

Section 1

Introduction and Potential Outcomes

Sooahn Shin

GOV 2003

Sept 9, 2021

Overview

- Logistics:
 - Section: Thur 3:00 - 4:15 pm @ K262 ~~@K105~~
 - TF Office Hours: Mon 1:30 - 2:30 / Thur 4:30 - 5:30 pm @ TBD
 - **Pset 1 released!** Due at 11:59 pm (ET) on Sept 15
 - We encourage you to share your questions on Ed.
 - By September 17: Find a collaborator for the project (check the open thread for finding partners on Ed).
- Today's topics:
 1. Identification and estimation
 2. Example: Political canvassing


Identification and Estimation

- The fundamental problem of causal inference (Holland 1986)
 - We only observe one potential outcome per unit
→ How do we infer the missing potential outcomes (= counterfactual)?

- Identification (*definition* of causal effects)
 - Assumptions for defining effects: e.g., SUTVA
 - Estimands (= Quantity of Interest): e.g., Sample Average Treatment Effect (SATE)

consistency
no interference
among units

- Estimation (*learning* from observed outcomes)



	sample	population
treated & control group	SATE	PATE
treated group	SATT	PATT

Example: Political canvassing¹

- Study of n voters $\begin{cases} n_1 \text{ canvassed} \\ n_0 \text{ " " "} \end{cases}$
 - n_1 are canvassed
 - $n_0 = n - n_1$ are not canvassed
- For each voter $i \in \{1, 2, \dots, n\}$, observe:

2. [outcome

1. [selection

$Y_i = \begin{cases} 1 & \text{if cand. A} \\ 0 & \text{" " B} \end{cases}$

$S_i = \begin{cases} 1 & \text{if } i \text{ turned out} \\ 0 & \text{o.w.} \end{cases}$

$D_i = \begin{cases} 1 & \text{if canvassed} \\ 0 & \text{o.w.} \end{cases}$
- Vote choice (observed outcome): $Y_i = 1$ if voter i cast ballot for candidate A, and 0 if the voter cast ballot for candidate B.
- Turnout (observed selection): $S_i = 1$ if voter i turned out, and 0 otherwise.
- Canvassing (treatment): $D_i = 1$ if canvassed, and 0 otherwise.
- Causal question: does canvassing (D_i) affect vote choice (Y_i)?
- Selection on samples:
 - canvassing may affect turnout (S_i), and
 - we only observe the vote choices of the voters who turned out
 \leadsto post-treatment bias $E[Y_i | D_i=1, S_i=1] - E[Y_i | D_i=0, S_i=1]$

¹Example adapted from 2021S STAT286/GOV2003 Review Question 1

Potential Outcomes and Principal Stratification

Data:

i	Age	Gender	D_i	S_i	Y_i
1	30	F	0	1	1
2	20	F	1	1	0
3	40	M	0	0	
4	25	F	1	0	
	\vdots	\vdots	\vdots	\vdots	\vdots

potential

$S_i(1)$	$S_i(0)$	$S_i(1) - S_i(0)$
2	1	2
1	2	2
2	0	2
0	2	2
\vdots	\vdots	\vdots

1. $D_i \rightarrow S_i$

- S_i : **Observed** turnout
- $S_i(d)$ for $d \in \{0, 1\}$: **Potential** turnout

- Recall the "consistency" assumption: $S_i = S_i(d)$ if $D_i = d$
(no hidden versions of treatment) *counter example: variation of amount/level*
- If canvassed [$D_i = d$], the potential turnout when the voter is canvassed [$S_i(d)$] is the observed turnout [S_i] *S: selection*

S_2 : selection

- We have four principal strata defined by $(S_i(0), S_i(1))$

- (1,1): turning out regardless of the canvassing
- (0,1): turning out only when being canvassed
- (1,0): turning out only when not being canvassed
- (0,0): never turning out

first entry $\rightarrow d=0$ $\begin{cases} S_i(0)=0 \\ S_i(0)=1 \end{cases}$ or 2
second entry $\rightarrow d=1$ $\begin{cases} S_i(1)=0 \\ S_i(1)=1 \end{cases}$ or 2

$$(s_i(0), s_i(1))$$

potential S_i given control potential S_i given treatment

Potential Outcomes and Principal Stratification

2. Vote choice does not exist if a voter i does not turn out

- Y_i : **Observed** vote choice

"selection"

- $Y_i(d, s)$ for $d, s \in \{0, 1\}$: **Potential** vote choice

$$Y_i \mid \underbrace{S_i=1}, D_i=d$$

- $Y_i(1, 0)$ and $Y_i(0, 0)$ are not well defined

$$\cancel{Y_i \mid S_i=0, D_i=d}$$

- $Y_i(1, 0)$: Potential vote choice if the voter is canvassed and didn't turn out \leadsto does not exist
- $Y_i(0, 0)$: Potential vote choice if the voter is not canvassed and didn't turn out \leadsto does not exist

• These two are different

potential vote choice $Y_i(d, s)$

principal strata $(S_i(0), S_i(1))$

Estimands (Quantity of Interest) ^{research question}

- Suppose effect of interest is the effect among those who turn out regardless of the treatment.

- What is the individual causal effect of canvassing on voting for candidate A among always turnout?

* This is only defined for voters who always turnout

$Y_i(1,0)$
 $Y_i(0,0)$ } → contradicts w/
the definition
of always turnout

$$\{ Y_i(1,1) - Y_i(0,1) \}$$

$(S_i(0), S_i(1))$ ind. causal effect

$$\begin{aligned} (0,0) &\rightarrow Y_i(1,0) - Y_i(0,0) \\ (0,1) &\rightarrow Y_i(1,1) - Y_i(0,1) \\ (1,0) &\rightarrow Y_i(1,0) - Y_i(0,1) \\ (1,1) &\rightarrow Y_i(1,1) - Y_i(0,1) \end{aligned}$$

- What is the population average treatment effect of canvassing on voting for candidate A among always turnout?

* $Y_i(1,0)$
 $Y_i(0,0)$ } not well-defined!

$$\mathbb{E}[Y_i(1,1) - Y_i(0,1) \mid (S_i(0), S_i(1)) = (1,1)]$$

Estimands

- Vote share for candidate A = $\frac{\text{Number of votes for A}}{\text{Number of those who turn out}}$
- What is the group-level causal effect of canvassing on candidate A's vote share (among n voters in the study)? *for all samples*

$$Z(1) - Z(0) \text{ where } Z(t) = \frac{\sum_{i=1}^n Y_i(t) S_i(t)}{\sum_{i=1}^n S_i(t)} \text{ for } t \in \{0, 1\}$$

\downarrow new quantity \downarrow everyone not canvassed

\uparrow everyone canvassed

Q. why not $\sum_{i=1}^n Y_i(t)$ for numerator?

If $S_i(t)=0 \rightarrow Y_i(t)$ not well defined

\leadsto Use $Y_i(t) S_i(t)$

so that if $S_i(t)=0 \rightarrow Y_i(t) S_i(t)=0$

$=1 \rightarrow \quad \quad = Y_i(t)$

• $\sum_{i=1}^n S_i(t) = \# \text{ turnout} \rightarrow \text{denominator}$

• $\sum_{i=1}^n Y_i(t) S_i(t) = \# \text{ votes for A among those turned out}$
 $\rightarrow \text{numerator}$