Econ 272/Mgtecon 607 - Section 2

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Announcements
2. Discussion Questions
3. Clustered Experiments
4. Clustering in Sampling and in Assignment
5. Practice Problems

Announcements

- Remember to email me about whether you wish to take the final exam, and if you require any special accommodations for it
  - Unless you are an Econ PhD, in which case you have no choice :)
  - If you are a predoc and wish to waive the class next year, you must take the exam as well

- Within-group correlations of a single variable try to answer the question: does *my* value going up mean that *your* value is more likely to go up/down?
  - $Corr(Y_i, Y_j)$ for $i \neq j$ can be calculated by constructing pairs (combinatorially) of observations from the data
  - Recall the definition/equation for correlation - how can it be simplified in this context?

$$Corr(Y_i, Y_j) = \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)Var(Y_j)}} = \frac{Cov(\tilde{Y}_i, \tilde{Y}_j)}{\cancel{Var(\tilde{Y}_i)}} = \underbrace{E[\tilde{Y}_i\tilde{Y}_j]}_{\frac{1}{\binom{N_g}{2}}\sum_{i \neq j}\tilde{Y}_i\tilde{Y}_j} - \cancel{E[\tilde{Y}_i]}E[\tilde{Y}_j]$$

$$\tilde{Y}_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

- Simplified equations for LZ/EHW variance estimators on slides (Lecture 4, Slide 4) are missing a factor of $1/N$

A note on terminology

- In the course, we will often use the term "population" to refer to the experimental sample since we are mainly thinking about design-based inference
- We will therefore refer to the broader population from which the experimental sample is drawn as the "super-population"
- We may occasionally refer to our experimental sample as the "sample", for example when we want to distinguish between sample and population ATEs
- It is confusing, I will try to be as explicit as possible

From last time: nontrivial sharp nulls

Test statistic for sharp null: $\underbrace{Y_i(1)}_{\tilde{Y}_i(1)} = \underbrace{a \cdot Y_i(0) + b}_{\tilde{Y}_i(0)}$

$$T = \bar{\tilde{Y}}_i(1) - \bar{\tilde{Y}}_i(0)$$

$$= \bar{Y}_i(1) - a\,\bar{Y}_i(0) - b$$

Discussion Questions - Clustered Randomized Experiments

1. What is the difference between clustered sampling and clustered assignment?
2. What is the thought experiment underlying clustered standard errors? When should you (not) use them?
3. What is the difference between a clustered randomized experiment and a stratified randomized experiment?
4. Should we prefer clustered or non-clustered randomized experiments? Why are we studying both?

Discussion Questions - Clustering in Sampling and in Assignment

1. What are the traditional frameworks for thinking about clustering standard errors? What are their drawbacks?
2. What do we mean by design-based clustering?
3. Does the data tell us whether we need to cluster our standard errors or not?
4. What is the new variance that takes into account clustering in sampling and in assignment? How does it compare to the Neyman/robust variance and the cluster-robust variance (Liang-Zeger)?
5. How can we estimate this variance?

1. Announcements
2. Discussion Questions
3. Clustered Experiments → Clustered Assignment
4. Clustering in Sampling and in Assignment
5. Practice Problems

Why cluster?

Spillovers, network effects → SUTVA violations

My outcome depends on your treatment

Idea: If spillovers occur in specific clusters
treating clusters as units.

Clustered Sampling vs. Assignment

- **Clustered Sampling:**

control

unsampled

treat

sample

C_1, C_2, C_3

Pick 2

- **Clustered Assignment:**

helps spillovers

treat

control

Notation

- G: Clusters
- $G_i \in \{1, ..., G\}$: cluster assignment for unit i
- N_g: # units in cluster g $(N = \sum_g N_g)$

$= \text{ if } N_g = n \ \forall g$
if TE constant

- **Estimands**:

$$\underbrace{\tau_{avg} = \frac{1}{N} \sum_i Y_i(1) - Y_i(0)}_{\text{overall avg. ATE}} \qquad \underbrace{\tau_{cluster} = \frac{1}{G} \sum_g \left(\frac{1}{N_g} \sum_i Y_i(1) - Y_i(0) \right)}_{\text{overall cluster-level ATE}}$$

- **Estimators**:

$$\underbrace{\hat{\tau}_{avg} = \bar{Y}_1 - \bar{Y}_0}_{\text{hard to estimate properly}} \qquad \hat{\tau}_{cluster} = \frac{1}{G_1} \sum_g \left(\frac{1}{N_g} \sum_g V_i Y_i \right) - \frac{1}{G_0} \sum_g \left(\frac{1}{N_g} \sum_g (1-V_i) Y_i \right)$$

Both estimators are unbiased

Variance: $\hat{\tau}_{cluster}$

$$\rightarrow V_{neyman} = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{N}$$

- **True Variance**:

$$Var(\hat{\tau}_{cluster}) = \frac{S_0^2}{G_0} + \frac{S_1^2}{G_1} - \frac{S_{01}^2}{G}$$

- **Estimator**:

$$\hat{Var}(\hat{\tau}_{cluster}) = \frac{\hat{S}_0^2}{G_0} + \frac{\hat{S}_1^2}{G_1}$$

$$S_0^2 = \frac{1}{G-1} \sum_g \left(\bar{Y}_g(0) - \overline{Y_g(0)}\right)^2$$

Challenge: treatment assignment not indep.

- **True Variance**: Asymptotics: Sequence of pops. $k = 1, 2, \ldots$
  In pop. k, sample all units, assign cluster g to treatment w.p. p
  Keep N_g fixed, let $G_k, N_k \to \infty$:
  $$\sqrt{N_k}\left(\hat{\tau}_{pop} - \tau_{pop}\right) \xrightarrow{d} N(0, V_k)$$
  eqn on slide 24

- **Estimator**: Conservative estimate var: $Y_i = \alpha + \beta W_i + \varepsilon_i$

$$\hat{V}_{cluster} = \left(\frac{1}{N}\sum \dot{W}_i \dot{W}_i^T\right)^{-1} \left(\frac{1}{N}\sum_g \sum_i (\varepsilon_i \dot{W}_i)^2\right)\left(\quad\right)^{-1}$$

$$\hat{V}_{L2} \quad \left(\geq V_{EHW}\right)$$

$$\dot{W} = W_i - \frac{1}{N}\sum_{j=1}^{N} W_j$$

Hypothesis Testing

Inference based on rand. distn. induced by

cluster assignment rather than indiv.

Typically rely asymptotics

But, Fisher exact & SIM work, but units of

analysis are clusters

2 "Flavors" of Error / Variability

1. Sampling – how sample constructed
2. Assignment – how treatment assigned

Both needed for inference in causal settings

→ Abadie et. al. (2023) (general framework)

(Even More) Notation

Sequences of populations indexed by $k = 1, 2, 3, \ldots$

sampling

- g_k: # clusters in k
- $n_{g,k}$: # units in cluster g in popn k
- $q_k \in (0, 1]$: cluster sampling probability in popn k $\begin{cases} q_k = 1 : \text{random sampling} \\ q_k < 1 : \text{clustered sampling} \end{cases}$
- p_k: prob. a unit is sampled from cluster
- $A_{k,g}$: prob. unit in cluster g in popn. k is treated: $\boxed{A_{k,g} \sim (\mu_k, \sigma_k^2)}$

Not all clusters need be sampled
Not all units need same treatment within cluster
} Neither C.S. + C.A.

Clustering Regimes

$\underset{\text{sampling}}{q_k}$, $\underbrace{\sigma^2_k, \mu_k}_{\text{assignment}}$

		Assignment, $A_{k,g}$		
		Random	**Partially Clustered**	**Clustered**
Sampling q_k	**Random**	$q_k=1$, $\sigma^2_k=0$	$q_k=1$, $0<\sigma^2_k<\mu_k(1-\mu_k)$	$q_k=1$, $\sigma^2_k=\mu_k(1-\mu_k)$
	Clustered	$q_k<1$, $\sigma^2_k=0$	$q_k<1$,	$q_k<1$,

Special Cases

- Random sampling of clusters from a large population of clusters $\to q_k$ small
- Random sampling from a large population $\to q_k = 1$, p_k small
  - Completely random assignment $A_{k,g} = A_k$ $(\sigma^2_k = 0)$
  - Clustered random assignment $A_{k,g} \in \{0, 1\}$
- What can we test?

Cannot test $q_k = 1$ vs. < 1

Can test $\sigma^2_k = 0$ vs. ≥ 0

Causal Cluster Variance (CCV) - Big Idea

Assume DiM converges: $\sqrt{N_k}\,(\hat{\tau}_k - \tau_k) \xrightarrow{d} N(0, v_k^{CCV})$

$$V_k^{CCV} = V_k^{CCV}(q_k, \sigma_k^2, p_k, \mu_k)$$

Special cases:

$$V_k^{EHW} = V_k^{CCV}(q_k = 1, \sigma_k^2 = 0)$$

$$V_k^{LZ} - V_k^{CCV} \geq 0 \rightarrow \text{LZ conservative}$$

$$\boxed{\text{Use } \hat{\tau}^{DiM} \text{ to est. } \tau_{pop}}$$

True Variance by Regime

R_{hr} $\hat{\tau}^{DiM} \to \tau_{pop}$

		Assignment, $A_{k,g}$		
		Random	**Partially Clustered**	**Clustered**
Sampling q_k	**Random**	EHW	CCV	LZ
	Clustered	CCV	CCV	CCV

$\geq$ LZ R_{hr} q_k small

If q_k small, EHW $\approx$ LZ $\approx$ CCV

CCV - Estimation

$\sigma^2_{\tau} \to 0$, $\hat{V}^{ccv} \to \hat{V}^{EHW}$ $\quad$ $\sigma^2_{\tau} \to \mu_k(\tau \mu_\tau)$, $\hat{V}^{ccv} \to \hat{V}^{L2}$

or

$q_k \to 0$

① Analytic: $\hat{V}^{ccv}_k = \hat{q}_k \cdot \hat{V}^{CCV}_k(1) + (1-\hat{q}_k)\hat{V}^{L2}_k$

knowledge on the # of clusters in sample

② Bootstrap: TSCB

		Assignment, $A_{k,g}$		
		Random	Partially Clustered	Clustered
Sampling q_k	Random	$\hat{V}^{\text{Ehw}}$	$\hat{V}^{\text{ccv}} \to \hat{q}_k = 1 \to \hat{V}^{\text{ccv}}(1)$	$\hat{V}^{\text{LZ}}$
	Clustered	$\hat{V}^{\text{ccv}}$	$\hat{V}^{\text{ccv}}$	$\hat{V}^{\text{ccv}}$

Note: the choice of estimator really matters here!! Standard errors can vary widely. TSCB works conservative

LZ: if wanted unobserved clusters or trying to extrapolate

EHW: if drawing conclusions local to sample

CCV: if in between, or if you have clustered sampling

1. Announcements
2. Discussion Questions
3. Clustered Experiments
4. Clustering in Sampling and in Assignment
5. Practice Problems

Practice Final 2016, #1

Suppose California is setting up a new job training program. To understand the impact a pilot program is set up where 10% of eligible individuals in 20 counties in California are randomly assigned to the new program. The outcome of interest is labor market status six months from randomization. Let Y_i denote the outcome, let $W_i \in \{0, 1\}$ denote the treatment, and let S_i denote the county.

exact p

(a) Describe how you could test the null hypothesis that there is no effect of the program whatsoever. Discuss all the choices you make in this implementation.

(b) Describe how you would estimate the average effect of the program. $\hat{\tau}^{DiM}$

(c) Describe how you would estimate the variance of this estimator, with and without clustering. $\hat{V}^{LZ}$

(d) Should you cluster here? What are the arguments for or against?

Final 2017, #2

Suppose we conduct a randomized experiment on a random sample of the US population. We assign the treatment randomly at the state level. The dataset observed by the econometrician includes individual-level outcomes, location (state) for each individual, and a treatment indicator that corresponds to the state.

(a) How would you estimate the average effect of the treatment?

(b) How would you estimate the variance of the estimator?

Final 2019, #1

Suppose we conduct a randomized experiment on a random sample of the US population. We assign the treatment randomly at the state level. The dataset observed by the econometrician includes individual-level outcomes, location (state) for each individual, and a binary treatment indicator that that is the same within each state.

(a) Suppose you estimated the average treatment effect as the difference in means by treatment status, give an expression for the variance of the estimator, and how you could estimate the unknown components of that variance?

(b) Suppose you used the Neyman variance estimator that ignored the state-level randomization and that was based on individual level randomization. Would you expect that to over or under estimate the true variance?