tau} \mid \mathbf{R}\}\} \\ &= E_{sp}\{\tau_{fs}\} \\ &= \tau_{sp}, \end{aligned} \tag{10}$$

The sampling variance of $\hat{\tau}$ over sampling from the super-population equals

$$\begin{aligned} \text{Var}_{sp}(\hat{\tau}) &= E_{sp}\{\text{Var}(\hat{\tau} \mid \mathbf{R})\} + \text{Var}_{sp}\{E(\hat{\tau} \mid \mathbf{R})\} \\ &= E_{sp}\{\text{Var}(\hat{\tau} \mid \mathbf{R})\} + \text{Var}_{sp}(\tau_{fs}) \\ &= E_{sp}\{\text{Var}(\hat{\tau} \mid \mathbf{R})\} + \frac{V_\tau}{n} \\ &= - \end{aligned} \tag{11}$$

3.5.3 IID and SRS

connection of above S and R

IID: sample with replacement

SRS: sample without replacement

SATE/PATE

continuous or countable

Can we inference on the superpopulation if we treat Y(1),Y(0) as iid sample?

ex: With Large sample, for example the Randomized controlled trial in IT company, what can we assume, the sample means the population? Is the RCT more like CRE or BRE?

4 Super Population and Observational View

When we talked about the assumption of assignment mechanism in randomized experiment, note that the 'assumption' is actually a 'fact' because we designed the experiment. This is not the case when it comes to observational studies, where we actually do not control and know the treatment assignment mechanism.

[In this case, observations are just realizations, just a sample of targeting population.]

In observational point of view, we view the observed units as observed samples from an unknown population, which we make assumptions and inference of. This is more like the traditional statistical analysis, with sample is IID and stochastic.

$$\{Z_i, Y_i(1), Y_i(0)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \{Z, Y(1), Y(0)\}. \tag{12}$$

We can factorize the joint distribution as

$$\Pr(Z, Y(1), Y(0)) = \Pr(Y(1), Y(0)) \times \Pr(Z \mid Y(1), Y(0)) \tag{13}$$

where

(1) $\Pr(Y(1), Y(0))$ is the outcome distribution.

(2) $\Pr(Z \mid Y(1), Y(0))$ is the treatment distribution conditional on the outcome.

What we observed as realization, Y, can be seemed as a nonparametric function of $\{Z, Y(1), Y(0)\}$, and is a random variable.

$$Y = ZY(1) + (1 - Z)Y(0), \tag{14}$$

$$\{Z_i, Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \{Z, ZY(1) + (1 - Z)Y(0)\}. \tag{15}$$

4.1 Set up: Population, Overlap Assumption and Population Estimands

Summarize:

- target population: super population with distribution $\{Z, Y(1), Y(0)\}$
- statistical assumption: iid samples of $\{Z_i, Y_i(1), Y_i(0)\}_{i=1}^n$ + unconfoundedness + overlap
- target quantity/estimand: population average treatment effect

$$\tau = E[Y(1) - Y(0)] = E(Y(1)) - E(Y(0)) \tag{16}$$

4.2 DIM estimator

Before we make estimating-necessary assumption of the joint distribution, let's see the DIM estimator in observational view, and derive its mean. We mainly consider two questions: 1. can it be an estimator of τ? 2. if it can, is it unbiased? when is it unbiased?

$$\begin{aligned} E\big[\hat{\tau}^{DIM}\big] &= E\left[\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)Y_i\right] \\ &= E\big[\bar{Y}_1\big] - E\big[\bar{Y}_0\big] \\ &= \frac{1}{n_1} n_1 E\big[Y \mid Z=1\big] + \frac{1}{n_0} n_0 E\big[Y \mid Z=0\big] \\ &= E\big[Y \mid Z=1\big] - E\big[Y \mid Z=0\big] \end{aligned} \tag{17}$$

Since $\tau = E(Y(1)) - E(Y(0))$, an intuitive equation to make $E(\hat{\tau}^{DIM}) = \tau$ is to satisfy: $E(Y \mid Z=z) = E(Y(z))$. So what assumption can make this happen?

4.3 random assumption

In this chapter, we introduce an assumption and its use for estimation[demonstrate what can be derived under it for estimation]. This assumption may be too strong and not plausible for observational studies, but it can shows us ahead how we deal with the i.i.d observational case, and later in the following chapter we will discuss a more plausible assumption.

Assumption 1 (Independence of Y(0),Y(1) and Z)**.**

$$Y(z) \perp\!\!\!\perp Z \quad \textit{for} \quad z = 0, 1$$

Since Z is a binary random variable, we can also express this assumption as:

$$Pr(Y(z) \mid Z=1) = Pr(Y(z) \mid Z=0), \quad z = 0, 1 \tag{18}$$

Under Assumption 1, we have:

- distribution

  For $z = 0, 1$,

$$\begin{aligned} Pr(Y \mid Z=z) &= Pr(ZY(1) + (1-Z)Y(0) \mid Z=z) \\ &= Pr(zY(1) + (1-z)Y(0) \mid Z=z) \\ &= Pr(Y(z) \mid Z=z) \\ &= Pr(Y(z)) \quad (\text{given } Y(z) \perp\!\!\!\perp Z) \end{aligned} \tag{19}$$

  Which means the observable distribution of Y in treatment arm $Z=z$ equals the distribution of the potential outcome $Y(z)$.

- identifiabiliy

  The notion of nonparametrically identifiability does not appear frequently in classic statistics where we typically can observe fully what we are interested of. But for causal inference with observational studies, it is important to check whether the parameter of interest(e.g. $\tau = E(Y(1) - Y(0))$) is identifiable, because it often depends on some unobserved random variables(a missing data question). Identification can be explained in this way: A

parameter θ is nonparametrically identifiable if it can be written as a function of the distribution of the observed data without any parametric model assumptions[1].

In the sense of identification, $E(Y(z))$ is nonparametrically identifiable when Assumption 1 holds because

$$E(Y \mid Z = z) = E(Y(z)) \tag{20}$$

Hence, any **linear** combination of $Y(1), Y(0)$ is identifiable. For example, nonparametrical-identifiabiliy is satisfied for $\tau = Y(1) - Y(0)$ with

$$\begin{aligned} E(Y(1) - Y(0)) &= E(Y(1)) - E(Y(0)) \\ &= E(Y \mid Z = 1) - E(Y \mid Z = 0) \end{aligned} \tag{21}$$

- estimation

  Although we cannot observe $Y(z)$ for every unit in causal analysis, the population average $E(Y(z))$ is estimable from the sample dataset through $E(Y \mid Z = z)$. That means, identifiability of $E(Y(z))$(following Assumption 1) allows us to estimate $E(Y(z))$.

  For example, a method-of-moments estimator for $E(Y \mid Z = z)$ (and thus for $E(Y(z))$) is

$$\begin{aligned} \widehat{E}(Y(z)) &= \widehat{E}(Y \mid Z = z) \\ &= \frac{1}{n_z} \sum_{i=1}^{n} \mathbf{1}_{\{Z_i = z\}} Y_i, \quad z = 0, 1 \end{aligned} \tag{22}$$

  Also, any linear combinations of $\{Y(1), Y(0)\}$ can be estimated, such as $\tau = E(Y(1) - Y(0))$, which we care about a lot in causal inference.

Assumption 2 (Independence of {Y(0),Y(1)} pair and Z)**.**

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z$$

A stronger version of assumption requires the joint distribution of $\{Y(1), Y(0)\}$ is independent of Z. And this assumption is needed if the parameter of interest is the causaleffects on other scales (for example, distribution, quantile, or some transformation of the outcome). Assumption 2 can derive 1, so it is also used as an usual assumption.[...]

But assumption of independence is so strong for observational study Generally for real world, the circumstances where assignment mechanism is independent of potential outcomes are rare. It is like an imposed randomized experiment assumption, and unrealistic for the observational studies, where the assignment mechanism is often related to potential outcomes.

In next section, we impose a plausible assumption of unconfoundedness, and then think of more ways to get a better estimation, with the information of covariates X.

4.4 An implicit assumption and sutle biasedness of DIM

Before we get our observed data, we can think a question: is there a chance of observing a dataset with only treatment group or control group and not being able to estimate causal estimands τ?

Assumption 3 (Overlap(without X))**.**

$$0 < Pr(Z = 1) < 1$$

- If $P(Z = 1) = 0$ or $P(Z = 1) = 1$, then we will get a sample of $n_1 = 0$ or $n_0 = 0$ w.p. 1 and end up with no treatment or control group.
- If $P(Z = 1) \in (0, 1)$, then we have a prob of ending up with no treatment or control group as

$$Pr(n_1 \text{ or } n_0 \text{ is } 0) = (1 - Pr(Z = 1))^n + Pr(Z = 1)^n$$

So even if the true distribution shows $Pr(Z = 1) = 0.5$, we still have a tiny prob ending up an extreme cause. This is a subtle point because this can be ruled out after we observed the data, but we can always think of this question before we get our data.

So if we define the DIM estimator as:

Let n_1 be the number of units in treatment

$$\hat{\tau} = \begin{cases} \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)Y_i & \text{if } n_1 = 0 \\ \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i & \text{if } n_0 = 0 \\ \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)Y_i & \text{if } n_1 \neq 0,\ n_0 \neq 0 \end{cases} \tag{23}$$

We can derive the mean pf DIM estimator as

$$\begin{aligned}
E(\hat{\tau}) &= \sum_{k=0}^{n} E(\hat{\tau} \mid n_1 = k) \\
&= E(\frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)Y_i \mid n_1 = 0)Pr(n_1 = 0) \\
&+ E(\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i \mid n_1 = n)Pr(n_1 = n) \\
&+ E(\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)Y_i \mid n_1 \in [1, n-1])Pr(n_1 \in [1, n-1]) \\
&= -E(Y \mid Z = 0)Pr(Z = 0)^n \\
&+ E(Y \mid Z = 1)Pr(Z = 1)^n \\
&+ (E(Y \mid Z = 1) - E(Y \mid Z = 0))(1 - Pr(Z = 0)^n - Pr(Z = 1)^n) \\
&= E(Y \mid Z = 1)(1 - Pr(Z = 0)^n) - E(Y \mid Z = 0)(1 - Pr(Z = 1)^n) \\
&= E(Y \mid Z = 1) - E(Y \mid Z = 0) + (-(E(Y \mid Z = 1)Pr(Z = 0)^n + E(Y \mid Z = 0)Pr(Z = 1)^n) \\
&= E(Y(1) - Y(0)) + (-E(Y(1)Pr(Z = 0)^n + E(Y(0))Pr(Z = 1)^n) \quad (\text{given } Y(z) \perp\!\!\!\perp Z)
\end{aligned} \tag{24}$$

So in the setting of random assignment, DIM estimator is biased but asymptotically unbiased.

4.5 An unbiased but not-able-to-estimate estimator IPW

Another form of estimating ATE is motivated by a fact:

$$E(Y(1)) = E(\frac{ZY}{Pr(Z=1)}), \quad E(Y(0)) = E(\frac{(1-Z)Y}{1-Pr(Z=1)}) \tag{25}$$

Proof of one side(similar for the other):

$$\begin{aligned} E(\frac{ZY}{Pr(Z=1)}) &= E(\frac{Z^2Y(1) + Z(1-Z)Y(0)}{Pr(Z=1)}) \\ &= E(\frac{\mathbf{1}_{\{Z=1\}}Y(1)}{Pr(Z=1)}) \\ &= E(Y(1))E(\frac{\mathbf{1}_{\{Z=1\}}}{Pr(Z=1)}) \quad (\text{given } Y(1) \perp\!\!\!\perp Z) \\ &= E(Y(1)) \end{aligned} \tag{26}$$

So we can define a moment-based estimator, which is IPW estimator.

$$\hat{\tau}^{ipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{Pr(Z=1)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-Pr(Z=1)}, \tag{27}$$

If we know the true $Pr(Z=1)$, it is an unbiased estimator.

4.6 xx

From the above section, the average effect τ can be estimated, and a simple estimator we think of is the DIM. But, under the settings of observational study, is this parameter unbiased? When it is, when it is not?

$$\begin{aligned} E\big[\hat{\tau}^{DIM}\big] &= E\left[\frac{1}{n_1}\sum_{i=1}^{n}Z_iY_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)Y_i\right] \\ &= E\big[\bar{Y}_1\big] - E\big[\bar{Y}_0\big] \\ &= E\big[Y \mid Z=1\big] - E\big[Y \mid Z=0\big] \\ &= E\big[Y(1)\big] - E\big[Y(0)\big]. \\ &= E() \end{aligned} \tag{28}$$

$$E\big[\hat{\tau}^{DIM}\big] = \tau + \big(E\big[Y(1) \mid Z=1\big] - E(Y(1))\big) - \big(E\big[Y(0) \mid Z=0\big] - E(Y(0))\big) \tag{29}$$

It seems that DIM is unbiased only when $\big(E\big[Y(1) \mid Z=1\big] - E[Y(1)]\big) - \big(E\big[Y(0) \mid Z=0\big] - E[Y(0)]\big) = 0$, which can be the case when: (1) $Z \perp\!\!\!\perp (Y(1), Y(0))$ so that $E\big[Y(1) \mid Z=1\big] = E[Y(1)]$, $E\big[Y(0) \mid Z=0\big] = E[Y(0)]$;(2)...consider U

But generally, the circumstances listed before are rare and we need to think of more ways to get a better estimation, with the information of covariates X. This is what we will talk about in the next note.

5 What do we mean when we combine RCT and OB data

We assume that the two dataset come from the SAME distribution of $\{X, Z, Y(1), Y(0)\} \sim P$. (This can be checked from the data!]

6 Some other discussions

6.1 what we assume, and what we observe

- definity of randomized study, because assumption is fact

- overlap and covariate balance(the relation of plausible(?) assumption and plausible(?) sample, intertextuality) [search 'balancing score' in DPbook]

6.2 An Unusual way of defining random assignment

See the data we get is a sample of X,Y,Z where Z is independent of Y, X, U and Z is random 0,1 with P(Z=1)=.., P(Z=0)=..

The things learned in stats4235.

6.3 Designing a Randomized Experiment

can we better design a randomized experiment with some information known?(when we know X and we can do stratified sampling, when we have results from OB data we collected, Bayesian view for example we know the mean of some parameter(FanLi's work))

6.4 multi-treatment levels

7 Summary

A proofs

The unbiasedness of $\hat{\tau}$ follows from the representation

$$\hat{\tau}^{DIM} = n_1^{-1}\sum_{i=1}^{n} Z_i Y_i - n_0^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i = n_1^{-1}\sum_{i=1}^{n} Z_i Y_i(1) - n_0^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i(0)$$

and the linearity of the expectation:

$$\begin{aligned}
E(\hat{\tau}^{DIM}) &= E\left\{ n_1^{-1}\sum_{i=1}^{n} Z_i Y_i(1) - n_0^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i(0)\right\} \\
&= n_1^{-1}\sum_{i=1}^{n} E(Z_i)Y_i(1) - n_0^{-1}\sum_{i=1}^{n} E(1-Z_i)Y_i(0) \\
&= n_1^{-1}\sum_{i=1}^{n} \frac{n_1}{n} Y_i(1) - n_0^{-1}\sum_{i=1}^{n} \frac{n_0}{n} Y_i(0) \\
&= n^{-1}\sum_{i=1}^{n} Y_i(1) - n^{-1}\sum_{i=1}^{n} Y_i(0) \\
&= \tau.
\end{aligned}$$

The variance of $\hat{\tau}$ follows from Lemma C.2 of simple random sampling:

$$\begin{aligned}
\mathrm{var}(\hat{\tau}^{DIM}) &= \frac{n_1 n_0}{n(n-1)}\sum_{i=1}^{n}\left\{ \frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0} - \frac{\bar{Y}(1)}{n_1} - \frac{\bar{Y}(0)}{n_0}\right\}^2 \\
&= \frac{n_1 n_0}{n(n-1)}\left[\frac{1}{n_1^2}\sum_{i=1}^{n}\{Y_i(1)-\bar{Y}(1)\}^2 + \frac{1}{n_0^2}\sum_{i=1}^{n}\{Y_i(0)-\bar{Y}(0)\}^2 \right. \\
&\qquad \left. + \frac{2}{n_1 n_0}\sum_{i=1}^{n}\{Y_i(1)-\bar{Y}(1)\}\{Y_i(0)-\bar{Y}(0)\}\right] \\
&= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{2}{n} S(1,0).
\end{aligned}$$

From Lemma 4.1, we can also write the variance as

$$\begin{aligned}
\mathrm{var}(\hat{\tau}^{DIM}) &= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{1}{n}\{S^2(1) + S^2(0) - S^2(\tau)\} \\
&= \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}.
\end{aligned}$$

References

[1] Peng Ding. *A first course in causal inference*. Chapman and Hall/CRC, 2024.

[2] Peng Ding, Xinran Li, and Luke W Miratrix. "Bridging finite and super population causal inference". In: *Journal of Causal Inference* 5.2 (2017), p. 20160027.

[3] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[4] Xinran Li and Peng Ding. "General forms of finite population central limit theorems with applications to causal inference". In: *Journal of the American Statistical Association* 112.520 (2017), pp. 1759–1769.