# 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

#### 1. Announcements

2. Discussion Questions

Clustered Experiments

4. Clustering in Sampling and in Assignmen

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

### Hints for Problem Set 2 (Due April 13, 11pm)

- 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?  $Corr(Y_i, Y_j) = Corr(Y_i, Y_j)$   $Vor(Y_i) Vor(Y_i)$   $Vor(Y_i) Vor(Y_i) Vor(Y_i)$   $Vor(Y_i) Vor(Y_i) Vor(Y_i)$   $Vor(Y_i) Vor(Y_i) Vor(Y_i)$   $Vor(Y_i) Vor(Y_i) Vor(Y_i)$
- 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: 
$$Y_i(1) = a \cdot Y_i(0) + b$$

$$\overline{Y_i(1)} \qquad \overline{Y_i(0)} \qquad \overline{Y_i(0)}$$

$$\overline{Y_i(1)} \qquad \overline{Y_i(0)} \qquad \overline{Y_i(0$$

#### 1. Announcements

#### 2. Discussion Questions

3. Clustered Experiments

4. Clustering in Sampling and in Assignmen

5. Practice Problems

### Discussion Questions - Clustered Randomized Experiments

- 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

- 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

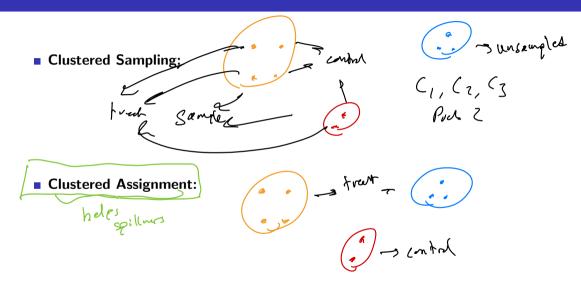
4. Clustering in Sampling and in Assignment

5. Practice Problems

# Why cluster?

Spillovers, neturn ellers -> SUTVA violeting My outcome depense on your treatment Idea: It spillerers over in speeche clusters treating dusters as worts.

### Clustered Sampling vs. Assignment



#### Notation

- G: Clustes
- $lackbox{0.5}{\ } G_i \in \{1,...,G\}$ : Cluster assignment for unt i
- Ng: Hums a cluster g (N= ENg)

#### Point Estimation

Estimators:

nators:

$$\hat{C}_{qq} = \frac{1}{4} - \frac{1}{4} \qquad \hat{C}_{1} \qquad \hat{C}_{1} \qquad \hat{C}_{2} \qquad \hat{C}_{3} \qquad \hat{C}_{4} \qquad \hat{C}_{5} \qquad \hat$$

#### Variance: $\hat{ au}_{ extit{cluster}}$

True Variance:

$$V_{cr}(\hat{\tau}_{uv}) = \frac{S_0^3}{6_0} + \frac{S_1^2}{6_1} - \frac{S_{01}^3}{6}$$

Estimator:

$$Var(\hat{z}_{uar}) = \frac{S_0}{60} + \frac{S_1}{61}$$

$$S_1^2 = \frac{1}{6-1} \underbrace{S(\overline{Y}_g(u) - \overline{Y}_g(u))^2}$$

## Variance: $\hat{ au}_{pop}$

Challage: trebuer assignment and onder.

True Variance: Asymptotis: Seque of gans. 621,2,...

En eren. k, surde all unes, assign char of to trums v. 2 p

Var Ng host, let 62, Nk > 0:

Nk (0, Vx) 24

■ Estimator: Congeriatre estable ven: Y: = 2+ PV, TE:

$$\hat{V}_{\text{cluster}} = \left(\frac{1}{2} \times i_{1} i_{2} \right)^{-1} \left(\frac{1}{2} \times i_{2} \times i_{3} i_{4} \right)^{-1} \\
\hat{V}_{12} = \left(\frac{1}{2} \times i_{1} i_{2} \right)^{-1} \left(\frac{1}{2} \times i_{2} \times i_{3} i_{4} i_{5} \right)^{-1} \\
\hat{V}_{12} = \left(\frac{1}{2} \times i_{1} i_{2} \right)^{-1} \left(\frac{1}{2} \times i_{2} \times i_{3} i_{4} i_{5} i_{5}$$

## Hypothesis Testing

Infere based on rant. Look subject by Typically rely asymptotics. But, Peter example SIM ware, but was al analysis on clusters

1. Announcements

2. Discussion Questions

3. Clustered Experiments

4. Clustering in Sampling and in Assignment

5. Practice Problems

# 2 "Flavors" of Error ( Variable)

I Sampling - has sample constituted

2 Assynwer - how treatmen assigned

Dors muded for where in card sellings -> Abada et. al. (2013) (general francis)

## (Even More) Notation

Sequences of populations indexed by k = 1, 2, 3, ...

- gk: H clushes m le
- ng,k: A une n char g a gran le
- - Pk: gods. a unt is sarded lime Wheter

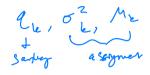
     Ak,g: gods. une on char of on gogs. h is tracks: [Akg ~ (Mb, o'k)]

Not all closure need be sounded of Neer C.S.

Not all chosure need some freshood with char S

C.A.

### Clustering Regimes



		Assignment, $A_{k,g}$			
		Random	Pa	rtially Clustered	Clustered
	Random	9=1,07,20	9,21,		The Correlation
Sampling		2		0602 CN2 (1-M2)	1
$q_k$	Clustered	GEC1 , 5 20	9661,		9621

### Special Cases

- Random sampling from a large population → 1, =1, % Smull

   Completely random assignment  $A_{k,j} = A_k$  ( $\sigma^2$ )
  - Clustered random assignment  $A_{e,g} \in \{0,1\}$
- What can we test?

# Causal Cluster Variance (CCV) - Big Idea

			Assignment, $A_{k,g}$	
		Random	Partially Clustered	Clustered
	Random	124~	LCU	1_2
Sampling $q_k$	Clustered	CCV	αv	cu

Jf Pic Small, Eltha 2 12 2 CCU

#### **CCV** - Estimation

@ Portstrag: TSCB

### Variance Estimation by Regime

	Assignment, $A_{k,g}$		
Random Partia	Illy Clustered Clustered		
	~ = ê621 - Î"(1) ÎL		
Sampling $q_k$ Clustered $\hat{V}$	cev jac		

Note: the choice of estimator really matters here!! Standard errors can vary widely. TSUS LZ: I would unred class on tyg he estable EHV: It drang lonelos land to some CIV: if subshum, or it you he climed souplay

1. Announcements

2. Discussion Questions

3. Clustered Experiments

4. Clustering in Sampling and in Assignmen

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.

- (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.
- (c) Describe how you would estimate the variance of this estimator, with and without clustering.
- (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?