Econ 272 / MGTECON 607: Intermediate Econometrics III – Section 2
April 10, 2025

1 Outline Today

1. Problem Set 2
2. Review
3. Practice Problems

2 Clarifications and Hints for Problem Set 2

For questions 1(e) and 1(f): You are asked to calculate within-cluster correlations of a single variable. This is effectively trying to answer the question: is *my* value somehow related to *your* value? Consider the case of the outcome variable Y. We can calculate the correlation $Corr(Y_i, Y_j)$ for $i \neq j$:

$$Corr(Y_i, Y_j) = \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)Var(Y_j)}} = \frac{Cov(Y_i, Y_j)}{Var(Y_i)}$$

Where the last equality comes from the fact that Y_i, Y_j are drawn from the same distribution and so have the same variance. Now, to make this calculation easier, let's instead examine the correlations between the normalized values of the outcome:

$$\tilde{Y}_i \equiv \frac{Y_i - \bar{Y}}{\sigma_Y} \rightarrow Corr(\tilde{Y}_i, \tilde{Y}_j) = E[\tilde{Y}_i \tilde{Y}_j]$$

Where the last equality comes from the definition of covariance and the fact that these normalized $\tilde{Y}$ have mean zero and unit variance. We can then estimate this by taking an average of this product over all pairs of units in a given cluster (with $i \neq j$).

Note: At some point in the problem set, you may wish to use the simplified LZ/EHW estimators from the slides (Lecture 4 slide #4), which correspond to the case of binary $W \in \{0, 1\}$. As written, those equations correspond to the equations of the variances of the corresponding limiting distributions, not as the direct equations for estimation. If you wish to actually estimate a variance term for a given ATE estimator using either of these equations, you must include an additional factor of $1/N$ out front. With this additional factor, they are directly comparable to the full matrix-notation versions on the previous slide.

3 Review of Concepts From Class

You should be able to answer the following questions:

3.1 Questions

3.1.1 Lecture 3: 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?

3.1.2 Lecture 4: 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?

3.2 Big Picture

Last week, we introduced randomized experiments and talked about the simplest experimental design: the completely randomized experiment. We then moved on to think about the role of covariates in experiments, for design and for analysis purposes. This week, we focused on clustered experiments. Clustered experiments are useful when we think SUTVA is violated in an experiment. That is, what if our treatment cannot guarantee that there is no interference between individuals, e.g. assigning teachers to classrooms. For a high-level overview of the last two weeks, see Figure 1. On Monday, we focused on the details of the different estimands we could be considering in the case of clustered experiments, τ^{pop} and $\tau^{cluster}$, as well as their respective estimators, true variances of the estimators and the estimated variances. On Wednesday, we left our experimental framework for a lecture and spent time talking through the case where we observe clustering in sampling and in assignment. We talked about new variance estimators that remain valid when we are neither in the case of random nor clustered assignment, which can often be the case in observational settings. Table 1 gives a broad overview of the possible combinations for clustering on the sample and assignment level and when the different combinations were discussed. In general, what we covered this week in the Wednesday lecture, covers a whole range of different scenarios and, in particular, the clustered sampling-random assignment case and random sampling-clustered assignment are two limiting cases.

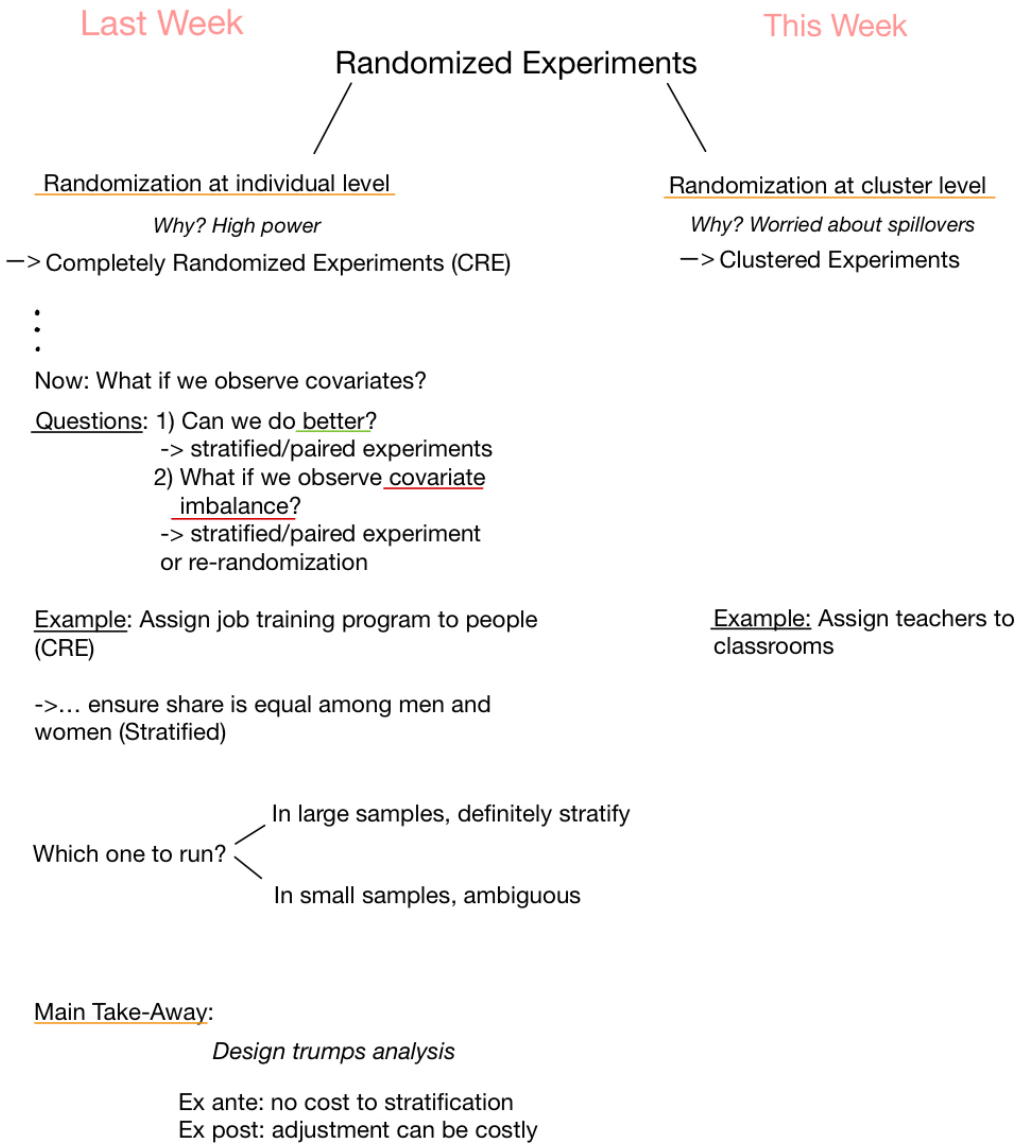

Figure 1: Overview Randomized Experiments

3.3 Clustered Experiments

In this section, we consider clustered experiments. First we discuss the difference between clustered sampling and clustered assignment since these two problems are often conflated in practice. Then we will talk about our three different areas of interest that we also discussed in the other experimental designs: point estimation, inference and hypothesis testing.

3.3.1 Clustered Sampling vs Clustered Assignment

When talking about clustered experiments, it is important to distinguish between clustered *assignment* and clustered *sampling*. While these two clustering levels are often conflated, they are actually

very different problems. However, the solutions to both are similar! Below, I give the definition of both:

Clustered Sampling: Sample M clusters out of the G clusters at random and sample a fraction p of the units from the sampled clusters. Examples include household surveys.

Clustered Assignment: Select M clusters at random and assign all units from those clusters to the active treatment and assign units from remaining $G - M$ clusters to control. Examples include teachers randomly assigned to classrooms.

This leaves us with four possible combinations:

		Assignment		
		Random	**Partially Clustered**	***Clustered***
Sampling	***Random***	*Lecture 1*	*Lecture 4*	*Lecture 3/4*
	Clustered	*Lecture 4*	*Lecture 4*	*Lecture 4*

Table 1: Possible Combinations for Clustering on Sample and Assignment Level

In Lecture 1, we ignored the sampling and took the units in the sample as given. Note that our analysis would not change if we assumed the sampling was clustered or not if all we are interested in is the effect on the sampled units and we are not concerned about any other uncertainty about unobserved clusters, etc..

Intuition: Assume we have three clusters, C_1, C_2 and C_3. For clustered sampling, intuitively the choice we make is that we pick for example 2 of the 3 clusters. Each cluster has two potential outcomes, $C_1(0), C_1(1), C_2(0), C_2(1), C_3(0)$ and $C_3(1)$. For the clustered assignment, we now intuitively pick 2 out of these 6 options.

Recommendation in practice: If you are concerned about spillover effects, clustered assignment will help with that.

3.3.2 Set-Up

For today, we focus on the clustered assignment case, where we randomly select clusters out of a super-population for clusters and **assign the whole cluster to treatment or control**. We randomly select clusters out of a super-population. Let us introduce some more *new* notation:

- G clusters
- $G_i \in \{1, ..., G\}$ is the cluster indicator, i.e. which cluster does individual i belong to
- N_g number of units in cluster g (i.e. $N = \sum_g N_g$)
- ATE = Average Treatment Effect

We still assume we are interested in a finite population of size N. The probability for an individual(or a cluster) of being treated is $p = \frac{G_1}{G}\ \forall i$, where G_1 denotes the number of clusters receiving treatment overall. Note that we still view our potential outcomes as fixed.

3.3.3 Point Estimation

Estimands. There are two estimands of interest now: (1) the estimand for the whole population and (2) the cluster specific estimand,

$$\text{Overall population ATE } \tau_{pop} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

$$\text{Overall cluster specific ATE } \tau_{cluster} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{N_g}\sum_{i:G_i=g}(Y_i(1) - Y_i(0))$$

Note: $\tau_{pop} = \tau_{cluster}$ if (1) the treatment effect is constant or (2) if $N_g = \frac{N}{G}\ \forall g$ (equal cluster sizes). While we can think of $\tau_{cluster}$ as a completely randomized experiment with the units being clusters instead of units, we can think of τ_{pop} as a completely randomized experiment with clusters as unit, but we are interested in a weighted average treatment effect instead of the average effect with weights proportional to cluster sizes.

The first estimand is more interesting, but it is harder to get a precise estimate of it in practice. We can precisely estimate the second one though. Generally, there are a few **issues** that arise with clustered randomized experiments:

- A loss of precision/power relative to completely randomized experiment (effectively less units)
- Choices regarding statistics and estimands matter, especially if there is a lot of variation in N_g!

Recommendation in practice: If the clusters you consider are all of approximately similar size, the difference between the inidividual level estimand and cluster level estimand will not be super big, which means that doing the simple analysis where the unit of analysis are the clusters will be sufficient.

Estimators. We find the following estimators for our two estimands of interest:

$$\text{Overall population ATE } \hat{\tau}_{pop} = \bar{Y}_1 - \bar{Y}_0 = \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i} - \frac{\sum_{i=1}^{N}(1-W_i)Y_i}{\sum_{i=1}^{N}(1-W_i)}$$

$$\begin{aligned}\text{Overall cluster specific ATE } \hat{\tau}_{cluster} &= \frac{1}{G_1}\sum_{g=1}^{G}\frac{1}{N_g}\sum_{i:G_i=g} W_i Y_i - \frac{1}{G_0}\sum_{g=1}^{G}\frac{1}{N_g}\sum_{i:G_i=g}(1-W_i)Y_i \\ &= \frac{1}{G_1}\sum_{g}\bar{W}_g\bar{Y}_g - \frac{1}{G_0}\sum_{g}(1-\bar{W}_g)\bar{Y}_g,\end{aligned}$$

where $\bar{Y}_g = \frac{1}{N_g}\sum_{i:G_i=g} Y_i$ and $\bar{W}_g = \frac{1}{N_g}\sum_{i:G_i=g} W_i$. This resembles the completely randomized experiment difference-in-means estimator. In fact, we can treat this case just like a completely randomized experiment, but with clusters as units of analysis instead of individuals.
Properties. Both estimators are unbiased for their estimands.

3.3.4 Variance

We inspect the two estimands for point estimation separately here. We start with $\hat{\tau}_{cluster}$.

True variance. Since we can think about the setting for this estimator as a completely randomized experiment with the clusters being the units of analysis, we can use the results from lecture 1 here.Thus, our true exact variance for this estimand is

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

with $S_w^2 = \frac{1}{G-1}\sum_g (\bar{Y}_g(w) - \overline{\bar{Y}_g(w)})^2$.

Estimator. Then the estimated (conservative) variance is

$$\widehat{V(\hat{\tau}_{cluster})} = \frac{s_0^2}{G_0} + \frac{s_1^2}{G_1},$$

with $s_w^2 = \frac{1}{G_w-1}\sum_{g:\bar{W}_g=w} (\bar{Y}_g - \overline{\bar{Y}_{gw}})^2$.

Note. This variance estimator is very close to the standard robust variance estimator for the regression models, based on using the units as clusters.

Next, we turn to $\hat{\tau}_{pop}$. Note that the standard robust variance estimator is not appropriate here because the treatment assignments are not independent. Exact variance calculations are not feasible here either since the number of treated and control units $N_1 = \sum_{i=1}^N W_i$ and $N_0 = \sum_{i=1}^N (1 - W_i)$ are stochastic (we assign a fixed number of clusters to treatment, but we don't fix N_1 and N_0).

True variance. We instead turn to asymptotics and consider a sequence of populations, $k = 1, 2, ...$ In population k, we sample all the units and assign cluster g to treatment with probability p. We estimate the average effect τ_{pop} with $\hat{\tau}_{pop} = \bar{Y}_1 - \bar{Y}_0$. We let the number of clusters G_k and the number of units N_k increase in the sequence of population, however keep the number of units per cluster, N_g fixed. We find that

$$\sqrt{N_k}\big(\frac{\hat{\tau}_{pop} - \tau_{pop}}{\sqrt{v_k}}\big) \to^d \mathcal{N}(0,1),$$

where v_k denotes the variance and is given in a long formula in the slides (slide 24).

Estimator. We can estimate this cluster variance conservatively using the regression $Y_i = \alpha + \tau W_i + \varepsilon_i$ and obtaining

$$\widehat{V}_{cluster} = \left(\frac{1}{N}\sum_{i=1}^N \dot{W}_i\dot{W}_i^T\right)^{-1} \left(\frac{1}{N}\sum_{g=1}^G \left(\sum_{i:G_i=g} \hat{\epsilon}_i\dot{W}_i\right)^2\right) \left(\frac{1}{N}\sum_{i=1}^N \dot{W}_i\dot{W}_i^T\right)^{-1} \equiv \hat{V}_{LZ}$$

where $\dot{W}_i = W_i - \frac{1}{N}\sum_{j=1}^{N} W_j$. This is also known as the Liang-Zeger variance. If we consider a population with clustered sampling, we can use this variance if we want our standard errors to capture the uncertainty for the broader super-population.

3.3.5 Hypothesis Testing

Inference is now based on the randomization distribution induced by the clustered assignment instead of individual assignment. Getting exact results is often difficult or sometimes even impossible, so we oftentimes rely on asymptotic results here, where a sequence of populations is considered with the number of clusters getting large.

Fisher Exact p-Values. In principal, we can still use Fisher's exact p-values for inference. This is a straightforward extension of the p-value calculations in completely randomized experiments. Remember that now the main unit of analysis are clusters instead of individuals. We still need to make two choices:

1. Choose a sharp null, e.g. no treatment effect whatsoever.
2. Choosing an appropriate statistic.

Note. The main issue in this case is the choice of the statistic. There are two natural choices

1. $T_{ave} = \frac{\sum_i W_i Y_i}{\sum_i W_i} - \frac{\sum_i (1-W_i) Y_i}{\sum_i (1-W_i)}$
2. $T_{clusters-ave} = \frac{1}{G_1}\sum_g \frac{\sum_{i:G_i=g} W_i Y_i}{N_g} - \frac{1}{G_0}\sum_g \frac{\sum_{i:G_i=g}(1-W_i) Y_i}{N_g}$

These are both equivalent when $N_g = \frac{N}{G}\ \forall g$, i.e. equal cluster sizes. Whenever there is substantial variation in N_g, there can be a big difference in power between the two statistics!

3.4 Clustering in Sampling and in Assignment

The lecture slides and these section notes are mainly based on the paper "When should we adjust our standard errors for clustering" (Abadie et al., 2023) [1]. If you want to know more details, you should definitely read through the paper. The main idea of the new variance and its respective estimators in this paper is that the authors introduce a framework that is **close to the traditional sampling framework that is oftentimes used to justify clustering (more on that in a bit) while explicitly incorporating the design component** that accounts for between-clusters variation in treatment assignments. Recall, by design component we mean incorporating information about how treatment was assigned. That is, completely at random like in an individually randomized experiment or clustered like in a clustered experiment or, what is new here, something in between, which we also refer to as "partially clustered". This partially clustered case is why we shift focus now to **observational settings** for a lecture, so we take a slight detour and leave our area of experimental design.

The motivation for a new framework stems from the fact that the authors want to highlight three common misconceptions about clustering adjustments:

[1] `https://academic.oup.com/qje/article/138/1/1/6750017`

1. When to cluster depends on the presence of a nonzero correlation between residuals for units belonging to the same cluster

2. There is no harm in using clustering adjustments when they are not required

3. The researcher only has two choices: either fully adjust for clustering and use the cluster standard errors, or not adjust the standard errors at all and use the robust standard errors

3.4.1 Traditional + New, Design-Based Frameworks

Traditional econometrics frameworks that hope to shed light on the question of whether to cluster your standard errors or not are motivated either by (1) model-based assumptions or (2) the sampling mechanism. For the first framework, researchers need to make assumptions about the error component structure of a model for the outcome variable in order to know when to adjust their standard errors for clustering and thus rely on a model-based econometric framework. For example, one might assume a random effects model, with random effects at the state-level. In the repeated sampling thought experiment, state random effects will now be drawn from their distributions in each new sample, which in turn then requires us to cluster our standard errors.

For the second framework, we think about the sampling mechanism in two stages: first, clusters are selected at random from an infinite population and, second, units are sampled at random from the sampled clusters (or you can also keep all the units in a cluster). While this is appropriate for a lot of applications, the authors argue that this is typically not a strategy for how a lot of data sets in economics are generated because researchers will oftentimes observe the whole population of clusters, or at least a large fraction of it – for example, all of the US – and then this second framework does not apply.

Neither of these traditional frameworks incorporate the design aspect of clustering which makes them inappropriate for inference on treatment effects. The new, design-based framework the authors introduces leans on the framework based on sampling mechanism, but also accounts for variation in treatment assignment. Thus, it nests the traditional case of clustered sampling and the case of clustered assignment in experiments as special cases. One important novel result is that this framework allows for intermediate cases as well where treatment assignment may depend on the cluster, but not perfectly and there is still variation in treatment assignment within clusters. Moreover, the researcher will not have to take a stand on the error component structure of a model for the outcome variable to calculate the standard errors. All the relevant variability here of the estimator around its estimand, the average treatment effect, comes from the sampling mechanism – how was the sample extracted from the population – and the assignment mechanism – how was treatment assigned. Overall, there are three sources of sampling variation in this framework that can lead to variation in the estimates: (1) variation across samples in which units are observed in each cluster, (2) potentially variation in which clusters are observed and (3) variation in the treatment assignment across the units. How much these components matter for the variance of the estimators of the average treatment effect will depend on (i) the sampling process, (ii) the assignment process and (iii) the heterogeneity in the treatment effects across clusters.

Note. This has important implication for how we can use data to learn about whether clustering adjustments are necessary. In this design-based framework, the data will **not** be informative about the need to adjust for clustering in the sampling process, however it **will** be informative about the need to adjust for clustering in the assignment.

3.4.2 Notation

We formally introduce new notation for the asymptotics. Like last week, we are thinking about sequences of populations indexed by $k = 1, 2, 3, ...$

- For each population k, there are
  - g_k clusters
  - $n_{k,g}$ units in cluster g, so $n_k = \sum_{g=1}^{g_k} n_{k,g}$ units total in the population
- Sampling: We can split the sampling process into two parts, one parameter that controls the fration of clusters in the population that is sampled and one that controls the fraction of units in the population that is sampled. More precisely,
  - Cluster g is sampled with probability $q_k \in (0, 1]$
    - Random sampling: $q_k = 1$
    - Clustered sampling: $q_k < 1$
  - Units from the sampled clusters are sampled with probability p_k
- Assignment:
  - For each cluster, a (random) probability $A_{k,g}$ is drawn randomly from a distribution, $\sim (\mu_k, \sigma_k^2)$
    - Random assignment: $\sigma_k^2 = 0$
    - Clustered assignment: $\sigma_k^2 = \mu_k(1 - \mu_k)$
  - Units in cluster g are assigned to the treatment with probability $A_{k,g}$
- We still observe treatment assignment vector $W_{k,i} \in \{0, 1\}$ and outcome $Y_{k,i} = y_{k,i}(W_{k,i})$

If we go back to our table about the different possibilities of clustering, we can now put some mathematical notation into it. You can find the table in Figure 2.

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

Table 2: Possible Combinations for Clustering on Sample and Assignment Level in Terms of q_k and σ_k^2

Estimand. Throughout, we are assuming that we are interested in the population average effect

$$\tau = \frac{1}{n_k} \sum_{i=1}^{n_k} (y_i(T) - y_i(C)).$$

Estimator. We will also consider the difference in means estimator throughout as an appropriate estimator for the population average treatment effect

$$\hat{\tau}^{DiM} = \bar{Y}_T - \bar{Y}_C.$$

Note. Population quantities will be denoted by lower case letters, e.g. n_k, g_k, while sample quantities will be denoted by upper case letters, e.g. N_k, G_k.

There are a number of special cases that this framework nests:

- Random sampling of clusters from a large population of clusters $->$ q_k is small
- Random sampling from a large population $->$ $q_k = 1, p_k$ is small
  - Completely random assignment $->$ $A_{k,g} = A_k \ \forall g$ or $\sigma_k^2 = 0$
    - We discussed these cases in lectures 1 and 2
    - Here, we use robust standard errors: Neyman/EHW
  - Clustered random assignment $->$ $A_{k,g} \in \{0,1\} \ \forall k, g$ or $\sigma_k^2 = \mu_k(1-\mu_k)$
    - We discussed these cases in lectures 3 and 4
    - Here we use cluster-robust standard errors: Liang-Zeger

Note. We can make our previous statement about when the data is informative about clustering a bit more precise with our new notation. We **can** test whether $\sigma_k^2 = 0$ or $\sigma_k^2 > 0$, which will be informative about whether there has been clustered assignment. However, we **cannot** say anything about whether $q_k = 1$ or $q_k < 1$ (clustered sampling) from the data because we don't know from the data if we observed the entire population of clusters or not.

3.4.3 New Variance: *Causal Cluster Variance (CCV)*

In this section, we state the new proposed variance – the causal cluster variance – and show its relationship to conventional variances like the robust variance and the cluster-adjusted variance. Let g_{ki} denote which cluster a unit is from. We are focused on the case where $A_{k,g}$ has a distribution with a positive variance with support that is different from $\{0,1\}$. Before we go into the variance formula, we define the residuals

$$\varepsilon_{ki}(C) = y_{ki}(C) - \frac{1}{n_k}\sum_{h=1}^{n_k} y_{kj}(C) \qquad \varepsilon_{ki}(T) = y_{ki}(T) - \frac{1}{n_k}\sum_{h=1}^{n_k} y_{kj}(T).$$

The authors show that the difference in means estimator, centered around the population average effect and scaled by the new appropriate standard deviation, is converging to a standard normal distribution:

$$\sqrt{N_k}(\hat{\tau}_k - \tau_k)/v_k^{\frac{1}{2}} \to^d \mathcal{N}(0,1),$$

where

$$v_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right) - p_k\,\frac{1}{n_k}\sum_{i=1}^{n_k}(\varepsilon_{k,i}(1)-\varepsilon_{k,i}(0))^2 \qquad (1)$$

$$+\,p_k\,(1-q_k)\,\frac{1}{n_k}\sum_{g=1}^{g_k}\left(\sum_{i:G_{k,i}=g}(\varepsilon_{k,i}(1)-\varepsilon_{k,i}(0))\right)^2 \qquad (2)$$

$$-\,p_k\,\sigma_k^2\,\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right) \qquad (3.1)$$

$$+\,p_k\,\sigma_k^2\,\frac{1}{n_k}\sum_{g=1}^{g_k}\left(\sum_{i:G_{k,i}=g}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right)\right)^2 \qquad (3.2)$$

This is a really huge and annoying formula, so let's try to break it up into smaller parts and relate it back to terms we know. We will slowly build up the variance components by moving through the table of the different possible combinations for sampling and assignment.

Random sampling and random assignment. This random-random scenario corresponds to $q_k = 1$ and $\sigma_k^2 = 0$. Thus, the variance simplifies to

$$v_k(q_k=1,\sigma_k^2=0) = \underbrace{\frac{1}{N_k}\sum_{i=1}^{n_k}\left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k}\right)}_{=v_k^{EHW}\text{ : estimand for robust variance}} - \underbrace{p_k\,\frac{1}{N_k}\sum_{i=1}^{n_k}(\varepsilon_{k,i}(1)-\varepsilon_{k,i}(0))^2}_{\text{finite-sample correction}} = (1)\text{ in }v_k.$$

The first term is the estimand corresponding to the robust variance estimator, EHW, while the second term refers to the finite-sample correction. This finite-sample correction term vanishes if there is no heterogeneity in treatment effects, i.e. $\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0) = y_{k,i}(1) - y_{k,i}(0) - \tau_k = 0$ or if the sample is a small fraction of the population ($p_k \approx 0$).

Clustered sampling and random assignment. This clustered-random scenario corresponds to $q_k < 1$ and $\sigma_k^2 = 0$. Thus, the variance increases by

$$v_k(q_k<1,\sigma_k^2=0) = v_k(q_k=1,\sigma^2=0) + \underbrace{p_k\,(1-q_k)\,\frac{1}{n_k}\sum_{g=1}^{g_k}\left(\sum_{i:G_{k,i}=g}(\varepsilon_{k,i}(1)-\varepsilon_{k,i}(0))\right)^2}_{(2)\text{ in }v_k}.$$

Note that we can rewrite (2) in terms of treatment effects,

$$p_k\,(1-q_k)\,\frac{1}{n_l}\sum_{g=1}^{g_k} n_{k,g}^2\,(\tau_{k,g}-\tau_k)^2.$$

Thus, this term vanishes if there is no heterogeneity in the average treatment effects across clusters. While the sample is informative about the heterogeneity in treatment effects, it is **not** informative about the value of q_k. This is something that comes from the context and the researcher needs to decide.

Clustered sampling and (partially) clustered assignment. This clustered-clustered scenario corresponds to the case where $q_k < 1$ and $\sigma_k^2 > 0$ (but might not be $\sigma_k^2 = \mu_k(1-\mu_k)$ necessarily. Now, given that we want to take this dependence into account in the variance formula, the variance has two additional terms

$$\begin{aligned} v_k(q_k < 1, \sigma_k^2 > 0) = v_k(q_k < 1, \sigma_k^2 = 0) \\ \underbrace{- p_k \ \sigma_k^2 \ \frac{1}{n_k} \ \sum_{i=1}^{n_k} \ \left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k} \right)}_{(3.1) \text{ in } v_k} \\ \underbrace{+ p_k \ \sigma_k^2 \ \frac{1}{n_k} \sum_{g=1}^{g_k} \ \left(\sum_{i:G_{k,i}=g} \ \left(\frac{\varepsilon_{k,i}^2(1)}{\mu_k} + \frac{\varepsilon_{k,i}^2(0)}{1-\mu_k} \right) \right)^2}_{(3.2) \text{ in } v_k} \end{aligned}$$

The sign of this expression depends on the amount of variation in potential outcomes that can be explained by the clusters. Here, the sample can potentially be information about the need for clustering.

In general, the different variance components of v_k can be of different orders of magnitudes. Which term may dominate the variance depends on (i) the magnitude of p_k, (ii) presence of clustering in sampling, (iii) presence of clustering in assignment and (iv) heterogeneity in the potential outcomes.

3.4.4 Relationship to the Robust and the Cluster Standard Errors

We can compare the newly derived variance to the limit versions of the robust and the cluster robust variances. We start off by comparing it to the robust variance estimand, v_k^{EHW}. This corresponds approximately to the first part of v_k and the case of random sampling, random assignment when we ignore the finite sample correction:

$$v_k^{EHW} \approx v_k(q_k = 1, \sigma_k^2 = 0).$$

One can show that in general the difference between the estimands, $v_k^{EHW} - v_k$, can be positive or negative, thus the robust variance estimator can remain invalid even in large samples as it might underestimate the true variance.

Next, we want to compare the variance v_k to the estimand corresponding to the Liang-Zeger cluster robust variance estimator, $v_k^{cluster}$. The formula for that is given in the lecture slides. The important take-away is that the difference is

$$v_k^{cluster} - v_k = p_k \ q_k \ \frac{1}{n_k} \ \sum_{g=1}^{g_k} \ \left(\sum_{i:G_{ki}=g} \ (\varepsilon_{k,i}(1) - \varepsilon_{k,i}(0)) \right)^2.$$

Note that this difference is always non-negative, meaning the cluster-robust variance can be **conservative**, but cannot underestimate the true variance v_k. When the expected fraction of clusters in the sample, q_k, is small, or when the average treatment effect is almost constant between clusters, $v_k^{cluster} \approx v_k$. v_k also approaches $v_k^{cluster}$ as $\sigma_k^2 \to \mu_k(1-\mu_k)$ (clustered assignment). As a result,

		Assignment		
		Random	**Partially Clustered**	***Clustered***
Sampling	***Random***	$\frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{N} / v_k^{EHW}$	v_k	$v_k^{cluster}$
	Clustered	($\approx v_k^{cluster}$ if q_k small) v_k	v_k	v_k

Table 3: True variance for $\hat{\tau}^{DiM}$ (individual) for all possible combinations for clustering on sample and assignment level

this variance bridges the gap between the extreme cases of EHW and LZ.

Figure 3 summarizes the appropriate true variance for all different cases.
Note. If p_k is small enough, v_k^{EHW} and $v_k^{cluster}$ are approximately equal to v_k. This is because the clustering in sampling or assignment does not matter much since the probability that two sample units are from the same cluster is small.

3.4.5 Estimator

We are interested in finding an appropriate variance estimator for our missing scenarios. That is, when both sampling and assignment are clustered and, particularly valuable in observational settings, we are in the partially clustered case, where $0 < \sigma_k^2 < \mu_k(1-\mu_k)$. We have found in the variance derivation formula that neither the robust variance estimator nor the Liang-Zeger one will be appropriate: if σ_k^2 is close to 0, the variance estimator should be close to $\hat{V}^{EHW}$, if σ_k^2 is close to $\mu_k(1-\mu_k)$, then the variance estimator should be close to $\hat{V}^{cluster}$. The paper proposes two new variance estimators: (1) analytic result, the *causal cluster variance estimator*, and (2) a re-sampling based (bootstrap) variance estimator. Note that if q_k is close to zero, both proposed variance estimators will be close to $\hat{V}^{cluster}$. Moreover, the proposed re-sampling variance estimator is not defined when there is clustered assignment. To be effective, the estimators require a large number of both treated and control units per cluster!

The analytic estimator, $\hat{V}_k^{CCV}$, is defined as follows:

$$\hat{V}_k^{CCV} = \hat{q}_k \times \hat{V}_k^{CVV}(1) + (1-\hat{q}_k) \times \hat{V}_k^{cluster},$$

where

$$\hat{V}_k^{CCV}(1) = \frac{1}{N_k\ \bar{W}_k^2(1-\bar{W}_k)^2}\ \sum_{g=1}^{g_k}\ \Bigg[\frac{1}{\bar{Z}_k^2}\Bigg(\sum_{i:G_{k,i}=g} R_{k,i}Z_{k,i} \\ \times\ ((W_{k,i}-\bar{W}_k)\hat{U}^*_{k,i} - (\hat{\tau}^*_{k,m}-\hat{\tau}^*_k)\bar{W}_k\ (1-\bar{W}_k))\Bigg)^2 \\ -\ \frac{1-\bar{Z}_k}{\bar{Z}_k^2}\ \sum_{i:G_{k,i}=g} R_{k,i}Z_{k,i}\Bigg((W_{k,i}-\bar{W}_k)\hat{U}^*_{k,i} \\ -\ (\hat{\tau}^*_{k,m}-\hat{\tau}_k\bar{W}_k(1-\bar{W}_k))\Bigg)^2\Bigg] \\ +\ (1-p_k)\ \sum_{g=1}^{g_k}\ \frac{\bar{N}_{k,m}}{N_k}\ (\hat{\tau}_{k,m}-\hat{\tau}_k)^2,$$

where $\bar{N}_{k,m}$ refers to the size of the sample in cluster m. We split the data into two subsamples and $Z_{k,i} \in \{0,1\}$ denotes the indicator whether unit i belongs in the second subsample. $\bar{Z}_k$ then denotes the mean of $Z_{k,i}$. Using the subsample with $Z_{k,i}=0$, we get estimates $\hat{\tau}^*_{k,m}$, $\hat{\alpha}^*_k$ and $\hat{\tau}^*_k$ for $\tau_{k,m}, \alpha_k, \tau_k$ respectively. For the observations with $Z_{k,i}=1$, we then calculate the fitted residuals: $\hat{U}^*_{k,i} = Y_{k,i} - \hat{\alpha}^*_k - \hat{\tau}^*_k\ W_{k,i}$. Note that a simple split here is enough to get a consistent estimator. For a more efficient one, we can use sample splitting. $\hat{V}_k^{CCV}(1)$ here refers to the variance estimator, where $q_k = 1$, so all clusters are sampled. Estimating $\hat{q}_k$ requires the knowledge of the total number of clusters in the population.

The second way to obtain a valid variance estimator is through bootstrapping. More precisely, the variance estimator is a two-stage-cluster-bootstrap (**TSCB**). It consists of two resampling stages, plus a few additional steps. You can find the algorithm in the lecture slides (slide 22).

Table 4 provides an overview for all the different variance estimators we have introduced in the last two weeks and when each one is appropriately used.

		Assignment		
		Random	**Partially Clustered**	***Clustered***
Sampling	***Random***	Neyman/Robust $-> \hat{V}_k^{EHW}$	CCV or TSCB $-> \hat{V}_k^{CCV}(1)$	Liang-Zeger $-> \hat{V}_k^{cluster} = \hat{V}_k^{LZ}$
	Clustered	($\approx$ Liang-Zeger $-> \hat{V}_k^{cluster}$) CCV or TSCB $-> \hat{V}_k^{CCV}$	CCV or TSCB $-> \hat{V}_k^{CCV}$	CCV (TSCB **not** available) $-> \hat{V}_k^{CCV}$

Table 4: Variance estimators for $\hat{\tau}^{DiM}$ (individual) for all possible combinations for clustering on sample and assignment level

4 Practice Problems

4.1 Clustered Randomized Experiments I (Practice Final Exam 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?

Solution Sketch:
Note that in the first lecture we just took the sample at face-value. We assumed the potential outcomes were fixed and so we can answer questions (a)-(c) first part with everything we did last week.
Moreover, there is no single correct solution to many of these questions, but some solutions are more insightful or lead to more efficient or robust analysis. The solution here only gives some suggested points your solution could/should address.

(a) In principle, we could look at the difference in means estimator and estimate its variance to calculate a t-statistic and compare it to a normal distribution. However, in a randomized experiment and with a sharp null (oftentimes we interpret "*no effect whatsover*" as a sharp null), we can use the Fisher's exact p-values in this scenario. To calculate the p-values, we need to make two choices: (1) sharp null hypothesis and (2) test statistic. According to the prompt, we are asked to test for no effect of the program whatsoever, thus, we define our null hypothesis as follows:

$$H_0 : \ Y_i(1) = Y_i(0) \ \forall i$$

We choose the difference-in-means test statistic, $T = \frac{\sum_i W_i Y_i}{\sum_i W_i} - \frac{\sum_i (1-W_i) Y_i}{\sum_i (1-W_i)}$. Next, we obtain the randomization distribution of T for all possible treatment assignments (or a subset thereof depending on how many individuals and treated units there are). Lastly, we obtain the p-value by counting how many of the test statistics are as extreme as the observed one (in absolute terms).
If you read the question as describing a stratified randomized experiment ($\mathcal{S}_i$ being the strata here), you could also take that into account in the Fisher test, e.g. calculating your statistic within county (stratum) and taking some (potentially weighted) average of the test statistics.
In principle, you know/can try to derive either the asymptotic (in sample size) distribution or its finite sample distribution based on randomization as the only source of randomness of the test statistic. In practice, I'd recommend doing randomization inference and estimating the exact randomization distribution of the test statistic by drawing random samples from the randomization distribution and calculating the test statistic for each of those, just like in problem set 1. Make sure you know how to describe randomization inference in words – if you have enough time on the exam, calling it randomization inference or permutation inference **and** describing it is better than just mentioning it.

(b)

- Difference-in-means estimator: $\hat{\tau}^{DiM} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$.
- Stratified randomized experiment if you read the question that way

(c) *Without clustering:*

- Use appropriate formulas for the estimators mentioned above (Neyman variance, Stratified variance, ...)

With clustering:

- Clustered standard errors: Liang-Zeger variance (see slide 25 of slide deck 3 for formula)

(d) The answer is that it depends. We don't have clustered assignment, but clustered sampling. Now you as the researcher have to make a choice. If you are interested in the average effect for these specific 20 counties and your repeated sampling thought experiment is to simply re-assign treatment, do not cluster. However, if you are interested in the average effect for all of California, you should use clustered standard errors.

4.2 Clustered Randomized Experiments II (Final Exam 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?

Solution Sketch:
Note that there is no single correct solution to many of these questions, but some solutions are more insightful or lead to more efficient or robust analysis. The solution here only gives some suggested points your solution could/should address.

(a) Note that the assignment is clustered here because it only varies at the state level. Depending on which estimand we are interested in (entire population or cluster averages), we could run a simple difference-in-means ($\hat{\tau}_{pop}$) as in the first week or we could average the outcomes at the cluster level and take the difference of treated and control cluster averages ($\hat{\tau}_{cluster}$).

(b) Given that treatment is assigned at the state level, we would want to use the clustered variance here.

- For $\hat{\tau}_{pop}$: we can use the Liang-Zeger variance estimator (see slide 25 of slide deck 3 for formula).
- For $\hat{\tau}_{cluster}$: we can use the Neyman variance, where we treat the cluster as the unit of analysis (see notes for formula).

4.3 Clustered Randomized Experiments III (Final Exam 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?

Solution Sketch:
Note that there is no single correct solution to many of these questions, but some solutions are more insightful or lead to more efficient or robust analysis. The solution here only gives some suggested points your solution could/should address.

(a)

- If you interpret the "difference in means" as the difference in individual means, we are effectively estimating $\hat{\tau}_{pop}$. From the lectures, we know that we can derive an exact variance formula for v_k (which is very long and complicated, but it is stated on slide 24 of slide deck 3. We can estimate this variance by the Liang-Zeger variance, which is given on slide 25 of slide deck 3.

- If you interpret the "difference in means" as the difference in cluster means, we are estimating $\hat{\tau}_{cluster}$. Here, the true variance is given by

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

  with $S_w^2 = \frac{1}{G-1}\sum_g (\bar{Y}_g(w) - \overline{\bar{Y}_g(w)})^2$. We drop the last term as it is unobserved in practice and use the conservative variance estimator

$$\widehat{V(\hat{\tau}_{cluster})} = \frac{s_0^2}{G_0} + \frac{s_1^2}{G_1},$$

  with $s_w^2 = \frac{1}{G_w - 1}\sum_{g:\bar{W}_g = w} (\bar{Y}_g - \overline{\bar{Y}_{gw}})^2$.

(b) If the outcomes are positively correlated within each state, the clustered standard error will be larger and we will thus understate the true variance if we use the Neyman variance estimator that ignores the state-level randomization.