Design-based efficiency for analyzing cluster-randomized experiments

by

Yeng Xiong

B.S., University of Tennessee at Martin, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Abstract

Cluster randomized experiments (CREs) have three defining features: (i) treatments are randomized to clusters, or groups of units, rather than units themselves, (ii) clusters are formed *a priori* to experimentation and without researcher intervention, and (iii) the research objective and analysis is still centered on units. CREs are common, particularly for intervention studies in public health and political science. Yet, despite their growing popularity, there is still ongoing debate, even among the experts, on their analysis and design methodologies. We focus on design-based estimators for measuring both the population average treatment effect (PATE) and the standard error (SE) under the Neyman-Rubin potential outcomes framework.

The inherent disparity between the experimental and observational units in CREs can lead to some analytical and design challenges—for example, bias, large variability, and/or lack of location invariance. Moreover, randomizing treatments to clusters is known to be less efficient than randomizing to individual units. Conventionally, clusters in CREs are sampled using simple random sampling. Stratifying or pair-matching clusters based on important covariates can improve precision on estimation.

We instead propose a different sampling scheme: sampling with probability proportional to size without replacement. This modification leads to a Horvitz-Thompson estimator (HT-PPS) of PATE that can accommodate the clustering structure in CREs without having to compromise on desirable statistical properties. We then derive a conservative estimator for the variance of our HT-PPS estimator. We also synthesize the myriad perspectives on designing CREs and produce recommendations on the best design practices. Finally, we introduce our R package *analyzeCRE* that implements the theoretical work in this dissertation and provide a guide on how to execute the functions for analyzing and designing CREs.

Design-based efficiency for analyzing cluster-randomized experiments

by

Yeng Xiong

B.S., University of Tennessee at Martin, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Approved by:

Major Professor
Michael Higgins

Copyright

Abstract

Cluster randomized experiments (CREs) have three defining features: (i) treatments are randomized to clusters, or groups of units, rather than units themselves, (ii) clusters are formed *a priori* to experimentation and without researcher intervention, and (iii) the research objective and analysis is still centered on units. CREs are common, particularly for intervention studies in public health and political science. Yet, despite their growing popularity, there is still ongoing debate, even among the experts, on their analysis and design methodologies. We focus on design-based estimators for measuring both the population average treatment effect (PATE) and the standard error (SE) under the Neyman-Rubin potential outcomes framework.

The inherent disparity between the experimental and observational units in CREs can lead to some analytical and design challenges—for example, bias, large variability, and/or lack of location invariance. Moreover, randomizing treatments to clusters is known to be less efficient than randomizing to individual units. Conventionally, clusters in CREs are sampled using simple random sampling. Stratifying or pair-matching clusters based on important covariates can improve precision on estimation.

We instead propose a different sampling scheme: sampling with probability proportional to size without replacement. This modification leads to a Horvitz-Thompson estimator (HT-PPS) of PATE that can accommodate the clustering structure in CREs without having to compromise on desirable statistical properties. We then derive a conservative estimator for the variance of our HT-PPS estimator. We also synthesize the myriad perspectives on designing CREs and produce recommendations on the best design practices. Finally, we introduce our R package *analyzeCRE* that implements the theoretical work in this dissertation and provide a guide on how to execute the functions for analyzing and designing CREs.

Table of Contents

List of Figures

List of Tables

List of Notations

This is a list of common notations used throughout the dissertation. The / separates the notations for the non-stratified and stratified/match-paired methods. Capitalization indicates random variables.

y_{kct}/y_{kcut} Potential outcome for unit k in cluster c under treatment t / in stratum u

t Treatment: $t = 1$ for treated, $t = 0$ for control

δ/δ_u Population average treatment effect (PATE) / for stratum u

n/n_u Number of units in population / in stratum u

m Number of strata/pairs

ℓ/ℓ_u Number of clusters in the population / in stratum u

n_c/n_{cu} Number of units in cluster c / in stratum u

S_u Indicator variable for sampling pair u; only used in match-paired sampling

$\#m$ Number of pairs sampled; only used in match-paired sampling

S_c/S_{cu} Indicator variable for sampling cluster / in stratum u

s/s_u Number of sampled clusters / in stratum u

T_{ct}/T_{cut} Indicator variable for treating cluster / in stratum u

$\#T_t/\#T_{ut}$ Number of sampled clusters given treatment t / in stratum u

S_{kc}/S_{kcu} Indicator variable for sampling unit k in cluster c / in stratum u

s_c/s_{cu} Number of sampled units in cluster c if c is sampled / in stratum u

$\#N_t/\#N_{ut}$ Number of units given treatment t / in stratum u

$\pi_{uu'}$ Probability pairs u and u' are sampled together

$\pi_{cc'}$ Probability clusters c and c' are sampled together

π_{kct}/π_{kcut} Probability unit (k, c) is observed under treatment t / in stratum u

Acknowledgments

Completing this dissertation has been one of the most challenging endeavor I have taken, and it would not have been possible without my community of professors, friends, and family. While there are many, I would like to mention a distinguished few here to express my gratitude.

First and foremost, my major professor, Dr. Mike Higgins, for being a *most* patient adviser and, most importantly, encouraging friend. I have learned so much under your guidance, and this dissertation is a significant, but certainly not full, testament of it.

My committee members: Dr. Christopher Vahl, Dr. Gyuhyeong Goh, Dr. Alissandra Stoyan, and Dr. Majid Jaberi-Douraki. Thank you for the kind advice and helpful suggestions that further strengthened my work and my knowledge.

The K-State Statistics Department. For the past seven years, you continuously challenge me to be a better version of myself. I owe my current and future success to you.

My friends Behnaz Moradi-jamei, Heman Shakeri, and Jiena Gu. You constantly inspire me to work harder and to strive for more.

My family. During such a difficult and divisive time, you still find the strength to support me and relieve me of stress and responsibility. I hope I have made you proud.

Finally, thank you Adam Burkhart, for always loving me unconditionally. Even when I did, you never gave up on me. I could not have gotten here without your love and sacrifice.

Dedication

To

my mom, Mai True Thao, for inspiring me,

my dad, Sheng Xiong, for encouraging me,

my younger siblings, Kia, Xee, Sue, Yer, Rosey, and Noah

for letting me be first in more ways than one.

Chapter 1

Introduction and Overview

1.1 Research thesis

Cluster-randomized experiments (CREs), in which groups of individuals are randomized to treatments, are becoming a common choice, particularly for intervention studies in areas such as public health and political science. Yet, despite their growing popularity, there is still ongoing debate on how to analyze and design them. Even among experts in the field, there is no consensus on the best approaches; the advice of one expert often runs counter to another. Worse still, the different techniques can lead to conflicting inferential interpretations. The purpose of this dissertation then is to investigate the current methodologies of CRE analysis and design and to propose a new, competitive design-based estimator for the population average treatment effect (PATE).

Randomized experiments are essential to scientific progression because they have the ability to unbiasedly recover the average treatment effect (ATE) free from making distributional assumptions. Staying true to this tradition, we work under Neyman's potential outcomes framework for causal inference (Splawa-Neyman et al., 1923). This framework bases inference solely on the randomization pattern developed naturally from sampling and treatment assignment so it is considered design-based. Conventionally, clusters in CREs are sampled using simple random sample without replacement (SRSWOR), and treatments

are assigned completely at random to clusters. Units within clusters may also be sampled using SRSWOR, and so estimators of PATE needs to account for all the different stages of randomization. It should also be noted that the experimental units are not necessarily the observational units. This can cause estimators to have undesirable properties such as bias, large variability, or not being location-invariant. Taking advantage that bigger clusters yield larger cluster totals, we suggest that clusters be sampled without replacement but with probability proportional to (cluster) size (PPSWOR). This modification leads to a Horvitz-Thompson estimator (HT-PPS) that is unbiased and location-invariant and has comparable variability to its SRSWOR counterparts. We also derive two conservative estimates of the variance for our HT-PPS estimator.

Randomizing treatments to clusters is known to have greater estimation variability than randomizing to individual units. Stratifying on or pairing up clusters based on important covariates can improve estimation precision, but it is unclear on which is better. Martin et al. (1993) cautions against using matched pairs, advising that it should only be used when there are more than 10 pairs of clusters, whereas Imai et al. (2009) vehemently advocate matching, declaring that it should always be used when possible. Matched pairs generally provide greater efficiency than stratification, but those against it are concerned about its drawbacks. Among them are the possible loss of degrees of freedom and limitation in estimating both the within-pair treatment variability and the intracluster correlation (ICC), a numeric summary of the clustering effect. The ICC plays a vital part in efficiency assessment, and so, is important in power calculations and sample size determination. Clearly, there is a need to synthesis the myriad perspectives on designing CREs and produce a guideline on the best practices.

In summary, our intent is to survey the current strategies in analyzing and designing CREs. From this, we will offer a new estimator of the PATE and establish its statistical properties, including a conservative estimation of its variance. We will also give recommendations on designing CREs to improve efficiency. The rest of this chapter will review some critical concepts that will aid in this purpose. The concepts are on CREs, causal inference under potential outcomes, and statistical properties of estimators.

1.2 Cluster-randomized experiments

Cluster-randomized experiments have three defining features: (i) treatments are randomized to clusters, or groups of units, rather than units themselves, (ii) known clusters are formed *a priori* to experimentation and without researcher intervention, and (iii) the research objective and analysis are still centered on units. Choices of units and clusters may include people in villages, students in classrooms, and family members in households. Due to the disparity between the treatment and analysis entities in CREs, they are known to be less precise than other experimental designs. Nonetheless, they are gaining prevalence, particularly among political studies [Gerber and Green (2000), Gerber and Green (2001), Green et al. (2003), Wantchekon (2003), Imai (2005), Gerber and Green (2005), King et al. (2007), Hansen and Bowers (2009), Beath et al. (2013), Avdeenko and Gilligan (2015)], public health [Donner (1998), Donner and Klar (2004), Kerry et al. (2005), Hayes and Moulton (2009)), psychology (Bruce et al. (2004), Small et al. (2008), Paluck (2009), Raver et al. (2009), Cilliers et al. (2016)], education [Springer et al. (2010), Carlson et al. (2017)], economic studies [Manning et al. (1987), Hidrobo et al. (2014)] and medical trials [Kramer et al. (2008), Grandes et al. (2009)]. This section sought to explain why CREs are an attractive alternative to other designs despite their deficiency and to clarify the possible estimands of interest.

1.2.1 Reasons for clustering

Common experimental schemes, such as completely randomized and blocked designs, randomly allocate treatments to individuals, but for certain fields, this may not be easy or even feasible to implement. For example, suppose we want to compare the efficacy of different teaching methods. Randomization to individuals requires that each student is randomly assigned to a teaching method. A more intuitive design choice would be to apply the treatments at the class level instead. Thus, CREs are often used for logistical efficiency.

CREs are also appealing as they deter interference (also known as contamination or treatment spillover) between units. Unit interference occurs when the altercation with a treated unit influences a control unit's response. Estimation of treatment effect will then

be biased towards a smaller size. Since units in a cluster are given the same treatment in CREs, unit interference is not a concern. On the other hand, cluster interference—units in a control cluster may be exposed to the treatment through relationships with units in a treated cluster—is a problematic possibility. Consider the teaching example: if a student in the treated classroom tutors a student from the control classroom, there is interference. Choosing clusters that are geographically far apart can abate this complication.

On the contrary, allowing unit interference in CREs is a great setting for measuring indirect treatment effects (Hayes and Moulton, 2009; Hitchings et al., 2018). To do so, the treated clusters would be further randomly divided in which some of the units will receive the treatment and some will not. Since these units are in the same cluster, interference is likely to happen. The difference between the untreated units in the treated clusters and the control units measures the indirect treatment effect. Areas like public health are interested in this quantity because the units exposed but not given treatment can still be beneficial in preventing the spread of diseases. Our focus, though, as are most researchers, is on the direct treatment effect measured through the average treatment effect.

1.2.2 Estimands

Average treatment effects can generically be expressed as average differences between the outcomes under treatment minus those under control:

$$\delta = \mathbb{E}(Y(1) - Y(0)) \tag{1.1}$$

where $Y(1)$ and $Y(0)$ are the outcomes for treated and control, respectively. Since CREs may include up to two stages of sampling—sampling clusters and sampling units within clusters, there are four causal quantities of interest: sample average treatment effect (SATE), cluster average treatment effect (CATE), unit average treatment effect (UATE), and population average treatment effect (PATE).

The defining differences between these four estimands are whether or not the clusters

Quantities	Clusters	Units within clusters	Inferential target
SATE	Observed	Observed	Observed sample
CATE	Observed	Sampled	Population within observed clusters
UATE	Sampled	Observed	Observed units within the population of clusters
PATE	Sampled	Sampled	Population

Table 1.1: CREs have four possible estimands, which determines the body of units to whom the results can be generalized. This table is a replica from Imai et al. (2009).

are randomly sampled from a population of clusters and if the units within the clusters are randomly sampled. In other words, these estimands depend on the units over which the expectation in eq. (1.1) is taken. For SATE, it is over the observed sample of clusters and units in those clusters. CATE looks at a random sample of units from the observed clusters whereas UATE is for the observed units within sampled clusters. The last estimand, PATE, is over the population of units as the clusters and units within clusters are both sampled, and hence, the most generalizable estimand. The external validity of the result highly depends on these estimands, and thus, stating them is critical to analyzing CREs. This discussion is outlined in Table 1.1. In our setup, we assume that researchers are able to sample from a population of clusters and sample units within the chosen clusters so PATE is the appropriate estimand. We will use the Neyman potential outcomes framework, elaborated in the next section, to analyze and estimate PATE.

1.3 Causal inference under potential outcomes

For research studies that center on questions of a causal nature, Fisher demonstrated that conducting a statistical experiment is undoubtedly the simplest and best approach. In his pivotal treatise on experimental design, he proclaimed that the treatment randomization present in experiments is the foundational premise for causal inference (Box, 1980; Fisher,

1926; Rubin, 2005). Thus, experiments are often hailed as the "gold standard." Like Fisher, Neyman also recognized the important relationship between randomizing treatments and causal inference. From it, he developed the potential outcomes framework and the Neyman-Rubin Causal Model (NRCM). They serve as a perceptive paradigm to study causality and the role that randomized experiments play in it. In this section, we start with a description of the potential outcomes framework and the NRCM. We will then state the necessary assumptions required for causal inference and explain how experiments can help achieve those assumptions.

1.3.1 Neyman-Rubin causal model

Suppose there are two treatment conditions: treatment and control, denoted $t = 1$ and $t = 0$, respectively. For a finite population of n units, define a treatment indicator, T_i, such that $T_i = 1$ if unit i received the treatment and $T_i = 0$ if unit i received the control. Let y_{i1} be the potential outcome of the ith unit exposed to treatment and y_{i0} be the potential outcome of the *same* unit exposed to control. Neyman's structure of potential outcomes is acutely intuitive for studying causal effects because:

> The values of post-exposure variables are potentially affected by the particular cause, t or c, to which the unit is exposed. This is nothing less than the statement that causes have effects, which is the very heart of the notion of causation. For the model to represent faithfully this state of affairs we need not a single variable, Y, to represent a response, but two variables, Y_t and Y_c, to represent two potential outcomes (Holland, 1986).

Hence, in this framework, the effect of the treatment on unit i is characterized as

$$\delta_i = y_{i1} - y_{i0} \tag{1.2}$$

and the average unit treatment effect (ATE) as

$$\delta = \sum_{i=1}^{n} \frac{\delta_i}{n} = \sum_{i=1}^{n} \frac{y_{i1} - y_{i0}}{n}. \tag{1.3}$$

Unfortunately, as a unit can only be treated or control, only one potential outcome is ever truly observed, a fact famously formalized as the *fundamental problem of causal inference* (Holland, 1986). The unobserved potential outcome for a unit is sometimes referred to as a *counterfactual*. The observed response of the ith unit can then be modeled by the Neyman-Rubin Causal Model (NRCM)

$$Y_i = y_{i1}T_i + y_{i0}(1 - T_i), \tag{1.4}$$

where the potential outcomes y_{i1} and y_{i0} are nonrandom. The response Y_i is y_{i1} if unit i received the treatment and y_{i0} if given control. The model inherently recognizes that randomness in the responses is a reflection of the design structure. There is no need for any distributional assumption.

To bypass the missing outcome problem, we shift the attention from unit treatment effect to the ATE:

$$\delta = \mathbb{E}(y_{i1} - y_{i0}) \tag{1.5}$$

$$= \mathbb{E}(y_{i1}) - \mathbb{E}(y_{i0}). \tag{1.6}$$

First, note that eq. (1.5) is equivalent to the algebraic form in eq. (1.3). Second, though the change in the two expectations seem negligible, the second indicates that the observed responses for *different* units can be used to find the ATE. Table 1.2 provides a summary of this discussion for $n = 4$. To determine the ATE for the four units:

$$\delta = \frac{1}{4}\left[(y_{11} - y_{10}) + \cdots + (y_{41} - y_{40})\right] \tag{1.7}$$

$$= \frac{1}{4}\left[(y_{11} + \cdots + y_{41}) - (y_{10} + \cdots + y_{40})\right]. \tag{1.8}$$

From the observed side of the table, we can see that this cannot be calculated due to the

	Potential			Observed		
Unit	Trt	Cont	δ_i	T_i	Y_i	δ_i
1	y_{11}	y_{10}	$y_{11} - y_{10}$	1	y_{11}	$y_{11}-?$
2	y_{21}	y_{20}	$y_{21} - y_{20}$	1	y_{21}	$y_{21}-?$
3	y_{31}	y_{30}	$y_{31} - y_{30}$	0	y_{30}	$? - y_{30}$
4	y_{41}	y_{40}	$y_{41} - y_{40}$	0	y_{40}	$? - y_{40}$
ATE		$\delta =$	$\frac{1}{4}[(y_{11} - y_{10}) + \cdots + (y_{41} - y_{40})]$		$\hat{\delta} =$	$\frac{1}{2}(y_{11} + y_{21}) - \frac{1}{2}(y_{30} + y_{40})$

Table 1.2: A unit's treatment effect, δ_i, is the difference between a unit's potential outcomes under treatment and control. Since a unit can only receive one treatment condition, δ_i is not estimable, but the average treatment effect, δ, is.

missing potential outcomes, but eq. (1.6) and (1.8) provides insight on how to estimate it:

$$\hat{\delta} = \frac{1}{2}(y_{11} + y_{21}) - \frac{1}{2}(y_{30} + y_{40}). \tag{1.9}$$

1.3.2 Assumptions

Causal inference is unattainable without assumptions, but it is crucial that they be clearly stated and justified (Rubin, 2005). In lieu of this, we proffer the assumptions needed for the NRCM and estimation of the ATE. Chief among them is the stable unit treatment value affect (SUTVA).

Assumption 1.3.1 (SUTVA)**.** *SUTVA stipulates that (i) each treatment level has only one form, and (ii) a unit's response is only affected by its own treatment allocation.*

The first part implies that regardless of how a unit is given treatment t, the response will be y_{it} and likewise for control. In other words, the study design does not influence the potential outcomes or the causal effect. The second specifies that there is no interference between units. This is clearly apparent in model 1.4 because a unit's response is a function of only its treatment indicator and not of any other unit's.

No interference is a strong condition that may be difficult to meet, even in an experimental

setting. Randomization alone cannot guarantee it since individuals undergoing different treatments may interact. Relaxing this assumption is a new developing research area [see Hudgens and Halloran (2008), Eckles et al. (2016), Aronow and Samii (2017), Jagadeesan et al. (2017), and Sussman and Airoldi (2017)]. As mentioned in Section 1.2.1, conducting a CRE, in which a group of units that are close together (and so susceptible to interference) is given the same treatment, is another viable solution.

In addition, the following two assumptions are needed:

Assumption 1.3.2 (Ignorability)**.** *For all covariates* $\mathbf{X}$*,* $P(T_i|y_{i1}, y_{i0}, \mathbf{X}) = P(T_i|Y_i, \mathbf{X})$*. That is, treatment assignment is independent of potential outcomes.*

Assumption 1.3.3 (Common support)**.** *For all covariates* $\mathbf{X}$*,* $0 < P(T_i|Y_i, \mathbf{X}) < 1$*. Specifically, each unit has some probability of getting treatment or control.*

Together, ignorability and common support constitute the property of strong ignorability or unconfoundedness. Ultimately, they ensure the formation of comparable treatment groups and unbiased estimation of the ATE. To illustrate:

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(y_{i1}|T_i = 1, \mathbf{X}) - \mathbb{E}(y_{i0}|T_i = 0, \mathbf{X})) &= \mathbb{E}(\mathbb{E}(y_{i1}|T_i = 1, \mathbf{X})) - \mathbb{E}(\mathbb{E}(y_{i0}|T_i = 0, \mathbf{X})) \\
&= \mathbb{E}(\mathbb{E}(y_{i1}|\mathbf{X})) - \mathbb{E}(\mathbb{E}(y_{i0}|\mathbf{X})) \\
&= \mathbb{E}(y_{i1}) - \mathbb{E}(y_{i0}) \\
&= \delta.
\end{aligned}$$

The first equality holds from common support, and the second from ignorability.

Randomized experiments guarantee both ignorability and common support since random allocation of treatments is done without knowledge of any information on the units, including the potential responses. Moreover, it assures all covariates, observed and unobserved, are balanced between the two treatment groups so common support is achieved. This does not necessarily mean that the groups are equal. They can differ, but the key is that this difference is due to chance, and statistics is well-equipped to quantify the uncertainty inherently present.

Lastly, we suppose that treatment randomization is assignment symmetric (Miratrix et al., 2013).

Assumption 1.3.4 (Assignment Symmetry)**.** *Let s be the sample size and $\#T_t$ be the number of treated entities. Conditional on $\#T_t$, all $\binom{s}{\#T_t}$ treatment assignments are equally likely.*

Common experimental designs such as complete randomization satisfies assignment symmetry. Block designs does too if randomization is independent across blocks. However, randomization that results in unequal treatment probabilities for units is not, generally, assignment symmetric. Since we suppose a complete randomization of treatments to clusters in CREs, this assumption is met for clusters, though, not for units, i.e., all combinations of treated clusters are equally likely but not for all combinations of treated units.

1.3.3 Estimation error in experiments

Error is an intrinsic truth in estimating the ATE, and so understanding what can lead to estimation error is an important task to undertake. The estimation error, ϵ, can be decomposed into four components:

$$\epsilon = \delta - \hat{\delta} = \epsilon_{S_x} + \epsilon_{S_u} + \epsilon_{T_x} + \epsilon_{T_u}, \tag{1.10}$$

where ϵ_{S_x} and ϵ_{S_u} are the errors from the sampling selection due to observed and unobserved covariates, respectively, and ϵ_{T_x} and ϵ_{T_u} are the errors from treatment imbalance, also, due to observed and unobserved covariates (Imai et al., 2007). Observed covariates are easily measurable (e.g., income, education level, severity of illness) whereas unobserved covariates are not as naked (e.g., job satisfaction, motivation, mental health). The sample selection error refers to these differences between the population and sample, and treatment imbalance error refers to these differences between the treated and control groups in the sample. Since experiments are able to satisfy the strong ignorability assumption, both ϵ_{T_x} and ϵ_{T_u} are, on average, negated. This is the strength and beauty behind experiments.

Eliminating, on average, ϵ_{S_x} and ϵ_{S_u} allows researchers to extend the result from the

sample to the general population. One way to achieve this is to use a probability-based sampling method to select individuals from the population. Traditional experimental analysis assumes that the sample is randomly selected from an infinitely large population. Under this assumption, units can be treated as mutually independent. However, since we are working with a finite population, sampling one unit will directly affect another unit. Therefore, our research consists of a combination of methodologies from sampling and experimental design pedagogies.

1.4 Desirable statistical properties of estimators

In choosing the best estimator, we will judge based on the statistical properties bias, variance, and location-invariance. We will now define these terms. Suppose there is an estimator $\hat{p}$ of parameter p.

Definition 1 (Bias)**.** *The bias of an estimator $\hat{p}$ is*

$$bias(\hat{p}) = \mathbb{E}(\hat{p}) - p. \tag{1.11}$$

If $bias(\hat{p}) = 0$, we say, the estimator $\hat{p}$ is unbiased for p.

Definition 2 (Variance)**.** *The variance of an estimator $\hat{p}$ is*

$$var(\hat{p}) = E\left([\hat{p} - E(\hat{p})]^2\right). \tag{1.12}$$

Taking a square root of $var(\hat{p})$ gives the standard error for $\hat{p}$.

Definition 3 (Mean squared error)**.** *The mean squared error of an estimator $\hat{p}$ is*

$$MSE(\hat{p}) = E[(\hat{p} - p)^2] = var(\hat{p}) + bias(\hat{p})^2. \tag{1.13}$$

Unbiasedness ensures that the estimator is measuring what it purports to, namely, the parameter, and the variance determines the precision of the estimator. The mean squared

error, accounting for both bias and variability, is often used as a comparison measurement.

Definition 4 (Location transformation)**.** *A location transformation of outcomes, y_i, occurs when*

$$y_i^* = a + by_i, \quad \textit{for all } \ i = 1, \ldots, n, \tag{1.14}$$

where a is a constant and $b = 1$.

Definition 5 (Location-invariant)**.** *An estimator, $\hat{p}$, is location-invariant if it does not change despite a location shift in the potential outcomes:*

$$\hat{p}(a + y_i) = \hat{p}(y_i), \quad \textit{for all } \ i = 1, \ldots, n. \tag{1.15}$$

Intuitively, any conversion involving a linear shift should not affect the PATE estimation. Location-invariance will guarantee this.

1.5 Organization of dissertation

We have thus far surveyed important assumptions and theories that are constructive to our goal of estimating PATE for CREs. Chapter 2 presents our proposed estimator, Horvitz-Thompson under probability-proportional-to-size sampling and investigates its statistical properties. We also examine variance estimation, providing two different conservative variance estimators. Then we extend our estimator to the case in which clusters may also be stratified at either the cluster-level, unit-level, or both. Using simulated data from the National Solidarity Programme (NSP) (Beath et al., 2013), we compare the HT-PPS estimator to its simple random sampling counterparts, validating it as a worthy contender for estimating PATE in CREs. Chapter 3 presents a discussion on which CRE estimation method is the best at efficiently using covariate information, either in the analysis or the design, for estimating both PATE and the standard error. From this and simulation based on the NSP data, we give recommendations on the best practices of designing and analyzing CREs.

Chapter 4 introduces our R package *analyzeCRE* and gives a guide on how to use it. The dissertation ends with concluding remarks and future work in Chapter 5.

Chapter 2

The Benefits of Probability-Proportional-to-Size Sampling in Cluster-Randomized Experiments

2.1 Introduction

Frequently in experiments, treatment is randomized across clusters, or groups, of units of interest instead of the units themselves. These are referred to as *cluster-randomized experiments* (CREs). Clusters of units are often formed *a priori* to the design of the experiment and without researcher intervention. Estimation of treatment effects is more precise when treatment is randomized across units (Cornfield, 1978); hence, logistical issues (rather than increased precision of treatment effect estimates) motivate the randomization of treatment across clusters. Reasons for such randomization include addressing issues with the ethicality, legality, or feasibility of randomizing treatment across units, reducing risk of treatment contamination, and mimicking the implementation of a proposed program (e.g. an educational

intervention) (Donner, 1998; Donner and Klar, 2004; Hayes and Moulton, 2009). Common settings for cluster-randomized experiments include: testing an educational intervention that is implemented within classrooms (Raver et al., 2009); evaluating efficacy of a health intervention that is implemented within clinics or medical practices (Bruce et al., 2004; Imai et al., 2009; King et al., 2007; Small et al., 2008); measuring increases in compliance and turnout from mailers sent to households (Gerber and Green, 2000); and identifying effects of interventions implemented within villages or other geographic regions (Beath et al., 2013; Paluck, 2009; Wantchekon, 2003).

To estimate and perform inference on the *population average treatment effect* (PATE), a CRE will require at least two stages of sampling: sampling clusters from a larger population of clusters (e.g. a sample of villages within a country) and sampling individual units from each of the sampled clusters—samples may be comprised of the entire sampling frame. After a sample of clusters is obtained, but before units are sampled within each cluster, treatment is allocated across sampled clusters. Researchers often improve the precision of treatment effect estimates by drawing a stratified sample of clusters and/or blocking sampled clusters before treatment assignment (Gail et al., 1996; Hansen et al., 2014; Hayes and Moulton, 2009; Imai et al., 2009; Imbens, 2011; Lewsey, 2004). When researchers are interested in heterogeneous treatment effects across subpopulations of interest, within-cluster samples may also be stratified (for an example, see Kerry et al. (2005)).

When clusters are sampled using simple random sampling (SRS) or stratified random sampling (StRS), current estimators of the PATE have undesirable properties. The unbiased Horvitz-Thompson (HT-SRS) estimator (Horvitz and Thompson, 1952) is not invariant to location shifts of responses, which inflates its variance. The location-invariant difference-in-means (DIM) estimator will be biased when treatment effects are correlated with cluster *sizes*—the number of units contained within each cluster (Middleton and Aronow, 2015). Thus, this estimator is only unbiased in special cases such as under sharp null of no unit-level treatment effect (Hansen et al., 2014) or when clusters are blocked or stratified exactly

on cluster sizes (Donner and Klar, 2004; Imai et al., 2009). Moreover, as we will show, when within-cluster samples are not drawn proportional to the cluster size, DIM may estimate a quantity different from the PATE. In fact, the only current estimator of the PATE that is both unbiased and location-invariant is the DesR estimator (DR) (Middleton and Aronow, 2015), which requires the introduction of an additional parameter; however, estimating this parameter will induce bias in the estimator.

We propose an adjustment in the *design* of the experiment—as opposed to adjusting weights of estimators after the experiment— for differences in cluster sizes: to sample clusters with *probability proportional to size* (PPS) (Cochran, 1977; Hansen and Hurwitz, 1943; Lohr, 2010). We show that, under this sampling scheme, the Horvitz-Thompson estimator (HT-PPS) is both unbiased and location invariant.

The paper is organized as follows: Section 2.2 introduces notation. Section 2.2.7 demonstrates problems with HT-SRS, DIM, and DR estimators of PATE under SRS of clusters. Section 2.3 demonstrates that the HT-PPS estimator is both unbiased and location-invariant under PPS-without-replacement sampling of clusters, gives standard errors and estimates of standard errors for HT-PPS, and shows equivalence of HT-PPS and DIM (under PPS) estimators when within-cluster sample sizes are the same across clusters. Section 2.4 extends results to the case where the sample of clusters and the within-cluster sample of units are stratified. Section 2.5 gives simulations on a data example, which shows that the HT-PPS estimator has the smallest mean squared error compared to the other estimators. This is due to the HT-PPS estimator being as efficient as the DIM estimator and being unbiased. It also shows that the estimated variance is conservative for the variability of HT-PPS estimator.

2.2 Notation, assumptions, and preliminaries

We consider a finite population of n units partitioned into ℓ clusters. Clusters are numbered 1 through ℓ. Let n_c denote the number of units within cluster c. Suppose units are ordered

in some way within each cluster; let (k, c) denote the k^{th} unit in cluster c. We now introduce sampling and treatment assignment notation in the order in which they are performed in a CRE.

2.2.1 Sampling clusters

A total of s clusters are sampled; we assume s is fixed and chosen by the researcher. Let S_c denote a cluster sampling indicator; $S_c = 1$ if and only if cluster c is contained in the sample.

$$S_c = \begin{cases} 1, & \text{cluster } c \text{ is sampled,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

By definition, $\sum_{c=1}^{\ell} S_c = s$.

2.2.2 Treatment assignment

Each of the s sampled clusters is assigned to either treatment or control. Let T_{ct} denote a treatment indicator; $T_{ct} = 1$ if and only if cluster c receives treatment $t \in \{0, 1\}$.

$$T_{ct} = \begin{cases} 1, & \text{cluster } c \text{ receives treatment } t, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

We define $T_{ct} = 0$ when $S_c = 0$. Let $\#T_t$ denote the number of clusters that receive treatment t.

We suppose that treatment assignment is *symmetric* across sampled clusters (Miratrix et al., 2013). That is, conditioned on the number of treated clusters $\#T_t$, each of the $\binom{s}{\#T_t}$ possible treatment assignments is equally likely. Symmetric treatment assignment implies

that, for any treatment $t \in \{0, 1\}$ and distinct clusters c, c':

$$\mathbb{E}\left(T_{ct} | \, \mathbf{S}\right) = \frac{\#T_t}{s}, \tag{2.3}$$

$$\mathbb{E}\left(T_{ct}T_{c't} | \, \mathbf{S}\right) = \frac{\#T_t(\#T_t - 1)}{s(s-1)}. \tag{2.4}$$

where $\mathbf{S} = (S_1, S_2, \ldots, S_n)$ denote a random set of cluster sampling indicator variables under a sampling design. Complete randomization is a special case of symmetric treatment assignment. When the sample of clusters is stratified, symmetric treatment assignment also requires independence of treatment assignment across strata, which is discussed in Section 2.4.

2.2.3 Within-cluster sampling

After treatment is assigned across clusters, a SRS of s_c units is drawn within each sampled cluster c. This sample is drawn independently of treatment assignment and independently across clusters. We assume that these sample sizes are non-random and do not depend on the set of clusters sampled. Let S_{kc} denote unit sampling indicator; $S_{kc} = 1$ if and only if the k^{th} unit in cluster c is sampled.

$$S_{kc} = \begin{cases} 1, & \text{unit } (k, c) \text{ is sampled,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

We define $S_{kc} = 0$ when $S_c = 0$. By definition, $\sum_{k=1}^{n_c} S_{kc} = s_c$.

2.2.4 Model of response: Neyman-Rubin Causal Model

Let y_{kct} denote the *potential outcome* of unit (k, c) given treatment t—the value of unit (k, c) we would have observed had that unit received treatment t. Note that y_{kct} is known if and only if that unit is sampled and receives treatment t (i.e., $S_c T_{ct} S_{kc} = 1$). Potential outcomes

are assumed to be nonrandom. Let $\mathbf{y} = (y_{kct})_{k=1,\, c=1,\, t=0}^{n_c\ \ell\ 1}$ denote the vector of potential outcomes.

Let Y_{kc} denote the observed response of unit (k, c) had that unit been sampled. We assume responses follow the Neyman-Rubin Causal Model (NRCM) (Holland, 1986; Rubin, 1974; Splawa-Neyman et al., 1923):

$$\begin{aligned} Y_{kc} &= y_{kc1}T_{c1} + y_{kc0}T_{c0} \\ &= y_{kc1}T_{c1} + y_{kc0}(1 - T_{c1}). \end{aligned} \tag{2.6}$$

Inherent in this model is the *stable-unit treatment value assumption* (SUTVA), which is often referred to as the *no-interference assumption*; the value of Y_{kc} only depends on the treatment assigned to cluster c and is not affected by the treatment assignment of any other cluster c'. Observe that, since each cluster receives a single treatment condition, this assumption only needs to hold across sampled clusters and does not need to hold for units within each cluster.

2.2.5 Parameter of interest

Our quantity of interest is the *population average treatment effect* (PATE):

$$\delta = \delta(\mathbf{y}) \equiv \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kc1} - y_{kc0}}{n} = \mu_1 - \mu_0, \tag{2.7}$$

where

$$\mu_t = \mu_t(\mathbf{y}) \equiv \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kct}}{n}. \tag{2.8}$$

denotes the *population mean for treatment* t. Let

$$\mu_{ct} \equiv \sum_{k=1}^{n_c} \frac{y_{kct}}{n_c} \tag{2.9}$$

denote the *population mean of cluster c for treatment t*. We can write the population mean as:

$$\mu_t = \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kct}}{n} = \sum_{c=1}^{\ell} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kct}}{n_c} = \sum_{c=1}^{\ell} \frac{n_c}{n} \mu_{ct}. \tag{2.10}$$

We then define

$$\sigma_{ct}^2 \equiv \sum_{k=1}^{n_c} \frac{(y_{kct} - \mu_{ct})^2}{n_c - 1}, \tag{2.11}$$

$$\sigma_{t,bet}^2 \equiv \sum_{c=1}^{\ell} \frac{n_c}{n} (\mu_{ct} - \mu_t)^2, \tag{2.12}$$

respectively, as the variance of potential outcomes within cluster c under treatment t and as the weighted across-cluster variance for treatment t.

2.2.6 Properties of estimators

A function of potential outcomes f is *monotonically increasing* if $f(\mathbf{y}^*) \geq f(\mathbf{y})$ whenever

$$y_{kct}^* \geq y_{kct}, \quad \text{for all } \; k \in \{1, \ldots, n_c\}, \; c \in \{1, \ldots, \ell\}, \; t \in \{0, 1\}. \tag{2.13}$$

A transformation of potential outcomes $\mathbf{y} \to \mathbf{y}^*$ is *linear* if, for constants a, b:

$$y_{kct}^* = a + b y_{kct}, \quad \text{for all } \; k \in \{1, \ldots, n_c\}, \; c \in \{1, \ldots, \ell\}, \; t \in \{0, 1\}. \tag{2.14}$$

For simplicity, we may write this as $\mathbf{y}^* = a + b\mathbf{y}$. A *location transformation* or *shift* is a linear transformation in which $b = 1$.

Observe that the population mean is a monotone increasing function that is linear in potential outcomes,

$$\mu_t(a + b\mathbf{y}) = a + b\mu_t(\mathbf{y}), \tag{2.15}$$

whereas the PATE is *location-invariant*—that is, the value does not change given a location

shift of potential outcomes,

$$\delta(a + \mathbf{y}) = \delta(\mathbf{y}). \tag{2.16}$$

2.2.7 Methods for estimating PATE under SRS of clusters

In CREs, clusters are typically sampled using SRS, and the common estimators under this sampling procedure include the Horvitz-Thompson (HT-SRS), the difference-in-means (DIM), and the Des-Raj (DR) estimators. The HT-SRS estimator weights each unit's outcome with the inverse of the probability that the unit is treated and selected. Therefore, it is unbiased, which recommends it as an appropriate estimator of PATE, but Imai et al. (2009) shows that it can be criticized on two counts. The first is that the estimator is known to have huge variability since it does not account for varying cluster size. Larger clusters will have greater sums of responses whereas smaller clusters will have smaller sums. The second being that it is not location-invariant. This poses a dilemma for variance calculation since the variance will change as a changes. The HT-SRS estimator will only be location-invariant when the number of treated *units* (not clusters) is equal to the number of units assigned to control, something of which researchers cannot control.

The DIM estimator is the difference between the sample means for treated and control units. The estimator, being elegantly simple, is favored among many researchers. Furthermore, contrary to the HT-SRS estimator, it is efficient and invariant to location shifts. However, Middleton and Aronow (2015) shows that it is biased in CREs. In actuality, the DIM estimator will be unbiased only when treatment effects are not correlated with cluster sizes and when within-cluster sample sizes are proportional to cluster sizes.

Middleton and Aronow (2015) instead advocates the DR estimator, which adds a regression component on cluster size to the HT-SRS estimator. This helps alleviate the two criticisms on HT-SRS, but unfortunately, the solution itself poses a problem. Estimating the regression coefficient will biased the estimator. Having an estimate of the coefficient prior to the experiment will eliminate the bias, but this is often not feasible. Aronow and Middleton

(2013) expands the DR estimator to allow for additional covariates, but the same issue still persists. In Appendices B–E, we prove the discussed shortcomings of these estimators.

2.3 Estimation of PATE under PPS sampling

Cluster size plays an important role in efficiently estimating the PATE in CREs. Both the DIM and DR estimators give each cluster an equal chance of being selected, regardless of cluster size, but account for it during the estimation stage. Staying true to the design-based philosophy of the Neyman-Rubin model, we advise instead to change the cluster sampling scheme to probability-proportional-to-size sampling (PPS), which can accommodate varying cluster sizes when sampling. Under PPS, we derive the HT estimator, which is unbiased, location-invariant, and efficient.

2.3.1 PPS sampling of clusters

To be precise, we define a PPS sample with s draws as any sample in which the probability of any cluster c of being sampled is $n_c s/n$. While, generally, PPS samples can be drawn with or without replacement, we focus exclusively on PPS samples drawn without replacement (PPSWOR), where the number of unique clusters sampled are fixed. This allows researchers to have greater control in designing a CRE under a budget constraint. A PPSWOR sampling scheme requires each cluster to contain no more than n/s units.

Drawing a PPSWOR sample is a deceptively unintuitive task and quite a bit of work has been devoted to efficient and/or exact selection of PPSWOR samples (Berger and Till, 2009; Brewer and Hanif, 1982; Hanurav, 1967; Sinha, 1973; Vijayan, 1968). Unlike SRS or sampling with replacement, PPSWOR sampling schemes are not uniquely defined solely by the property that the marginal probability of sampling a cluster is $n_c s/n$. Instead, for each pair of unique clusters c, c' a PPSWOR sampling scheme requires knowing the joint probability $\pi_{cc'}$ of having both of these clusters included in the sample. To reduce variance

in estimators, it is useful to choose a sampling scheme such that

$$P(S_c = 1)P(S_{c'} = 1) = \frac{n_c n_{c'} s^2}{n^2} \geq \pi_{cc'} > 0. \tag{2.17}$$

Sunter (1986, 1977) provides list-sequential methods for drawing a PPSWOR sample of general size n satisfying (2.17).

2.3.2 Horvitz-Thompson estimator under PPS sampling

We define the *Horvitz-Thompson estimator under PPS sampling (HT-PPS) for the population mean under treatment* t as:

$$\hat{\mu}_{t,\text{HT,PPS}} = \hat{\mu}_{t,\text{HT,PPS}}(\mathbf{y}) \equiv \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c}. \tag{2.18}$$

In words, this estimate is obtained by finding each cluster c that receives treatment t, computing the average response within each of these clusters, and then taking the average of these within-cluster averages. The *HT-PPS estimator for the PATE* is the difference of the HT-PPS estimator for the population mean under treatment and under control:

$$\hat{\delta}_{\text{HT-PPS}} = \hat{\mu}_{1,\text{HT-PPS}} - \hat{\mu}_{0,\text{HT-PPS}}. \tag{2.19}$$

Note that if a mean estimator is linear in potential outcomes, then the PATE estimator consisting of the mean estimators will be location-invariant. The HT-PPS estimator for the population mean is linear in potential outcomes, which is formally stated in the following lemma:

Lemma 6. *Suppose that clusters are sampled according to PPSWOR sampling, and suppose that treatment is symmetric across clusters. Then:*

$$\hat{\mu}_{t,\text{HT, PPS}}(a + \mathbf{y}) = a + \hat{\mu}_{t,\text{HT,PPS}}(\mathbf{y}). \tag{2.20}$$

The location invariance, unbiasedness, and variance of the HT-PPS estimator for PATE is then provided in the following theorem:

Theorem 7. *Suppose that clusters are sampled according to PPSWOR sampling, and suppose that treatment is symmetric across clusters. Then:*

$$\hat{\delta}_{HT,PPS}(a+\mathbf{y}) = \hat{\delta}_{HT,PPS}(\mathbf{y}) \tag{2.21}$$

$$\mathbb{E}\left(\hat{\delta}_{HT,PPS}\right) = \delta, \tag{2.22}$$

$$\begin{aligned} Var\left(\hat{\delta}_{HT,PPS}\right) = & \sum_{t=0}^{1}\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)\sigma^2_{t,bet} + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left(\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mu_t^2\right)\right] \\ & + \sum_{t=0}^{1}\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\left(1-\frac{s_c}{n_c}\right)\frac{\sigma^2_{ct}}{s_c} \\ & - 2\sum_{c=1}^{\ell}\sum_{c'\neq c}\left[\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2}\right]\mu_{c1}\mu_{c'0} + 2\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{c1}\mu_{c0}. \end{aligned} \tag{2.23}$$

The standard error for the HT-PPS estimator of PATE is then the square root of eq. (2.23). A proof of the lemma and theorem is given in Appendix A.

A PPS sample naturally gives larger clusters a greater probability of being selected. Hence, the sample will be biased towards larger clusters. However, the HT-PPS estimator takes this into consideration as weights when estimating the PATE, thereby, eliminating the bias. Moreover, if the same number of units are sampled from each cluster (say, $\#u$), this will give each treated (controlled) unit in the population an equal probability of being sampled, which does not hold for a SRS of clusters:

$$P(S_c T_{ct} S_{kc} = 1|\text{PPS}) = \frac{\#T_t \#u}{n} \tag{2.24}$$

$$P(S_c T_{ct} S_{kc} = 1|\text{SRS}) = \frac{\#T_t \#u}{\ell n_c}. \tag{2.25}$$

Under this condition, then, the HT-PPS estimator and the DIM estimator (given a PPS sample of clusters) will be the same.

2.3.3 Variance estimator for HT-PPS estimator

Since

$$\text{Var}(\hat{\delta}) = \text{Var}(\hat{\mu}_1) + \text{Var}(\hat{\mu}_0) - 2\text{Cov}(\hat{\mu}_1, \hat{\mu}_0), \tag{2.26}$$

estimating each of the three components will give an estimator for the variance of the HT-PPS estimator for the PATE. The Sen-Yates-Grundy (SYG) variance estimator is an unbiased estimator for the first two parts involving the sampling variance of $\hat{\mu}_t$ (Lohr, 2010). On the contrary, since the last term of eq. (2.23) requires clusters being both treated and controlled, there is no unbiased estimator for the covariance between $\hat{\mu}_1$ and $\hat{\mu}_0$. Consequently, the variance of the HT-PPS estimator cannot be unbiasedly estimated, but a conservative bound is instead provided:

$$\begin{aligned}\widehat{\text{Var}}_{\text{C}}(\hat{\delta}_{\text{HT,PPS}}) = \frac{1}{2}\sum_{t=0}^{1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left[\frac{s(s-1)}{\pi_{cc'}\#T_t(\#T_t-1)}\frac{n_c n_{c'}}{n^2} - \frac{1}{\#T_t^2}\right] S_c T_{ct} S_{c'} T_{c't} \left(\hat{\mu}_{ct} - \hat{\mu}_{c't}\right)^2 \\ - 2\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2}\right]\frac{s(s-1)}{\pi_{cc'}}\frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0}\hat{\mu}_{c1}\hat{\mu}_{c0} \\ + \sum_{t=0}^{1}\sum_{c=1}^{\ell}\frac{S_c T_{ct}}{\#T_t}\frac{n_c^2}{n^2}\hat{\mu}_{ct}^2\end{aligned} \tag{2.27}$$

where $\hat{\mu}_{ct} = \sum_{k=1}^{n_c} y_{kct} S_{kc}/s_c$ is the within-cluster sample mean for t. Note that the first term is SYG variance estimator for μ_t, and the last two terms make up the covariance bound. In Appendix A.5.2, eq. (2.27) is shown to be positively biased for eq. (2.23). Taking the square root of eq. (2.27) will give the estimated standard error of the PATE estimator.

Estimating the variance requires knowledge about the $\pi_{cc'}$ under the specific sampling procedure used to obtain a PPS of clusters, but this is rarely given in practice. Therefore, the $\pi_{cc'}$ needs to be estimated too. This can be achieved using analytical approximations (Berger and Till, 2009; Lohr, 2010) or Monte Carlo simulations (Fattorini, 2009). These approximations, though, could further bias the variance estimator. Thus, we proffer another

estimator that does not depend on the $\pi_{cc'}$:

$$\widehat{\text{Var}}_{\text{WR}}(\hat{\delta}_{\text{HT,PPS}}) = \sum_{t=0}^{1}\left(1-\frac{1}{\#T_t}\right)^{-1}\sum_{c=1}^{\ell}\frac{S_c T_{ct}}{\#T_t^2}\left(\hat{\mu}_{ct}-\hat{\mu}_t\right)^2 + \sum_{t=0}^{1}\sum_{c=1}^{\ell}\frac{S_c T_{ct}}{\#T_t}\frac{n_c}{n}\hat{\mu}_{ct}^2. \tag{2.28}$$

This is the PPS with-replacement variance estimator if clusters are independently sampled and treated. So long as the PPSWOR sampling scheme meets condition 2.17, the WR variance estimator will be larger than the estimator in eq. (2.27) assuming the $\pi_{cc'}$ are known.

2.4 Allowing for stratification

Current literature recommends stratifying and/or blocking on cluster size to further reduce sampling variability. Since PPS sampling already incorporate size variation, stratifying on other prognostic cluster covariates, rather than cluster size, can drastically improve estimation. For example, villages may be stratified based on whether they are in a rural/urban environment or based on the villages' geographic region. Suppose that the population of ℓ clusters are partitioned into m strata based on a categorical cluster characteristic (or a discretized numerical one). Cluster sampling and treatment assignment is done within each stratum and independently across strata. The cluster-stratified HT-PPS estimator is then defined (without the use of indicator variables) as

$$\hat{\delta}_{\text{CS,HT,PPS}} = \sum_{u=1}^{m}\frac{n_u}{n}\left[\sum_{\substack{c\in u,\\ t=1}}^{\#T_1}\frac{1}{\#T_1}\sum_{k=1}^{s_c}\frac{y_{kcu1}}{s_c} - \sum_{\substack{c'\in u,\\ t=0}}^{\#T_0}\frac{1}{\#T_0}\sum_{k^*=1}^{s_{c'}}\frac{y_{k^*c'u0}}{s_{c'}}\right] \tag{2.29}$$

$$= \sum_{u=1}^{m}\frac{n_u}{n}\hat{\delta}_{u,\text{HT,PPS}} \tag{2.30}$$

where n_u is the population of units in stratum u. The statistical properties for the cluster-stratified HT-PPS estimator can be easily derived from Theorem 7.

Theorem 8. *Suppose clusters are first stratified. Suppose also that clusters are sampled with*

PPSWOR and treatments are randomized within stratum and independently across strata. Then:

$$\mathbb{E}\left(\hat{\delta}_{\text{CS,HT,PPS}}\right) = \delta \tag{2.31}$$

$$Var\left(\hat{\delta}_{\text{CS,HT,PPS}}\right) = \sum_{u=1}^{m} \frac{n_u^2}{n^2} Var\left(\hat{\delta}_{u,\text{HT,PPS}}\right) \tag{2.32}$$

$$\hat{\delta}_{\text{CS,HT,PPS}}(a + \mathbf{y}) = \hat{\delta}_{\text{CS,HT,PPS}}(\mathbf{y}). \tag{2.33}$$

Plugging eq. (2.27) or (2.28) into eq. (2.32) will give a conservative estimate of the sampling variability for the cluster-stratified HT-PPS estimator.

Stratification may be applied to units within clusters instead of on clusters. In this setting, the n_c units in cluster c are divided into q_c strata with n_v units in each stratum. A SRS sample of s_v units is taken. The unit-stratified HT-PPS estimator is

$$\hat{\delta}_{\text{US,HT,PPS}} = \sum_{\substack{c=1,\\ t=1}}^{\#T_1} \frac{1}{\#T_1} \sum_{v \in c}^{q_c} \frac{n_v}{n_c} \sum_{k \in v}^{s_v} \frac{y_{kvc1}}{s_v} - \sum_{\substack{c=1,\\ t=0}}^{\#T_0} \frac{1}{\#T_0} \sum_{v \in c}^{q_c} \frac{n_v}{n_c} \sum_{k \in v}^{s_v} \frac{y_{kvc0}}{s_v}. \tag{2.34}$$

Since the stratification is on the units within a cluster, we need to only adjust the within-cluster variance in Theorem 7 to get the statistical properties for the unit-stratified HT-PPS estimator. Hence,

$$\text{Var}(\hat{\mu}_{ct}) = \frac{1}{\#T_t} \sum_{c=1}^{\ell} \frac{n_c}{n} \left(1 - \frac{s_c}{n_c}\right) \frac{\sigma_{ct}^2}{s_c} \tag{2.35}$$

will instead be

$$\text{Var}(\hat{\mu}_{ct}) = \frac{1}{\#T_t} \sum_{c=1}^{\ell} \frac{n_c}{n} \sum_{v \in c}^{q_c} \frac{n_v^2}{n_c^2} \left(1 - \frac{s_v}{n_v}\right) \frac{\sigma_{ct}^2}{s_v}. \tag{2.36}$$

Similarly, eq. (2.27) and (2.28) are still conservative estimates of the sampling variability, but $\hat{\mu}_{ct}$ will instead be the stratified estimator of the within-cluster sample mean for t. Naturally, if stratification is desired at both the cluster- and unit-levels, combining eq. (2.29) and eq. (2.34) will give an unbiased estimator for the PATE.

2.5 Data example

Beath et al. (2013) perform an experiment in Afghanistan to investigate whether development programs with mandatory women contribution can change villagers' perspectives on women's political participation. Five hundred villages, ranging from sizes 60 to 9000, were sampled and matched into pairs. Within each pair, one village is randomly assigned to receive the National Solidarity Programme (NSP), and the other village serves as a control to receive the NSP after the experiment. The NSP creates a community development council and provides grants for village development projects. The council is then responsible for distributing the grants among the projects. However, the NSP stipulates that half of the council must be women and at least one of the projects must be a priority for the women. After two years, ten head-of-household men and ten head-of-household women from each village are selected for a follow-up survey. Respondents are asked whether women should have equal decision making in the village council.

We perform Monte Carlo simulations to compare the HT-PPS estimator to its SRS counterparts. We generate the potential outcomes from a fitted LOWESS line of the NSP data and then mimic a simplified experiment in which clusters are randomly sampled using either PPS or SRS. The R package *TeachingSampling* is used to perform Sunter's PPSWOR sampling. We vary the number of sampled clusters from 20 to 200. Treatments are assigned completely at random to the sampled clusters. For ease, we fix the number of treated clusters to be half of the sampled clusters, but this will not drastically change the theoretical results.

The PATE is then estimated with the HT-PPS, DIM, HT-SRS, biased DR, and Hájek estimators. For the DR estimator, θ is optimally estimated as described in Middleton and Aronow (2015) using the simulated sample data. The Hájek (HJ) estimator (see (Hajek, 1971)) is a ratio estimator similar to the DIM. It estimates the population mean for treatment t as a ratio of the estimated treated (controlled) cluster total over the total number of treated (controlled) units in the sampled clusters. The other estimators are as described in Section 2.2.7.

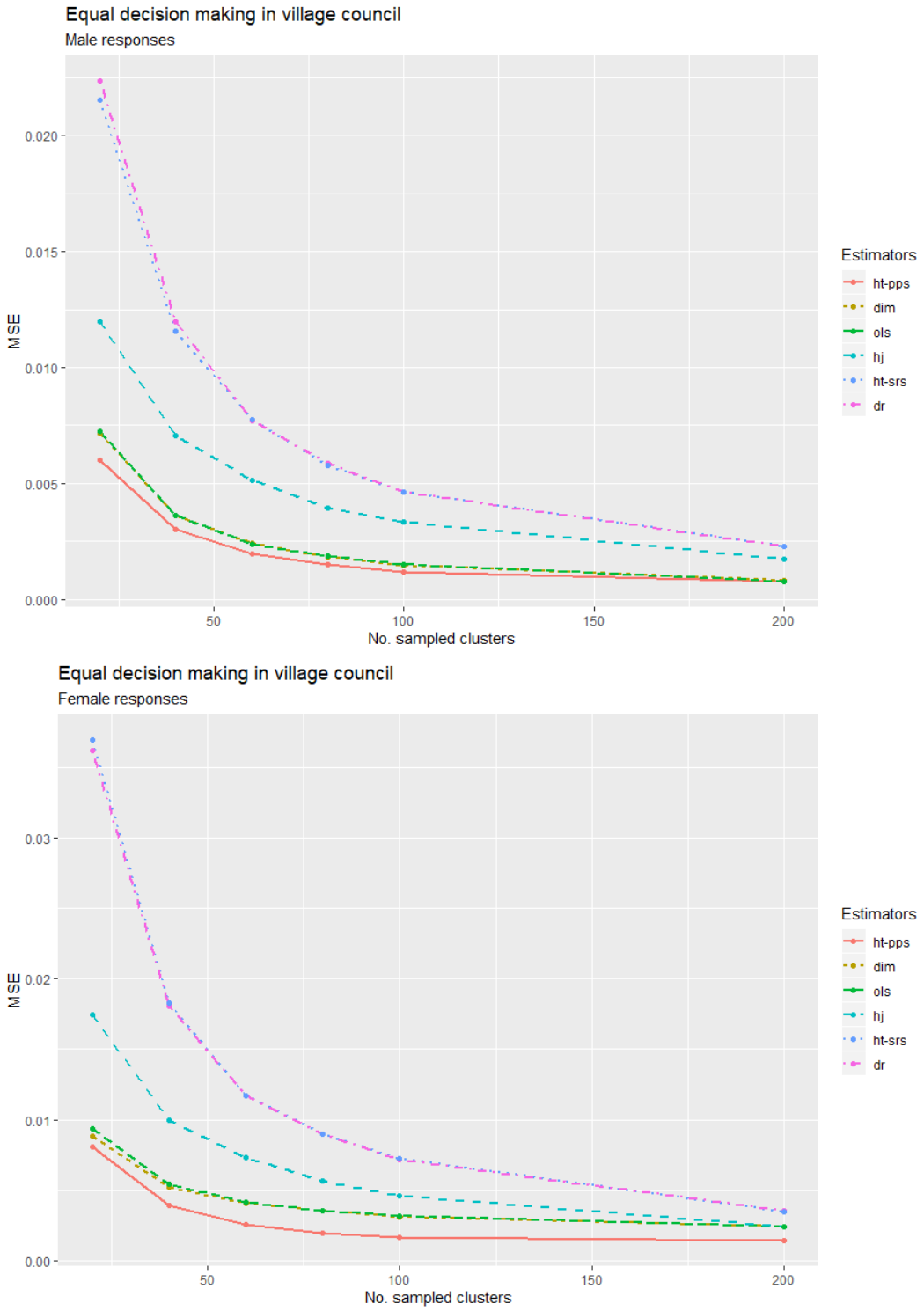

Figure 2.1: The number of sampled clusters are 20, 40, 60, 80, 100, and 200. Results are based on 10,000 simulations. Our estimator, HT-PPS (red and solid), performs better than the SRS estimators.

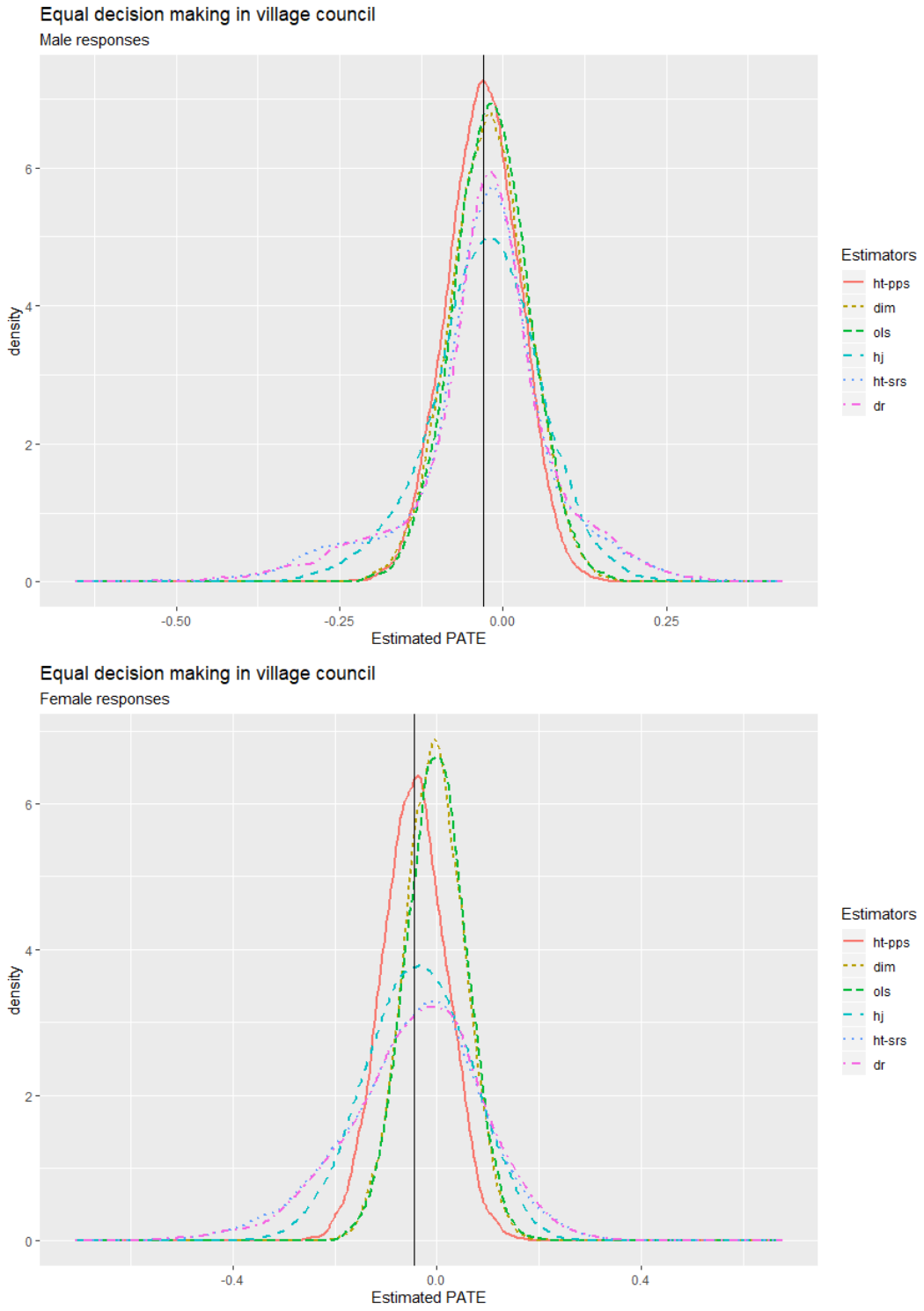

Figure 2.2: Results based on 10,000 simulations of sampling 40 clusters. The solid vertical line is the PATE (-0.0302 for male, -0.0448 for female). Our estimator, HT-PPS (red and solid), is unbiased and as efficient as the DIM.

Male responses				
	SYG SE est		WR SE est	
No. sampled clusters	Est. var.	Samp. var.	Est. var.	Samp. var.
20	1.626	0.066	0.063	2.57E-4
40	1.749	0.021	0.048	5.77E-5
60	-	-	0.041	2.41E-5
80	-	-	0.036	1.29E-5
100	-	-	0.033	8.43E-6
200	-	-	0.025	2.14E-6
Female responses				
	SYG SE est		WR SE est	
No. sampled clusters	Est. var.	Samp. var.	Est. var.	Samp. var.
20	2.046	0.035	0.045	3.28E-4
40	2.096	0.013	0.032	1.13E-4
60	-	-	0.030	5.29E-4
80	-	-	0.029	2.82E-5
100	-	-	0.028	1.57E-5
200	-	-	0.025	2.32E-6

Table 2.1: Results based on 10,000 simulations. SE cannot be estimated for approximate PPSWOR samples, but even for exact PPSWOR, the SYG WOR SE estimates are larger than the WR SE estimates, possibly from the increased bias in estimating the joint cluster inclusion probabilites $\pi_{cc'}$.

Figure 2.1 compares the MSE of the estimators as the number of sampled clusters are varied. To get an exact PPSWOR sample of the NSP data, the number of sampled clusters can be at most 45, and thus, the samples of sizes 60 and up are only approximately PPS. Even so, the HT-PPS estimator performs best out of all the estimators, including the omnipresent DIM, across all sample sizes of clusters. Figure 2.2 gives a more thorough comparison of the sampling distributions for the PATE estimators when 40 clusters are sampled.

In addition, Monte Carlo simulations are done to examine the performance of estimating

the sampling variance of the HT-PPS estimator. The $\pi_{cc'}$ are estimated using analytical approximations. Table 2.1 provides statistics (estimated variance and true sampling variability) on the variance estimation as the number of sampled clusters are varied [1].

2.6 Conclusion

Experiments are the "gold standard" for investigating causal relationships, but traditionally, the causal inference is limited to the convenience sample recruited for the experiment. Often, though, researchers prefer to generalize to individuals beyond those in the sample. This then requires a random sample from the population of interest. Since populations are naturally structured in groups, it is easier to sampled groups, rather than individuals, to be in experiments; thus, cluster-randomized experiments are a fitting design choice.

On the other hand, the multi-level constitution of CREs poses analytical adversities. Much of the difficulties arises from unequal cluster sizes. If clusters contain the same number of units, all estimators discussed would be the same, and the idea of choosing the "best" estimator would be nonexistent. However, varying cluster sizes are intrinsic to CREs. Hence, in this paper, we account for cluster sizes by sampling clusters with probability proportional to size. Estimating PATE can then be done with the HT-PPS estimator. The HT-PPS estimator is an attractive alternative to SRS-based estimators since it is intuitive, unbiased, efficient, and location-invariant. We also derive two conservative variance estimators for the sampling variability of the HT-PPS estimator.

Stratification and blocking can still be used to further reduce the sampling variability, but with PPS sampling, other more important covariates can be used instead of cluster size. We have done some work on how stratification may affect the HT-PPS estimator, but we plan to expand on it to include blocking.

[1] Some of the estimated SEs for the PPSWOR designs are negative, but performance analysis is based only on the positive values.

Chapter 3

Best Practices for Cluster-Randomized Experiments

3.1 Introduction

The primary objective of a randomized experiment is to estimate the average causal effect of a treatment, but an equally important secondary objective is to estimate the standard error. The standard error gives an efficiency measure on how well the treatment effect was estimated and is pivotal in the planning and inferential stages. Hence, the goal, if possible, should be to simultaneously estimate both the treatment effect and the standard error efficiently. While most methodologists agree that use of prognostically important covariates will help minimize the standard error, there is no consensus on which method will be the most reductive. This is especially true in the case of cluster-randomized experiments (CREs), in which treatments are randomized to group of units instead of straight to individual units. These kind of experiments are commonly implemented when treatment randomization is more practical or cost-effective at the cluster level (e.g., teaching methods are randomized to classes instead of individual students). However, having a cluster as the experimental entity leads to a significant loss in efficiency as compared to unit randomized experiments. This increase

in estimation variability comes from responses being considerably more different between clusters, particularly when cluster size—number of individuals in a cluster—varies greatly.

Under the Neyman-Rubin model of potential outcomes, both treatment and standard error estimation are affected by choice of estimator and experimental design. Some authors may choose to keep the design simple and incorporate covariate information directly into the estimator instead. These include the Hajék (HJ), difference-in-means (DIM), and Des-Raj (DR) estimators. Both the HJ and DIM utilize covariate information as ratio estimators, and the DR as a regression estimator. On the design side, CRE experts generally recommends stratification, blocking, or pair matching to improve treatment estimation. Xiong and Higgins (2020) also showed that sampling clusters proportional to cluster size is another attractive design alternative. These design methods use a covariate to group similar clusters for either sampling or treatment randomization. The Horvitz-Thompson estimator, which integrates the design into the estimator, is then typically used to estimate the average treatment effect.

Previous design-based research treated these two aspects separately, either comparing the benefits of the different designs (see, Donner (1998); Donner and Klar (2004); Imai et al. (2009); Imbens (2011)) or the estimators (see, Middleton and Aronow (2015); Xiong and Higgins (2020)). Moreover, while some of these works do discuss standard error estimation, it is usually in terms of the highlighted method only. In general, exploring how these various methods may affect their standard error estimation is understudied. Seeing how all of them serve the same purpose, we now synthesize them into one comparison to see if there is a best method for both treatment estimation and standard error estimation.

Using simulated data based on the National Solidarity Programme (Beath et al., 2013), we find that HT under PPSWOR sampling is the best at estimating both the average treatment effect and the standard error, even with a small number of sampled clusters. This is only method that seems to simultaneously minimize the variance for both the PATE and SE estimation. The other methods tend to have an opposing effect. Those methods (e.g., strati-

fication, matched-pairs, DIM, HJ and OLS with clustered SE) that are helpful in estimating PATE tend to have more error in estimating the SE.

This paper is organized as follows. Section 3.2 explains the potential outcome framework under which we are working and defines the specific average treatment effect we are measuring. Section 3.3 expresses the design-based estimators in general form and provides a short comparison, and section 3.4 describes the different designs common for CREs . Section 3.5 addresses the challenges of standard error estimations. Section 3.6 presents the simulation results, leading us to our conclusions on designing and analyzing CREs in Section 3.7.

3.2 Potential outcomes framework

Suppose there is a finite population of n units grouped into ℓ clusters. This partition of units is done without researcher intervention. Allowing cluster size, the number of units in a cluster, to vary, let n_c be the number of units in cluster c. Denote the k^{th} unit in cluster c then as (k, c). Each unit (k, c) has a potential outcome for treatment t, representing the value of the unit *if* it receives treatment t, designated as y_{kct}. Here, $t = 1$ for treated and $t = 0$ for controlled. We use "treatment" or "treated" to sometimes refer to $t = 1$ specifically and sometimes in general to $t \in \{0, 1\}$, but the case should be clear contextually.

To assess the efficacy of a treatment, researchers are often interested in the population average treatment effect (PATE),

$$\delta = \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kc1} - y_{kc0}}{n} = \mu_1 - \mu_0, \tag{3.1}$$

where

$$\mu_t = \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kct}}{n}. \tag{3.2}$$

However, as a unit is either given treatment or control and not both, only y_{kc1} or y_{kc0} can be known for each unit. In CREs, this is entirely dependent on the treatment assignment of

the unit's cluster, which is formally summarized in the Neyman-Rubin model of responses. Let T_{ct} be the cluster treatment indicator variable—that is, $T_{ct} = 1$ if cluster c is assigned treatment t and $T_{ct} = 0$ if not. A unit's response, following the Neyman-Rubin model, can then be written as:

$$Y_{kc} = y_{kc1}T_{c1} + y_{kc0}T_{c0} \tag{3.3}$$

$$= y_{kc1}T_{c1} + y_{kc0}(1 - T_{c1}). \tag{3.4}$$

The model is nonparametric as responses do not have any distributional assumptions. Variations in responses instead are assumed to arise from the experimental design. However, within the model is the stable unit treatment value assumption (SUTVA), which necessitates that a unit's response is not affected by another unit's treatment assignment. This assumption needs only hold for units across clusters and not within clusters in CREs since all units within the same cluster are given the same treatment.

We also assume that treatments are symmetrically assigned (Miratrix et al., 2013). That is, every combination of assigning the $\#T_t$ treated clusters is equally likely. Completely randomizing treatments to clusters satisfies the treatment symmetry assumption. If clusters are stratified, randomizing treatment within stratum but independently across strata also satisfies this assumption.

3.3 Estimation of PATE

Since the PATE requires that both potential outcomes be known for each unit, the PATE is not observable and must be estimated. Estimating it with a CRE involves at least three stages of randomization: sampling clusters, assigning treatments, and sampling units within clusters. To represent each of the randomization stage, we define three indicator variables. Let S_c be the cluster sampling indicator, T_{ct} be the cluster treatment indicator, and S_{kc} be the unit sampling indicator. A unit's response is then only observed if it is sampled

from a treated cluster (i.e., $S_c T_{ct} S_{kc} = 1$), and the probability of this occurring, denoted as $\pi_{kct} = P(S_c T_{ct} S_{kc} = 1)$, plays an important role in estimating the PATE efficiently. Here, we list the most common estimators used to analyze CREs and provide a summary of their merits.

3.3.1 Horvitz-Thompson estimator

The Horvitz-Thompson (HT) estimator of the PATE uses π_{kct} to inversely weight each unit's response (Horvitz and Thompson, 1952). The HT estimator can be written as

$$\hat{\delta}_{\text{HT}} = \frac{1}{n} \sum_{c=1}^{n} \sum_{k=1}^{n_c} \frac{y_{kc1} S_c T_{c1} S_{kc}}{\pi_{kc1}} - \frac{1}{n} \sum_{c^*=1}^{n} \sum_{k^*=1}^{n_{c^*}} \frac{y_{k^*c^*0} S_{c^*} T_{c^*0} S_{k^*c^*}}{\pi_{k^*c^*0}}. \tag{3.5}$$

From eq. (3.5), it is clear to see that the HT estimator is unbiased for any design in which the inclusion probabilities are known, but it can be very imprecise under simple random sampling (SRS). Thus, it is most commonly used in more economic designs to offset the inefficiency in this estimator. The HT estimator is also, surprisingly, not location-invariant under SRS, and so, any transformation with a constant shift in potential outcomes will affect the PATE estimation.

3.3.2 Ratio estimators

A ratio estimator has the form

$$\hat{\delta}_{\text{R}} = \frac{\sum_{c=1}^{\ell} \sum_{k=1}^{n_c} w_{kc1} y_{kc1} S_c T_{c1} S_{kc}}{\sum_{c=1}^{\ell} \sum_{k=1}^{n_c} w_{kc1} S_c T_{c1} S_{kc}} - \frac{\sum_{c^*=1}^{\ell} \sum_{k^*=1}^{n_{c^*}} w_{k^*c^*0} y_{k^*c^*0} S_{c^*} T_{c^*0} S_{k^*c^*}}{\sum_{c^*=1}^{\ell} \sum_{k^*=1}^{n_{c^*}} w_{k^*c^*0} S_{c^*} T_{c^*0} S_{k^*c^*}} \tag{3.6}$$

where w_{kct} is the weight for each response. Defining $w_{kct} = 1/\pi_{kct}$ will give the Hajék (HJ) estimator (Hajek, 1971) and $w_{kct} = 1$ will give the difference-in-means (DIM) estimator. Both estimators are biased, but they tend to vary less, even with as simple a design as SRS. Ratio estimators are particularly useful when the population size n is not a known quantity.

3.3.3 Des-Raj estimator

The Des-Raj (DR) estimator adds a regression component to support the design-based estimation in the HT estimator (Middleton and Aronow, 2015):

$$\begin{aligned}\hat{\delta}_{\text{DR}} =& \frac{1}{n}\sum_{c=1}^{n}\sum_{k=1}^{n_c}\frac{y_{kc1}S_cT_{c1}S_{kc}}{\pi_{kc1}} - \frac{\theta}{n}\sum_{c=1}^{\ell}\frac{x_{c1}S_cT_{c1}}{\pi_{c1}} \\ &- \left(\frac{1}{n}\sum_{c^*=1}^{n}\sum_{k^*=1}^{n_{c^*}}\frac{y_{k^*c^*0}S_{c^*}T_{c^*0}S_{k^*c^*}}{\pi_{k^*c^*0}} - \frac{\theta}{n}\sum_{c^*=1}^{\ell}\frac{x_{c^*t}S_cT_{c0}}{\pi_{c^*0}}\right)\end{aligned} \tag{3.7}$$

where θ is a regression coefficient of the cluster covariate x_{ct} and π_{ct} is the inclusion probability of cluster given treatment t. The DR estimator is then able to discharge some of the inefficiency from the HT estimator and still retain its unbiasedness *if* θ is known or has been previously estimated. Unfortunately, this is rarely the case, particularly since estimation of the regression coefficients require knowledge about treatment assignment. Estimating θ with the sample data, though, will biased the estimator. The trade-off then may be insignificant.

3.4 CRE designs

Implementing a CRE design equates to specifying a method for each randomization stage: sampling clusters, assigning treatment, and sampling units within clusters. While the last two stages can affect estimation of PATE, especially the second stage, our focus is on the first stage. The most common approach is to sample clusters via simple random sampling without replacement (SRSWOR), but sampling schemes such as stratified sampling, match-paired sampling, or probability-proportional-to-size without replacement sampling (PPSWOR), incorporate covariates that can improve estimation of PATE. Here, we explicitly describe these different designs and present the inclusion probabilities under each of them. Plugging in π_{kct} into the estimators in Section 3.3 will give the specific form.

3.4.1 Simple random sample of clusters

From the population of ℓ clusters, a SRSWOR of fixed size s is taken. The sample of clusters are then randomly divided into two treatment groups. Let $\#T_t$ represent the number of clusters given treatment t. From a total of n_c units in a treated cluster c, a SRSWOR of s_c units is taken. It is possible to have this sample of units consists of all units in the cluster, but since units are more similar within clusters, it is more cost effective to sample units from different clusters. Therefore, under a SRSWOR of clusters,

$$\pi_{kct} = \frac{\#T_t}{\ell}\frac{s_c}{n_c}. \tag{3.8}$$

A SRSWOR of clusters will randomly select a cluster, irrespective of how large or small it is, with equal probability. As such, the HT estimator under this design tends to suffer from large variability. Often, alternative estimators (e.g., DIM, HJ, and DR) are preferred under this design since they can alleviate these issues, but they are not without their own complications.

3.4.2 PPS sample of clusters

If the unit inclusion probability π_{kct} are proportional to the outcomes y_{kct}, there will be no variability between the cluster means. With no knowledge of the outcomes though, the probability are instead chosen to be proportional to a covariate closely related to outcomes for a similar effect. Cluster size is known to be a crucial covariate to consider and include when analyzing CREs as it frequently serves as a substitute for other more predictive, albeit unknown, covariates. Therefore, accounting for varying cluster size when sampling them will help to improve estimation efficiency. A probability-proportional-to-size without-replacement (PPSWOR) sample will select a cluster with a probability of $n_c s/n$, where s is a fixed sample size of clusters. Once clusters are sampled, treatments are then completely randomized to them, and within-cluster units can be sampled with SRSWOR. The inclusion

probabilities for this design are

$$\pi_{kct} = \frac{\#T_t s_c}{n}. \tag{3.9}$$

Sunter (1977) and Sunter (1986) outline two procedures to get a PPSWOR sample of a general fixed size. The Sunter sampling methods will give exact PPSWOR sample if all clusters have no more than n/s units; otherwise, the methods will give an approximate sample that can incur estimation bias. While there are other possible methods, the sample size is either random or limited to two; both of which are impractical for planning purposes.

3.4.3 Stratified random sample of clusters

Rather than use cluster size to directly sample clusters, stratified random sampling uses the information to group clusters of similar size into strata. Suppose there are m strata with ℓ_u clusters and n_u units in stratum u. In a stratified CRE, all three randomization stages occur within each stratum and independently across strata. Within each stratum, clusters are sampled with SRSWOR, treatments are completely randomized to sampled clusters, and units are randomly sampled from sampled clusters. In other words, a CRE with a SRSWOR of clusters are independently conducted in each stratum. Therefore, the inclusion probabilities needs to be defined within each stratum u as

$$\pi_{kcut} = \frac{\#T_{ut}}{\ell_u} \frac{s_{cu}}{n_{cu}}, \tag{3.10}$$

where each notation is as noted in Section 3.4.1 but adjusted to be within each stratum. The HT estimator is then given by

$$\hat{\delta}_{\text{ST}} = \sum_{u=1}^{m} \frac{n_u}{n} \hat{\delta}_u \tag{3.11}$$

where the stratum HT estimator, $\hat{\delta}_u$ is of the form in eq. (3.5) but weighted by the stratum inclusion probabilities in eq. (3.10).

3.4.4 Matched-pair random sample of clusters

A match-paired sample of clusters is structured similar to a stratified random sample. Using a covariate like cluster size, pairs of clusters with the most similar covariate values are grouped together. Define m as the number of pairs and n_u as the number of units in pair u. However, unlike in stratified random sample, clusters are not randomly sampled from each stratum. Instead, pairs of clusters are randomly sampled together. Assume the $\#m$ selected pairs are chosen with SRSWOR. Then treatments are randomized within pair such that one of the clusters is given treatment and the other is given control, and lastly, units are randomly sample within clusters. Thus, we need both the inclusion probability for a pair u and the inclusion probabilities for units in the selected pair, which are respectively

$$\pi_u = \frac{\#m}{m}, \tag{3.12}$$

$$\pi_{kcut} = \frac{1}{2}\frac{s_{cu}}{n_{cu}}. \tag{3.13}$$

Adjusting for the two-step sampling, the HT estimator for the matched-pair case is

$$\hat{\delta}_{\text{MP}} = \sum_{u=1}^{m} \frac{n_u}{n}\frac{\hat{\delta}_u S_u}{\pi_u} \tag{3.14}$$

where S_u is the sampling indicator for pair u, $\hat{\delta}_u$ is the HT estimator in eq. (3.5) for pair u, weighted by eq. (3.13).

3.4.5 Note on treatment randomization and unit sampling

The previous designs only alter the cluster sampling scheme to allow for inclusion of covariate information while the cluster treatment randomization and unit sampling methods are kept simple throughout. It is possible to also change them so that they may account for covariate information too. For the treatment randomization stage, blocking is often used in such situations. To be clear, blocking is distinguished from stratified sampling in that

blocking is done on a *sample*, not on the population. In actuality, in both the stratified and matched-pair sampling, since treatments are randomized within each stratum/pair, this would be considered blocking. While this is not necessary and treatments can be randomized completely at random to clusters sampled from either stratified or matched-pair sampling, it seems silly to not also utilize this available information to balance treatment randomization.

Traditional design theory usually makes some existential assumption about blocks, either that they are sample from a population of blocks (i.e., blocks are random) or that they are a census of this block population (i.e., blocks are fixed). The stratified sampling in Section 3.4.3 falls in the latter case whereas the matched-pair sampling in Section 3.4.4 is of the former. In practice, though, it may be more reasonable to randomly sample with SRSWOR and then perform blocking or matching on the sample units. For example, with increased computation power, optimal blocks can be formed with distance metrics rather than categorical factors. However, analyzing the PATE in this case is more challenging since the blocks depend on which clusters are sampled. Conditioning on the sample of clusters may help ease the analysis, though this requires more in-depth research.

If there are covariates taken at the unit level, they can be used to stratified the units for more precise estimation of responses within clusters. Estimation can easily be modified to integrate this additional intricacy [see (Xiong and Higgins, 2020)]. On the contrary, this will not help improve estimation of responses between clusters, which is often more problematic.

3.4.6 Debates about design

Known to have larger variability than with unclustered data, CREs in practice frequently implement blocking to eliminate some inefficiency. In general, CRE methodologists agree that some form of blocking does help but dispute about the best form: blocking with more than two clusters per block or blocking with just two clusters per block. The second one is better referred to as matched-pair, which is similar in nature to the matched-pair sampling described in Section 3.4.4. The precision gain in blocking comes from the increased treatment

balance of similar clusters, and so logically, the matched-pair blocking would produce the most balance and thus the greatest gain. Imai et al. (2009) vehemently advocates that "pair matching should be used whenever feasible." Nonetheless, there are some estimation constraint on the SE that led authors, most notably Donner (1998); Donner and Klar (2004); Martin et al. (1993) and Imbens (2011), to advise against the matched-pairs design and instead endorse the general blocked design.

3.5 Estimation of SE

Estimation of SE in CREs must take care to capture the clustered nature of responses. Otherwise, the estimated SE will be negatively biased, invalidating any inferential conclusions drawn. Under the design paradigm, variances of estimators are derived directly from the design so SE estimation naturally accounts for clustering and does not need to rely on modeling assumptions or external adjustments. The general variance for any estimator in Section 3.4 is

$$\text{Var}\left(\hat{\delta}\right) = \Big[\text{Var}\left(\hat{\mu}_1\right) + \text{Var}(\hat{\mu}_0) - 2\text{Cov}(\hat{\mu}_1, \hat{\mu}_0)\Big]. \tag{3.15}$$

Taking the square root of this expression will yield the true SE of an estimator. Estimation of this quantity is based on a method-of-moments approach, in which each term in the variance is estimated with its sample counterpart. The first two terms can usually be estimated, though they have their challenges too, but the last term is not estimable. We first start with a discussion of these issues under SRSWOR and PPSWOR and then broaden it to the stratified and matched-pair cases. Because these formulae are rather lengthy, we do not list them here but instead give their equation reference number. Full form of the variance estimators are provided in the Appendices.

3.5.1 Challenges with estimating the variance of $\hat{\mu}_t$

The $\text{Var}(\hat{\mu}_t)$ terms for the HT estimator can be written in a closed form and estimated with the Sen-Yates-Grundy (SYG) variance estimator. The SYG variance estimator is unbiased under both SRSWOR [eq. (B.28)] and PPSWOR [eq. (A.33)] if the joint cluster inclusion probabilities, $\pi_{cc'}$, are known. The $\pi_{cc'}$ are easily calculated under SRSWOR but calculating them exactly under PPSWOR, if at all possible, can be computationally cumbersome since they depend on the specific PPSWOR sampling algorithm. Certainly, the $\pi_{cc'}$ can either be estimated with analytical approximations (Berger and Till, 2009; Lohr, 2010) or Monte Carlo simulations (Fattorini, 2009), but they tend to work only when an exact PPSWOR sample is collected. Even then, the approximation will bias the estimation, and without knowledge of the $\pi_{cc'}$, it may be hard to say in which the direction the bias is. To work around the $\pi_{cc'}$, the with-replacement (WR) variance estimator [eq. (A.34)] can be used instead as conservative estimates.

Both the DIM and HJ estimators (under SRSWOR) contain random quantities in their denominator, so writing the $\text{Var}(\hat{\mu}_t)$ terms in a closed form is not easily possible. Instead, the two estimators are linearized using the Taylor series, and the variance estimators are derived based on this approximation [eq. (C.19) for DIM and eq. (D.13) for HJ]. As for the DR estimator, since it is a variant of the HT estimator, the SYG variance estimator will also be unbiased under SRSWOR [eq. (E.12)].

3.5.2 Covariance bound

To show why the covariance term is not estimable,

$$\text{Cov}(\hat{\mu}_1, \hat{\mu}_0) = \mathbb{E}(\hat{\mu}_1\hat{\mu}_0) - \mu_1\mu_0 \tag{3.16}$$

$$= \mathbb{E}(\hat{\mu}_1\hat{\mu}_0) - \frac{1}{n^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}\sum_{k=1}^{n_c}\sum_{k^*=1}^{n_{c'}} y_{kc1}y_{k^*c'0} - \frac{1}{n^2}\sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kc1}y_{kc0}. \tag{3.17}$$

Note that the last component of the covariance requires that cluster c receive both treatments, which is not observable. Using Young's inequality, we can bound it with

$$\begin{aligned}\mathrm{Cov}_C(\hat{\mu}_1, \hat{\mu}_0) =& \mathbb{E}(\hat{\mu}_1\hat{\mu}_0) - \frac{1}{n^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}\sum_{k=1}^{n_c}\sum_{k^*=1}^{n_{c'}} y_{kc1}y_{k^*c'0} \\ &- \frac{1}{n^2}\sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kc1}^2 - \frac{1}{n^2}\sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kc0}^2. \end{aligned} \tag{3.18}$$

This covariance bound can be unbiasedly estimated with sample statistics, and using it in the SE estimation provides a conservative estimate. This bound is used in all SE estimation.

3.5.3 Extension to stratified and matched-pair

All of the previous discourse on variance estimation still pertains to the stratified and matched-pair cases but now the estimation is completed within stratum/pair. Both the stratified and matched-pair sampling consists of an additional grouping level above clustering that needs to also be accounted for in the variance estimation. As all strata are observed and all randomization is conducted within each stratum, the variance for stratification is simply

$$\mathrm{Var}\left(\hat{\delta}_{\mathrm{ST}}\right) = \sum_{u=1}^{m}\frac{n_u^2}{n^2}\mathrm{Var}\left(\hat{\delta}_u\right) \tag{3.19}$$

where $\mathrm{Var}\left(\hat{\delta}_u\right)$ would be the same as in eq. (3.15). Thus, all the components can be estimated as mentioned in Sections 3.5.1 and 3.5.2. The matched-pair variance, on the other hand, is not as straight-forward.

Since pairs of clusters are the sampling units rather than individual clusters, there is an additional variance term to reflect this:

$$\begin{aligned}\mathrm{Var}\left(\hat{\delta}_{\mathrm{MP}}\right) =& \sum_{u=1}^{m}\frac{n_u^2}{n^2}\frac{\mathrm{Var}\left(\hat{\delta}_u\right)}{\pi_u} \\ &+ \sum_{u=1}^{m}\frac{n_u^2}{n^2}\frac{\delta_u^2}{\pi_u}(1-\pi_u) + \sum_{u=1}^{m}\sum_{u'\neq u}\frac{n_u n_{u'}}{n^2}\frac{\delta_u\delta_{u'}}{\pi_u\pi_{u'}}\left(\pi_{uu'} - \pi_u\pi_{u'}\right), \end{aligned} \tag{3.20}$$

where $\pi_{uu'}$ is the probability that any two pairs are sampled together. The first term measures the within-pair variability while the last two terms correspond to the across-pair variability. From eq. (3.15), estimating Var $\left(\hat{\delta}_u\right)$ involves estimating the mean treatment variability, which is not possible with the matched-pair design since the mean cannot be determined from just one treated cluster. This is the main concern that critics of the matched-pairs design cautioned against. Since there are multiple treated clusters within a stratum/block, this is not a problem that the stratified/blocked design have. To circumvent this problem, (Imai et al., 2009) changes the treatment randomization from occurring within pairs to directly on the pairs. Note that the matched-pair sampling method in Section 3.4.4 does not include this adaptation.

3.6 Simulation results

The National Solidarity Programme (NSP) is a community program to promote development in rural Afghanistan villages and part of a political experiment to gauge its effect on democratic processes. Among the objectives of this study is evaluating how the NSP may change perspective on women's political participation. Five hundred villages, ranging from sizes 60 to 9000, were sampled and matched into pairs. Within each pair, one village is randomly assigned to enroll into the NSP, and the other village serves as a control to enroll after the experiment. As part of the program, a community council is elected to oversee and budget projects. However, half of the council must be women and at least one of the projects must be a priority for the women. After two years, ten head-of-household men and ten head-of-household women from each village are selected for a follow-up survey. Respondents are asked whether women should have equal decision making in the village council.

We conduct Monte Carlo simulations to compare the performance of each design. Potential outcomes are generated from a fitted LOWESS line of the NSP data before we run the simulated experiments as outlined in Section 3.4. For PPSWOR sampling, the R package

TeachingSampling is used to perform Sunter's PPSWOR sampling. For stratified sampling, the villages are divided into three different groups depending on cluster size (i.e., < 500, $500 - 1500$, ≥ 1500); the number of sampled clusters from each stratum is chosen to be as close to the Neyman allocation as possible. For matched-pair sampling, clusters are also paired based on the Euclidean distance of cluster size. For the male data, the average distance between clusters in a pair is 20.51 and for the female data, it is 15.86. We vary the number of sampled clusters from 20 to 200. For ease, the number of treated clusters are fixed to be half of the sampled clusters, but this will not drastically change the theoretical results. The HT estimator is then used to estimate the PATE under SRSWOR, PPSWOR, stratified, and matched-pair sampling. We also include the other estimators, DIM, HJ, and DR, under SRSWOR of clusters. Adhering to realistic practices, the regression coefficient for the DR estimator is estimated with the sample data, thereby biasing the estimator for PATE. The OLS estimator with clustered SE, which is a popular choice in political studies and is used in (Beath et al., 2013), is also considered.

Figure 3.1 presents the simulation results for estimating the PATE among male and female responses. Across the sample sizes considered, the PPSWOR design performs the best, having the lowest MSE. Therefore, PPSWOR sampling is significantly beneficial in CREs, even more so than the standard design methods, especially when the number of clusters available for experimentation is small. Note also that after sampling 45 clusters, the PPSWOR samples are no longer exact, and yet, the bias accrued from approximating the PPSWOR sample does not negatively impact performance. This is likely because the largest cluster is only of 9000 units. A main hindrance of PPSWOR is the calculation of the joint cluster inclusion probability needed for SE estimation. In our simulation, we were not able to calculate the estimated PPSWOR SE for sizes greater than 40, even with analytical approximation to the joint inclusion probability. However, with the WR variance estimator, this is no longer a deterrent. Figure 3.2 shows that, with the use of the with-replacement variance estimator, the PPSWOR design gives the smallest variance for a conservative SE

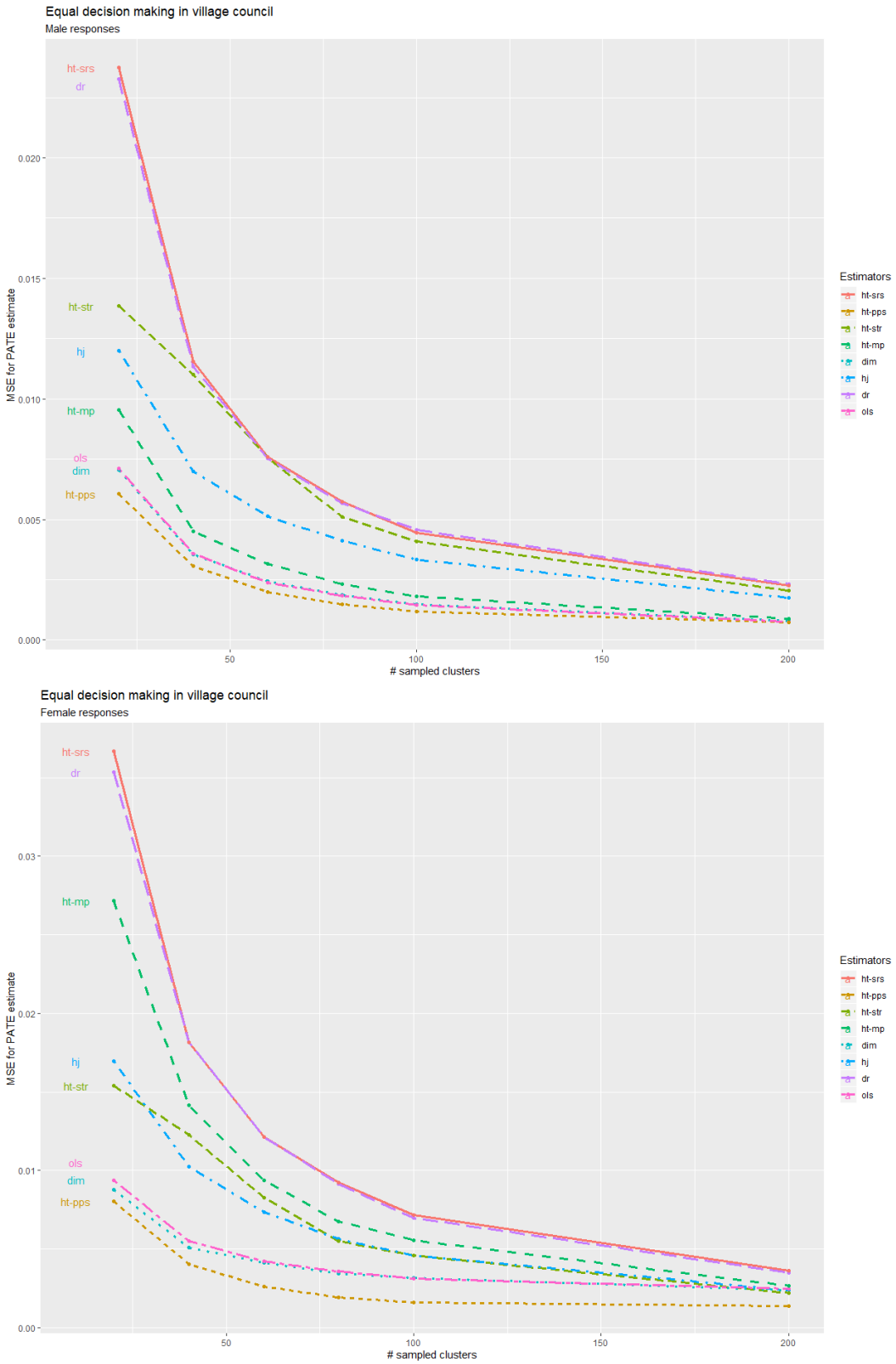

Figure 3.1: The number of sampled clusters are 20, 40, 60, 80, 100, and 200. Results are based on 10,000 simulations. The PPSWOR estimator have the lowest MSE for estimating PATE, despite the bias from being approximately PPS after sampling 40 clusters.

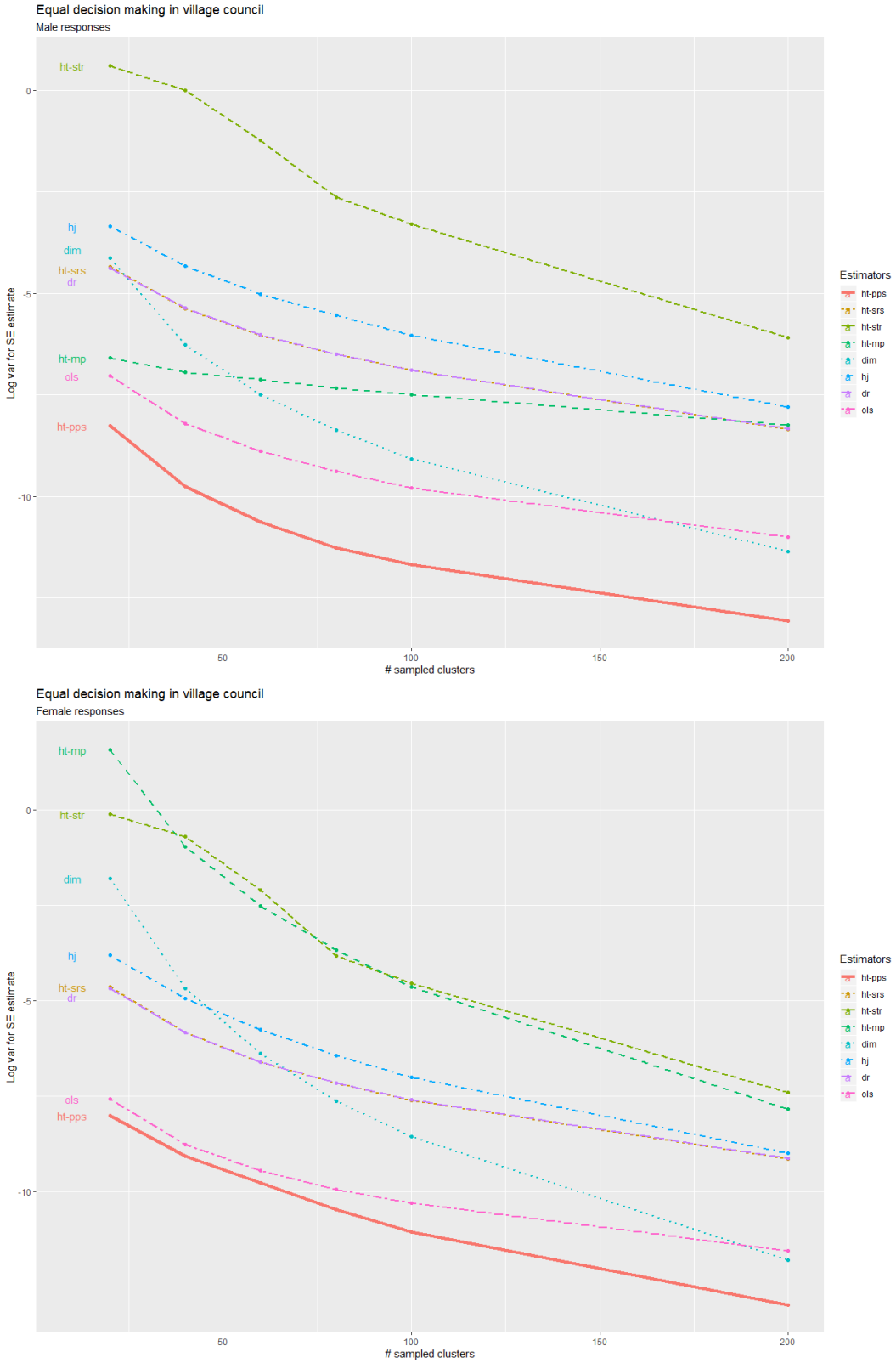

Figure 3.2: The number of sampled clusters are 20, 40, 60, 80, 100, and 200. Results are based on 10,000 simulations. Note that the OLS SE estimates are based on the robust clustered SE estimator. Despite being based on with-replacement sampling, the PPSWOR SE estimator is the most efficient, ensuring valid and more precise inferences.

estimation [1]. This ensures valid and more precise inferences compared to all other methods considered. The DIM and OLS estimators are also effective for PATE estimation albeit biased. On the contrary, the estimated SE for these estimators are more varied than for PPSWOR. Hence, while they are able to estimate the PATE efficiently (almost as good as PPSWOR), being able to get a good measure of that efficiency is lost.

3.7 Conclusion

For its logistical ease, CREs suffer from statistical inefficiency as compared to traditional unit-randomized experiments. To combat this, covariate information is employed to refine treatment estimation, and there has been numerous approaches proposed to use it productively. These approaches can be either implemented in the design (stratification, matched-pairs, or PPSWOR) or in the estimator (DIM, HJ, DR, or OLS). In our study, we evaluate these methods for estimating the PATE and the SE in CREs under the Neyman-Rubin model of potential outcomes. Based on our simulation results, we give these recommendations for designing and analyzing CREs.

- With a smaller number of clusters recruited for experimentation, choosing the best approach for effectively using covariate information is more imminent. With a larger number of clusters, the differential gain among the approaches are less pronounced when estimating PATE but not so for SE estimation. Hence, even with a large sample size of clusters, there are still reasons to carefully choose the design and analysis methods.

- Given the covariate information of cluster size, PPSWOR is the most efficient among all the included methods at estimating the PATE and SE. Therefore, it can guarantee the smallest error in estimating the PATE and be able to give a good measure of that

[1] Some of the estimated SEs for the PPSWOR designs are negative, but performance analysis is based only on the positive values.

error. However, this only holds if the sample is close to being exactly PPS. The less exact the sample is, the more bias the method will gain. In such cases, though, it is recommended that clusters be stratified first based on another covariate and then PPSWOR sampling may be conducted within strata.

- The OLS estimator is still a popular choice amongst researchers, despite its well-known drawbacks such as biased treatment estimates and negatively biased estimated SE. Clustered SEs are then often used to avoid giving under-estimated SE. While it certainly does give conservative estimates, it may be sacrificing too much precision.

- To fully answer the question of stratification/block versus matched-pair, there still needs to be more in depth research.

CREs have grown in popularity, and so the need for a guide on the best methods for designing and analyzing these experiments is more necessary. We certainly do not claim to have achieved this endeavor, but this paper strives to further advance it.

Chapter 4

analyzeCRE: an R package for analyzing cluster-randomized experiments using potential outcomes

4.1 Introduction

Cluster-randomized experiments (CRE) are gaining prevalence, particularly among political studies [Hansen and Bowers (2009), Beath et al. (2013), Avdeenko and Gilligan (2015)], public health [Hayes and Moulton (2009)), psychology (Paluck (2009), Raver et al. (2009), Cilliers et al. (2016)], education [Springer et al. (2010), Carlson et al. (2017)], economic studies [Hidrobo et al. (2014)] and medical trials [Grandes et al. (2009)]. With such a growth in popularity, there is a need to create a tool to aid in analyzing these type of experiments. To this end, the R package *analyzeCRE* is developed to analyze the population average treatment effect (PATE) for CREs under the Neyman-Rubin model of potential outcomes. To our knowledge, there are currently only two packages, *estimatr* (Blair et al., 2020) and *experiment* (Imai et al., 2019), that provides CRE analysis from a non-parametric, design-based framework.

Under a non-parametric framework, randomness in responses are assumed due to the experimental design. Therefore, analysis of the PATE requires that clusters be sampled from a population of clusters, treatment are assigned at random to sampled clusters, and units are sampled from the population of units within sampled clusters. Typically, clusters are randomly chosen with simple random sampling (SRS) to be in the CRE and then estimated with either the Horvitz-Thompson (HT) estimator or the difference-in-means (DIM) estimator. However, these estimators are less than ideal since they either are significantly inefficient (i.e., HT) or biased (i.e., DIM). Alternatively, Xiong and Higgins (2020) demonstrates that sampling the clusters with probability proportional to (cluster) size can lead to efficient and unbiased estimation of the PATE. The proposed package *analyzeCRE* is an analytic companion to this body of theoretical research. It also include the standard estimators of HT, DIM, Hajék (HJ) and Des-Raj (DR) under a SRS or stratified sampling of clusters. As such it is is a more comprehensive estimation tool compared to the other two available packages. The *estimatr* package only have the HT and DIM estimators, and the *experiment* package is limited to the matched-pairs design of CREs.

This chapter aims to showcase the functionality of the *analyzeCRE* package for estimating the PATE in CREs and to promote the applicability of the PPSWOR sampling scheme in CREs. This paper is organized as follows: in Section 4.2, a review of the the Neyman-Rubin causal model and a summary of the different estimators are given. In Section 4.3, the functions under the package are specified with a description of how to implement them. Finally, in Section 4.4, a simulation example about teachers workload is provided to illustrate the ease and capability of the *analyzeCRE* package.

4.2 Theoretical framework

Suppose there is a finite population of n units grouped into ℓ clusters. This partition of units is done without researcher intervention. Allowing cluster size, the number of units in

a cluster, to vary, let n_c be the number of units in cluster c. Denote the k^{th} unit in cluster c then as (k, c). Each unit (k, c) has a potential outcome for treatment t, representing the value of the unit *if* it receives treatment t, designated as y_{kct}. Here, $t = 1$ for treated and $t = 0$ for controlled. We use "treatment" or "treated" to sometimes refer to $t = 1$ specifically and sometimes in general to $t \in \{0, 1\}$, but the case should be clear contextually.

To assess the efficacy of a treatment, researchers are often interested in the population average treatment effect (PATE),

$$\delta = \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kc1} - y_{kc0}}{n} = \mu_1 - \mu_0, \tag{4.1}$$

where

$$\mu_t = \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kct}}{n}. \tag{4.2}$$

However, as a unit is either given treatment or control and not both, only y_{kc1} or y_{kc0} can be known for each unit. In CREs, this is entirely dependent on the treatment assignment of the unit's cluster, which is formally summarized in the Neyman-Rubin model of responses. Let T_{ct} be the cluster treatment indicator variable—that is, $T_{ct} = 1$ if cluster c is assigned treatment t and $T_{ct} = 0$ if not. A unit's response, following the Neyman-Rubin model, can then be written as:

$$Y_{kc} = y_{kc1} T_{c1} + y_{kc0} T_{c0} \tag{4.3}$$

$$= y_{kc1} T_{c1} + y_{kc0}(1 - T_{c1}). \tag{4.4}$$

Since the PATE requires that both potential outcomes be known for each unit, the PATE is not observable and must be estimated. Estimating it with a CRE involves at least three stages of randomization: sampling clusters, assigning treatments, and sampling units within clusters. To represent each of the randomization stage, we define three indicator variables. Let S_c be the cluster sampling indicator, T_{ct} be the cluster treatment indicator, and S_{kc} be the

unit sampling indicator. Estimation of the PATE then depends on how the randomization is specified in each stage. We primarily allow the cluster sampling scheme to change and assume that treatments are randomized completely at random to sampled clusters (except for stratified sampling, in which treatments are assigned within stratum) and that units are sampled within sampled clusters via SRS. Denote $\#T_t$ as the number of clusters given treatment t and s_c as the number of units sampled from cluster c.

HT estimator under PPSWOR

Cluster size is known to be a crucial covariate to consider and include when analyzing CREs as it frequently serves as a substitute for other more predictive, albeit unknown, covariates. Therefore, accounting for varying cluster size when sampling them will help to improve estimation efficiency. A probability-proportional-to-size without-replacement (PPSWOR) sample will select a cluster with a probability of $n_c s/n$, where s is a fixed sample size of clusters. The HT estimator under PPSWOR sampling, referred to as HT-PPS, for PATE is then

$$\hat{\delta}_{\text{HT,PPS}} = \sum_{c=1}^{\ell} \frac{S_c T_{c1}}{\#T_1} \sum_{k=1}^{n_c} \frac{y_{kc1} S_{kc}}{s_c} - \sum_{c^*=1}^{\ell} \frac{S_{c^*} T_{c^*0}}{\#T_0} \sum_{k^*=1}^{n_{c^*}} \frac{y_{k^*c^*0} S_{k^*c^*}}{s_{c^*}}. \tag{4.5}$$

The HT-PPS estimator is unbiased and precise in estimating PATE.

However, it is usually criticized for two things: (i) there is not an approach that can get an exact PPSWOR sample of clusters, and (ii) the variance estimation depends on the joint cluster inclusion probabilities $\pi_{cc'}$ for $c' \neq c$, which are not always readily available. For the first criticism, Sunter (1977) and Sunter (1986) outline two procedures to get a PPSWOR sample of a general fixed size. The Sunter sampling methods will give exact PPSWOR sample if all clusters have no more than n/s units; otherwise, the methods will give an approximate sample that can incur estimation bias. For the second, the with-replacement (WR) is supplied, which do not need the $\pi_{cc'}$. The SYG PPSWOR variance estimator is also given, though this will require the $\pi_{cc'}$. Exact formulation of these variance estimators are given in Appendix A.

HT estimator under SRSWOR

The HT estimator assuming clusters are sampled with SRSWOR (HT-SRS) for PATE is defined as:

$$\hat{\delta}_{\text{HT,SRS}} = \ell \sum_{c=1}^{\ell} \frac{S_c T_{c1}}{\#T_1} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kc1} S_{kc}}{s_c} - \ell \sum_{c^*=1}^{\ell} \frac{S_{c^*} T_{c^*0}}{\#T_0} \frac{n_{c^*}}{n} \sum_{k^*=1}^{n_{c^*}} \frac{y_{k^*c^*0} S_{k^*c^*}}{s_{c^*}}. \tag{4.6}$$

Unbiasedness of the HT-SRS estimator surely recommends it as an appropriate estimator of PATE, but it can be criticized on two counts. The first is that the estimator is known to have huge variability since it does not account for varying cluster size. Larger clusters will have greater sums of responses whereas smaller clusters will have smaller sums. The second is that it is not location-invariant. The HT-SRS estimator will only be location-invariant when the number of treated *units* (not clusters) is equal to the number of units assigned to control, something of which researchers cannot control. Lack of location-invariance will not affect the unbiasedness of the estimator, even for linearly transformed outcomes, but this problem presents itself in the variance calculation. The variance will change as a changes. The conservative SYG variance estimator for the HT-SRS estimator is given in Appendix B.

Des-Raj estimator under SRSWOR

The DR estimator under a SRSWOR of clusters for PATE can be written as (Middleton and Aronow, 2015):

$$\begin{aligned}\hat{\delta}_{\text{DR,SRS}} =& \ell \sum_{c=1}^{\ell} \frac{S_c T_{c1}}{\#T_1} \frac{n_c}{n} \left[\sum_{k=1}^{n_c} \frac{y_{kc1} S_{kc}}{s_c} - \frac{\theta}{n_c} \left(n_c - \frac{n}{\ell} \right) \right] \\ & - \ell \sum_{c^*=1}^{\ell} \frac{S_{c^*} T_{c^*0}}{\#T_0} \frac{n_{c^*}}{n} \left[\sum_{k^*=1}^{n_{c^*}} \frac{y_{k^*c^*0} S_{k^*c^*}}{s_{c^*}} - \frac{\theta}{n_{c^*}} \left(n_{c^*} - \frac{n}{\ell} \right) \right]\end{aligned} \tag{4.7}$$

where θ is a regression coefficient. Note that the first part of the DR estimator is the HT-SRS estimator. To alleviate the faults of the HT-SRS esimator but retain unbiasedness, the DR estimator includes a regression component on cluster size. The DR estimator is

also generally less varied than the HT-SRS as the cluster size covariate directly explains for the variability intrinsic to the HT-SRS estimator. Unfortunately, the solution itself poses a problem. Estimating the regression coefficient θ will biased the estimator. Having an estimate of θ prior to the experiment will eliminate the bias, but this is often not feasible. The conservative SYG variance estimator for the DR estimator is given in Appendix E.

Difference-in-means estimator under SRSWOR

The DIM estimator under SRSWOR sampling of clusters for PATE is:

$$\hat{\delta}_{\text{DIM,SRS}} = \frac{\sum_{c=1}^{\ell} S_c T_{c1} \sum_{k=1}^{n_c} y_{kc1} S_{kc}}{\sum_{c=1}^{\ell} S_c T_{c1} s_c} - \frac{\sum_{c^*=1}^{\ell} S_{c^*} T_{c^*0} \sum_{k^*=1}^{n_{c^*}} y_{k^*c^*0} S_{k^*c^*}}{\sum_{c^*=1}^{\ell} S_{c^*} T_{c0} s_{c^*}}. \tag{4.8}$$

The DIM estimator, being elegantly simple, is favored among many CRE experiments. Furthermore, contrary to the HT-SRS estimator, it is efficient and invariant to location shifts. However, it is biased, i.e., it does not estimate the PATE. The variance for the DIM estimator is also difficult to write in a closed form. Therefore, a Taylor series linear approximation is applied and the variance estimator is based on that approximation. The conservative variance estimator for the DIM estimator is given in Appendix C.

Hajék estimator under SRSWOR

The HJ estimator under SRSWOR sampling of clusters for PATE is (Hajek, 1971):

$$\hat{\delta}_{\text{DIM,SRS}} = \frac{\sum_{c=1}^{\ell} S_c T_{c1} \frac{n_c}{s_c} \sum_{k=1}^{n_c} y_{kc1} S_{kc}}{\sum_{c=1}^{\ell} S_c T_{c1} n_c} - \frac{\sum_{c^*=1}^{\ell} S_{c^*} T_{c^*0} \frac{n_{c^*}}{s_{c^*}} \sum_{k^*=1}^{n_{c^*}} y_{k^*c^*0} S_{k^*c^*}}{\sum_{c^*=1}^{\ell} S_{c^*} T_{c0} n_{c^*}}. \tag{4.9}$$

Similar to the DIM estimator, the HJ estimator is also biased for PATE, and the variance estimation is also based on Taylor series approximation. The conservative variance estimator for the HJ estimator is given in Appendix D.

Estimation under stratification

Suppose there are m strata with ℓ_u clusters and n_u units in stratum u. Stratum can be defined based on a categorical or discretized numeric variable. In a stratified CRE, all three randomization stages occur within each stratum and independently across strata. In other words, a CRE is independently conducted in each stratum. Estimation of the PATE under stratification is then a straight-forward application of the above methods and is given by

$$\hat{\delta}_{\text{ST}} = \sum_{u=1}^{m} \frac{n_u}{n} \hat{\delta}_u \tag{4.10}$$

$$\widehat{\text{Var}}\left(\hat{\delta}_{\text{ST}}\right) = \sum_{u=1}^{m} \frac{n_u^2}{n^2} \widehat{\text{Var}}\left(\hat{\delta}_u\right) \tag{4.11}$$

where the stratum estimator, $\hat{\delta}_u$, can take on any of the forms we already mentioned above, depending on the design used and estimator chosen. Currently, only the HT-SRS is supported with stratification.

4.3 *analyzeCRE* functions

The *analyzeCRE* R package contains the most common design-based estimators for analyzing CREs as outlined in Section 4.2. Table 4.1 list the functions available in the package and provides a simple description of each. We will then explain in more detail how to implement these functions.

The first group in Table 4.1 are the estimation functions. Based on how clusters are sampled and treatments are randomized to clusters, one of these functions can be selected to provide estimates of the PATE and the variance. Note that for `STRTHTvar()`, treatment must be randomized within the defined strata/pairs. If units are sampled within clusters, they must be sampled with SRSWOR for all functions. The functions require that the observed data be summarized already by clusters. Then the input data for the estimation functions are:

Function	Description
`PPSest()`	Calculates the estimated PATE and variance using the HT estimator assuming that clusters are sampled via PPSWOR
`HTest()`	Calculates the estimated PATE and variance using the HT estimator assuming that clusters are sampled via SRSWOR
`STRTHTest()`	Calculates the estimated PATE and variance using the HT estimator assuming that clusters are sampled via SRSWOR within defined strata and treatments are assigned within strata/block
`MPHTest`	Calculates the estimated PATE and variance using the HT estimator assuming that pairs of clusters are sampled via SRSWOR and treatments are assigned within pairs
`HJest()`	Calculates the estimated PATE and variance using the Hajék estimator assuming that clusters are sampled via SRSWOR
`DIMest()`	Calculates the estimated PATE and variance using the DIM estimator assuming that clusters are sampled via SRSWOR
`DRest()`	Calculates the estimated PATE and variance using the DR estimator assuming that clusters are sampled via SRSWOR
`getSampleClust()`	Samples clusters via Sunter's PPSWOR, SRSWOR, or stratified sampling
`getTrtClust()`	Conducts treatment randomization completely-at-random, within defined blocks, within pairs, or on pairs
`getWithinSample()`	Conducts within cluster sampling of units via SRSWOR
`getSampleData()`	Combines the `getSampleClust()`, `getTrtClust()`, and `getWithinSample()` to get sample data

Table 4.1: Summary of functions in the *analyzeCRE* package

- `n.clust`: number of clusters in the population
- `n.units`: number of units in the population; not necessary for `DIMvar()` and `HJvar()`
- `clustersum`: sum of responses for each sampled cluster
- `clustervar`: variance of responses within each sampled cluster
- trtindex: treatment indicator for each sampled cluster; t=1 for treated, t=0 for control
- `clustersize`: number of units in each sampled cluster
- `withinsampsize`: number of sampled units from each sampled cluster
- `ppsvartype`: indication of which variance estimator should be used when estimating the HT-PPS variance; choose from "wor" (SYG without-replacement) or the default "wr" (with-replacement); only need for `PPSvar()`
- `jointprob`: an $\ell \times \ell$ matrix of joint cluster inclusion probabilities; only need if `ppsvartype` is selected as "wor"
- `opt.theta`: optimum theta for the DR estimator if known; defaults to NULL and will be estimated with sample data; only need for `DRvar()`.

The estimation functions will output a labeled vector of size 2, in which the first entry is the PATE estimate and the second is the SE estimate.

Unlike other packages, this package provides two different variance estimators for the HT-PPS estimator. The WR variance estimators do not require the joint cluster inclusion probabilities. Therefore, users can still get an estimate for the variance even if they do not have the joint probabilities, which is the more realistic case. However, users can still opt to use the SYG without-replacement variance estimator if they are fortunate enough to have the joint probabilities or wish to approximate them. R packages such as *jipApprox* (Sichera, 2019) can be used to approximate the joint probabilities analytically or through simulations.

While the focus of the package are the analysis functions, we also include functions that help with the design side. After all, successful analysis is a reflection of designing proficiently. These functions are in the second group of Table 4.1. The `getSampleClust()`, `getTrtClust()`, and `getWithinSample()` functions constitutes the randomization stages of a CRE. They can be employed individually or together in the `getSampleData()` function. The inputs for the latter function (the previous ones have the same inputs) are:

- `clusterid`: an $\ell \times 1$ vector labeling each cluster
- `clustsize`: an $\ell \times 1$ vector indicating the number of units in each cluster
- `samp.method`: choose between "SRS," "PPS" (using Sunter's algorithm), or "STRTSRS-BLK" (SRSWOR of clusters within strata); default is "SRS."
- `stratum`: name of categorical variable used for stratification; only needed if `samp.method` is selected as "STRTSRS-BLK"
- `c.stratsize`: an $m \times 1$ vector indicating the number of clusters in each stratum; only needed if `samp.method` selected as "STRTSRS-BLK"
- `strat.sampsize`: an $m \times 1$ vector indicating the number of clusters to be sampled from each stratum; only needed if `samp.method` selected as "STRTSRS-BLK"
- `no.sampclust`: number of clusters to be sampled
- `no.trtclust`: number of treated clusters; defaults to half of `no.sampclust`
- `perc`: a proportion indicating what percentage of the cluster units should be sample; default is `perc=1` in which all units are sampled from a cluster
- `withinsampsize`: (optional) an $\ell \times 1$ vector indicating how many units to sample from each cluster; alternative to `perc`
- `unitid`: (optional) an $n \times 1$ vector labeling units in the population

The function(s) will randomly select clusters based on the method specified in `samp.method`, randomly assign treatment to the sampled clusters, and randomly select units from each sampled clusters. It will then output a data frame of sampled units designating their cluster, the treatment assignment of cluster, and the cluster size.

4.4 Example

The database *teachers* contains information about the workload of teachers in the Maricopa County, AZ public schools in 1994. These databases are available in the R package *SDaA* (Verbeke, 2014). In this example, the 310 teachers are the units of interest and the 31 schools serve as the clusters. Suppose that a PPSWOR sample of 26 schools, with half to receive a treatment, is desired for a cluster-randomized experiment. The function `getSampleData` can help with setting up this design:

```
> teachers = SDaA::teachers
> clustersize = table(teachers$school)
> schoollab = sort(unique(teachers$school))
> sampledata = getSampleData(clusterid = schoollab,
+     clustsize = clustersize, no.sampclust = 26,
+     samp.method = "PPS")
> sampledata[1:6,]
  clusterid.unit unitid clustsize sampclustindex trtclustindex
               2      2         4              1             0
               2      3         4              1             0
               2      4         4              1             0
               2      5         4              1             0
               3      6         7              1             1
               3      7         7              1             1
```

This output excerpt informs the user that schools 2 and 3 are sampled, and that school 2 is control while school 3 is treated. In this case, all units were sampled from each treated cluster, so all four teachers (unitid 2-5) from school 2 are included.

Suppose that the treatment is expected to reduce the number of hours teachers are working at school. To represent this, we simulate the potential outcomes as

```
> set.seed(838996)
> rows.na = which(is.na(teachers$hrwork))
> teachers = teachers[-rows.na,]
> schoollab = sort(unique(teachers$school))
> clustersize = as.numeric(table(teachers$school))
> trtresp =  0.894*teachers$hrwork - 3 + rnorm(n = 307)
> contresp = teachers$hrwork + rnorm(307)
> pate = mean(trtresp - contresp)
> pate
[1] -6.505042
```

The first two lines eliminates teachers who did not report the number of hours they worked, leaving 307 teachers in our supposed population. Assuming that there is treatment heterogeneity, the simulated average treatment effect for the population is -6.51. It is reasonable to think that the response of hours worked might be correlated to the number of teachers at a school, and so doing a CRE with PPSWOR sampling of clusters can produce a more efficient estimate. Running the code

```
> ppsest = PPSvar(n.clust = 31, n.units = 307,
+    clustersum = sampledata$clustsum,
+    clustervar = sampledata$clustvar, trtindex = sampledata$trtindex,
+    clustersize = sampledata$clustersize,
+    withinsampsize = sampledata$withinsampsize)
```

```
> ppsest
     PATE.EST     SE.EST
     -7.078663    2.379460
```

gives a PATE estimate of -7.1 with an estimated standard error of 2.38. This estimate is comparable to the true simulated PATE of -6.51.

4.5 Conclusion

Cluster-randomized experiments, though logistically easier to implement and more cost-effective, are challenging analytically. Standard methods available are either biased or drastically inefficient. On the other hand, sampling clusters with probability-proportional-to-size can alleviate these undesirable properties. The HT estimator of PATE under PPSWOR sampling has been proven to be unbiased and as efficient as the favored difference-in-means estimator. It has also been shown that the WR variance estimator, while conservative, will give more precise standard error estimates without the need for the joint cluster inclusion probabilities.

The few existing R packages that focus on analyzing CREs do not have the HT-PPS as an option to estimate PATE, or if they do, they do not provide an alternative variance estimator. Either the user must specify the joint probabilities or the variance will be calculated under strong assumptions such as the constant treatment effect assumption. The package *analyzeCRE* is created to fill this gap. The package also include many of the standard design-based methods, but it is the authors' hope that with the *analyzeCRE* package, researchers will be more inclined to implement PPSWOR sampling in CREs.

Chapter 5

Concluding Remarks

5.1 Summary of dissertation

Experiments, in the traditional sense, usually involve a limited number of experimental units in a strictly controlled environment. However, with advances in technology and, we hope, increased understanding in causal inference, researchers are constantly pushing these limitations further back. In one such area is social science and public health, where experiments can be widely conducted, even at a national level. Such experiments, referred to as either field or social experiments, typically require the use of cluster-randomized experiments (CREs), in which the treatments are randomized not to individual units but to a community (commonly called cluster) of units. This not only help with logistic but also with budgeting. Certainly, CREs do not fit the traditional definition of statistical experiments, and thus, conventional methodologies will not be appropriate for analyzing them. For example, ordinary least squares (OLS) is established with the supposition that treatments are assigned to units. Using these methods will seriously underestimate the standard error, leading to spurious inferences, since CREs are known to be less efficient. Consequently, analyzing CREs are not as forthright.

Many methodologists have addressed this problem, using tools from both the parametric

and non-parametric sides. In this dissertation, we favor the non-parametric approach and work under the Neyman-Rubin causal model of potential outcomes. Under this model, randomness in responses are a direct result of experimental design, which is felicitous for estimating the population average treatment effect in CREs. Being able to quantify the average causal effect of a treatment from a sample and then projecting it to the population involves the use of randomization in the CRE design. Explicitly, it entails randomly sampling clusters from a population of clusters, randomly assigning treatments to sampled clusters, and randomly selecting units from the within cluster population of units.

The standard CRE design would choose clusters randomly with a simple random sample (SRS) scheme, assign treatments completely at random to sampled clusters, and sample units via SRS within sampled clusters. Then either a Horvitz-Thompson (HT-SRS) estimator or a difference-in-means (DIM) estimator, incorporating the SRS design, is used to estimate the PATE. However, these methods are less than ideal as the HT-SRS estimator is significantly inefficient and the DIM estimator is biased. Other approaches have also been suggested to help alleviate the large variability in the unbiased HT-SRS estimator. These include design changes such as stratifying/blocking clusters and pair matching clusters or analytical alterations such as the Des-Raj estimator.

The purpose of this dissertation is to render a thorough discourse on the design and analysis of CREs and to proffer an alternative sampling design. In Chapter 2, we propose that clusters be sampled with the probability-proportional-to-size (PPS) procedure instead of SRS. The HT estimator under this design, referred to as HT-PPS, maintains the unbiasedness of the estimator while decreasing the inefficiency that it suffers under SRS, all while staying genuine to the design-based framework. That being said, the HT-PPS estimator comes with its own challenges. The first being that getting an exact PPS without-replacement sample of clusters is difficult when some clusters dominate cluster sizes (e.g., large metropolises). The second being that variance estimation depends on knowledge of the joint cluster inclusion probabilities, which is frequently not possible. Confronting these

two issues, we advocate using Sunter's list-sequential algorithm to get a PPSWOR sample of clusters and present a variance estimators for the HT-PPS estimator that is free from the joint inclusion probabilities.

Chapter 2 serves as an introduction to the HT-PPS estimator, but in Chapter 3 we place the HT-PPS estimator in scrutiny, juxtaposing it with the other favored design methods: stratification and pair matching. There is an ongoing, and somewhat heated, debate about the merits of stratifying/blocking versus pair matching. A secondary goal is to get some insight on this question while the overall objective is to compare choices of designs and estimators to see which is the most effective at incorporating covariate information. While we did not get any additional intuition on which is the best between stratifying/blocking and pair matching, we may have settled the question by proclaiming that PPSWOR is the best, outperforming both in terms of having the lowest MSE when estimating PATE and the lowest variance when estimating the SE. As a matter of fact, the DIM is the only estimator that responds similarly to the HT-PPS for PATE estimation, but it is not as proficient in estimating its SE.

Finally, in Chapter 4, we present a guide to our R package *analyzeCRE*. This package is a complementary companion to the theoretical work completed in this dissertation. It contains functions that will execute estimation of both the PATE and variance for the HT-PPS, HT-SRS, DIM, HJ, and DR and stratified estimators. With this freely available software and a vignette on how to utilize it, the HT-PPS will hopefully be more tractable, and researchers will be more persuaded to adopt the HT-PPS estimator in place of the DIM and OLS (with clustered SE) estimators.

5.2 Future Research

Blocking is a common approach to improve estimation of treatment effects. In fact, it may be the most common since hardly any experiment is ever conducted without some

form of blocking, especially in CREs with its given lack of efficiency. However, the theory of blocking always supposes this existential population of blocks, from which a sample of blocks are taken. In practice, though, researchers do not sample blocks, and yet, they are interested in generalizing results beyond the blocks they do have. It would make more sense for researchers to sample clusters and then grouped similar clusters into a block based on a covariate. With the growth in computation power, it is becoming the norm to optimally form blocks using a distance metric on multiple covariates. Therefore, analysis of this design needs to account for the increased error in sampling clusters instead of blocks and the increased precision of optimal blocking. Besides blocking, one can also apply Neyman's allocation to the within-cluster sample size to minimize the total variability.

Even with the SYG and with-replacement SE estimators for the PPSWOR design, it is possible to get undefined SE estimates (from negative variance estimates). Further adjustments are needed then to prevent frequent undefined SE estimates.

Lastly, the work in this dissertation focuses solely on estimation, but the natural next step would be inferential analysis. This includes hypothesis testing, power calculations, and sample size computations. Central to all of these areas is the intracluster correlation (ICC), which measures how similar responses are within cluster. Unfortunately, the ICC is difficult to calculate when cluster size varies. There is still much research that needs to be done in these areas, particularly when applied to CREs.

Bibliography

P. M. Aronow and J. Middleton. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1), 2013. doi: 10.1515/jci-2012-0009.

P. M. Aronow and C. Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017. doi: 10.1214/16-AOAS1005. URL https://projecteuclid.org/euclid.aoas/1514430272.

A. Avdeenko and M. J. Gilligan. International interventions to build social capital: Evidence from a field experiment in Sudan. *American Political Science Review*, 109(3):427449, 2015. doi: 10.1017/S0003055415000210.

A. Beath, F. Christina, and R. Enikolopov. Empowering women through development aid: Evidence from a field experiment in Afghanistan. *American Political Science Review*, 107 (03):540 – 557, 2013. doi: 10.1017/S0003055413000270. URL http://www.jstor.org/stable/43654923.

Y. G. Berger and Y. Till. *Sampling with Unequal Probabilities*, volume 29A, chapter 2, pages 39 – 54. Elsevier B. V., 2009.

G. Blair, J. Cooper, A. Coppock, M. Humphreys, and L. Sonnet. *estimatr: Fast Estimators for Design-Based Inference*, 2020. URL https://CRAN.R-project.org/package=estimatr. R package version 0.22.0.

Joan Fisher Box. R. A. Fisher and the design of experiments, 1922-1926. *The American*

Statistician, 34(1):1–7, 1980. ISSN 00031305. URL http://www.jstor.org/stable/2682986.

K. R. W. Brewer and M. Hanif. *Sampling with Unequal Probabilities (Lecture Notes in Statistics)*. Springer, 1982.

M. L. Bruce, T. R. T. Have, C. F. Reynolds, H. C. Schulberg, B. H. Mulsant, G. K. Brown, G. J. McAvay, J. L. Pearson, and G. S. Alexopoulos. Reducing sucidal ideation and depression symtoms in depressed older primary care patients: A randomized controlled trial. *JAMA*, 291(09):1081 – 1091, 2004. doi: 10.1001/jama.291.9.1081. URL https://jamanetwork.com/journals/jama/fullarticle/198310.

D. Carlson, G. Borman, and M. Robinson. A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, 33(3):378–398, July 2017. doi: 10.3102/0162373711412765.

J. Cilliers, O. Dube, and B. Siddiqi. Reconciling after civil conflict increases social capital but decreases individual well-being. *Science*, 352(6287):787–794, 2016. ISSN 0036-8075. doi: 10.1126/science.aad9682. URL http://science.sciencemag.org/content/352/6287/787.

W. G. Cochran. *Sampling Techniques*. Wiley, 1977.

J. Cornfield. Randomized by group: A formal analysis. *American Journal of Epidemiology*, 108(02):100 – 102, 1978. doi: 10.1093/oxfordjournals.aje.a112592.

A. Donner. Some aspects of the design and analysis of cluster randomization trials. *Journal of the Royal Statistical Society*, 47(01):95 – 113, 1998. doi: 10.1111/1467-9876.00100.

A. Donner and N. Klar. Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, 94(03):416 – 422, 2004. doi: 10.2105/AJPH.94.3.416.

D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2016. doi: 10.1515/jci-2015-0021.

Lorenzo Fattorini. An adaptive algorithm for estimating inclusion probabilities and performing the Horvitz–Thompson criterion in complex designs. *Computational Statistics*, 24(4):623, Mar 2009. ISSN 1613-9658. doi: 10.1007/s00180-009-0149-9. URL https://doi.org/10.1007/s00180-009-0149-9.

Sir Fisher, R. A. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513, 1926. URL http://hdl.handle.net/2440/15191. (Appears in *Fisher: Collected papers relating to statistical and mathematical theory and applications*).

M. H. Gail, S. D. Mark, R. J. Carroll, S. B. Green, and D. Pee. On design considerations and randomizaton-based inference for community intervention trials. *Statistics in Medicine*, 15(11):1069 – 1092, 1996. doi: 10.1002/(SICI)1097-0258(19960615)15:11⟨1069::AID-SIM220⟩3.0.CO;2-Q.

A. S. Gerber and D. P. Green. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, 94(3):653 – 663, September 2000. doi: 10.2307/2585837.

A. S. Gerber and D. P. Green. Do phone calls increase voter turnout?: A field experiment. *Public Opinion Quarterly*, 65(1):75–85, 2001. doi: 10.1177/0002716205278445.

A. S. Gerber and D. P. Green. Correction to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005). *American Political Science Review*, 99(2):301–313, May 2005. doi: 10.1017/S000305540505166X.

G. Grandes, A. Sanchez, R. O. Sanchez-Pinilla, J. Torcal, I. Montoya, K. Lizarraga, J. Serra, and PEPAF Group. Effectiveness of physical activity advice and prescription by physicians

in routine primary care: A cluster randomized trial. *Arch Intern Med.*, 169(7):694–701, April 2009. doi: 10.1001/archinternmed.2009.23.

D. P. Green, A. S. Gerber, and D. W. Nickerson. Getting out the vote in local elections: results from six door-to-door canvassing experiments. *The Journal of Politics*, 65(4):1083–1096, November 2003. doi: 10.1111/1468-2508.t01-1-00126.

J. Hajek. Discussion of 'An essay on the logical foundations of survey sampling, part one,' by D. Basu. In V. P. Godambe and D. A. Sprout, editors, *Foundations of Statistical Inference*, page 236. Holt, Rhinehart, and Winston, 1971.

B. B. Hansen and J. Bowers. Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statisical Association*, 104(487):873–885, 2009. doi: 10.1198/jasa.2009.ap06589.

B. B. Hansen, P. R. Rosenbaum, and D. S. Small. Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Society*, 109(505), 2014. doi: 10.1080/01621459.2013.863157.

M. H Hansen and W. N. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(04):333 – 362, 1943. doi: 10.1214/aoms/1177731356.

T. V. Hanurav. Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society*, pages 374 – 391, 1967.

R. Hayes and L. Moulton. *Cluster Randomised Trials*. Chapman & Hall/CRC, 2009.

M. Hidrobo, J. Hoddinott, A. Peterman, A. Margolies, and V. Moreira. Cash, food, or vouchers? Evidence from a randomized experiment in northern Ecuador. *Journal of Development Economics*, 107:144–156, March 2014. doi: 10.1016/j.jdeveco.2013.11.009.

Matt D T Hitchings, Marc Lipsitch, Rui Wang, and Steven E Bellan. Competing effects of indirect protection and clustering on the power of cluster-randomized controlled vaccine

trials. *American Journal of Epidemiology*, 187(8):1763–1771, 2018. doi: 10.1093/aje/kwy047. URL http://dx.doi.org/10.1093/aje/kwy047.

P. W. Holland. Statisics and causal inference. *American Statistical Association*, 81(396):945 – 960, 1986.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663 – 685, 1952. doi: 10.1080/01621459.1952.10483446.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of American Statisical Association*, 103(482):832–842, 2008. doi: 10.1198/016214508000000292.

K. Imai. Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, May 2005. doi: 10.1017/S0003055405051658.

K. Imai, G. King, and E. Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171 (2):481–502, 12 2007.

K. Imai, G. King, and C. Nall. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science*, 24(01):29 – 53, 2009. doi: 10.1214/08-STS274.

K. Imai, Z. Jiang, and M. Li. *experiment: R Package for Designing and Analyzing Randomized Experiments*, 2019. URL https://CRAN.R-project.org/package=experiment. R package version 1.2.0.

G. W. Imbens. Experimental design for unit and cluster randomized trials. 2011. URL http://cyrussamii.com/wp-content/uploads/2011/06/Imbens_June_8_paper.pdf.

R. Jagadeesan, N. Pillai, and A. Volfovsky. Designs for estimating the treatment effect in networks with interference. *ArXiv e-prints*, May 2017.

S. M. Kerry, F. P. Cappuccio, L. Emmett, J. Plange-Rhule, and J. B. Eastwood. Reducing selection bias in a cluster randomized trial in West African villages. *Clinical Trials*, 02 (02):125 – 129, 2005. doi: 10.1191/1740774505cn074oa.

G. King, E. Gakidou, N. Ravishankar, R. T. Moore, J. Lakin, M. Vargas, M. M. Tllez-Rojo, vila J. E. H., M. H. vila, and H. H. Llamas. A "politically robust" experiment design for public policy evaluation, with application to the Mexican universal health insurance program. *Journal of Policy Analysis and Management*, 26(03):479 – 506, 2007. doi: 10.1002/pam.20279.

M. S. Kramer, F. Aboud, E. Mironova, and et. al. Breastfeeding and child cognitive development: New evidence from a large randomized trial. *Arch Gen Psychiatry*, 65(5):578–584, May 2008. doi: 10.1001/archpsyc.65.5.578.

J. D. Lewsey. Comparing completely and stratified randomized design in cluster randomized trials when the stratifying factor is the cluster size: A simulation study. *Statistics in Medicine*, 23(06):897 – 905, 2004. doi: 10.1002/sim.1665.

S. L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, 2nd edition, 2010.

W. G. Manning, J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz. Health insurance and the demand for medical care: Evidence from a randomized experiment. *The American Economic Review*, 77(3):251–277, June 1987. URL http://www.jstor.org/stable/1804094.

D. C. Martin, P. Dieher, E. B. Perrin, and T. D. Koepsell. The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine*, 12(3-4): 329–338, February 1993. doi: 10.1002/sim.4780120315.

J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6:39 – 75, 2015. doi: 10.1515/spp-2013-0002.

L. W. Miratrix, J. S. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of Royal Statisical Society*, 75(02):369 – 396, 2013. doi: 10.1111/j.1467-9868.2012.01048.x.

E. L. Paluck. Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology*, 96(03):574, 2009. doi: 10.1037/a0011989.

C. C. Raver, S. M. Jones, C. Li-Grining, F. Zhai, M. W. Metzger, and B. Soloman. Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 77(02):302, 2009. doi: 10.1037/a0015302.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(05):688, 1974. doi: 10.1037/h0037350.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880.

R. Sichera. *jipApprox: Approximate Inclusion Probabilities for Survey Sampling*, 2019. URL https://CRAN.R-project.org/package=jipApprox. R package version 0.1.2.

B. K. Sinha. On sampling schemes to realize preassigned sets of inclusion probabilities of first two orders. *Calcutta Statistical Association Bulletin*, 22:89 – 110, 1973. doi: 10.1177/0008068319740103.

D. S. Small, T. R. T. Have, and P. R. Rosenbaum. Randomization inference in a group-randomized trial of treatments for depession: Covariate adjustment, noncompliance, and

quantile effects. *Journal of the American Statistical Association*, 103(481):271 – 279, 2008. doi: 10.1198/016214507000000897.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4): 465–472, 1923. ISSN 08834237. URL http://www.jstor.org/stable/2245382. (Translated in 1990).

M. G. Springer, D. Ballou, L. S. Hamilton, V. Le, J. R. Lockwood, D. F. McCaffrey, M. Pepper, and B. Stecher. Teacher pay for performance: Experimental evidence from the project on incentives in teaching. Technical report, Vanderbilt University, 2010. URL https://www.rand.org/pubs/reprints/RP1416.html.

A. Sunter. Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54(01):33 – 50, 1986. doi: 10.2307/1403257.

A. B. Sunter. List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26(3):261–268, 1977.

D. L. Sussman and E. M. Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *ArXiv e-prints*, February 2017.

T. Verbeke. *SDaA: Sampling: Design and Analysis*, 2014. URL https://CRAN.R-project.org/package=SDaA. R package version 0.1-3.

K. Vijayan. An exact πps sampling scheme-generalization of a method of Hanurav. *Journal of the Royal Statistical Society*, pages 556 – 566, 1968.

L. Wantchekon. Clientelism and voting behavior: Evidence from a field experiment in Benin. *World Politics*, 55(03):399 – 422, 2003. doi: 10.1353/wp.2003.0018.

Y. Xiong and M. J. Higgins. The benefits of probability-proportional-to-size sampling in cluster-randomized experiments, 2020.

Appendix A

Properties of HT-PPS estimator

Let S_c be the indicator variable for sampling clusters, T_{ct} be the assignment indicator for treatment t, and S_{kc} be the indicator for sampling units. We begin by computing expectations, variances, and covariances of these indicators under PPSWOR sampling of clusters. These will be helpful in examining the properties of the HT-PPS estimator for PATE.

A.1 Indicator properties under PPS

Define $\pi_{cc'} \equiv E(S_c S_{c'}) = P(S_c = 1, S_{c'} = 1)$ as the probability of sampling both cluster c and c'.

$$\mathbb{E}(S_c|\#T_t) = \frac{n_c s}{n} \tag{A.1}$$

$$\begin{aligned}
\mathbb{E}(S_c^2 T_{ct}^2|\#T_t) &= \mathbb{E}(S_c T_{ct}|\#T_t) = \mathbb{E}(S_c \mathbb{E}(T_{ct}|\mathbf{S})|\#T_t) \\
&= \mathbb{E}\left(S_c \frac{\#T_t}{s} \middle| \#T_t \right) = \frac{\#T_t}{s}\mathbb{E}(S_c|\#T_t) \\
&= = \frac{n_c \#T_t}{n}
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
\mathbb{E}(S_c S_{c'} T_{ct} T_{c't}|\#T_t) &= \mathbb{E}(S_c S_{c'} \mathbb{E}(T_{ct} T_{c't}|\mathbf{S})|\#T_t) \\
&= \mathbb{E}\left(S_c S_{c'} \frac{\#T_t}{s} \frac{\#T_t - 1}{s-1} \middle| \#T_t \right) = \frac{\#T_t(\#T_t)}{s(s-1)}\mathbb{E}(S_c S_{c'}|\#T_t)
\end{aligned}$$

$$= \frac{\#T_t(\#T_t - 1)}{s(s-1)}\pi_{cc'} \tag{A.3}$$

$$\begin{aligned}\mathbb{E}(S_c S_{c'} T_{c1} T_{c'0} | \#T_1, \#T_0) &= \mathbb{E}(S_c S_{c'} \mathbb{E}(T_{c1} T_{c'0} | \mathbf{S}) | \#T_1, \#T_0) \\ &= \frac{\#T_1}{s}\frac{\#T_0}{s-1}\mathbb{E}(S_c S_{c'} | \#T_1, \#T_0) \\ &= \frac{\#T_1 \#T_0}{s(s-1)}\pi_{cc'} \end{aligned} \tag{A.4}$$

$$\mathbb{E}\left(\frac{S_c^2 T_{ct}^2}{\#T_t^2}\right) = \mathbb{E}\left(\frac{1}{\#T_t^2}\mathbb{E}(S_c T_{ct} | \#T_t)\right) = \frac{n_c}{n}\mathbb{E}\left(\frac{1}{\#T_t}\right) \tag{A.5}$$

$$\begin{aligned}\mathbb{E}\left(\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right) &= \mathbb{E}\left(\frac{1}{\#T_t^2}\mathbb{E}(S_c S_{c'} T_{ct} T_{c't} | \#T_t)\right) \\ &= \frac{\pi_{cc'}}{s(s-1)}\mathbb{E}\left(1 - \frac{1}{\#T_t}\right) \end{aligned} \tag{A.6}$$

$$\mathrm{Var}(S_c T_{ct} | \#T_t) = \frac{n_c \#T_t}{n}\left(1 - \frac{n_c \#T_t}{n}\right) \tag{A.7}$$

Conditional on the sampled clusters, units are sampled within a cluster using simple random sampling, and this secondary sampling stage is independent of cluster treatment assignment. Thus, the expectation of within-cluster sampling indicators are independent of the cluster treatment indicators. Moreover, within-cluster samples are drawn independently across clusters, and so for distinct units k and k' in the same cluster c or distinct units k and k^* in different clusters c and c':

$$\mathbb{E}\left(S_{kc} | \mathbf{S}\right) = \frac{s_c}{n_c} \tag{A.8}$$

$$\mathbb{E}\left(S_{kc} S_{k'c} | \mathbf{S}\right) = \frac{s_c(s_c - 1)}{n_c(n_c - 1)} \tag{A.9}$$

$$\mathbb{E}\left(S_{kc} S_{k^*c'} | \mathbf{S}\right) = \frac{s_c s_{c'}}{n_c n_{c'}} \tag{A.10}$$

$$\mathrm{Var}(S_{kc} | \mathbf{S}) = \frac{s_c}{n_c}\left(1 - \frac{s_c}{n_c}\right) \tag{A.11}$$

$$\begin{aligned}\mathrm{Cov}(S_{kc}, S_{k'c} | \mathbf{S}) &= \mathbb{E}(S_{kc} S_{k'c} | \mathbf{S}) - \mathbb{E}(S_{kc} | \mathbf{S})\mathbb{E}(S_{k'c} | \mathbf{S}) \\ &= -\frac{s_c}{n_c}\frac{1}{n_c - 1}\left(1 - \frac{s_c}{n_c}\right) \end{aligned} \tag{A.12}$$

A.2 Location invariance of HT-PPS estimator for PATE

Since,

$$\begin{aligned}\hat{\mu}_{t,\mathrm{HT,PPS}}(a+\mathbf{y}) &= \sum_{c=1}^{\ell}\frac{S_c T_{ct}}{\#T_t}\sum_{k=1}^{n_c}\frac{(a+y_{kct})S_{kc}}{s_c} \\ &= a\sum_{c=1}^{\ell}\frac{S_c T_{c1}}{\#T_1}\sum_{k=1}^{n_c}\frac{S_{kc}}{s_c} + \sum_{c=1}^{\ell}\frac{S_c T_{c1}}{\#T_1}\sum_{k=1}^{n_c}\frac{y_{kc1}S_{kc}}{s_c} \\ &= a + \hat{\mu}_{t,\mathrm{HT,PPS}}(\mathbf{y}) \end{aligned} \tag{A.13}$$

the HT-PPS estimator for PATE is location-invariant:

$$\hat{\delta}_{\mathrm{HT,PPS}}(a+\mathbf{y}) = \hat{\mu}_{1,\mathrm{HT,PPS}}(a+\mathbf{y}) - \hat{\mu}_{0,\mathrm{HT,PPS}}(a+\mathbf{y}) = \hat{\delta}_{\mathrm{HT,PPS}}(\mathbf{y}). \tag{A.14}$$

A.3 Expectation of HT-PPS estimator for PATE

$$\begin{aligned}\mathbb{E}(\hat{\delta}_{\text{HT-PPS}}) &= \mathbb{E}\left(\sum_{c=1}^{\ell}\frac{S_c T_{c1}}{\#T_1}\sum_{k=1}^{n_c}\frac{y_{kc1}S_{kc}}{s_c} - \sum_{c^*=1}^{\ell}\frac{S_{c^*}T_{c^*0}}{\#T_0}\sum_{k^*=1}^{n_{c^*}}\frac{y_{k^*c^*0}S_{k^*c^*}}{s_{c^*}}\right) \\ &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\frac{y_{kc1}}{s_c}\mathbb{E}\left(\frac{S_c T_{c1}S_{kc}}{\#T_1}\right) - \sum_{c^*=1}^{\ell}\sum_{k^*=1}^{n_{c^*}}\frac{y_{k^*c^*0}}{s_{c^*}}\mathbb{E}\left(\frac{S_{c^*}T_{c^*0}S_{k^*c^*}}{\#T_0}\right) \\ &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\frac{y_{kc1}}{s_c}\mathbb{E}\left(S_c\mathbb{E}\left(\frac{T_{c1}}{\#T_1}\middle|\mathbf{S}\right)\mathbb{E}(S_{kc}|\mathbf{S})\right) \\ &\quad - \sum_{c^*=1}^{\ell}\sum_{k^*=1}^{n_{c^*}}\frac{y_{k^*c^*0}}{s_{c^*}}\mathbb{E}\left(S_{c^*}\mathbb{E}\left(\frac{T_{c^*0}}{\#T_0}\middle|\mathbf{S}\right)\mathbb{E}(S_{k^*c^*}|\mathbf{S})\right) \\ &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\frac{y_{kc1}}{n_c s}\mathbb{E}\left(S_c\right) - \sum_{c^*=1}^{\ell}\sum_{k^*=1}^{n_{c^*}}\frac{y_{k^*c^*0}}{n_{c^*}s}\mathbb{E}\left(S_{c^*}\right) \\ &= \sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{c1} - \sum_{c^*=1}^{\ell}\frac{n_{c^*}}{n}\mu_{c^*0} \\ &= \mu_1 - \mu_0 = \delta. \end{aligned} \tag{A.15}$$

A.4 Variance of HT-PPS estimator for PATE

From the property

$$\mathrm{Var}(\hat{\delta}) = \mathrm{Var}(\hat{\mu}_1 - \hat{\mu}_0) = \mathrm{Var}(\hat{\mu}_1) + \mathrm{Var}(\hat{\mu}_0) - 2\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_0) \tag{A.16}$$

each term is expanded upon to derive the variance of the HT-PPS estimator for PATE and obtain a variance estimator.

A.4.1 Variance of HT-PPS estimator for population mean

Using the law of total variance,

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{t,\mathrm{HT,PPS}}) &= \mathrm{Var}\left(\sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c}\right) \\
&= \mathrm{Var}\left[\mathbb{E}\left(\sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right] \\
&\quad + \mathbb{E}\left[\mathrm{Var}\left(\sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right].
\end{aligned} \tag{A.17}$$

The first terms can be further simplified:

$$\begin{aligned}
&\mathrm{Var}\left[\mathbb{E}\left(\sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right] = \mathrm{Var}\left[\sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \sum_{k=1}^{n_c} \frac{y_{kct}}{s_c} \mathbb{E}(S_{kc} | \mathbf{S}, \mathbf{T})\right] \\
&= \sum_{c=1}^{\ell} \mathrm{Var}\left(\mu_{ct} \frac{S_c T_{ct}}{\#T_t}\right) + \sum_{c=1}^{\ell} \sum_{c' \neq c} \mathrm{Cov}\left(\mu_{ct} \frac{S_c T_{ct}}{\#T_t}, \mu_{c't} \frac{S_{c'} T_{c't}}{\#T_t}\right) \\
&= \sum_{c=1}^{\ell} \mu_{ct}^2 \mathrm{Var}\left(\frac{S_c T_{ct}}{\#T_t}\right) + \sum_{c=1}^{\ell} \sum_{c' \neq c} \mu_{ct} \mu_{c't} \mathrm{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \frac{S_{c'} T_{c't}}{\#T_t}\right) \\
&= \sum_{c=1}^{\ell} \mu_{ct}^2 \left[\mathbb{E}\left(\frac{S_c T_{ct}}{\#T_t}\right) - \mathbb{E}\left(\frac{S_c T_{ct}}{\#T_t}\right)^2\right] \\
&\quad + \sum_{c=1}^{\ell} \sum_{c' \neq c} \mu_{ct} \mu_{c't} \left[\mathbb{E}\left(\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right) - \mathbb{E}\left(\frac{S_c T_{ct}}{\#T_t}\right) \mathbb{E}\left(\frac{S_{c'} T_{c't}}{\#T_t}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2 - \sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{n_c^2}{n^2}\mu_{ct}^2 \\
&\quad + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{n_c^2}{n^2}\mu_{ct}\mu_{c't} \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2 + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mu_t^2 \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2 - \mathbb{E}\left(\frac{1}{\#T_t}\right)\mu_t^2 \\
&\quad + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\mu_t^2 \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2 - \mu_t^2\right] + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mu_t^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\frac{n_c}{n}(\mu_{ct}^2 - 2\mu_t\mu_{ct} + \mu_t^2)\right] + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mu_t^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\frac{n_c}{n}(\mu_{ct} - \mu_t)^2\right] + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mu_t^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sigma_{t,bet}^2 + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't} - \mu_t^2\right] \qquad \text{(A.18)}
\end{aligned}
$$

where $\sigma_{t,bet}^2$ is the weighted variance of cluster means. Simplifying the second term:

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}\left(\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\sum_{k=1}^{n_c}\frac{y_{kct}S_{kc}}{s_c}\middle|\mathbf{S},\mathbf{T}\right)\right] = \sum_{c=1}^{\ell}\mathrm{Var}(\hat{\mu}_{ct}|\mathbf{S},\mathbf{T})\mathbb{E}\left(\frac{S_cT_{ct}}{\#T_t^2}\right) \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mathrm{Var}(\hat{\mu}_{ct}) \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\left[\mathrm{Var}\left(\sum_{k=1}^{n_c}\frac{y_{kct}S_{kc}}{s_c}\right) + \mathrm{Cov}\left(\sum_{k=1}^{n_c}\frac{y_{kct}S_{kc}}{s_c}, \sum_{k'\neq k}\frac{y_{k'ct}S_{k'c}}{s_c}\right)\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\left[\sum_{k=1}^{n_c}\frac{y_{kct}^2}{s_cn_c}\left(1-\frac{s_c}{n_c}\right) - \sum_{k=1}^{n_c}\sum_{k'\neq k}\frac{y_{kct}y_{k'ct}}{s_cn_c(n_c-1)}\left(1-\frac{s_c}{n_c}\right)\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{1}{n(n_c-1)s_c}\left(1-\frac{s_c}{n_c}\right)\left[(n_c-1)\sum_{k=1}^{n_c}y_{kct}^2 - \sum_{c=1}^{n_c}\sum_{c\neq c'}y_{kct}y_{kc't}\right]
\end{aligned}
$$

$$= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{1}{n(n_c-1)s_c}\left(1-\frac{s_c}{n_c}\right)\left[(n_c-1)\sum_{k=1}^{n_c}y_{kct}^2-\sum_{k=1}^{n_c}\sum_{k'=1}^{n_c}y_{kct}y_{k'ct}+\sum_{k=1}^{n_c}y_{kct}^2\right]$$

$$= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{1}{n(n_c-1)s_c}\left(1-\frac{s_c}{n_c}\right)\left[n_c\sum_{k=1}^{n_c}y_{kct}^2-\left(\sum_{k=1}^{n_c}y_{kct}\right)^2\right]$$

$$= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\left(1-\frac{s_c}{n_c}\right)\frac{\sigma_{ct}^2}{s_c}. \tag{A.19}$$

where σ_{ct}^2 is the within-cluster population variance of potential outcomes under treatment t. The variance for the HT-PPS mean estimator is then

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{t,\mathrm{HT,\ SRS}}) &= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\sum_{c\neq c'}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't}-\mu_t^2\\ &\quad+\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mathrm{Var}(\hat{\mu}_{ct}) \qquad (A.20)\\ &= \mathbb{E}\left(\frac{1}{\#T_t}\right)\sigma_{t,bet}^2+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't}-\mu_t^2\right]\\ &\quad+\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\left(1-\frac{s_c}{n_c}\right)\frac{\sigma_{ct}^2}{s_c}. \qquad (A.21)\end{aligned}$$

A.4.2 Covariance of HT-PPS estimator for population means

Note that:

$$\begin{aligned}\hat{\mu}_{1,\mathrm{HT,PPS}}\hat{\mu}_{0,\mathrm{HT,PPS}} &= \left(\sum_{c=1}^{\ell}\frac{S_cT_{c1}}{\#T_1}\sum_{k=1}^{n_c}\frac{y_{kc1}S_{kc}}{s_c}\right)\left(\sum_{c'=1}^{\ell}\frac{S_{c'}T_{c'0}}{\#T_0}\sum_{k^*=1}^{n_{c'}}\frac{y_{k*c'0}S_{k*c'}}{s_{c'}}\right)\\ &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\sum_{c'=1}^{\ell}\sum_{k^*=1}^{n_{c'}}\frac{y_{kc1}y_{k^*c'0}}{s_cs_{c'}}\frac{S_cT_{c1}S_{kc}S_{c'}T_{c'0}S_{k^*c'}}{\#T_1\#T_0}\\ &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\sum_{c'\neq c}^{\ell}\sum_{k^*=1}^{n_{c'}}\frac{y_{kc1}y_{k^*c'0}}{s_cs_{c'}}\frac{S_cT_{c1}S_{kc}S_{c'}T_{c'0}S_{k^*c'}}{\#T_1\#T_0}. \qquad (A.22)\end{aligned}$$

The last equality comes from the fact that a cluster can only be given one treatment. Therefore,

$$
\begin{aligned}
&\mathrm{Cov}(\hat{\mu}_{1,\mathrm{HT,PPS}}, \hat{\mu}_{0,\mathrm{HT,PPS}}) = \mathbb{E}(\hat{\mu}_{1,\mathrm{HT,PPS}}\hat{\mu}_{0,\mathrm{HT,PPS}}) - \mathbb{E}(\hat{\mu}_{1,\mathrm{HT,PPS}})\mathbb{E}(\hat{\mu}_{0,\mathrm{HT,PPS}}) \\
&= \mathbb{E}\left(\sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\sum_{c'\neq c}\sum_{k^*=1}^{n_{c'}} \frac{y_{kc1}y_{k^*c'0}}{s_c s_{c'}} \frac{S_c T_{c1} S_{kc} S_{c'} T_{c'0} S_{k^*c'}}{\#T_1 \#T_0}\right) - \mu_1\mu_0 \\
&= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\sum_{c'\neq c}\sum_{k^*=1}^{n_{c'}} \frac{y_{kc1}y_{k^*c'0}}{s_c s_{c'}} \mathbb{E}\left[\mathbb{E}\left(\frac{S_c T_{c1} S_{kc} S_{c'} T_{c'0} S_{k^*c'}}{\#T_1 \#T_0}|\mathbf{S}\right)\right] - \mu_1\mu_0 \\
&= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\sum_{c'\neq c}\sum_{k^*=1}^{n_{c'}} \frac{y_{kc1}y_{k^*c'0}}{s_c s_{c'}} \mathbb{E}\left[S_c S_{c'}\mathbb{E}\left(\frac{T_{c1}T_{c'0}}{\#T_1\#T_0}\middle|\mathbf{S}\right)\mathbb{E}(S_{kc}S_{k^*c'}|\mathbf{S})\right] - \mu_1\mu_0 \\
&= \frac{1}{s(s-1)}\sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\sum_{c'\neq c}\sum_{k^*=1}^{n_{c'}} \frac{y_{kc1}y_{k^*c'0}}{n_c n_{c'}}\mathbb{E}\left(S_c S_{c'}\right) - \mu_1\mu_0 \\
&= \frac{1}{s(s-1)}\sum_{c=1}^{\ell}\sum_{c'\neq c}\pi_{cc'}\mu_{c1}\mu_{c'0} - \mu_1\mu_0 \\
&= \sum_{c=1}^{\ell}\sum_{c'\neq c}\left[\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2}\right]\mu_{c1}\mu_{c'0} - \sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{c1}\mu_{c0}.
\end{aligned} \tag{A.23}
$$

A.5 SYG variance estimator for HT-PPS

To get a variance estimator of $\hat{\delta}$, we need to obtain estimators for each term in eq. (A.16).

A.5.1 SYG variance estimator for population mean

The SYG variance estimator is

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\mu}_t) = &\frac{1}{2}\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}\#T_t(\#T_t-1)}\frac{n_c n_{c'}}{n^2} - \frac{1}{\#T_t^2}\right] S_c T_{ct} S_{c'} T_{c't}(\hat{\mu}_{ct} - \hat{\mu}_{c't})^2 \\
&+ \sum_{c=1}^{\ell}\frac{S_c T_{ct}}{\#T_t}\frac{n_c}{n}\widehat{\mathrm{Var}}(\hat{\mu}_{ct})
\end{aligned} \tag{A.24}
$$

where

$$\widehat{\mathrm{Var}}(\hat{\mu}_{ct}) = \left(1 - \frac{s_c}{n_c}\right) \frac{\hat{\sigma}^2_{ct}}{s_c}. \tag{A.25}$$

The $\hat{\sigma}^2_{ct}$ is the sample variance of outcomes, which is unbiased for the population variance σ^2_{ct}. We will now show that the SYG variance is unbiased for $\mathrm{Var}(\hat{\mu}_t)$. This requires the following:

$$\sum_{c' \neq c}^{\ell} n_{c'} = n - n_c \tag{A.26}$$

and

$$\sum_{c' \neq c}^{\ell} \pi_{cc'} = \sum_{c' \neq c}^{\ell} E(S_c S_{c'}) = E[S_c(s - S_c)] = \frac{n_c s}{n}(s-1). \tag{A.27}$$

Therefore, the expectation is

$$\begin{aligned}
\mathbb{E}\left(\widehat{\mathrm{Var}}(\hat{\mu}_t)\right) &= \mathbb{E}\left(\frac{1}{2}\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}}\frac{n_c n_{c'}}{n^2}\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t(\#T_t - 1)} - \frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right][\hat{\mu}_{ct} - \hat{\mu}_{c't}]^2 \right. \\
&\quad \left. + \sum_{c=1}^{\ell}\frac{n_c}{n}\frac{S_c T_{ct}}{\#T_t}\widehat{\mathrm{Var}}(\hat{\mu}_{ct})\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{1}{2}\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}}\frac{n_c n_{c'}}{n^2}\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t(\#T_t - 1)} - \frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right][\hat{\mu}_{ct} - \hat{\mu}_{c't}]^2 \middle| \mathbf{S}, \mathbf{T}\right)\right) \\
&\quad + \mathbb{E}\left(\mathbb{E}\left(\sum_{c=1}^{\ell}\frac{n_c}{n}\frac{S_c T_{ct}}{\#T_t}\widehat{\mathrm{Var}}(\hat{\mu}_{ct}) \middle| \mathbf{S}, \mathbf{T}\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}}\frac{n_c n_{c'}}{n^2}\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t(\#T_t - 1)} - \frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right][\hat{\mu}^2_{ct} - \hat{\mu}_{ct}\hat{\mu}_{c't}] \middle| \mathbf{S}, \mathbf{T}\right)\right) \\
&\quad + \mathbb{E}\left(\sum_{c=1}^{\ell}\frac{n_c}{n}\frac{S_c T_{ct}}{\#T_t}\mathrm{Var}(\hat{\mu}_{ct})\right) \\
&= \mathbb{E}\left(\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}}\frac{n_c n_{c'}}{n^2}\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t(\#T_t - 1)} - \frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right][\mu^2_{ct} + \mathrm{Var}(\hat{\mu}_{ct}) - \mu_{ct}\mu_{c't}]\right) \\
&\quad + \mathbb{E}\left(\sum_{c=1}^{\ell}\frac{n_c}{n}\frac{S_c T_{ct}}{\#T_t}\mathrm{Var}(\hat{\mu}_{ct})\right) \\
&= \sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}}\frac{n_c n_{c'}}{n^2}\mathbb{E}\left(\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t(\#T_t - 1)}\right) - \mathbb{E}\left(\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right)\right][\mu^2_{ct} + \mathrm{Var}(\hat{\mu}_{ct}) - \mu_{ct}\mu_{c't}]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{c=1}^{\ell}\frac{n_c}{n}\mathrm{Var}(\hat{\mu}_{ct})\mathbb{E}\left(\frac{S_cT_{ct}}{\#T_t}\right)\\
&=\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{s(s-1)}{\pi_{cc'}}\frac{n_cn_{c'}}{n^2}\mathbb{E}\left(\frac{1}{\#T_t(\#T_t-1)}\mathbb{E}\left(S_cS_{c'}T_{ct}T_{c't}|\,\#T_t\right)\right)-\mathbb{E}\left(\frac{1}{\#T_t^2}\mathbb{E}\left(S_cS_{c'}T_{ct}T_{c't}|\,\#T_t\right)\right)\right]\\
&\quad\cdot\left[\mu_{ct}^2+\mathrm{Var}(\hat{\mu}_{ct})-\mu_{ct}\mu_{c't}\right]+\sum_{c=1}^{\ell}\frac{n_c}{n}\mathrm{Var}(\hat{\mu}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\mathbb{E}\left(S_cT_{ct}|\#T_t\right)\right)\\
&=\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{n_cn_{c'}}{n^2}-\frac{\pi_{cc'}}{s(s-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right]\left(\mu_{ct}^2+\mathrm{Var}(\hat{\mu}_{ct})-\mu_{ct}\mu_{c't}\right)+\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mathrm{Var}(\hat{\mu}_{ct})\\
&=\sum_{c=1}^{\ell}\sum_{c'\neq c}\left[\frac{n_cn_{c'}}{n^2}-\frac{\pi_{cc'}}{s(s-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right]\left(\mu_{ct}^2+\mathrm{Var}(\hat{\mu}_{ct})\right)\\
&\quad-\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{n_cn_{c'}}{n^2}-\frac{\pi_{cc'}}{s(s-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right]\mu_{ct}\mu_{c't}+\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mathrm{Var}(\hat{\mu}_{ct})\\
&=\sum_{c=1}^{\ell}\left[\frac{n_c}{n^2}\sum_{c'\neq c}n_{c'}-\frac{1}{s(s-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c'\neq c}\pi_{cc'}\right]\left(\mu_{ct}^2+\mathrm{Var}(\hat{\mu}_{ct})\right)\\
&\quad-\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{n_cn_{c'}}{n^2}-\frac{\pi_{cc'}}{s(s-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right]\mu_{ct}\mu_{c't}+\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mathrm{Var}(\hat{\mu}_{ct})\\
&=\sum_{c=1}^{\ell}\left[\frac{n_c}{n}\left(1-\frac{n_c}{n}\right)-\frac{n_c}{n}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right]\left(\mu_{ct}^2+\mathrm{Var}(\hat{\mu}_{ct})\right)\\
&\quad-\sum_{c=1}^{\ell}\sum_{c\neq c'}\left[\frac{n_cn_{c'}}{n^2}-\frac{\pi_{cc'}}{s(s-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right]\mu_{ct}\mu_{c't}+\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mathrm{Var}(\hat{\mu}_{ct})\\
&=\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2-\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{ct}^2-\sum_{c=1}^{\ell}\sum_{c\neq c'}\frac{n_cn_{c'}}{n^2}\mu_{ct}\mu_{c't}\\
&\quad+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\sum_{c\neq c'}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't}+\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mathrm{Var}(\hat{\mu}_{ct})\\
&=\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}^2+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\sum_{c\neq c'}\frac{\pi_{cc'}}{s(s-1)}\mu_{ct}\mu_{c't}-\mu_t^2\\
&\quad+\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}\frac{n_c}{n}\mathrm{Var}(\hat{\mu}_{ct}).
\end{aligned}
\tag{A.28}
$$

This is equal to eq. (A.20).

A.5.2 Covariance bound

The covariance term in eq. (A.16) is not estimable because it requires that a cluster is given treatment and control [note the last term in eq. (A.23)]. Therefore, the covariance needs to be bounded. An estimator of this bound is given by

$$\begin{aligned}\widehat{\mathrm{Cov}}_C(\hat{\mu}_1, \hat{\mu}_0) = &\sum_{c=1}^{\ell} \sum_{c' \neq c} \left[1 - \frac{n_c n_{c'}}{n^2} \frac{s(s-1)}{\pi_{cc'}}\right] \frac{S_c T_{ct} S_{c'} T_{c't}}{\#T_1 \#T_0} \hat{\mu}_{c1} \hat{\mu}_{c'0} \\ &- \frac{1}{2} \sum_{t=0}^{1} \sum_{c=1}^{\ell} \frac{n_c}{n} \frac{S_c T_{ct}}{\#T_t} \left[\hat{\mu}_{ct}^2 - \widehat{\mathrm{Var}}(\hat{\mu}_{ct})\right]. \end{aligned} \tag{A.29}$$

Taking expectation of the estimated covariance bound, we get:

$$\begin{aligned}
\mathbb{E}\left(\widehat{\mathrm{Cov}}_C(\hat{\mu}_1, \hat{\mu}_0)\right) &= \mathbb{E}\left[\mathbb{E}\left(\sum_{c=1}^{\ell} \sum_{c' \neq c} \left[1 - \frac{n_c n_{c'}}{n^2} \frac{s(s-1)}{\pi_{cc'}}\right] \frac{S_c T_{ct} S_{c'} T_{c't}}{\#T_1 \#T_0} \hat{\mu}_{c1} \hat{\mu}_{c'0} \middle| \mathbf{S}, \mathbf{T}\right)\right] \\
&- \mathbb{E}\left[\mathbb{E}\left(\frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \frac{S_c T_{c1}}{\#T_1} \hat{\mu}_{c1}^2 \middle| \mathbf{S}, \mathbf{T}\right)\right] - \mathbb{E}\left[\mathbb{E}\left(\frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \frac{S_c T_{c0}}{\#T_0} \hat{\mu}_{c0}^2 \middle| \mathbf{S}, \mathbf{T}\right)\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left(\frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \frac{S_c T_{c1}}{\#T_1} \widehat{\mathrm{Var}}(\hat{\mu}_{c1}) \middle| \mathbf{S}, \mathbf{T}\right)\right] + \mathbb{E}\left[\mathbb{E}\left(\frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \frac{S_c T_{c0}}{\#T_0} \widehat{\mathrm{Var}}(\hat{\mu}_{c0}) \middle| \mathbf{S}, \mathbf{T}\right)\right] \\
&= \sum_{c=1}^{\ell} \sum_{c' \neq c} \left[1 - \frac{n_c n_{c'}}{n^2} \frac{s(s-1)}{\pi_{cc'}}\right] \mu_{c1} \mu_{c'0} \mathbb{E}\left(\frac{S_c T_{ct} S_{c'} T_{c't}}{\#T_1 \#T_0}\right) \\
&- \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \left[\mu_{c1}^2 + \mathrm{Var}(\hat{\mu}_{c1})\right] \mathbb{E}\left(\frac{S_c T_{c1}}{\#T_1}\right) - \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \left[\mu_{c0}^2 + \mathrm{Var}(\hat{\mu}_{c1})\right] \mathbb{E}\left(\frac{S_c T_{c0}}{\#T_0}\right) \\
&+ \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \mathrm{Var}(\hat{\mu}_{c1}) \mathbb{E}\left(\frac{S_c T_{c1}}{\#T_1}\right) + \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \mathrm{Var}(\hat{\mu}_{c0}) \mathbb{E}\left(\frac{S_c T_{c0}}{\#T_0}\right) \\
&= \sum_{c=1}^{\ell} \sum_{c' \neq c} \left[1 - \frac{n_c n_{c'}}{n^2} \frac{s(s-1)}{\pi_{cc'}}\right] \mu_{c1} \mu_{c'0} \mathbb{E}\left[\frac{1}{\#T_1 \#T_0} \mathbb{E}\left(S_c S_{c'} T_{ct} T_{c't} | \#T_1, \#T_0\right)\right] \\
&- \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \left[\mu_{c1}^2 + \mathrm{Var}(\hat{\mu}_{c1})\right] \mathbb{E}\left[\frac{1}{\#T_1} \mathbb{E}\left(S_c T_{c1} | \#T_1\right)\right] \\
&- \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \left[\mu_{c0}^2 + \mathrm{Var}(\hat{\mu}_{c1})\right] \mathbb{E}\left[\frac{1}{\#T_0} \mathbb{E}\left(S_c T_{c0} | \#T_0\right)\right] \\
&+ \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \mathrm{Var}(\hat{\mu}_{c1}) \mathbb{E}\left[\frac{1}{\#T_1} \mathbb{E}\left(S_c T_{c1} | \#T_1\right)\right] + \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c}{n} \mathrm{Var}(\hat{\mu}_{c0}) \mathbb{E}\left[\frac{1}{\#T_0} \mathbb{E}\left(S_c T_{c0} | \#T_0\right)\right]
\end{aligned}$$

$$= \sum_{c=1}^{\ell} \sum_{c' \neq c} \left[\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2} \right] \mu_{c1} \mu_{c'0} - \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{c1}^2 - \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{c0}^2. \tag{A.30}$$

We next show that eq. (A.30) is no larger than eq. (A.23), using Young's inequality.

Lemma 9 (Young's Inequality)**.** *If a, b are nonnegative real numbers and p, q are positive real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{A.31}$$

Take $p = q = 2$, then

$$\begin{aligned}
\mathrm{Cov}(\hat{\mu}_{1,\text{HT-PPS}}, \hat{\mu}_{0,\text{HT-PPS}}) &= \sum_{c=1}^{\ell} \sum_{c \neq c'} \left(\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2} \right) \mu_{c1} \mu_{c0} - \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{c1} \mu_{c0} \\
&\geq \sum_{c=1}^{\ell} \sum_{c \neq c'} \left(\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2} \right) \mu_{c1} \mu_{c0} \\
&\quad - \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{c1}^2 - \frac{1}{2} \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{c0}^2 \\
&= \mathrm{Cov}_C(\hat{\mu}_{1,\text{HT-PPS}}, \hat{\mu}_{0,\text{HT-PPS}}).
\end{aligned} \tag{A.32}$$

A.5.3 SYG variance estimator of HT-PPS for PATE

From eq. (A.28) and eq. (A.32), we see that

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\delta}_{\text{HT,PPS}}) = \frac{1}{2} \sum_{t=0}^{1} \sum_{c=1}^{\ell} \sum_{c' \neq c} \left[\frac{s(s-1)}{\pi_{cc'} \#T_t (\#T_t - 1)} \frac{n_c n_{c'}}{n^2} - \frac{1}{\#T_t^2} \right] S_c T_{ct} S_{c'} T_{c't} \left(\hat{\mu}_{ct} - \hat{\mu}_{c't} \right)^2 \\
+ \sum_{t=0}^{1} \sum_{c=1}^{\ell} \frac{n_c}{n} \frac{S_c T_{ct}}{\#T_t} \hat{\mu}_{ct}^2 \\
- 2 \sum_{c=1}^{\ell} \sum_{c \neq c'} \left[\frac{\pi_{cc'}}{s(s-1)} - \frac{n_c n_{c'}}{n^2} \right] \frac{s(s-1)}{\pi_{cc'}} \frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0} \hat{\mu}_{c1} \hat{\mu}_{c0}
\end{aligned} \tag{A.33}$$

is a conservative bound for $\mathrm{Var}(\hat{\delta}_{\text{HT,PPS}})$.

A.6 With-replacement variance estimator of HT-PPS for PATE

Since the variance estimator in eq. (A.33) requires $\pi_{cc'}$, we also give an alternative estimator that do not need the $\pi_{cc'}$. We use the same covariance bound as in the SYG variance estimator but assume that the clusters are independently sampled and treated.

$$\begin{aligned}\widehat{\mathrm{Var}}_{\mathrm{WR}}\left(\hat{\delta}_{\mathrm{HT,\ PPS}}\right) = &\sum_{t=0}^{1}\sum_{c=1}^{\ell}\frac{n_c}{n}\frac{S_c T_{ct}}{\#T_t}\left(\hat{\mu}_{ct}-\hat{\mu}_t\right)^2 \\ &-\frac{2}{s}\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0}\mu_{c1}\mu_{c'0} + \sum_{t=0}^{1}\sum_{c=1}^{\ell}\frac{n_c}{n}\frac{S_c T_{ct}}{\#T_t}\left[\hat{\mu}_{ct}^2 - \widehat{\mathrm{Var}}(\hat{\mu}_{ct})\right]\end{aligned} \tag{A.34}$$

Appendix B

Properties of the HT-SRS estimator

Let S_c be the indicator variable for sampling clusters, T_{ct} be the assignment indicator for treatment t, and S_{kc} be the indicator for sampling units. We begin by computing expectations, variances, and covariances of these indicators under SRSWOR sampling of clusters. These will be helpful in examining the properties of the SRSWOR estimators for PATE.

B.1 Useful indicator expectations under SRS

Let $\mathbf{S} = (S_1, S_2, \ldots, S_n)$ denote a random set of cluster sampling indicator variables under SRS. For any distinct clusters c and c' and distinct treatments $t = 1$ and $t = 0$, the following expectations hold under complete randomization of treatment to units, provided that $S_c = 1$ (and when applicable, $S_{c'} = 1$):

$$\mathbb{E}(S_c|\#T_t) = \frac{s}{\ell} \tag{B.1}$$

$$\mathbb{E}(S_c S_{c'}|\#T_t) = \frac{s}{\ell}\frac{s-1}{\ell-1} \tag{B.2}$$

$$\begin{aligned}\mathbb{E}\left(S_c^2 T_{ct}^2|\#T_t\right) &= \mathbb{E}\left(S_c T_{ct}|\#T_t\right) = \mathbb{E}\left(S_c \mathbb{E}\left(T_{ct}|\,\mathbf{S}\right)|\#T_t\right) \\ &= \frac{\#T_t}{s}\mathbb{E}\left(S_c|\#T_t\right) \\ &= \frac{\#T_t}{\ell}\end{aligned} \tag{B.3}$$

$$\begin{aligned}
\mathbb{E}(S_c S_{c'} T_{ct} T_{c't} | \#T_t) &= \mathbb{E}(S_c S_{c'} \mathbb{E}(T_{ct} T_{c't} | \mathbf{S}) | \#T_t) \\
&= \frac{\#T_t}{s} \frac{\#T_t - 1}{s-1} \mathbb{E}(S_c S_{c'} | \#T_t) \\
&= \frac{\#T_t(\#T_t - 1)}{\ell(\ell-1)} \qquad \text{(B.4)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(S_c S_{c'} T_{c1} T_{c'0} | \#T_1, \#T_0) &= \mathbb{E}(S_c S_{c'} \mathbb{E}(T_{c1} T_{c'0} | \mathbf{S}) | \#T_1, \#T_0) \\
&= \frac{\#T_1}{s} \frac{\#T_0}{s-1} \mathbb{E}(S_c S_{c'} | \#T_1, \#T_0) \\
&= \frac{\#T_1 \#T_0}{\ell(\ell-1)} \qquad \text{(B.5)}
\end{aligned}$$

$$\mathbb{E}\left(\frac{S_c^2 T_{ct}^2}{\#T_t^2}\right) = \mathbb{E}\left(\frac{1}{\#T_t^2} \mathbb{E}(S_c^2 T_{ct}^2 | \#T_t)\right) = \frac{1}{\ell} \mathbb{E}\left(\frac{1}{\#T_t}\right) \qquad \text{(B.6)}$$

$$\text{Var}\left(S_c T_{ct} | \#T_t\right) = \frac{\#T_t}{\ell}\left(1 - \frac{\#T_t}{\ell}\right) \qquad \text{(B.7)}$$

$$\begin{aligned}
\text{Var}\left(\frac{S_c T_{ct}}{\#T_t}\right) &= \text{Var}\left(\frac{1}{\#T_t} \mathbb{E}(S_c T_{ct} | \#T_t)\right) + \mathbb{E}\left(\frac{1}{\#T_t^2} \text{Var}(S_c T_{ct} | \#T_t)\right) \\
&= \frac{1}{\ell} \mathbb{E}\left(\frac{1}{\#T_t}\right) - \frac{1}{\ell^2} \qquad \text{(B.8)}
\end{aligned}$$

$$\begin{aligned}
\text{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \frac{S_{c'} T_{c't}}{\#T_t}\right) &= \text{Cov}\left(\frac{1}{\#T_t} \mathbb{E}(S_c T_{ct} | \#T_t), \frac{1}{\#T_t} \mathbb{E}(S_{c'} T_{c't} | \#T_t)\right) \\
&\quad + \mathbb{E}\left(\frac{1}{\#T_t^2} \text{Cov}(S_c T_{ct}, S_{c'} T_{c't} | \#T_t)\right) \\
&= \mathbb{E}\left(\frac{1}{\#T_t^2}\left[\mathbb{E}(S_c S_{c'} T_{ct} T_{c't} | \#T_t) - \mathbb{E}(S_c T_{ct} | \#T_t) \mathbb{E}(S_{c'} T_{c't} | \#T_t)\right]\right) \\
&= \frac{1}{\ell(\ell-1)} \mathbb{E}\left(1 - \frac{1}{\#T_t}\right) - \frac{1}{\ell^2} \qquad \text{(B.9)}
\end{aligned}$$

$$\begin{aligned}
\text{Var}\left(\frac{S_c T_{c1} T_{c'0}}{\#T_1 \#T_0}\right) &= \mathbb{E}\left(\frac{1}{\#T_1 \#T_0} \mathbb{E}(S_c^2 T_{c1} T_{c'0} | \#T_1, \#T_0)\right) \\
&\quad - \mathbb{E}\left(\frac{1}{\#T_1} \mathbb{E}(S_c T_{c1} | \#T_1)\right) \mathbb{E}\left(\frac{1}{\#T_0} \mathbb{E}(S_c T_{c0} | \#T_0)\right) \\
&= -\frac{1}{\ell^2} \qquad \text{(B.10)}
\end{aligned}$$

$$\begin{aligned}
\text{Cov}\left(\frac{S_c T_{c1}}{\#T_1}, \frac{S_{c'} T_{c'0}}{\#T_0}\right) &= \mathbb{E}\left(\frac{1}{\#T_1 \#T_0} \mathbb{E}(S_c S_{c'} T_{c1} T_{c'0} | \#T_1, \#T_0)\right) \\
&\quad - \mathbb{E}\left(\frac{1}{\#T_1} \mathbb{E}(S_c T_{c1} | \#T_1)\right) \mathbb{E}\left(\frac{1}{\#T_0} \mathbb{E}(S_{c'} T_{c'0} | \#T_0)\right) \\
&= \frac{1}{\ell(\ell-1)} - \frac{1}{\ell^2} \qquad \text{(B.11)}
\end{aligned}$$

Conditional on the sampled clusters, units are sampled within a cluster using simple random sampling, and this secondary sampling stage is independent of cluster treatment assignment. Thus, the expectation of within-cluster sampling indicators are independent of the cluster treatment indicators. Moreover, within-cluster samples are drawn independently across clusters, and so for distinct units k and k' in the same cluster c or distinct units k and k^* in different clusters c and c':

$$\mathbb{E}\left(S_{kc}|\mathbf{S}\right) = \frac{s_c}{n_c} \tag{B.12}$$

$$\mathbb{E}\left(S_{kc}S_{k'c}|\mathbf{S}\right) = \frac{s_c(s_c-1)}{n_c(n_c-1)} \tag{B.13}$$

$$\mathbb{E}\left(S_{kc}S_{k^*c'}|\mathbf{S}\right) = \frac{s_c s_{c'}}{n_c n_{c'}} \tag{B.14}$$

$$\mathrm{Var}(S_{kc}|\mathbf{S}) = \frac{s_c}{n_c}\left(1-\frac{s_c}{n_c}\right) \tag{B.15}$$

$$\begin{aligned}\mathrm{Cov}(S_{kc}, S_{k'c}|\mathbf{S}) &= \mathbb{E}(S_{kc}S_{k'c}|\mathbf{S}) - \mathbb{E}(S_{kc}|\mathbf{S})\mathbb{E}(S_{k'c}|\mathbf{S}) \\ &= -\frac{s_c}{n_c}\frac{1}{n_c-1}\left(1-\frac{s_c}{n_c}\right)\end{aligned} \tag{B.16}$$

B.2 Expectation of HT estimator for PATE

For any treatment t,

$$\begin{aligned}\mathbb{E}(\hat{\mu}_{t,\mathrm{HT,SRS}}) &= \mathbb{E}\left(\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{y_{kct}S_{kc}}{s_c}\right) \\ &= \ell\sum_{c=1}^{\ell}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{y_{kct}}{s_c}\mathbb{E}\left(\frac{S_cT_{ct}S_{kc}}{\#T_t}\right) \\ &= \ell\sum_{c=1}^{\ell}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{y_{kct}}{s_c}\mathbb{E}\left(\mathbb{E}\left(\frac{S_cT_{ct}S_{kc}}{\#T_t}\middle|\mathbf{S}\right)\right) \\ &= \ell\sum_{c=1}^{\ell}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{y_{kct}}{s_c}\mathbb{E}\left(S_c\mathbb{E}\left(\frac{T_{ct}}{\#T_t}\middle|\mathbf{S}\right)\mathbb{E}\left(S_{kc}|\mathbf{S}\right)\right) \\ &= \frac{\ell}{s}\sum_{c=1}^{\ell}\frac{1}{n}\sum_{k=1}^{n_c}y_{kct}\mathbb{E}(S_c)\end{aligned}$$

$$= \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \frac{y_{kct}}{n} = \mu_t. \tag{B.17}$$

Then $\hat{\delta}_{\text{HT,SRS}}$ is unbiased for δ:

$$\begin{aligned} \mathbb{E}(\left(\hat{\delta}_{\text{HT,SRS}}\right) &= \mathbb{E}(\hat{\mu}_{1,\text{HT,SRS}}) - \mathbb{E}(\hat{\mu}_{0,\text{HT,SRS}}) \\ &= \mu_1 - \mu_0 = \delta. \end{aligned} \tag{B.18}$$

B.3 Variance of HT-SRS for PATE

B.3.1 Variance of HT-SRS for population mean

By the law of total variance,

$$\begin{aligned} \text{Var}(\hat{\mu}_{t,\text{HT,SRS}}) = &\text{Var}\left[\mathbb{E}\left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right] \\ &+ \mathbb{E}\left[\text{Var}\left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right]. \end{aligned} \tag{B.19}$$

Let's focus on deriving and simplifying each term separately. The first term:

$$\begin{aligned} &\text{Var}\left[\mathbb{E}\left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right] = \text{Var}\left[\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \mathbb{E}\left(\sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} \middle| \mathbf{S}, \mathbf{T}\right)\right] \\ &= \text{Var}\left[\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \mu_{ct}\right] \\ &= \ell^2 \sum_{c=1}^{\ell} \sum_{c'=1}^{\ell} \frac{n_c n_{c'}}{n^2} \mu_{ct} \mu_{c't} \text{Cov}\left[\frac{S_c T_{ct}}{\#T_t}, \frac{S_{c'} T_{c't}}{\#T_t}\right] \\ &= \ell^2 \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{ct}^2 \text{Var}\left[\frac{S_c T_{ct}}{\#T_t}\right] + \ell^2 \sum_{c=1}^{\ell} \sum_{c' \neq c} \frac{n_c n_{c'}}{n^2} \mu_{ct} \mu_{c't} \text{Cov}\left[\frac{S_c T_{ct}}{\#T_t}, \frac{S_{c'} T_{c't}}{\#T_t}\right] \\ &= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mu_{ct}^2 + \mathbb{E}\left(1 - \frac{1}{\#T_t}\right) \frac{\ell}{\ell - 1} \sum_{c=1}^{\ell} \sum_{c' \neq c} \frac{n_c n_{c'}}{n^2} \mu_{ct} \mu_{c't} \end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\mu_{ct}^2 - \sum_{c=1}^{\ell}\sum_{c\neq c} \frac{n_c n_{c'}}{n^2}\mu_{ct}\mu_{c't} \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\mu_{ct}^2 + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{n_c n_{c'}}{n^2}\mu_{ct}\mu_{c't} - \mu_t^2 \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\mu_{ct}^2 - \mathbb{E}\left(\frac{1}{\#T_t}\right)\mu_t^2 \\
&\quad + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{n_c n_{c'}}{n^2}\mu_{ct}\mu_{c't} - \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\mu_t^2 \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left[\ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\mu_{ct}^2 - \mu_t^2\right] + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{n_c n_{c'}}{n^2}\mu_{ct}\mu_{c't} - \mu_t^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left[\ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\mu_{ct}^2 - \left(\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}\right)^2\right] \\
&\quad + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'=1}^{\ell}\frac{n_c n_{c'}}{n^2}\mu_{ct}\mu_{c't} - \frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{ct}^2 - \mu_t^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left[\ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\mu_{ct}^2 - \ell^2\left(\frac{1}{\ell}\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}\right)^2\right] + \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left[\frac{1}{\ell-1}\mu_t^2 - \frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{ct}^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell(\ell-1)\left[\frac{1}{\ell-1}\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{ct}^2 - \frac{\ell}{\ell-1}\left(\frac{1}{\ell}\sum_{c=1}^{\ell}\frac{n_c}{n}\mu_{ct}\right)^2\right] \\
&\quad - \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{1}{\ell-1}\left[\ell\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\mu_{ct}^2 - \mu_t^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell(\ell-1)\mathrm{Var}\left(\frac{n_c}{n}\mu_{ct}\right) - \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\ell\mathrm{Var}\left(\frac{n_c}{n}\mu_{ct}\right) \\
&= \ell\mathrm{Var}\left(\frac{n_c}{n}\mu_{ct}\right)\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)(\ell-1) - \mathbb{E}\left(1-\frac{1}{\#T_t}\right)\right] \\
&= \ell^2\left[\mathbb{E}\left(\frac{1}{\#T_t}\right) - \frac{1}{\ell}\right]\mathrm{Var}\left(\frac{n_c}{n}\mu_{ct}\right). \qquad \text{(B.20)}
\end{aligned}
$$

Since units are sampled independently across clusters, the second term is:

$$
\mathbb{E}\left[\mathrm{Var}\left(\ell\sum_{c=1}^{\ell}\frac{S_c T_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{y_{kct}S_{kc}}{s_c}\middle|\mathbf{S},\mathbf{T}\right)\right] = \mathbb{E}\left[\ell^2\sum_{c=1}^{\ell}\frac{S_c^2 T_{ct}^2}{\#T_t^2}\frac{n_c^2}{n^2}\mathrm{Var}\left(\hat{\mu}_{ct}|\,\mathbf{S},\mathbf{T}\right)\right]
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\ell^2 \sum_{c=1}^{\ell} \frac{S_c^2 T_{ct}^2}{\#T_t^2} \frac{n_c^2}{n^2} \mathrm{Var}(\hat{\mu}_{ct}|\mathbf{S},\mathbf{T})\right] \\
&= \ell^2 \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mathrm{Var}(\hat{\mu}_{ct}|\mathbf{S},\mathbf{T}) \mathbb{E}\left[\frac{S_c^2 T_{ct}^2}{\#T_t^2}\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \mathrm{Var}(\hat{\mu}_{ct}) \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \left[\mathrm{Var}\left(\sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c}\right) + \mathrm{Cov}\left(\sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c}, \sum_{k' \neq k} \frac{y_{k'ct} S_{k'c}}{s_c}\right)\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \left[\sum_{k=1}^{n_c} \frac{y_{kct}^2}{s_c^2} \frac{s_c}{n_c}\left(1 - \frac{s_c}{n_c}\right) - \sum_{k=1}^{n_c} \sum_{k' \neq k} \frac{y_{kct} y_{k'ct}}{s_c^2} \frac{s_c}{n_c(n_c-1)}\left(1 - \frac{s_c}{n_c}\right)\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c}{n^2 s_c}\left(1 - \frac{s_c}{n_c}\right) \frac{1}{(n_c-1)} \left[(n_c-1)\sum_{k=1}^{n_c} y_{kct}^2 - \sum_{k=1}^{n_c} \sum_{k' \neq k} y_{kct} y_{k'ct}\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c}{n^2 s_c}\left(1 - \frac{s_c}{n_c}\right) \frac{1}{(n_c-1)} \left[(n_c-1)\sum_{k=1}^{n_c} y_{kct}^2 - \sum_{k=1}^{n_c} \sum_{k'=1}^{n_c} y_{kct} y_{k'ct} + \sum_{k=1}^{n_c} y_{kct}^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c}{n^2 s_c}\left(1 - \frac{s_c}{n_c}\right) \frac{1}{(n_c-1)} \left[n_c \sum_{k=1}^{n_c} y_{kct}^2 - \left(\sum_{k=1}^{n_c} y_{kct}\right)^2\right] \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\left(1 - \frac{s_c}{n_c}\right) \frac{\mathrm{Var}(y_{kct})}{s_c}. \qquad \text{(B.21)}
\end{aligned}
$$

Thus, the variance of the HT estimator for μ_t is

$$
\mathrm{Var}(\hat{\mu}_{t,\mathrm{HT,SRS}}) = \ell^2 \left[\mathbb{E}\left(\frac{1}{\#T_t}\right) - \frac{1}{\ell}\right] \mathrm{Var}\left(\frac{n_c}{n}\mu_{ct}\right) + \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\left(1 - \frac{s_c}{n_c}\right) \frac{\mathrm{Var}(y_{kct})}{s_c} \qquad \text{(B.22)}
$$

where $\mathrm{Var}\left(\frac{n_c}{n}\mu_{ct}\right)$ is the population variance of weighted cluster means and $\mathrm{Var}(y_{kct})$ is the population variance of outcomes in a cluster.

B.3.2 Covariance of HT-SRS for population means

We derive the covariance term through the property:

$$
\mathrm{Cov}(\hat{\mu}_{1,\mathrm{HT,\ SRS}}) = \mathbb{E}(\hat{\mu}_{1,\mathrm{HT,\ SRS}} \hat{\mu}_{0,\mathrm{HT,\ SRS}}) - \mathbb{E}(\hat{\mu}_{1,\mathrm{HT,\ SRS}}) \mathbb{E}(\hat{\mu}_{0,\mathrm{HT,\ SRS}}). \qquad \text{(B.23)}
$$

Note that:

$$\begin{aligned}
\hat{\mu}_{1,\text{HT, SRS}}\hat{\mu}_{0,\text{HT, SRS}} &= \left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{c1}}{\#T_1} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kc1} S_{kc}}{s_c}\right)\left(\ell \sum_{c'=1}^{\ell} \frac{S_{c'} T_{c'0}}{\#T_0} \frac{n_{c'}}{n} \sum_{k^*=1}^{n_{c'}} \frac{y_{k^*c'0} S_{k^*c'}}{s_{c'}}\right) \\
&= \ell^2 \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \sum_{c'=1}^{\ell} \sum_{k^*=1}^{n_{c'}} \frac{n_c n_{c'}}{n^2} \frac{y_{kc1} y_{k^*c'0}}{s_c s_{c'}} \frac{S_c S_{c'} T_{c1} T_{c'0} S_{kc} S_{k^*c'}}{\#T_1 \#T_0} \\
&= \ell^2 \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \sum_{c'\neq c}^{\ell} \sum_{k^*=1}^{n_{c'}} \frac{n_c n_{c'}}{n^2} \frac{y_{kc1} y_{k^*c'0}}{s_c s_{c'}} \frac{S_c S_{c'} T_{c1} T_{c'0} S_{kc} S_{k^*c'}}{\#T_1 \#T_0}. \qquad \text{(B.24)}
\end{aligned}$$

The last equality comes from the fact that a cluster can only be given one treatment. Then:

$$\begin{aligned}
\mathbb{E}(\hat{\mu}_{1,\text{HT, SRS}}\hat{\mu}_{0,\text{HT, SRS}}) &= \mathbb{E}\left(\ell^2 \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \sum_{c'\neq c}^{\ell} \sum_{k^*=1}^{n_{c'}} \frac{n_c n_{c'}}{n^2} \frac{y_{kc1} y_{k^*c'0}}{s_c s_{c'}} \frac{S_c S_{c'} T_{c1} T_{c'0} S_{kc} S_{k^*c'}}{\#T_1 \#T_0}\right) \\
&= \ell^2 \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \sum_{c'\neq c}^{\ell} \sum_{k^*=1}^{n_{c'}} \frac{n_c n_{c'}}{n^2} \frac{y_{kc1} y_{k^*c'0}}{s_c s_{c'}} \mathbb{E}\left(\frac{S_c S_{c'} T_{c1} T_{c'0} S_{kc} S_{k^*c'}}{\#T_1 \#T_0}\right) \\
&= \ell^2 \sum_{c=1}^{\ell} \sum_{k=1}^{n_c} \sum_{c'\neq c}^{\ell} \sum_{k^*=1}^{n_{c'}} \frac{n_c n_{c'}}{n^2} \frac{y_{kc1} y_{k^*c'0}}{s_c s_{c'}} \mathbb{E}\left(\frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0} \mathbb{E}(S_{kc} S_{k^*c'} | \mathbf{S}, \mathbf{T})\right) \\
&= \ell^2 \sum_{c=1}^{\ell} \sum_{c'\neq c}^{\ell} \frac{n_c n_{c'}}{n^2} \mu_{c1} \mu_{c'0} \mathbb{E}\left(\frac{1}{\#T_1 \#T_0} \mathbb{E}\left(S_c S_{c'} T_{c1} T_{c'0} | \#T_1, \#T_0\right)\right) \\
&= \frac{\ell}{\ell - 1} \sum_{c=1}^{\ell} \sum_{c\neq c'}^{\ell} \frac{n_c n_{c'}}{n^2} \mu_{c1} \mu_{c'0}. \qquad \text{(B.25)}
\end{aligned}$$

Therefore:

$$\text{Cov}(\hat{\mu}_{1,\text{HT, SRS}}) = \frac{\ell}{\ell - 1} \sum_{c=1}^{\ell} \sum_{c\neq c'}^{\ell} \frac{n_c n_{c'}}{n^2} \mu_{c1} \mu_{c'0} - \mu_1 \mu_0. \qquad \text{(B.26)}$$

B.3.3 Variance of HT-SRS estimator for PATE

From (B.22) and (B.26), the variance of the Horvitz-Thompson estimator of PATE is

$$\begin{aligned}
\text{Var}\left(\hat{\delta}_{\text{HT,SRS}}\right) &= \text{Var}\left(\hat{\mu}_{1,\text{HT,SRS}}\right) + \text{Var}\left(\hat{\mu}_{0,\text{HT,SRS}}\right) - 2\text{Cov}(\hat{\mu}_{1,\text{HT,SRS}}, \hat{\mu}_{0,\text{HT,SRS}}) \\
&= \ell^2 \sum_{t=0}^{1} \left[\mathbb{E}\left(\frac{1}{\#T_t}\right) - \frac{1}{\ell}\right] \text{Var}\left(\frac{n_c}{n} \mu_{ct}\right)
\end{aligned}$$

$$+\sum_{t=0}^{1}\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\left(1-\frac{s_c}{n_c}\right)\frac{\mathrm{Var}(y_{kct})}{s_c}$$
$$-2\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c\neq c'}\frac{n_c n_{c'}}{n^2}\mu_{c1}\mu_{c'0}+2\mu_1\mu_0. \tag{B.27}$$

B.4 SYG variance estimator of HT-SRS for PATE

$$\widehat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{HT,\,SRS}}\right)=\frac{\ell^2}{2}\sum_{t=0}^{1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(1-\frac{\#T_t}{\ell}\right)\frac{S_cS_{c'}T_{ct}T_{c't}}{\#T_t^2(\#T_t-1)}\left(\frac{n_c\hat{\mu}_{ct}}{n}-\frac{n_{c'}\hat{\mu}_{c't}}{n}\right)^2$$
$$+\ell\sum_{t=0}^{1}\sum_{c=1}^{\ell}\frac{n_c^2}{n^2}\frac{S_cT_{ct}}{\#T_t}\hat{\mu}_{ct}^2-2\ell\sum_{c=1}^{\ell}\sum_{c'\neq c}\frac{n_cn_{c'}}{n^2}\frac{S_cS_{c'}T_{c1}T_{c'0}}{\#T_1\#T_0}\hat{\mu}_{c1}\hat{\mu}_{c'0} \tag{B.28}$$

B.5 Linear transforms on HT-SRS estimator

Observe that, for any constants a, b:

$$\begin{aligned}\hat{\mu}_{t,\mathrm{HT,SRS}}(a+b\mathbf{y})&=\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{(a+by_{kct})S_{kc}}{s_c}\\&=\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{aS_{kc}}{s_c}+\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{by_{kct}S_{kc}}{s_c}\\&=a\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{S_{kc}}{s_c}+b\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}\sum_{k=1}^{n_c}\frac{y_{kct}S_{kc}}{s_c}\\&=a\ell\sum_{c=1}^{\ell}\frac{S_cT_{ct}}{\#T_t}\frac{n_c}{n}+b\hat{\mu}_{t,\mathrm{HT,SRS}}(\mathbf{y})\\&=a\frac{\ell\#N_t}{n\#T_t}+b\hat{\mu}_{t,\mathrm{HT,SRS}}(\mathbf{y}).\end{aligned} \tag{B.29}$$

Hence, the estimated PATE for linearly transformed outcomes via HT-SRS is

$$\begin{aligned}\hat{\delta}_{\mathrm{HT,SRS}}(a+\mathbf{y})&=\hat{\mu}_{1,\mathrm{HT,SRS}}(a+\mathbf{y})-\hat{\mu}_{0,\mathrm{HT,SRS}}(a+\mathbf{y})\\&=\left(a\frac{\ell\#N_1}{n\#T_1}+\hat{\mu}_{1,\mathrm{HT,SRS}}(\mathbf{y})\right)-\left(a\frac{\ell\#N_0}{n\#T_0}+b\hat{\mu}_{0,\mathrm{HT,SRS}}(\mathbf{y})\right)\end{aligned}$$

$$= a\frac{\ell}{n}\left(\frac{\#N_1}{\#T_1} - \frac{\#N_0}{\#T_0}\right) + \hat{\delta}_{\text{HT,SRS}}(\mathbf{y}). \tag{B.30}$$

B.5.1 Variance of HT-SRS estimator for transformed outcomes

For linearly transformed potential outcomes, the variance will then be

$$\begin{aligned}
\text{Var}(\hat{\delta}_{\text{HT,SRS}}(a+\mathbf{y})) &= \text{Var}\left[a\frac{\ell}{n}\left(\frac{\#N_1}{\#T_1} - \frac{\#N_0}{\#T_0}\right) + \hat{\delta}_{\text{HT,SRS}}\right] \\
&= \text{Var}\left[a\frac{\ell}{n}\left(\frac{\#N_1}{\#T_1} - \frac{\#N_0}{\#T_0}\right)\right] + \text{Var}\left(\hat{\delta}_{\text{HT,SRS}}\right) \\
&\quad + 2\text{Cov}\left[a\frac{\ell}{n}\left(\frac{\#N_1}{\#T_1} - \frac{\#N_0}{\#T_0}\right), \hat{\delta}_{\text{HT,SRS}}\right] \\
&= a^2\left(\frac{\ell}{n}\right)^2 \text{Var}\left(\frac{\#N_1}{\#T_1} - \frac{\#N_0}{\#T_0}\right) + 2a\frac{\ell}{n}\text{Cov}\left(\frac{\#N_1}{\#T_1} - \frac{\#N_0}{\#T_0}, \hat{\delta}_{\text{HT,SRS}}\right) \\
&\quad + \text{Var}(\hat{\delta}_{\text{HT,SRS}}) \\
&= a^2\left(\frac{\ell}{n}\right)^2 \left[\text{Var}\left(\frac{\#N_1}{\#T_1}\right) + \text{Var}\left(\frac{\#N_0}{\#T_0}\right) - 2\text{Cov}\left(\frac{\#N_1}{\#T_1}, \frac{\#N_0}{\#T_0}\right)\right] \\
&\quad + 2a\frac{\ell}{n}\left[\text{Cov}\left(\frac{\#N_1}{\#T_1}, \hat{\delta}_{\text{HT,SRS}}\right) - \text{Cov}\left(\frac{\#N_0}{\#T_0}, \hat{\delta}_{\text{HT,SRS}}\right)\right] \\
&\quad + \text{Var}(\hat{\delta}_{\text{HT,SRS}}).
\end{aligned} \tag{B.31}$$

Appendix C

Properties of the DIM estimator

Let S_c be the indicator variable for sampling clusters, T_{ct} be the assignment indicator for treatment t, and S_{kc} be the indicator for sampling units. Suppose that $\#K = \sum_{c=1}^{\ell} s_c$ is the number of sampled units and that $\#K_t = \sum_{c=1}^{\ell} S_c T_{ct} s_c$ is the number of sampled units given treatment t. This section uses the indicator properties calculated in section B.1.

C.1 Linear transformation on DIM estimator for PATE

For any constants a, b,

$$\begin{aligned}
\hat{\mu}_{t,\text{DIM,SRS}}(a + b\mathbf{y}) &= \frac{\sum_{c=1}^{\ell} S_c T_{ct} \sum_{c=1}^{n_c} (a + b y_{kct}) S_{kc}}{\#K_t} \\
&= \frac{\sum_{c=1}^{\ell} S_c T_{ct} \sum_{c=1}^{n_c} a S_{kc}}{\#K_t} + \frac{\sum_{c=1}^{\ell} S_c T_{ct} \sum_{c=1}^{n_c} b y_{kct} S_{kc}}{\#K_t} \\
&= a \frac{\#K_t}{\#K_t} + b \frac{\sum_{c=1}^{\ell} S_c T_{ct} \sum_{c=1}^{n_c} y_{kct} S_{kc}}{\#K_t} \\
&= a + b \hat{\mu}_{t,\text{DIM,SRS}}(\mathbf{y}) \qquad \text{(C.1)}
\end{aligned}$$

Since $\hat{\mu}_{t,\text{DIM,SRS}}$ is linear,

$$\hat{\delta}_{\text{DIM,SRS}}(a + \mathbf{y}) = \hat{\mu}_{1,\text{DIM,SRS}}(a + \mathbf{y}) - \hat{\mu}_{0,\text{DIM,SRS}}(a + \mathbf{y}) = \hat{\delta}_{\text{DIM,SRS}}. \qquad \text{(C.2)}$$

C.2 Expectation of DIM estimator for PATE

Since the DIM estimator of μ_t is a ratio estimator, following Lohr (2010) and Middleton and Aronow (2015), we can use the following relationship to calculate the expectation of the DIM estimator. When both u and $v \geq 0$ are random, it can be shown that

$$\mathbb{E}\left(\frac{u}{v}\right) = \frac{1}{\mathbb{E}(v)}\left[\mathbb{E}(u) - \text{Cov}\left(\frac{u}{v}, v\right)\right]. \tag{C.3}$$

Given that,

$$\mathbb{E}\left(\sum_{c=1}^{\ell} S_c T_{ct} \sum_{k=1}^{n_c} y_{kct} S_{kc}\right) = \sum_{c=1}^{\ell}\sum_{k=1}^{n_c} \frac{\#T_t}{\ell}\frac{s_c}{n_c} y_{kct} \tag{C.4}$$

$$\mathbb{E}\left(\sum_{c=1}^{\ell} S_c T_{ct} s_c\right) = \sum_{c=1}^{\ell} \frac{\#T_t}{\ell} s_c, \tag{C.5}$$

the expectation of the DIM estimator for the population mean is

$$\begin{aligned}\mathbb{E}(\hat{\mu}_{t,\text{DIM, SRS}}) &= \mathbb{E}\left(\frac{\sum_{c=1}^{\ell} S_c T_{ct} \sum_{k=1}^{n_c} y_{kct} S_{kc}}{\sum_{c=1}^{\ell} S_c T_{ct} s_c}\right) \\ &= \frac{1}{\#K}\left[\sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\frac{s_c}{n_c} y_{kct} - \frac{\ell}{\#T_t}\text{Cov}\left(\hat{\mu}_{t,\text{DIM, SRS}}, \#K_t\right)\right]\end{aligned} \tag{C.6}$$

Thus, the DIM estimator is biased for PATE:

$$\begin{aligned}\mathbb{E}(\hat{\delta}_{\text{DIM,SRS}}) &= \mathbb{E}(\hat{\mu}_{1,\text{DIM,SRS}}) - \mathbb{E}(\hat{\mu}_{0,\text{DIM,SRS}}) \\ &= \frac{1}{\#K}\left[\sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\frac{s_c}{n_c} y_{kc1} - \frac{\ell}{\#T_1}\text{Cov}\left(\hat{\mu}_{1,\text{DIM, SRS}}, \#K_1\right)\right] \\ &\quad - \frac{1}{\#K}\left[\sum_{c^*=1}^{\ell}\sum_{k^*=1}^{n_{c^*}}\frac{s_{c^*}}{n_{c^*}} y_{k^*c^*0} - \frac{\ell}{\#T_0}\text{Cov}\left(\hat{\mu}_{0,\text{DIM, SRS}}, \#K_0\right)\right] \\ &= \frac{1}{\#K}\left[\sum_{c=1}^{\ell}\sum_{k=1}^{n_c}\frac{s_c}{n_c}(y_{kc1} - y_{kc0})\right]\end{aligned}$$

$$
\begin{aligned}
&- \frac{\ell}{\#K}\left[\frac{1}{\#T_1}\mathrm{Cov}\left(\hat{\mu}_{1,\mathrm{DIM,\ SRS}}, \#K_1\right) + \frac{1}{\#T_0}\mathrm{Cov}\left(\hat{\mu}_{0,\mathrm{DIM,\ SRS}}, \#K_0\right)\right] \\
&= \sum_{c=1}^{\ell} \frac{s_c}{\#K}(\mu_{c1} - \mu_{c0}) \\
&\quad - \frac{\ell}{\#K}\left[\frac{1}{\#T_1}\mathrm{Cov}\left(\hat{\mu}_{1,\mathrm{DIM,\ SRS}}, \#K_1\right) + \frac{1}{\#T_0}\mathrm{Cov}\left(\hat{\mu}_{0,\mathrm{DIM,\ SRS}}, \#K_0\right)\right]. \quad \text{(C.7)}
\end{aligned}
$$

The bias of the DIM estimator for PATE is

$$
\begin{aligned}
bias(\hat{\delta}_{\mathrm{DIM,SRS}}) &= \mathbb{E}(\hat{\delta}_{\mathrm{DIM,SRS}}) - \delta \\
&= \sum_{c=1}^{\ell}\left(\frac{s_c}{\#K} - \frac{n_c}{n}\right)(\mu_{c1} - \mu_{c0}) \\
&\quad - \frac{\ell}{\#K}\left[\frac{1}{\#T_1}\mathrm{Cov}\left(\hat{\mu}_{1,\mathrm{DIM,\ SRS}}, \#K_1\right) + \frac{1}{\#T_0}\mathrm{Cov}\left(\hat{\mu}_{0,\mathrm{DIM,\ SRS}}, \#K_0\right)\right].
\end{aligned} \quad \text{(C.8)}
$$

C.3 Variance of DIM estimator for PATE

C.3.1 Variance of DIM estimator for population mean

Let

$$\hat{\tau}^*_{ct} = \sum_{k=1}^{n_c} y_{kct} S_{kc}, \quad \text{(C.9)}$$

$$\tau^*_{ct} = \sum_{k=1}^{n_c} \frac{s_c}{n_c} y_{kct} \quad \text{(C.10)}$$

and

$$\mu^*_t = \sum_{c=1}^{\ell} \frac{s_c}{\#K}\mu_{ct}. \quad \text{(C.11)}$$

Using Taylor series expansion, the DIM estimator can be approximated by

$$\hat{\mu}_{t,\mathrm{DIM,\ SRS}} \approx \mu^*_t + \frac{\ell}{\#K}\sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t}\left(\hat{\tau}^*_{ct} - \mu^*_t s_c\right). \quad \text{(C.12)}$$

Then

$$
\begin{aligned}
&\mathrm{Var}(\hat{\mu}_{t,\mathrm{DIM,\ SRS}}) \\
&\approx \frac{1}{\#K^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\ell^2\mathrm{Cov}\left[\frac{S_cT_{ct}}{\#T_t}\left(\hat{\tau}^*_{ct}-\mu^*_t s_c\right),\frac{S_{c^*}T_{c^*t}}{\#T_t}\left(\hat{\tau}^*_{c^*t}-\mu^*_t s_{c^*}\right)\right] \\
&= \frac{1}{\#K^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\ell^2\mathrm{Cov}\left(\mathbb{E}\left[\left.\frac{S_cT_{ct}}{\#T_t}\left(\hat{\tau}^*_{ct}-\mu^*_t s_c\right)\right|\mathbf{S},\mathbf{T}\right],\mathbb{E}\left[\left.\frac{S_{c^*}T_{c^*t}}{\#T_t}\left(\hat{\tau}^*_{c^*t}-\mu^*_t s_{c^*}\right)\right|\mathbf{S},\mathbf{T}\right]\right) \\
&\quad+\frac{1}{\#K^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\ell^2\mathbb{E}\left(\mathrm{Cov}\left[\left.\frac{S_cT_{ct}}{\#T_t}\left(\hat{\tau}^*_{ct}-\mu^*_t s_c\right),\frac{S_{c^*}T_{c^*t}}{\#T_t}\left(\hat{\tau}^*_{c^*t}-\mu^*_t s_{c^*}\right)\right|\mathbf{S},\mathbf{T}\right]\right). \qquad \text{(C.13)}
\end{aligned}
$$

Focusing on the first term of (C.13):

$$
\begin{aligned}
&\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\ell^2\mathrm{Cov}\left(\mathbb{E}\left[\left.\frac{S_cT_{ct}}{\#T_t}\left(\hat{\tau}^*_{ct}-\mu^*_t s_c\right)\right|\mathbf{S},\mathbf{T}\right],\mathbb{E}\left[\left.\frac{S_{c^*}T_{c^*t}}{\#T_t}\left(\hat{\tau}^*_{c^*t}-\mu^*_t s_{c^*}\right)\right|\mathbf{S},\mathbf{T}\right]\right) \\
&=\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\ell^2\mathrm{Cov}\left[\frac{S_cT_{ct}}{\#T_t}\left(\tau^*_{ct}-\mu^*_t s_c\right),\frac{S_{c^*}T_{c^*t}}{\#T_t}\left(\tau^*_{c^*t}-\mu^*_t s_{c^*}\right)\right] \\
&=\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\ell^2\left(\tau^*_{ct}-\mu^*_t s_c\right)\left(\tau^*_{c^*t}-\mu^*_t s_{c^*}\right)\mathrm{Cov}\left(\frac{S_cT_{ct}}{\#T_t},\frac{S_{c^*}T_{c^*t}}{\#T_t}\right) \\
&=\sum_{c=1}^{\ell}\ell^2\left(\tau^*_{ct}-\mu^*_t s_c\right)^2\mathrm{Var}\left(\frac{S_cT_{ct}}{\#T_t}\right) \\
&\quad+\sum_{c=1}^{\ell}\sum_{c'\neq c}\ell^2\left(\tau^*_{ct}-\mu^*_t s_c\right)\left(\tau^*_{c't}-\mu^*_t s_{c'}\right)\mathrm{Cov}\left(\frac{S_cT_{ct}}{\#T_t},\frac{S_{c'}T_{c't}}{\#T_t}\right) \\
&=\sum_{c=1}^{\ell}\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell-1\right]\left(\tau^*_{ct}-\mu^*_t s_c\right)^2 \\
&\quad-\sum_{c=1}^{\ell}\sum_{c'\neq c}\left[\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{(\ell-1)}-1\right]\left(\tau^*_{ct}-\mu^*_t s_c\right)\left(\tau^*_{c't}-\mu^*_t s_{c'}\right) \\
&=\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell}\left(\tau^*_{ct}-\mu^*_t s_c\right)^2-\sum_{c=1}^{\ell}\left(\tau^*_{ct}-\mu^*_t s_c\right)^2 \\
&\quad+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau^*_{ct}-\mu^*_t s_c\right)\left(\tau^*_{c't}-\mu^*_t s_{c'}\right) \\
&\quad-\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau^*_{ct}-\mu^*_t s_c\right)\left(\tau^*_{c't}-\mu^*_t s_{c'}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)^2 \\
&\quad+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\left(\tau_{c't}^{*}-\mu_t^{*}s_{c'}\right)-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right]^2 \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\left(\ell\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)^2-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right]^2\right) \\
&\quad+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left(\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\left(\tau_{c't}^{*}-\mu_t^{*}s_{c'}\right)\right. \\
&\qquad\qquad\left.-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right]^2\right) \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell(\ell-1)\left(\frac{1}{\ell-1}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)^2-\frac{\ell}{\ell-1}\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right]^2\right) \\
&\quad+\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left(\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\left(\tau_{c't}^{*}-\mu_t^{*}s_{c'}\right)\right. \\
&\qquad\qquad\left.-\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)^2-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right]^2\right) \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell(\ell-1)\mathrm{Var}\left[\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right] \\
&\quad-\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\ell\left(\frac{1}{\ell-1}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)^2-\frac{\ell}{\ell-1}\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right]^2\right) \\
&= \ell^2\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)-\frac{1}{\ell}\right]\mathrm{Var}\left[\left(\tau_{ct}^{*}-\mu_t^{*}s_c\right)\right] \qquad \text{(C.14)}
\end{aligned}
$$

The second term of (C.13):

$$
\begin{aligned}
&\ell^2\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathbb{E}\left(\mathrm{Cov}\left[\frac{S_cT_{ct}}{\#T_t}\left(\hat{\tau}_{ct}^{*}-\mu_t^{*}s_c\right),\frac{S_{c^*}T_{c^*t}}{\#T_t}\left(\hat{\tau}_{c^*t}^{*}-\mu_t^{*}s_{c^*}\right)\middle|\mathbf{S},\mathbf{T}\right]\right) \\
&= \ell^2\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathbb{E}\left[\frac{S_cS_{c^*}T_{ct}T_{c^*t}}{\#T_t^2}\mathrm{Cov}\left(\hat{\tau}_{ct}^{*},\hat{\tau}_{c^*t}^{*}\right)\right] \\
&= \ell^2\sum_{c=1}^{\ell}\mathrm{Var}\left(\hat{\tau}_{ct}^{*}\right)\mathbb{E}\left(\frac{S_c^2T_{ct}^2}{\#T_t^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell}\mathrm{Var}\left(\sum_{k=1}^{n_c} y_{kct}S_{kc}\right) \\
&= \mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell} s_c\left(1-\frac{s_c}{n_c}\right)\mathrm{Var}(y_{kct}) \qquad \text{(C.15)}
\end{aligned}
$$

The variance of $\hat{\mu}_{t,\text{DIM,SRS}}$ is then

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}_{t,\text{DIM, SRS}}) \approx{}& \frac{\ell^2}{\#K^2}\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)-\frac{1}{\ell}\right]\mathrm{Var}\left[(\tau^*_{ct}-\mu^*_t s_c)\right] \\
&+\frac{\ell}{\#K^2}\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell} s_c\left(1-\frac{s_c}{n_c}\right)\mathrm{Var}(y_{kct}) \qquad \text{(C.16)}
\end{aligned}
$$

C.3.2 Covariance of DIM estimator for the population means

Based on the Taylor expansion in eq. (C.12),

$$
\begin{aligned}
&\mathrm{Cov}(\hat{\mu}_{1,\text{DIM, SRS}},\hat{\mu}_{0,\text{DIM, SRS}}) \\
&\approx \frac{\ell^2}{\#K^2}\mathrm{Cov}\left[\sum_{c=1}^{\ell}\frac{S_cT_{c1}}{\#T_1}\left(\hat{\tau}^*_{c1}-\mu^*_1 s_c\right),\sum_{c^*=1}^{\ell}\frac{S_{c^*}T_{c^*0}}{\#T_0}\left(\hat{\tau}^*_{c^*0}-\mu^*_0 s_{c^*}\right)\right] \\
&= \frac{\ell^2}{\#K^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathrm{Cov}\left(\mathbb{E}\left[\left.\frac{S_cT_{c1}}{\#T_1}\left(\hat{\tau}^*_{c1}-\mu^*_1 s_c\right)\right|\mathbf{S},\mathbf{T}\right],\mathbb{E}\left[\left.\frac{S_{c^*}T_{c^*0}}{\#T_0}\left(\hat{\tau}^*_{c^*0}-\mu^*_0 s_{c^*}\right)\right|\mathbf{S},\mathbf{T}\right]\right) \\
&\quad+\frac{\ell^2}{\#K^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathbb{E}\left(\mathrm{Cov}\left[\left.\frac{S_cT_{c1}}{\#T_1}\left(\hat{\tau}^*_{c1}-\mu^*_1 s_c\right),\frac{S_{c^*}T_{c^*0}}{\#T_0}\left(\hat{\tau}^*_{c^*0}-\mu^*_0 s_{c^*}\right)\right|\mathbf{S},\mathbf{T}\right]\right) \\
&= \frac{\ell^2}{\#K^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\left(\tau^*_{c1}-\mu^*_1 s_c\right)\left(\tau^*_{c^*0}-\mu^*_0 s_{c^*}\right)\mathrm{Cov}\left(\frac{S_cT_{c1}}{\#T_1},\frac{S_{c^*}T_{c^*0}}{\#T_0}\right) \\
&= \frac{\ell^2}{\#K^2}\sum_{c=1}^{\ell}\left(\tau^*_{c1}-\mu^*_1 s_c\right)\left(\tau^*_{c0}-\mu^*_0 s_c\right)\mathrm{Var}\left(\frac{S_cT_{c1}T_{c0}}{\#T_1\#T_0}\right) \\
&\qquad+\frac{\ell^2}{\#K^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau^*_{c1}-\mu^*_1 s_c\right)\left(\tau^*_{c'0}-\mu^*_0 s_{c'}\right)\mathrm{Cov}\left(\frac{S_cT_{c1}}{\#T_1},\frac{S_{c'}T_{c'0}}{\#T_0}\right) \\
&= \frac{1}{\#K^2}\frac{1}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau^*_{c1}-\mu^*_1 s_c\right)\left(\tau^*_{c'0}-\mu^*_0 s_{c'}\right)-\frac{1}{\#K^2}\sum_{c=1}^{\ell}\left(\tau^*_{c1}-\mu^*_1 s_c\right)\left(\tau^*_{c0}-\mu^*_0 s_c\right) \qquad \text{(C.17)}
\end{aligned}
$$

The third equality occurs because units are sampled independently across clusters and that units in the same cluster are given the same treatments (i. e., $\mathrm{Cov}(S_{kc}, S_{k^*c'}) = \mathrm{Cov}(S_{kc}, S_{k'c}) = 0$).

C.3.3 Variance of DIM estimator for PATE

From eq. (C.16) and eq. (C.17), the variance of the DIM estimator can be approximated by

$$\begin{aligned}
\mathrm{Var}(\hat{\delta}_{\mathrm{DIM,SRS}}) \approx & \frac{\ell^2}{\#K^2} \sum_{t=0}^{1} \left[\mathbb{E}\left(\frac{1}{\#T_t}\right) - \frac{1}{\ell} \right] \mathrm{Var}\left(\tau_{ct}^* - \mu_t^* s_c\right) \\
& + \frac{\ell}{\#K^2} \sum_{t=0}^{1} \mathbb{E}\left(\frac{1}{\#T_t}\right) \sum_{c=1}^{\ell} s_c \left(1 - \frac{s_c}{n_c}\right) \mathrm{Var}(y_{kct}) \\
& - \frac{2}{\#K^2} \frac{1}{\ell - 1} \sum_{c=1}^{\ell} \sum_{c' \neq c} \left(\tau_{c1}^* - \mu_1^* s_c\right) \left(\tau_{c0}^* - \mu_0^* s_c\right) \\
& + \frac{2}{\#K^2} \sum_{c=1}^{\ell} \left(\tau_{c1}^* - \mu_1^* s_c\right) \left(\tau_{c0}^* - \mu_0^* s_c\right). \qquad \text{(C.18)}
\end{aligned}$$

C.4 Variance estimator of DIM estimator for PATE

$$\begin{aligned}
\widehat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{DIM,\ SRS}}\right) = & \sum_{t=0}^{1} \frac{\ell^2}{\#K^2} \sum_{c=1}^{\ell} \left(1 - \frac{\#T_t}{\ell}\right) \frac{S_c T_{ct}}{\#T_t} \widehat{\mathrm{Var}}\left[\left(\hat{\tau}_{ct}^* - \hat{\mu}_t^* s_c\right)\right] \\
& - 2 \frac{\ell}{\#K^2} \sum_{c=1}^{\ell} \sum_{c' \neq c} \frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0} \left(\hat{\tau}_{c1}^* - \hat{\mu}_1^* s_c\right) \left(\hat{\tau}_{c'0}^* - \hat{\mu}_0^* s_{c'}\right) \\
& + \sum_{t=0}^{1} \frac{\ell}{\#K^2} \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \left(\hat{\tau}_{ct}^* - \hat{\mu}_t^* s_c\right)^2 \qquad \text{(C.19)}
\end{aligned}$$

where

$$\widehat{\mathrm{Var}}\left[\left(\tau_{ct}^* - \mu_t^* s_c\right)\right] = \frac{1}{\#T_t - 1} \sum_{c=1}^{\ell} S_c T_{ct} \left(\hat{\tau}_{ct}^* - \hat{\mu}_t^* s_c\right)^2. \qquad \text{(C.20)}$$

Appendix D

Properties of Hajék estimator

Let S_c be the indicator variable for sampling clusters, T_{ct} be the assignment indicator for treatment t, and S_{kc} be the indicator for sampling units. Suppose that $\#N = \sum_{c=1}^{\ell} n_c$ is the total number of units in sampled clusters and that $\#N_t = \sum_{c=1}^{\ell} S_c T_{ct} n_c$ is the number of units given treatment t. This section uses the indicator properties calculated in section B.1.

D.1 Linear transformation on the HJ estimator for PATE

For any constants a, b,

$$\begin{aligned}
\hat{\mu}_{t,\text{HJ,SRS}}(a + b\mathbf{y}) &= \frac{\sum_{c=1}^{\ell} S_c T_{ct} (\frac{n_c}{s_c}) \sum_{k=1}^{n_c} (a + b y_{kct}) S_{kc}}{\#N_t} \\
&= \frac{\sum_{c=1}^{\ell} S_c T_{ct} (\frac{n_c}{s_c}) \sum_{k=1}^{n_c} a S_{kc}}{\#N_t} + \frac{\sum_{c=1}^{\ell} S_c T_{ct} (\frac{n_c}{s_c}) \sum_{k=1}^{n_c} b y_{kct} S_{kc}}{\#N_t} \\
&= a \frac{\sum_{c=1}^{\ell} S_c T_{ct} (\frac{n_c}{s_c}) \sum_{k=1}^{n_c} S_{kc}}{\#N_t} + b \frac{\sum_{c=1}^{\ell} S_c T_{ct} (\frac{n_c}{s_c}) \sum_{k=1}^{n_c} y_{kct} S_{kc}}{\#N_t} \\
&= a \frac{\#N_t}{\#N_t} + b \hat{\mu}_{t,\text{HJ,SRS}}(\mathbf{y}) \\
&= a + b \hat{\mu}_{t,\text{HJ,SRS}}(\mathbf{y}) \qquad \text{(D.1)}
\end{aligned}$$

Since $\hat{\mu}_{t,\text{HJ,SRS}}$ is linear,

$$\begin{aligned}\hat{\delta}_{\text{HJ,SRS}}(a+\mathbf{y}) &= \hat{\mu}_{1,\text{HJ,SRS}}(a+\mathbf{y}) - \hat{\mu}_{0,\text{HJ,SRS}}(a+\mathbf{y}) \\ &= \hat{\delta}_{\text{HJ,SRS}}. \end{aligned} \tag{D.2}$$

D.2 Expectation of HJ estimator for PATE

The expectation of a ratio estimator is

$$\mathbb{E}\left(\frac{u}{v}\right) = \frac{1}{\mathbb{E}(v)}\left[\mathbb{E}(u) - \text{Cov}\left(\frac{u}{v}, v\right)\right]. \tag{D.3}$$

We first need to find the expectations of the numerator and denominator of the HJ estimator:

$$\begin{aligned}\mathbb{E}\left(\sum_{c=1}^{\ell} S_c T_{ct} \frac{n_c}{s_c} \sum_{k=1}^{n_c} y_{kct} S_{kc}\right) &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kct}\frac{n_c}{s_c}\mathbb{E}(S_c T_{ct} S_{kc}) \\ &= \sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kct}\frac{n_c}{s_c}\frac{\#T_t s_c}{\ell n_c} \\ &= \frac{\#T_t}{\ell}\sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kct} \end{aligned} \tag{D.4}$$

$$\begin{aligned}\mathbb{E}(\#N_t) = \mathbb{E}\left(\sum_{c=1}^{\ell} S_c T_{ct} n_c\right) &= \sum_{c=1}^{\ell} n_c \mathbb{E}(S_c T_{ct}) \\ &= \sum_{c=1}^{\ell} n_c \frac{\#T_t}{\ell} \\ &= \frac{\#T_t n}{\ell}. \end{aligned} \tag{D.5}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left(\hat{\mu}_{t,\text{HJ}}\right) &= \frac{1}{\mathbb{E}(\#N_t)}\left[\mathbb{E}\left(\sum_{c=1}^{\ell} S_c T_{ct}\frac{n_c}{s_c}\sum_{k=1}^{n_c} y_{kct} S_{kc}\right) - \text{Cov}\left(\hat{\mu}_t, \#N_t\right)\right] \\ &= \frac{\ell}{\#T_t n}\left[\frac{\#T_t}{\ell}\sum_{c=1}^{\ell}\sum_{k=1}^{n_c} y_{kct} - \text{Cov}\left(\hat{\mu}_t, \#N_t\right)\right] \end{aligned}$$

$$= \mu_t - \frac{\ell}{\#T_t n}\mathrm{Cov}\left(\hat{\mu}_t, \#N_t\right). \tag{D.6}$$

D.3 Variance of HJ estimator for PATE

D.3.1 Variance of HJ estimator for population mean

Since the numerator and denominator of the HJ estimator are random, finding the exact variance can be difficult. We can use Taylor series linearization to approximate it:

$$\hat{\mu}_{t,\mathrm{HJ}} \approx \mu_t + \frac{1}{\#N}\sum_{c=1}^{\ell}\left(\hat{\tau}_{ct} - \mu_t n_c\right)\frac{\ell}{\#T_t}S_c T_{ct}, \tag{D.7}$$

where $\hat{\tau}_{ct} = \sum_{c=1}^{n_c}\frac{n_c}{s_c}y_{kct}S_{kc}$. The variance of the HJ estimator is then

$$\mathrm{Var}\left(\hat{\mu}_{t,\mathrm{HJ}}\right) \approx \frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathrm{Cov}\left[\frac{\ell}{\#T_t}\left(\hat{\tau}_{ct} - \mu_t n_c\right)S_c T_{ct}, \frac{\ell}{\#T_t}\left(\hat{\tau}_{c^*t} - \mu_t n_{c^*}\right)S_{c^*}T_{c^*t}\right]. \tag{D.8}$$

By the law of total (co)variance,

$$\begin{aligned}
&\mathrm{Var}\left(\hat{\mu}_{t,\mathrm{HJ}}\right) \approx \\
&\quad\frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathrm{Cov}\left(\mathbb{E}\left[\frac{\ell}{\#T_t}\left(\hat{\tau}_{ct} - \mu_t n_c\right)S_c T_{ct}\middle|\mathbf{S},\mathbf{T}\right], \mathbb{E}\left[\frac{\ell}{\#T_t}\left(\hat{\tau}_{c^*t} - \mu_t n_{c^*}\right)S_{c^*}T_{c^*t}\middle|\mathbf{S},\mathbf{T}\right]\right) \\
&\qquad+\frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathbb{E}\left(\mathrm{Cov}\left[\frac{\ell}{\#T_t}\left(\hat{\tau}_{ct} - \mu_t n_c\right)S_c T_{ct}, \frac{\ell}{\#T_t}\left(\hat{\tau}_{c^*t} - \mu_t n_{c^*t}\right)S_{c^*}T_{c^*t}\middle|\mathbf{S},\mathbf{T}\right]\right) \\
&\quad= \frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathrm{Cov}\left(\frac{\ell}{\#T_t}(\tau_{ct} - \mu_t n_c)S_c T_{ct}, \frac{\ell}{\#T_t}(\tau_{c^*t} - \mu_t n_{c^*})S_{c^*}T_{c^*t}\right) \\
&\qquad+\frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathbb{E}\left(\frac{\ell^2}{\#T_t^2}\mathrm{Cov}\left[\left(\hat{\tau}_{ct} - \mu_t n_c\right), \left(\hat{\tau}_{c^*t} - \mu_t n_{c^*}\right)\middle|\mathbf{S},\mathbf{T}\right]S_c S_{c^*}T_{ct}T_{c^*t}\right) \\
&\quad= \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}(\tau_{ct} - \mu_t n_c)(\tau_{c^*t} - \mu_t n_{c^*})\mathrm{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \frac{S_{c^*}T_{c^*t}}{\#T_t}\right) \\
&\qquad+\frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathrm{Cov}(\hat{\tau}_{ct}, \hat{\tau}_{c^*t}|\mathbf{S},\mathbf{T})\mathbb{E}\left(\frac{S_c S_{c^*}T_{ct}T_{c^*t}}{\#T_t^2}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}(\tau_{ct}-\mu_t n_c)^2 \mathrm{Var}\left(\frac{S_c T_{ct}}{\#T_t}\right) \\
&\quad + \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}(\tau_{ct}-\mu_t n_c)(\tau_{c't}-\mu_t n_{c'})\mathrm{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \frac{S_{c'} T_{c't}}{\#T_t}\right) \\
&\quad + \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{S_c^2 T_{ct}^2}{\#T_t^2}\right) + \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}\mathrm{Cov}(\hat{\tau}_{ct},\hat{\tau}_{c't})\mathbb{E}\left(\frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2}\right).
\end{aligned} \tag{D.9}
$$

Since the cluster totals are independent across clusters, the last term of D.9 will be 0, and so

$$
\begin{aligned}
\mathrm{Var}\left(\hat{\mu}_{t,\mathrm{HJ}}\right) &\approx \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}(\tau_{ct}-\mu_t n_c)^2\left[\frac{1}{\ell}\mathbb{E}\left(\frac{1}{\#T_t}\right)-\frac{1}{\ell^2}\right] \\
&\quad - \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}(\tau_{ct}-\mu_t n_c)(\tau_{c't}-\mu_t n_{c'})\left[\frac{1}{\ell(\ell-1)}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)-\frac{1}{\ell^2}\right] \\
&\quad + \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\frac{1}{\ell}\mathbb{E}\left(\frac{1}{\#T_t}\right) \\
&= \frac{1}{\#N^2}\sum_{c=1}^{\ell}\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell-1\right](\tau_{ct}-\mu_t n_c)^2 \\
&\quad - \frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left[\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{(\ell-1)}-1\right](\tau_{ct}-\mu_t n_c)(\tau_{c't}-\mu_t n_{c'}) \\
&\quad + \frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right) \\
&= \frac{1}{\#N^2}\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell}(\tau_{ct}-\mu_t n_c)^2 - \frac{1}{\#N^2}\sum_{c=1}^{\ell}(\tau_{ct}-\mu_t n_c)^2 \\
&\quad + \frac{1}{\#N^2}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}(\tau_{ct}-\mu_t n_c)(\tau_{c't}-\mu_t n_{c'}) \\
&\quad - \frac{1}{\#N^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}(\tau_{ct}-\mu_t n_c)(\tau_{c't}-\mu_t n_{c'}) + \frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right) \\
&= \frac{1}{\#N^2}\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell\sum_{c=1}^{\ell}(\tau_{ct}-\mu_t n_c)^2 - \frac{\ell^2}{\#N^2}\left[\frac{1}{\ell}\sum_{c=1}^{\ell}(\tau_{ct}-\mu_t n_c)\right]^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{\#N^2}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{ct}-\mu_t n_c\right)\left(\tau_{c't}-\mu_t n_{c'}\right)\\
&+ \frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right)\\
&= \frac{1}{\#N^2}\mathbb{E}\left(\frac{1}{\#T_t}\right)\left(\ell\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)^2-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)\right]^2\right)\\
&+ \frac{1}{\#N^2}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left(\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{ct}-\mu_t n_c\right)\left(\tau_{c't}-\mu_t n_{c'}\right)\right.\\
&\qquad\qquad\left.-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)\right]^2\right)\\
&+ \frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right)\\
&= \frac{1}{\#N^2}\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell(\ell-1)\left(\frac{1}{\ell-1}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)^2-\frac{\ell}{\ell-1}\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)\right]^2\right)\\
&+ \frac{1}{\#N^2}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\left(\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)\left(\tau_{c't}-\mu_t n_{c'}\right)\right.\\
&\qquad\qquad\left.-\frac{\ell}{\ell-1}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)^2-\ell^2\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)\right]^2\right)\\
&+ \frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right)\\
&= \frac{1}{\#N^2}\mathbb{E}\left(\frac{1}{\#T_t}\right)\ell(\ell-1)\mathrm{Var}\left[\left(\tau_{ct}-\mu_t n_c\right)\right]\\
&- \frac{1}{\#N^2}\mathbb{E}\left(1-\frac{1}{\#T_t}\right)\ell\left(\frac{1}{\ell-1}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)^2-\frac{\ell}{\ell-1}\left[\frac{1}{\ell}\sum_{c=1}^{\ell}\left(\tau_{ct}-\mu_t n_c\right)\right]^2\right)\\
&+ \frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right)\\
&= \ell^2\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)-\frac{1}{\ell}\right]\mathrm{Var}\left[\left(\tau_{ct}-\mu_t n_c\right)\right]+\frac{\ell}{\#N^2}\sum_{c=1}^{\ell}\mathrm{Var}(\hat{\tau}_{ct})\mathbb{E}\left(\frac{1}{\#T_t}\right) \qquad \text{(D.10)}
\end{aligned}
$$

D.3.2 Covariance of HJ estimator for population mean

$$
\begin{aligned}
&\mathrm{Cov}(\hat{\mu}_{1,\mathrm{HJ}}, \hat{\mu}_{0,\mathrm{HJ}}) \\
&\approx \frac{\ell^2}{\#N^2}\mathrm{Cov}\left[\sum_{c=1}^{\ell}\frac{S_c T_{c1}}{\#T_1}\left(\hat{\tau}_{c1}-\mu_1 n_c\right), \sum_{c^*=1}^{\ell}\frac{S_{c^*} T_{c^*0}}{\#T_0}\left(\hat{\tau}_{c^*0}-\mu_0 n_{c^*}\right)\right] \\
&= \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathrm{Cov}\left(\mathbb{E}\left[\left.\frac{S_c T_{c1}}{\#T_1}\left(\hat{\tau}_{c1}-\mu_1 n_c\right)\right| \mathbf{S},\mathbf{T}\right], \mathbb{E}\left[\left.\frac{S_{c^*} T_{c^*0}}{\#T_0}\left(\hat{\tau}_{c^*0}-\mu_0 n_{c^*}\right)\right| \mathbf{S},\mathbf{T}\right]\right) \\
&\quad + \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\mathbb{E}\left(\mathrm{Cov}\left[\left.\frac{S_c T_{c1}}{\#T_1}\left(\hat{\tau}_{c1}-\mu_1 n_c\right), \frac{S_{c^*} T_{c^*0}}{\#T_0}\left(\hat{\tau}_{c^*0}-\mu_0 n_{c^*}\right)\right| \mathbf{S},\mathbf{T}\right]\right) \\
&= \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c^*=1}^{\ell}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c^*0}-\mu_0 n_{c^*}\right)\mathrm{Cov}\left(\frac{S_c T_{c1}}{\#T_1}, \frac{S_{c^*} T_{c^*0}}{\#T_0}\right) \\
&= \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c0}-\mu_0 n_c\right)\mathrm{Var}\left(\frac{S_c T_{c1} T_{c0}}{\#T_1 \#T_0}\right) \\
&\qquad + \frac{\ell^2}{\#N^2}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c'0}-\mu_0 n_{c'}\right)\mathrm{Cov}\left(\frac{S_c T_{c1}}{\#T_1}, \frac{S_{c'} T_{c'0}}{\#T_0}\right) \\
&= \frac{1}{\#N^2}\frac{1}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c'0}-\mu_0 n_{c'}\right) - \frac{1}{\#N^2}\sum_{c=1}^{\ell}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c0}-\mu_0 n_c\right)
\end{aligned}
\tag{D.11}
$$

D.3.3 Variance of HJ estimator for PATE

From eq. (D.10) and eq. (D.11), the variance of the DIM estimator can be approximated by

$$
\begin{aligned}
\mathrm{Var}(\hat{\delta}_{\mathrm{HJ,SRS}}) \approx\ & \frac{\ell^2}{\#N^2}\sum_{t=0}^{1}\left[\mathbb{E}\left(\frac{1}{\#T_t}\right)-\frac{1}{\ell}\right]\mathrm{Var}\left(\tau_{ct}-\mu_t n_c\right) \\
& + \frac{\ell}{\#N^2}\sum_{t=0}^{1}\mathbb{E}\left(\frac{1}{\#T_t}\right)\sum_{c=1}^{\ell}s_c\left(1-\frac{s_c}{n_c}\right)\mathrm{Var}(y_{kct}) \\
& - \frac{2}{\#N^2}\frac{1}{\ell-1}\sum_{c=1}^{\ell}\sum_{c'\neq c}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c0}-\mu_0 n_c\right) \\
& + \frac{2}{\#N^2}\sum_{c=1}^{\ell}\left(\tau_{c1}-\mu_1 n_c\right)\left(\tau_{c0}-\mu_0 n_c\right).
\end{aligned}
\tag{D.12}
$$

D.4 Variance estimator of HJ estimator for PATE

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{HJ,\ SRS}}\right) =& \sum_{t=0}^{1} \frac{\ell^2}{\#N^2} \sum_{c=1}^{\ell} \left(1 - \frac{\#T_t}{\ell}\right) \frac{S_c T_{ct}}{\#T_t} \widehat{\mathrm{Var}}\left[(\hat{\tau}_{ct} - \hat{\mu}_t s_c)\right] \\ & - 2\frac{\ell}{\#N^2} \sum_{c=1}^{\ell} \sum_{c' \neq c} \frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0} \left(\hat{\tau}_{c1} - \hat{\mu}_1 s_c\right) \left(\hat{\tau}_{c'0} - \hat{\mu}_0 s_{c'}\right) \\ & + \sum_{t=0}^{1} \frac{\ell}{\#N^2} \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \left(\hat{\tau}_{ct} - \hat{\mu}_t s_c\right)^2 \end{aligned} \tag{D.13}$$

where

$$\widehat{\mathrm{Var}}\left[(\tau_{ct} - \mu_t s_c)\right] = \frac{1}{\#T_t - 1} \sum_{c=1}^{\ell} S_c T_{ct} \left(\hat{\tau}_{ct} - \hat{\mu}_t s_c\right)^2 . \tag{D.14}$$

Appendix E

Properties of the DR estimator

Let S_c be the indicator variable for sampling clusters, T_{ct} be the assignment indicator for treatment t, and S_{kc} be the indicator for sampling units. This section uses the indicator properties calculated in section B.1.

E.1 Linear transforms on DR estimator for PATE

For any constants a, b,

$$\begin{aligned}
\hat{\mu}_{t,\text{DR,SRS}}(a+b\mathbf{y}) &= \ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \left[\sum_{k=1}^{n_c} \frac{(a+by_{kct})S_{kc}}{s_c} - \frac{(a+b\theta_t)}{n_c} \left(n_c - \frac{n}{\ell} \right) \right] \\
&= \ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \left[a - \frac{a}{n_c} \left(n_c - \frac{n}{\ell} \right) + \sum_{k=1}^{n_c} \frac{by_{kct} S_{kc}}{s_c} - \frac{b\theta_t}{n_c} \left(n_c - \frac{n}{\ell} \right) \right] \\
&= a \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} + b\hat{\mu}_{t,\text{DR,SRS}} \\
&= a + b\hat{\mu}_{t,\text{DR,SRS}}(\mathbf{y}) \qquad \text{(E.1)}
\end{aligned}$$

Since $\hat{\mu}_t$ is linear, $\hat{\delta}$ is location-invariant:

$$\hat{\delta}_{\text{DR,SRS}}(a+\mathbf{y}) = \hat{\mu}_{1,\text{DR,SRS}}(a+\mathbf{y}) - \hat{\mu}_{0,\text{DR,SRS}}(a+\mathbf{y}) = \hat{\delta}_{\text{DR,SRS}}. \qquad \text{(E.2)}$$

E.2 Expectation of DR estimator for population mean

From eq. (B.17),

$$\begin{aligned}
\mathbb{E}(\mu_{t,\mathrm{DR,SRS}}) &= \mathbb{E}\left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \left[\sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} - \frac{\theta_t}{n_c}\left(n_c - \frac{n}{\ell}\right)\right]\right) \\
&= \mathbb{E}\left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c}\right) - \mathbb{E}\left(\ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{\theta_t}{n}\left[n_c - \frac{n}{\ell}\right]\right) \\
&= \mu_t - \ell \sum_{c=1}^{\ell} \frac{\theta_t}{n}\left(n_c - \frac{n}{\ell}\right) \mathbb{E}\left(\frac{S_c T_{ct}}{\#T_t}\right) \\
&= \mu_t - \sum_{c=1}^{\ell} \frac{\theta_t}{n}\left(n_c - \frac{n}{\ell}\right) = \mu_t
\end{aligned} \tag{E.3}$$

Then $\hat{\delta}_{\mathrm{DR,SRS}}$ is unbiased for δ:

$$\begin{aligned}
\mathbb{E}(\left(\hat{\delta}_{\mathrm{DR,SRS}}\right) &= \mathbb{E}(\hat{\mu}_{1,\mathrm{DR,SRS}}) - \mathbb{E}(\hat{\mu}_{0,\mathrm{DR,SRS}}) \\
&= \mu_1 - \mu_0 = \delta.
\end{aligned} \tag{E.4}$$

E.3 Expectation of DR estimator with estimated θ

If θ is unknown and needs to be estimated with the data,

$$\begin{aligned}
\mathbb{E}(\hat{\mu}_{t,\mathrm{DR}}) &= \mu_t - \ell \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \mathbb{E}\left(\frac{S_c T_{ct} \hat{\theta}}{\#T_t}\right) \\
&= \mu_t - \ell \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \left[\mathrm{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \hat{\theta}\right) - \mathbb{E}\left(\frac{S_c T_{ct}}{\#T_t}\right) \mathbb{E}(\hat{\theta})\right] \\
&= \mu_t - \ell \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \mathrm{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \hat{\theta}\right) + \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \mathbb{E}(\hat{\theta}) \\
&= \mu_t - \ell \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \mathrm{Cov}\left(\frac{S_c T_{ct}}{\#T_t}, \hat{\theta}\right).
\end{aligned} \tag{E.5}$$

Therefore, the DR estimator will no longer be unbiased:

$$\mathbb{E}(\hat{\delta}_{\text{DR,SRS}}) = \delta - \ell \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \text{Cov}\left(\frac{S_c T_{c1}}{\#T_1}, \hat{\theta}\right) + \ell \sum_{c=1}^{\ell} \frac{1}{n}\left(n_c - \frac{n}{\ell}\right) \text{Cov}\left(\frac{S_c T_{c0}}{\#T_0}, \hat{\theta}\right). \tag{E.6}$$

E.4 Variance of DR estimator for PATE

E.4.1 Variance of DR estimator for population mean

Note that

$$\hat{\mu}_{t,\text{DR,SRS}} = \ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \left[\sum_{k=1}^{n_c} \frac{y_{kct} S_{kc}}{s_c} - \frac{\theta_t}{n_c}\left(n_c - \frac{n}{\ell}\right)\right] \tag{E.7}$$

$$= \ell \sum_{c=1}^{\ell} \frac{S_c T_{ct}}{\#T_t} \frac{n_c}{n} \hat{\tilde{\mu}}_{ct}, \tag{E.8}$$

which has the same form as the HT estimator. Hence, based on eq. (B.22), the variance of $\hat{\mu}_{t,\text{DR,SRS}}$ is

$$\text{Var}(\hat{\mu}_{t,\text{DR,SRS}}) = \ell^2 \left[\mathbb{E}\left(\frac{1}{\#T_t}\right) - \frac{1}{\ell}\right] \text{Var}\left(\frac{n_c}{n}\tilde{\mu}_{ct}\right) + \mathbb{E}\left(\frac{1}{\#T_t}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2}\left(1 - \frac{s_c}{n_c}\right) \frac{\text{Var}(y_{kct})}{s_c} \tag{E.9}$$

where $\text{Var}\left(\frac{n_c}{n}\tilde{\mu}_{ct}\right)$ is the population variance of the weighted transformed treatment means $\tilde{\mu}_{ct} = \mu_{ct} - \frac{\theta_t}{n_c}(n_c - \frac{n}{\ell})$ and $\text{Var}(y_{kct})$ is the population variance of within-cluster outcomes under treatment t.

E.4.2 Covariance of DR estimator for the population means

From eq. (B.26),

$$\mathrm{Cov}(\hat{\mu}_{1,\mathrm{DR,SRS}}, \hat{\mu}_{0,\mathrm{DR,SRS}}) = \frac{\ell}{\ell-1} \sum_{c=1}^{\ell} \sum_{c \neq c'} \frac{n_c n_{c'}}{n^2} \tilde{\mu}_{c1} \tilde{\mu}_{c'0} - \mu_1 \mu_0. \tag{E.10}$$

E.4.3 Variance of DR estimator for PATE

From eq. (E.9) and eq. (E.10),

$$\begin{aligned}
\mathrm{Var}(\hat{\delta}_{\mathrm{DR,SRS}}) = \ell^2 & \left[\mathbb{E}\left(\frac{1}{\#T_1}\right) - \frac{1}{\ell} \right] \mathrm{Var}\left(\frac{n_c}{n} \tilde{\mu}_{c1}\right) + \mathbb{E}\left(\frac{1}{\#T_1}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \left(1 - \frac{s_c}{n_c}\right) \frac{\mathrm{Var}(y_{kc1})}{s_c} \\
& + \ell^2 \left[\mathbb{E}\left(\frac{1}{\#T_0}\right) - \frac{1}{\ell} \right] \mathrm{Var}\left(\frac{n_c}{n} \tilde{\mu}_{c0}\right) + \mathbb{E}\left(\frac{1}{\#T_0}\right) \ell \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \left(1 - \frac{s_c}{n_c}\right) \frac{\mathrm{Var}(y_{kc0})}{s_c} \\
& - 2 \frac{\ell}{\ell-1} \sum_{c=1}^{\ell} \sum_{c \neq c'} \frac{n_c n_{c'}}{n^2} \tilde{\mu}_{c1} \tilde{\mu}_{c'0} + 2\mu_1 \mu_0.
\end{aligned} \tag{E.11}$$

E.5 SYG variance estimator of DR estimator for PATE

$$\begin{aligned}
\widehat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{DR,\ SRS}}\right) = & \frac{\ell^2}{2} \sum_{t=0}^{1} \sum_{c=1}^{\ell} \sum_{c' \neq c} \left(1 - \frac{\#T_t}{\ell}\right) \frac{S_c S_{c'} T_{ct} T_{c't}}{\#T_t^2 (\#T_t - 1)} \left(\frac{n_c \hat{\tilde{\mu}}_{ct}}{n} - \frac{n_{c'} \hat{\tilde{\mu}}_{c't}}{n} \right)^2 \\
& + \ell \sum_{t=0}^{1} \sum_{c=1}^{\ell} \frac{n_c^2}{n^2} \frac{S_c T_{ct}}{\#T_t} \hat{\tilde{\mu}}_{ct}^2 - 2\ell \sum_{c=1}^{\ell} \sum_{c' \neq c} \frac{n_c n_{c'}}{n^2} \frac{S_c S_{c'} T_{c1} T_{c'0}}{\#T_1 \#T_0} \hat{\tilde{\mu}}_{c1} \hat{\tilde{\mu}}_{c'0}
\end{aligned} \tag{E.12}$$

where $\tilde{\mu}_{ct} = \mu_{ct} - \frac{\theta_t}{n_c}(n_c - \frac{n}{\ell})$.

Appendix F

Simulation Results

Simulation results comparing PATE and SE estimations of HT-PPS, DIM, HT-MP, HT-STR, OLS, HT-SRS, HJ, and DR estimators. Note that for sizes 60+, the samples are approximately PPSWOR (hence, the increased bias) and that the SE estimations for the HT-PPS is based on the WR SE estimators.

Male Responses							
Est.	No. sampled clusters	PATE est.	PATE bias	PATE SE	PATE MSE	SE est.	SE var.
ht-pps	20	-0.032	-1.32e-3	0.077	5.94e-3	0.063	2.57e-4
	40	-0.030	-2.17e-4	0.054	2.94e-3	0.048	5.77e-5
	60	-0.030	7.17e-4	0.044	1.96e-3	0.041	2.41e-5
	80	-0.028	2.67e-3	0.038	1.48e-3	0.036	1.29e-5
	100	-0.025	5.34e-3	0.034	1.18e-3	0.033	8.43e-6
	200	-0.018	1.24e-2	0.024	7.18e-4	0.025	2.14e-6
dim	20	-0.021	9.06e-3	0.084	7.07e-3	0.650	1.59e-2
	40	-0.021	9.71e-3	0.059	3.54e-3	0.324	1.91e-3
	60	-0.019	1.09e-2	0.048	2.44e-3	0.217	5.60e-4
	80	-0.020	1.05e-2	0.042	1.86e-3	0.164	2.31e-4
	100	-0.020	1.05e-2	0.037	1.50e-3	0.132	1.14e-4
	200	-0.020	1.02e-2	0.026	7.96e-4	0.068	1.17e-5
ht-mp	20	-0.031	-7.71e-4	0.098	9.53e-3	0.043	1.38e-3
	40	-0.030	4.33e-4	0.067	4.52e-3	0.046	9.73e-4
	60	-0.030	4.71e-4	0.056	3.17e-3	0.049	8.02e-4
	80	-0.030	-1.27e-4	0.048	2.32e-6	0.050	6.58e-4
	100	-0.030	1.15e-4	0.043	1.82e-3	0.051	5.52e-4
	200	-0.030	2.21e-4	0.029	8.70e-4	0.054	2.63e-4

Table F.1: Simulation results for National Solidarity Program

Male Responses							
Est.	No. sampled clusters	PATE est.	PATE bias	PATE SE	PATE MSE	SE est.	SE var.
ht-str	20	-0.031	-1.12e-3	0.118	1.38e-2	2.916	1.822
	40	-0.030	8.12e-5	0.105	1.10e-2	1.476	9.90e-1
	60	-0.029	1.70e-3	0.087	7.55e-3	0.884	2.90e-1
	80	-0.030	-1.73e-4	0.072	5.12e-3	0.565	7.14e-2
	100	-0.030	3.46e-4	0.064	4.09e-3	0.430	3.70e-2
	200	-0.031	-8.61e-4	0.045	2.05e-3	0.161	2.28e-3
ols	20	-0.017	1.34e-2	0.083	7.12e-3	0.280	8.76e-4
	40	-0.020	1.03e-2	0.059	3.57e-3	0.238	2.74e-4
	60	-0.019	1.09e-2	0.048	2.38e-3	0.216	1.38e-4
	80	-0.019	1.14e-2	0.041	1.83e-3	0.202	8.43e-5
	100	-0.020	1.03e-2	0.037	1.45e-3	0.191	5.61e-5
	200	-0.020	1.01e-2	0.026	7.64e-4	0.161	1.66e-5
ht-srs	20	-0.029	.33e-3	0.154	2.38e-2	0.110	1.31e-2
	40	-0.028	2.54e-3	0.107	1.16e-2	0.088	4.64e-3
	60	-0.029	1.75e-3	0.087	7.59e-3	0.075	2.39e-3
	80	-0.032	-1.67e-3	0.076	5.75e-3	0.067	1.52e-3
	100	-0.031	-6.88e-4	0.067	4.45e-3	0.062	1.01e-3
	200	-0.030	5.54e-4	0.047	2.26e-3	0.046	2.36e-4

Table F.2: Simulation results for NSP continued

Male Responses							
Est.	No. sampled clusters	PATE est.	PATE bias	PATE SE	PATE MSE	SE est.	SE var.
hj	20	-0.024	6.34e-3	0.109	1.20e-2	0.568	3.52e-2
	40	-0.026	3.90e-3	0.084	7.00e-3	0.343	1.32e-2
	60	-0.026	4.13e-3	0.072	5.15e-3	0.251	6.59e-3
	80	-0.027	2.90-3	0.064	4.11e-3	0.203	3.94e-3
	100	-0.028	2.68e-3	0.058	3.33e-3	0.170	2.39e-3
	200	-0.029	1.00e-3	0.042	1.74e-3	0.097	4.13e-4
dr	20	-0.030	2.47e-4	0.153	2.33e-2	0.111	1.25e-2
	40	-0.030	-8.02e-5	0.107	1.13e-2	0.088	4.66e-3
	60	-0.031	-5.49e-4	0.087	7.54e-3	0.075	2.44e-3
	80	-0.030	7.95e-5	0.075	5.69e-3	0.068	1.51e-3
	100	-0.030	6.79e-6	0.068	4.58e-3	0.062	1.01e-3
	200	-0.030	1.07e-4	0.048	2.33e-3	0.047	2.40e-4

Table F.3: Simulation results for NSP continued

Female Responses							
Est.	No. sampled clusters	PATE est.	PATE bias	PATE SE	PATE MSE	SE est.	SE var.
ht-pps	20	-0.045	-1.32e-4	0.090	8.15e-3	0.045	3.28e-4
	40	-0.046	-7.67e-4	0.063	3.93e-3	0.032	1.13e-4
	60	-0.043	2.05e-3	0.051	2.64e-3	0.030	5.71e-5
	80	-0.040	4.63e-3	0.044	1.97e-3	0.029	2.82e-5
	100	-0.035	9.33e-3	0.039	1.61e-3	0.028	1.57e-5
	200	-0.019	2.56e-2	0.027	1.39e-3	0.025	2.32e-6
dim	20	-0.002	4.27e-2	0.084	8.87e-3	3.18	1.63e-1
	40	-0.003	4.12e-2	0.059	5.18e-3	1.079	9.24e-3
	60	-0.004	4.07e-2	0.049	4.06e-3	0.574	1.67e-3
	80	-0.003	4.12e-2	0.042	3.49e-3	0.366	4.85e-4
	100	-0.004	4.10e-2	0.038	3.12e-3	0.258	1.90e-4
	200	-0.004	4.11e-2	0.027	2.40e-3	0.085	7.53e-6
ht-mp	20	-0.0378	6.89e-3	0.165	2.71e-2	2.728	4.794
	40	-0.036	9.16e-3	0.119	1.42e-2	1.023	3.80e-1
	60	-0.037	7.40e-3	0.097	9.39e-3	0.574	7.96e-2
	80	-0.036	8.44e-3	0.082	6.73e-3	0.369	2.50e-2
	100	-0.036	8.19e-3	0.074	5.56e-3	0.262	9.59e-3
	200	-0.038	7.10e-3	0.051	2.63e-3	0.082	3.88e-4

Table F.4: Simulation results for NSP continued

Male Responses							
Est.	No. sampled clusters	PATE est.	PATE bias	PATE SE	PATE MSE	SE est.	SE var.
ht-str	20	-0.045	-5.64e-4	0.124	1.54e-2	3.352	8.85e-1
	40	-0.045	-9.16e-5	0.111	1.23e-2	1.710	4.92e-1
	60	-0.046	-9.84e-4	0.091	8.24e-3	1.003	1.20e-1
	80	-0.045	1.66e-4	0.074	5.49e-3	0.611	2.18e-2
	100	-0.048	-1.09e-4	0.068	4.58e-8	0.457	1.05e-2
	200	-0.045	6.05e-5	0.047	2.20e-3	0.163	6.03e-4
ols	20	0.001	4.57e-2	0.085	9.39e-3	0.283	5.13e-4
	40	-0.001	4.41e-2	0.060	5.523e-3	0.241	1.54e-4
	60	-0.001	4.36e-2	0.048	4.22e-3	0.218	7.78e-5
	80	-0.002	4.26e-2	0.042	3.58e-3	0.203	4.75e-5
	100	-0.003	4.15e-2	0.037	3.12e-3	0.193	3.34e-5
	200	-0.003	4.20e-2	0.026	2.46e-3	0.162	9.58e-6
ht-srs	20	-0.044	2.42e-4	0.192	3.67e-2	0.166	9.55e-3
	40	-0.044	6.85e-4	0.135	1.81e-2	0.123	2.89e-3
	60	-0.045	-2.25e-4	0.110	1.21e-2	0.103	1.34e-3
	80	-0.046	-8.69e-4	0.096	9.25e-3	0.091	7.74e-4
	100	-0.044	5.31e-4	0.085	7.14e-3	0.082	4.89e-4
	200	-0.045	1.03e-4	0.060	3.60e-3	0.059	1.06e-4

Table F.5: Simulation results for NSP continued

Male Responses							
Est.	No. sampled clusters	PATE est.	PATE bias	PATE SE	PATE MSE	SE est.	SE var.
hj	20	-0.033	20e-2	0.130	1.70e-2	0.698	2.22e-2
	40	-0.037	7.91e-3	0.101	1.02e-2	0.421	7.06e-3
	60	-0.040	5.13e-3	0.086	7.34e-3	0.306	3.19e-3
	80	-0.042	3.20e-3	0.075	5.63e-3	0.243	1.60e-3
	100	-0.043	2.21e-3	0.068	4.57e-3	0.203	9.06e-4
	200	-0.043	1.25e-3	0.049	2.38e-3	0.113	1.23e-4
dr	20	-0.044	9.61e-4	0.188	3.54e-2	0.171	9.24e-3
	40	-0.043	1.54e-3	0.135	1.82e-2	0.128	2.93e-3
	60	-0.046	-7.36e-4	0.110	1.21e-2	0.107	1.35e-3
	80	-0.046	-1.18e-3	0.096	9.12e-3	0.094	7.71e-4
	100	-0.046	-8.56e-4	0.083	6.97e-3	0.085	5.00e-4
	200	-0.046	-9.34e-4	0.059	3.46e-3	0.061	1.08e-4

Table F.6: Simulation results for NSP continued