Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments

Joel A. Middleton and Peter M. Aronow

March 3, 2013

Abstract

Many estimators of the average treatment effect, including the difference-in-means, may be biased when clusters of units are allocated to treatment. This bias remains even when the number of units within each cluster grows asymptotically large. In this paper, we propose simple, unbiased, location-invariant, and covariate-adjusted estimators of the average treatment effect in experiments with complete randomization of clusters, along with associated variance estimators. We then analyze a cluster-randomized field experiment on voter mobilization in the United States, demonstrating that the proposed estimators have precision that is comparable (if not superior) to that of existing biased estimators of the average treatment effect.

1 Introduction

In recent years, largely beginning with Freedman (2008a,b), researchers have been paying increased attention to the properties of treatment effect estimators for randomized experiments under a design-based model. Under the design-based model (Neyman, 1923, 1934; Sarndal, 1978), potential outcomes are fixed and the only source of stochasticity lies in the random administration of a treatment to a finite population. Importantly, Freedman (2008a) demonstrated that, under a such a model, regression adjustment is generally biased (though consistent) and may harm efficiency. Since, researchers have been seeking to derive methods that do not suffer from these problems (Lin, in press; Miratrix, Sekhon and Yu, in press) or to assess the operating characteristics of common model-based estimators (Humphreys, 2009; Samii and Aronow, 2012) under the design-based paradigm. However, this research has largely focused on experiments wherein treatment is randomized at the unit level.

Although extensively studied under the model-based paradigm (see, e.g., Donner and Klar, 2000), comparatively little attention has been based to designs with complete randomization of clusters under the design-based paradigm. The aforementioned estimators are not directly applicable to cluster-randomized designs. Even seemingly design-based estimators – such as the difference-in-means estimator – may suffer from bias even when all units have an equal probability of treatment assignment. Importantly, Middleton (2008) proves the bias of the difference-in-means estimator (and inconsistency under asymptotic scalings that entail a fixed number of clusters) for

completely randomized experiments with unequal cluster sizes. Similarly, Imai, King and Nall (2009) recognize the bias of the difference-in-means estimator and propose solutions that require altering the design of the experiment. The authors recommend pair matching on observables in order to reduce the amount of bias and variance that may result from standard analysis of cluster-randomized experiments. The closest analogue to our approach, however, may be found in Hansen and Bowers (2009), which proposes similar – though not necessarily unbiased – design-based estimators for cluster-randomized experiments with noncompliance. (Hansen and Bowers, 2008, also derives design-based balance tests for cluster-randomized experiments).

In this paper, we propose a simple and unbiased design-based estimator for the ATE when treatment has been completely randomized at the level of the cluster. Drawing from classical sampling theory, we then propose a natural extension to both improve efficiency and confer the property of location invariance: the Des Raj (1965) difference estimator, which remains unbiased even in small samples. For each of these estimators, we also derive two different variance estimators for the estimated ATE. We then examine a field experiment designed to assess the effect of voter mobilization in a United States presidential election, using randomization inference to assess the bias and standard errors (SEs) of a number of estimators under two different null hypotheses. Whereas many common treatment effect estimators, including the difference-in-means, ordinary least squares regression and random effects regression fail to unbiasedly recover the ATE, the proposed methods are unbiased estimators of the ATE that are comparable in efficiency to their biased alternatives.

2 Potential outcomes

The basis of our design-based approach is the model of potential outcomes introduced by Neyman (1923) and popularized by Rubin (1974). Define treatment indicator $D_i \in \{0, 1\}$ for units $i \in 1, 2, ..., N$ such that $D_i = 1$ when unit i receives the treatment and $D_i = 0$ otherwise. Assuming that the stable unit treatment value assumption (Rubin, 1978, 2005) holds, let Y_{1i} be the potential outcome if unit i is exposed to the treatment, and let Y_{0i} be the potential outcome if unit i is not exposed to the treatment. The observed experimental outcome Y_i may be expressed as a function of the potential outcomes and the assigned treatment: $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$. The causal effect of the treatment on unit i, τ_i, is defined as the difference between the two potential outcomes for unit i: $\tau_i \equiv Y_{1i} - Y_{0i}$. And, by definition, the ATE, denoted by Δ, is the average value of τ_i for all units i. Under this model, the only random component of the experiment is the allocation of units to treatment and control groups.

Since $\tau_i \equiv Y_{1i} - Y_{0i}$, the ATE is equivalently

$$\Delta = \frac{\sum_{i=1}^{N} (Y_{1i} - Y_{0i})}{N} = \frac{1}{N}\left[\sum_{i=1}^{N} Y_{1i} - \sum_{i=1}^{N} Y_{0i}\right] = \frac{1}{N}\left[Y_1^T - Y_0^T\right],$$

where Y_1^T is the sum of potential outcomes if in the treatment condition and Y_0^T is the sum of potential outcomes if in the control condition. An estimator of Δ can be constructed using estimators

of Y_0^T and Y_1^T:

$$\widehat{\Delta} = \frac{1}{N}\left[\widehat{Y_1^T} - \widehat{Y_0^T}\right], \tag{1}$$

where $\widehat{Y_1^T}$ is the estimated sum of potential outcomes under treatment and $\widehat{Y_0^T}$ is the estimated sum of potential outcomes under control.

Formally, the bias of an estimator is the difference between the expected value of the estimator (over all randomizations) and the true parameter of interest. If the estimators $\widehat{Y_0^T}$ and $\widehat{Y_1^T}$ are unbiased, the corresponding estimator of Δ is also unbiased since

$$\mathrm{E}[\widehat{\Delta}] = \frac{1}{N}\left[\mathrm{E}[\widehat{Y_1^T}] - \mathrm{E}[\widehat{Y_0^T}]\right] = \frac{1}{N}\left[Y_1^T - Y_0^T\right] = \Delta.$$

3 Properties of the difference-in-means estimator

In this section, we examine the properties of the difference-in-means estimator. We begin with a short derivation of the unbiasedness of the difference-in-means estimator under complete random assignment of units. We then articulate the source of the bias for the difference-in-means estimator when applied to a cluster randomized experiment and examine the asymptotic properties of the estimator.

3.1 Unbiased estimation of treatment effects under complete random assignment of units

Define N and n_t as integers such that $0 < n_t < N$. Complete random assignment of treatment implies that n_t, a fixed number, units are randomly allocated to be treated ($D_i = 1$) and the other $n_c = N - n_t$ are placed in the control group ($D_i = 0$). Define I_0 as the set of all i such that $D_i = 0$ and I_1 as the set of all i such that $D_i = 1$.

To derive an unbiased estimator of the ATE under complete random assignment, we can first posit estimators of Y_0^T and Y_1^T. Define an estimator of Y_0^T,

$$\widehat{Y_{0,S}^T} = \frac{N}{n_c}\sum_{i\in I_0} Y_{0i} = \frac{N}{n_c}\sum_{i\in I_0} Y_i \tag{2}$$

and, similarly, define an estimator of Y_1^T,

$$\widehat{Y_{1,S}^T} = \frac{N}{n_t}\sum_{i\in I_1} Y_{1i} = \frac{N}{n_t}\sum_{i\in I_1} Y_i. \tag{3}$$

It is trivial to show that the estimators in equations 2 and 3 are unbiased under complete random assignment:

$$\mathrm{E}[\widehat{Y_{0,S}^T}] = \mathrm{E}\left[\frac{N}{n_t}\sum_{i\in I_0} Y_i\right] = N\cdot\overline{Y_0} = Y_0^T, \tag{4}$$

where $\overline{Y_0}$ is the mean value of Y_{0i} over all i units (and is not an observable quantity). A proof for the unbiasedness of $\widehat{Y_{1,S}^T}$ is, likewise, trivial and directly follows the form of equation 4.

From equation 1, it follows that we may construct an estimator of Δ:

$$\widehat{\Delta_S} = \frac{1}{N}\left[\widehat{Y_{1,S}^T} - \widehat{Y_{0,S}^T}\right] = \frac{\sum\limits_{i \in I_1} Y_i}{n_t} - \frac{\sum\limits_{i \in I_0} Y_i}{n_c}, \tag{5}$$

where $\sum\limits_{i \in I_1} Y_i/n_t$ is the mean value of Y_i for all units assigned to treatment and $\sum\limits_{i \in I_0} Y_i/n_c$ is the mean value of Y_i for all units assigned to control. $\widehat{\Delta_S}$ is known as the difference-in-means estimator and is perhaps the most well-known estimator of treatment effects in randomized experiments.

3.2 Properties of the difference-in-means estimator under complete random assignment of clusters

Under clustered random assignment, the difference-in-means estimator is no longer generally unbiased, despite all individuals having the same probability of entering into each treatment condition. The unit of randomization is no longer the individual: instead, clusters (or groups of individuals) are assigned to treatment. While complete randomization of units may yield more efficient designs in principle, in practice, a number of settings may dictate clustered designs. Some examples include when treatment must be applied at the level of the cluster, when outcome measures are only available at the level of the cluster, or when unit interference (e.g., treatment synergies or spillover effects) is an important aspect of treatment.

Very often, in these settings where unit randomization is infeasible or undesirable, the researcher does not have control over the size of clusters (e.g., household, village). As a consequence, bias can arise in estimation because the number of individuals assigned to treatment is no longer a fixed quantity. We begin this section by deriving the bias associated with the difference-in-means estimator.

Formally, suppose each cluster $j = 1, 2, ..., M$ is assigned to either treatment or control. Define m_t and M as integers such that $0 < m_t < M$. Now m_t clusters are randomly assigned to treatment ($D_j = 1$) and the remaining $m_c = M - m_t$ clusters are assigned to control ($D_j = 0$). Define J_0 as the set of all j such that $D_j = 0$ and J_1 as the set of all j such that $D_j = 1$. Let Y_{0ij} be the response of the i^{th} individual in the j^{th} cluster if the cluster is assigned to control and let Y_{1ij} be the response of the i^{th} individual in the j^{th} cluster if the cluster is assigned to treatment. Let n_j be the number of individuals in the j^{th} cluster. Note that all individuals have the same probability m_t/M of entering treatment.

The estimators in equations 2 and 3 can be rewritten as $\widehat{Y_{1,S}^T} = N \sum\limits_{j \in J_1} \sum\limits_{i=1}^{n_j} Y_{ij} / \sum\limits_{j \in J_1} n_j$ and $\widehat{Y_{0,S}^T} =$

$N \sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{ij} / \sum_{j \in J_0} n_j$. The difference-in-means estimator in equation 5 can therefore be rewritten

$$\widehat{\Delta_S} = \frac{1}{N}\left[\widehat{Y_{1,S}^T} - \widehat{Y_{0,S}^T}\right] = \frac{\sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{ij}}{\sum_{j \in J_1} n_j} - \frac{\sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{ij}}{\sum_{j \in J_0} n_j}. \tag{6}$$

The double summations in the numerators make explicit that summation takes place across individuals in different clusters. In the denominators, the summations operate over clusters. While the estimator remains unchanged from equation 5, rewriting it in this way reveals a fundamental issue with its application.

The trouble with using the estimator in equation 6 is that the quantities $n_t = \sum_{j \in J_1} n_j$ and $n_c = \sum_{j \in J_0} n_j$ are no longer fixed numbers as they were in equation 5, but are now random variables. The total number of individuals in treatment and control now depends on the size of the particular clusters assigned to the experimental groups. To understand why this dependence is problematic, we need only examine equation 4: in the second line, the terms N/n_t and N/n_c may be moved to the outside of the expectation operator because they are fixed constants. When n_t and n_c are random variables, calculating the expectation is more involved. In general, for a ratio of two random variables u, v, (u/v),

$$\mathrm{E}\left[\frac{u}{v}\right] = \frac{1}{\mathrm{E}[v]}\left[\mathrm{E}[u] - \mathrm{Cov}\left(\frac{u}{v}, v\right)\right] \tag{7}$$

if $v > 0$ (Hartley and Ross, 1954). Because the difference-in-means estimator is the difference between two ratios of random variables we can use the result in equation 7 to derive the bias of the difference-in-means estimator in equation 6. Roughly following Middleton (2008),

$$\begin{aligned}
\mathrm{E}\left[\widehat{\Delta_S}\right] &= \mathrm{E}\left[\sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{1ij} / \sum_{j \in J_1} n_j\right] - \mathrm{E}\left[\sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{0ij} / \sum_{j \in J_0} n_j\right] \\
&= \frac{1}{N}\left[Y_1^T - Y_0^T\right] - \frac{M}{N}\left[\frac{1}{m_t}\mathrm{Cov}\left(\sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{1ij} / \sum_{j \in J_1} n_j, \sum_{j \in J_1} n_j\right)\right. \\
&\qquad \left. - \frac{1}{m_c}\mathrm{Cov}\left(\sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{0ij} / \sum_{j \in J_0} n_j, \sum_{j \in J_0} n_j\right)\right].
\end{aligned}$$

It follows that the bias, $\mathrm{E}\left[\widehat{\Delta_S}\right] - \Delta =$

$$-\frac{M}{N}\left[\frac{1}{m_t}\mathrm{Cov}\left(\sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{1ij} / \sum_{j \in J_1} n_j, \sum_{j \in J_1} n_j\right) - \frac{1}{m_c}\mathrm{Cov}\left(\sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{0ij} / \sum_{j \in J_0} n_j, \sum_{j \in J_0} n_j\right)\right]. \tag{8}$$

Inspection of this term reveals that if the size of the cluster is correlated with the potential outcomes in the cluster, the difference-in-means estimator is biased. In some special cases, there will be no bias, such as when the cluster size does not vary or when there is no covariance between cluster size and outcomes.

3.3 Asymptotic properties of the difference-in-means estimator with complete random assignment of clusters

In this section, we demonstrate two important facts about the difference-in-means estimator. First, in a proof adapted from Middleton (2008), we will show that the difference-in-means estimator is consistent as M grows. Second, we demonstrate that the difference-in-means estimator is not necessarily consistent as N grows. Consistency of a statistic under a finite population is defined given a sequence of h finite populations H where $M_h < M_{h+1}$, $m_{th} < m_{th+1}$ and $m_{ch} < m_{ch+1}$ for $h = 1, 2, 3, ...$ The estimator $\widehat{\Delta_S}$ is said to be a consistent estimator of Δ if $\widehat{\Delta_S} \xrightarrow{p} \Delta$ (converges in probability) as $h \longrightarrow \infty$.

To show that the difference-in-means estimator is consistent with large M, we follow Brewer (1979) in assuming that as $h \longrightarrow \infty$, the finite population H increases as follows: (1) the original population of M clusters is exactly copied $(h-1)$ times; (2) from each of the h copies, m_t clusters are allocated to treatment (such that $0 < m_t < M$) and the remaining $m_c = M - m_t$ are allocated to control; (3) the h subsets are collected in a single population of hM clusters, with hm_t clusters in treatment and $hm_c = hM - hm_t$ in control; and (4) $\widehat{\Delta_S}$ is defined as the difference-in-means estimator as in equation 5, only now summation takes place across all hm_c and hm_t clusters. A less restrictive set of assumptions is possible, but this setup is convenient because H is easy to visualize and moment assumptions are built-in. We express the estimator as, $\widehat{\Delta_S} = \sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{ij} / \sum_{j \in J_1} n_j - \sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{ij} / \sum_{j \in J_0} n_j$, where in this case J_1 is defined as the set of hm_t treatment clusters and J_0 is defined as the set of hm_c control clusters. As $h \longrightarrow \infty$, by the weak law of large numbers, $\sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{ij} \xrightarrow{p} Y_1^T \cdot \frac{hm_t}{hM}$, $\sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{ij} \xrightarrow{p} Y_0^T \cdot \frac{hm_c}{hM}$, $\sum_{j \in J_1} n_j \xrightarrow{p} N \cdot \frac{hm_t}{hM}$ and $\sum_{j \in J_0} n_j \xrightarrow{p} N \cdot \frac{hm_c}{hM}$. By Slutsky's theorem,

$$\widehat{\Delta_S} \xrightarrow{p} \frac{Y_1^T \cdot \frac{hm_t}{hM}}{N \cdot \frac{hm_t}{hM}} - \frac{Y_0^T \cdot \frac{hm_c}{hM}}{N \cdot \frac{hm_c}{hM}} = \frac{Y_1^T - Y_0^T}{N}. \tag{9}$$

This proves that the difference-in-means estimator is consistent as the number of clusters grows.

In the case where the size (rather than the number) of the clusters grows as $h \longrightarrow \infty$, the finite population H increases as follows: (1) m_t of the original clusters are allocated to treatment (such that $0 < m_t < M$) and the remaining $m_c = M - m_t$ are allocated to control; (2) the original population of M units is exactly copied $(h-1)$ times, but this time the h copies of a cluster are considered part of one supercluster; and (3) $\widehat{\Delta_S}$ is defined as the difference-in-means estimator as in equation 9, but now the inner summation takes place across all hn_j units in each cluster.

To show that the difference-in-means estimator is not necessarily consistent simply with large N, we express the estimator as,

$$\widehat{\Delta_S} = \frac{\sum_{j \in J_1} \sum_{i=1}^{hn_j} Y_{ij}}{\sum_{j \in J_1} hn_j} - \frac{\sum_{j \in J_0} \sum_{i=1}^{hn_j} Y_{ij}}{\sum_{j \in J_0} hn_j} = \frac{\sum_{j \in J_1} \sum_{i=1}^{n_j} Y_{ij}}{\sum_{j \in J_1} n_j} - \frac{\sum_{j \in J_0} \sum_{i=1}^{n_j} Y_{ij}}{\sum_{j \in J_0} n_j}. \tag{10}$$

As $h \longrightarrow \infty$, the estimate remains unchanged with large N if the number of clusters is fixed. This proves that the bias articulated in equation 8 is unmitigated for increasingly large clusters.

4 Unbiased estimation of treatment effects under complete random assignment of clusters

By understanding the bias as a problem fundamental to ratio estimation, we can circumvent the bias with an alternative design-based estimator. Notationally, it helps to clarify the task if we consider cluster *totals* - i.e., the sum of the responses of the individuals in each cluster. Define $Y_{0j}^T = \sum_{i=1}^{n_j} Y_{0ij}$ as the sum of responses of the individuals in the j^{th} cluster if assigned to control and $Y_{1j}^T = \sum_{i=1}^{n_j} Y_{1ij}$ as the sum of responses of the individuals in the j^{th} cluster if assigned to treatment. For each individual, only one of the two possible responses, Y_{0ij} or Y_{1ij}, may be observed and, since individuals are assigned to treatment conditions in clusters, for any given cluster, only one of the possible totals Y_{0j}^T or Y_{1j}^T, may be observed. The observed cluster total for cluster j, Y_j^T, may be expressed as: $Y_j^T = D_j Y_{1j}^T + (1 - D_j) Y_{0j}^T$.

Using this new notation, the ATE may be expressed as

$$\Delta = \frac{\sum_{j=1}^{M} \sum_{i=1}^{n_j} (Y_{1ij} - Y_{0ij})}{\sum_{j=1}^{M} \sum_{i=1}^{n_j} 1} = \frac{\sum_{j=1}^{M} Y_{1j}^T - \sum_{j=1}^{M} Y_{0j}^T}{\sum_{j=1}^{M} n_j} = \frac{1}{N} \left[Y_1^T - Y_0^T \right].$$

We can again construct an unbiased estimator for Δ through unbiased estimators of Y_0^T and Y_1^T. Following the logic of equation 4,

$$\widehat{Y_{0,HT}^T} = \frac{M}{m_c} \sum_{j \in J_0} Y_{0j}^T = \frac{M}{m_c} \sum_{j \in J_0} Y_j^T. \tag{11}$$

One can think of this estimator as estimating the average of the cluster totals (among control clusters) and then multiplying by the number of clusters M to get the estimated total for all units in the study. Likewise,

$$\widehat{Y_{1,HT}^T} = \frac{M}{m_t} \sum_{j \in J_1} Y_{1j}^T = \frac{M}{m_t} \sum_{j \in J_1} Y_j^T. \tag{12}$$

Following the same steps as equation 4, it is trivial to show that $\widehat{Y^T_{0,HT}}$ and $\widehat{Y^T_{1,HT}}$ are unbiased estimators of Y^T_0 and Y^T_1, respectively. The terms M/m_t and M/m_c are fixed; when taking the expectations of equations 11 and 12, they can be moved outside the expectation operator. Note that the random variables at the root of the ratio estimation problem above, n_t and n_c, do not appear in either estimator. From these two unbiased estimators, we may therefore construct an estimator of the ATE:

$$\widehat{\Delta_{HT}} = \frac{1}{N}\left[\widehat{Y^T_{1,HT}} - \widehat{Y^T_{0,HT}}\right] = \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} Y^T_j - \frac{1}{m_c}\sum_{j\in J_0} Y^T_j\right]. \tag{13}$$

We refer to this estimator as the Horvitz-Thompson (HT) estimator because it is a special case of the well-known estimator from sampling theory (Horvitz and Thompson, 1952; Chaudhuri and Stenger, 2005).

The HT estimator can be criticized on two grounds. First, as Imai, King and Nall (2009) suggest, this estimator is not location invariant. We offer a proof of the non-invariance of the HT estimator in section 4.1. Second, the HT estimator can be highly imprecise; cluster sums tend to vary a great deal because there are more individuals in some clusters than in others. In large clusters, totals may tend to be large and in small clusters, totals may tend to be smaller. In section 5.1, we will develop an estimator that addresses both these limitations.

4.1 Non-invariance of the Horvitz-Thompson estimator

To show that the estimator in equation 13 is not invariant to location shifts, let Y^*_{1ij} be a linear transformation of the treatment outcome for the i^{th} person in the j^{th} cluster such that $Y^*_{1ij} \equiv b_0 + b_1 \cdot Y_{1ij}$ and likewise, the control outcomes, $Y^*_{0ij} \equiv b_0 + b_1 \cdot Y_{0ij}$. Invariance to this transformation would imply that, when analyzing the transformed data, we achieve the relationship between the old estimate and new estimate such that

$$\widehat{\Delta}^*_{HT} = b_1 \cdot \widehat{\Delta}_{HT}, \tag{14}$$

i.e., the ATE estimated from linearly transformed outcomes will be equal to the ATE estimated from non-transformed outcomes multiplied by the scaling factor b_1. In Appendix A, we demonstrate that the HT estimator is not location-invariant because the estimate based on the transformed data will be

$$\widehat{\Delta}^*_{HT} = b_0 \cdot \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} n_j - \frac{1}{m_c}\sum_{j\in J_0} n_j\right] + b_1 \cdot \widehat{\Delta}_{HT}. \tag{15}$$

Unless $b_0 = 0$, the term on the left does not generally reduce to zero but instead varies across treatment assignments, so equation 15 is not generally equivalent to equation 14 for a given randomization. Note that, while a multiplicative scale change (e.g., transforming feet to inches) need not be a concern, a linear transformation that includes a location shift (e.g., reversing a binary indicator variable or transforming Fahrenheit to Celsius) will lead to a violation of invariance. For any given randomization, linearly transforming the data such that the intercept changes can yield substantively different estimates.

4.2 Deriving estimators of the variance of the Horvitz-Thompson estimator under complete random assignment of clusters

In our derivation of variances, we will follow the general formulations of Freedman, Pisani and Purves (1997), which follow from a long tradition dating from Neyman (1923). The variance of the estimator in equation 13 will be

$$\mathrm{V}(\widehat{\Delta}) = \frac{1}{N^2}\left[\mathrm{V}(\widehat{Y_0^T}) + \mathrm{V}(\widehat{Y_1^T}) - 2\mathrm{Cov}(\widehat{Y_0^T}, \widehat{Y_1^T})\right]. \tag{16}$$

This expression is the true, not estimated, variance. To construct an unbiased estimator of this variance, we must have unbiased estimators of each of the quantities in equation 16. While unbiased estimators may be constructed for $\mathrm{V}(\widehat{Y_0^T})$ and $\mathrm{V}(\widehat{Y_1^T})$, there does not generally exist an unbiased estimator for $\mathrm{Cov}(\widehat{Y_0^T}, \widehat{Y_1^T})$ because the joint distribution of potential outcomes is not observable.

We may, however, derive a generally conservative estimator of the variance. First, we derive the components of the true variance from equation 16. From the principles of finite population sampling,

$$\mathrm{V}\left(\widehat{Y_{0,HT}^T}\right) = \frac{M^2}{m_c}\left(\frac{M-m_c}{M-1}\right)\sigma^2(Y_{0j}^T),$$
$$\mathrm{V}\left(\widehat{Y_{1,HT}^T}\right) = \frac{M^2}{m_t}\left(\frac{M-m_t}{M-1}\right)\sigma^2(Y_{1j}^T),$$

and

$$\mathrm{Cov}(\widehat{Y_{0,HT}^T}, \widehat{Y_{1,HT}^T}) = -\frac{M^2}{M-1}\sigma(Y_{0j}^T, Y_{1j}^T),$$

where, given features v_j and w_j for $j \in 1, ..., M$, finite population variance $\sigma^2(v_j) = \frac{1}{M}\sum_{j=1}^{M}(v_j - \frac{1}{M}\sum_{j=1}^{M} v_j)^2$ and finite population covariance $\sigma(v_j, w_j) = \frac{1}{M}\sum_{j=1}^{M}(v_j - \frac{1}{M}\sum_{j=1}^{M} v_j)(w_j - \frac{1}{M}\sum_{j=1}^{M} w_j)$.

From equation 16,

$$\begin{aligned}
\mathrm{V}(\widehat{\Delta_{HT}}) &= \frac{1}{N^2}\left[\frac{M^2}{m_c}\left(\frac{M-m_c}{M-1}\right)\sigma^2(Y_{0j}^T) + \frac{M^2}{m_t}\left(\frac{M-m_t}{M-1}\right)\sigma^2(Y_{1j}^T)\right. \\
&\qquad \left. + \frac{2M^2}{M-1}\sigma(Y_{0j}^T, Y_{1j}^T)\right] \\
&= \frac{M^2}{N^2}\left[\frac{M}{M-1}\left[\frac{\sigma^2(Y_{0j}^T)}{m_c} + \frac{\sigma^2(Y_{1j}^T)}{m_t}\right]\right. \\
&\qquad \left. + \frac{1}{M-1}\left[2\sigma(Y_{0j}^T, Y_{1j}^T) - \sigma^2(Y_{0j}^T) - \sigma^2(Y_{1j}^T)\right]\right].
\end{aligned} \tag{17}$$

Since $2\sigma(Y_{0j}^T, Y_{1j}^T) - \sigma^2(Y_{0j}^T) - \sigma^2(Y_{1j}^T) \leq 0$, it follows that

$$\mathrm{V}(\widehat{\Delta_{HT}}) \leq \mathrm{V}_{\text{apx}}(\widehat{\Delta_{HT}}) = \frac{M^3}{N^2(M-1)}\left[\frac{\sigma^2(Y_{0j}^T)}{m_c} + \frac{\sigma^2(Y_{1j}^T)}{m_t}\right].$$

Substituting unbiased estimators of $\sigma^2(Y_{0j}^T)$ and $\sigma^2(Y_{1j}^T)$ (Cochran, 1977, Theorem 2.4), we may derive an unbiased estimator of the quantity $\mathrm{V}_{\text{apx}}(\widehat{\Delta_{HT}})$,

$$\widehat{\mathrm{V}}(\widehat{\Delta_{HT}}) = \frac{M^2}{N^2}\left[\frac{\sum\limits_{j\in J_0}\left(Y_j^T - \overline{Y_{cj}^T}\right)^2}{m_c(m_c-1)} + \frac{\sum\limits_{j\in J_1}\left(Y_j^T - \overline{Y_{tj}^T}\right)^2}{m_t(m_t-1)}\right],$$

where $\overline{Y_{mj}^T} = \sum\limits_{j\in J_0} Y_j^T/m_c$, the mean value of Y_j^T over all $j \in J_0$ and $\overline{Y_{tj}^T} = \sum\limits_{j\in J_1} Y_j^T/m_t$, the mean value of Y_j^T over all $j \in J_1$. (When M is large, researchers may encounter numerical problems computing M^2 and, later, M^4. This problem may be obviated by replacing M^2/N^2 with $(\frac{1}{M}\sum\limits_{j=1}^{M} n_j)^{-2}$, the reciprocal of the square of the average number of units per cluster.)

The bias of the variance estimator is always nonnegative, thus ensuring that the variance estimator is conservative. However, while $\widehat{\mathrm{V}}(\widehat{\Delta_{HT}})$ is conservative, it may also be imprecise. Another option for estimating the variance is to assume a sharp null hypothesis and either analytically or computationally calculate the variance of the estimator. One common sharp null hypothesis is that of the sharp null hypothesis of no treatment effect: $H_0 : \tau_i = 0, \forall i$. H_0 implies that the treatment has no effect whatsoever on the outcome, i.e., that both potential outcomes are identical: $Y_{0i} = Y_{1i} = Y_i$. When the sharp null hypothesis of no effect holds, we know two important facts: $\sigma^2(Y_{0j}) = \sigma^2(Y_{1j}) = \sigma^2(Y_j)$ and $\sigma(Y_{0j}, Y_{1j}) = \sigma^2(Y_j)$. By substituting in σ^2 into the last line of equation 17, we may calculate the true variance under this null hypothesis,

$$\begin{aligned}\mathrm{V}^N(\widehat{\Delta_{HT}}) &= \frac{M^2}{N^2}\left[\frac{M}{M-1}\left[\frac{\sigma^2(Y_j^T)}{m_c} + \frac{\sigma^2(Y_j^T)}{m_t}\right] + \frac{1}{M-1}\left[2\sigma^2(Y_j^T) - \sigma^2(Y_j^T) - \sigma^2(Y_j^T)\right]\right] \\ &= \frac{M^4\sigma^2(Y_j^T)}{N^2(M-1)m_c m_t}.\end{aligned}$$

Note that if the sharp null hypothesis holds, $\mathrm{V}^N(\widehat{\Delta_{HT}})$ is the *true* variance, which can be calculated from the data exactly. When the sharp null hypothesis does not necessarily hold, $\mathrm{V}^N(\widehat{\Delta_{HT}})$ may be construed as an estimator of $\mathrm{V}(\Delta_{HT})$. We therefore refer to a variance estimator constructed by assuming the sharp null hypothesis of no effect as $\widehat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$.

The primary benefit of using $\widehat{\mathrm{V}}^N(\Delta_{HT})$ is that it tends to be more stable than $\widehat{\mathrm{V}}(\widehat{\Delta_{HT}})$, particularly when either n_c or n_t is small, because it combines the variance of the treatment and control groups. In cases where $\widehat{\mathrm{V}}(\widehat{\Delta_{HT}})$ is imprecise, $\widehat{\mathrm{V}}^N(\Delta_{HT})$ may be preferable; highly imprecise standard errors may be downwardly biased even when the associated variance estimator is conservative. The square root is a concave function so, by Jensen's inequality, $\mathrm{E}\left[\widehat{\mathrm{V}}(\widehat{\Delta_{HT}})^{0.5}\right] \leq \left(\mathrm{E}\left[\widehat{\mathrm{V}}(\widehat{\Delta_{HT}})\right]\right)^{0.5}$. Since the estimates from $\widehat{\mathrm{V}}^N(\Delta_{HT})$ will tend to remain stable across randomizations, its use may therefore avoid the bias resulting from Jensen's inequality. However, when effect sizes are large, $\widehat{\mathrm{V}}^N(\Delta_{HT})$ will tend to overestimate the true sampling variability. Note that

computational approximations of exact bias and variance terms may be computed for any estimator under any given sharp null hypothesis using randomization inference, as detailed in section 6.3.

5 Difference estimators

In this section, we propose a simple extension to the Horvitz-Thompson estimator to improve the efficiency of the estimator as well as confer the important property of location invariance.

5.1 Des Raj difference estimator for cluster size

A major source of variability with the Horvitz-Thompson estimator is the variation in the number of individuals in each cluster. Clusters with large n_j will tend to have larger values of Y_j^T – that is, in real world situations, as clusters get larger, the sum of the outcomes for that cluster will also tend to get larger. We use the Des Raj (1965) difference estimator to reduce this variability. To derive the Des Raj difference estimator in this context, we will derive our estimates of the study population totals, Y_{0j}^T and Y_{0j}^T by "differencing" off some of the variability:

$$\widehat{Y_{0,R1}^T} = \frac{M}{m_c} \sum_{j \in J_0} \left(Y_j^T - k(n_j - N/M) \right), \tag{18}$$

where constant k is a *prior* estimate of the regression coefficient from a regression of Y_j^T on n_j and $(n_j - N/M)$ is the difference between the size of cluster j and the average cluster size. (A similar estimator is proposed by Hansen and Bowers (2009), differing primarily in that it contains a random denominator.) k is also roughly equivalent to an estimate of the average value of Y_{ij} for all units and does not have a causal interpretation. In section 5.3, we derive an exact expression for the optimal value of k, which depends on both potential outcomes and the specifics of the experimental design. Similarly,

$$\widehat{Y_{1,R1}^T} = \frac{M}{m_t} \sum_{j \in J_1} \left(Y_j^T - k(n_j - N/M) \right). \tag{19}$$

To develop an intuition about this method, note that it is equivalent to defining a new "differenced" variable U_j^T, where $U_j^T = Y_j^T - k(n_j - N/M)$ and conducting the analysis based on U_j^T instead of Y_j^T. So long as k is fixed before analysis, this strategy does not lead to bias because

$$k\mathrm{E}\left[n_j - N/M\right] = k \cdot 0 = 0. \tag{20}$$

It follows that the HT and Des Raj estimators have the same expected value. Since $\widehat{Y_{0,R1}^T}$ and $\widehat{Y_{1,R1}^T}$ are unbiased, it follows that the Des Raj estimator,

$$\widehat{\Delta_{R1}} = \frac{1}{N} \left[\widehat{Y_{1,R1}^T} - \widehat{Y_{0,R1}^T} \right],$$

is also unbiased. However, estimating k from the same data set can lead to bias, as we demonstrate in B.

Deriving a conservative estimator of the variance of the Des Raj estimator follows directly from section 4.2:

$$\widehat{\mathrm{V}}(\widehat{\Delta_{R1}}) = \frac{M^2}{N^2}\left[\frac{\sum\limits_{j\in J_0}\left(U_j^T - \overline{U_{cj}^T}\right)^2}{m_c(m_c-1)} + \frac{\sum\limits_{j\in J_1}\left(U_j^T - \overline{U_{tj}^T}\right)^2}{m_t(m_t-1)}\right],$$

where $\overline{U_{cj}^T} = \sum\limits_{j\in J_0} U_j^T/m_c$, the mean value of U_j^T in the control condition and $\overline{U_{tj}^T} = \sum\limits_{j\in J_1} U_j^T/m_t$, the mean value of U_j^T in the treatment condition. Similarly, from section 4.2, we may easily construct a variance estimator for $\widehat{\Delta_{R1}}$ by assuming the sharp null hypothesis of no treatment effect:

$$\widehat{\mathrm{V}}^N(\widehat{\Delta_{R1}}) = \frac{M^4\sigma^2(U_j^T)}{N^2(M-1)m_cm_t}.$$

5.2 Invariance of the Des Raj difference estimator

One benefit of the Des Raj estimator is that it has invariance to location transformation, regardless of the accuracy of the researcher's choice of k. In this section, we prove the invariance of the Des Raj estimator. When Y_{0ij} and Y_{1ij} are linearly transformed, k will also change: the same transformation must be applied to k as to Y_{0ij}^T and Y_{1ij}^T. Since k is on the same scale as the outcome variable, when the outcome variable is transformed, k will also be transformed:

$$k^* = (b_0 + b_1 \cdot k). \tag{21}$$

Using this new k^*, we may again define new differenced treatment outcomes,

$$\begin{aligned}
U_{1j}^{T*} &= Y_{1j}^{T*} - k^* \cdot (n_j - N/M) \\
&= \sum_{i=1}^{n_j}(b_0 + b_1 \cdot Y_{1ij}) - (b_0 + b_1 \cdot k)\cdot(n_j - N/M) \\
&= n_j \cdot b_0 + b_1 \cdot Y_{1j}^T - (b_0 + b_1 \cdot k)\cdot(n_j - N/M) \\
&= b_0 \cdot N/M + b_1 \cdot U_{1j}^T.
\end{aligned}$$

And, likewise, we may define new differenced control outcomes, $U_{0j}^{T*} = b_0 \cdot N/M + b_1 \cdot U_{0j}^T$. The estimate based on these transformed variables will be

$$\begin{aligned}
\widehat{\Delta_{R1}^*} &= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} U_j^{T*} - \frac{1}{m_c}\sum_{j\in J_0} U_j^{T*}\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\left(b_0 \cdot N/M + b_1 \cdot U_{1j}^T\right) - \frac{1}{m_c}\sum_{j\in J_0}\left(b_0 \cdot N/M + b_1 \cdot U_{1j}^T\right)\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}b_1\sum_{j\in J_1} U_{1j}^T - \frac{1}{m_c}b_1\sum_{j\in J_0} U_{0j}^T\right] \\
&= b_1 \cdot \widehat{\Delta_{R1}}.
\end{aligned} \tag{22}$$

The Des Raj estimator is therefore invariant to linear transformation because any linear transformation to the outcome will necessarily be reflected in k.

Note that the HT estimator may be considered a special case of the Des Raj estimator when $k = 0$. However, unlike the HT estimator, the explicit assumption that $k = 0$ ensures that when the scale of the outcome changes, the scale of k also changes. The non-invariance of the HT estimator may therefore be thought of as a failure to recognize the implicit assumption that $k = 0$ and to transform to k^* when the scale of the outcome changes.

5.3 Optimal selection of k

To derive the optimal value of k, we begin by noting that the variance of U_{0j}^T is

$$\begin{aligned}\sigma^2\left(U_{0j}^T\right) &= \frac{\sum_j \left(U_{0j}^T - \overline{U_{0j}^T}\right)^2}{M} \\ &= \frac{\sum_j \left(Y_{0j}^T - k\left(n_j - N/M\right) - \overline{Y_{0j}^T}\right)^2}{M} \\ &= \sigma^2\left(Y_{0j}^T\right) + k^2\sigma^2\left(n_j\right) - 2k\sigma\left(n_j, Y_{0j}^T\right),\end{aligned}$$

where $\overline{U_{0j}^T}$ is the mean value of U_{0j}^T over all j clusters. k_{optim_c}, the value of k that minimizes $\sigma^2\left(U_{0j}^T\right)$, can be found using simple optimization. Since the second derivative with respect to k, $2\sigma^2\left(n_j\right)$, must be positive, we may set the first derivative equal to zero and solve for k, so that

$$k_{optim_c} = \frac{\sigma\left(n_j, Y_{0j}^T\right)}{\sigma^2\left(n_j\right)}. \tag{23}$$

Equation 23 should look familiar to the reader: the best fitting k is the ordinary least squares coefficient.

Likewise, the optimal value of k for the potential outcomes under treatment is $k_{optim_t} = \sigma\left(n_j, Y_{1j}^T\right)/\sigma^2\left(n_j\right)$. Given that k_{optim_t} does not generally equal k_{optim_c}, a researcher could justifiably identify different values of k for treatment and control groups. In practice, however, this would require a great amount of prior knowledge (including knowledge about treatment effects); for this reason, a single value of k will typically be preferable. In Appendix C, we derive a single optimal value of k, $k_{optim*} = m_t k_{optim_c}/M + m_c k_{optim_t}/M$.

Unlike a structural parameter, the value of k_{optim*} will depend on the numbers of clusters that are assigned to treatment and to control. Perhaps counterintuitively, when there are fewer clusters in the control condition, k_{optim*} is more heavily weighted toward k_{optim_c}, the value of k that minimizes $\sigma^2\left(U_{0j}^T\right)$ (and vice versa). A simple intuition for this weighting is that the condition with fewer clusters will contribute more to the overall variance of the estimator; so, the greatest increase in precision comes from adjustments made to units in that condition.

The chosen value of k will reduce the variability of the Des Raj estimator, $\widehat{\Delta_{R1}}$, relative to the HT estimator when, for $k_{optim*} > 0$, $0 < k < 2k_{optim*}$ and, for $k_{optim*} < 0$, $0 > k > 2k_{optim*}$.

In other words, the Des Raj estimator will have better precision than the HT estimator unless the researcher picks a k with the wrong sign or chooses a k that is more than twice the magnitude of k_{optim*}. In practice, it is rare that the researcher would not be able to achieve improved precision through Des Raj estimation. k_{optim*} will tend to be close to the average outcome for all individuals; the researcher will usually have prior knowledge about the mean individual-level outcome.

Under the sharp null hypothesis of no treatment effect, $k_{optim*} = k_{optim_c} = k_{optim_t} = \sigma\left(n_j, Y_j^T\right)/\sigma^2\left(n_j\right)$, and thus the optimal k would be the ordinary least squares coefficient from regressing Y_j^T on n_j. *Prima facie*, the intuitive next step would be to try to estimate k from the data, utilizing ordinary least squares on the observed data (perhaps controlling for D_j). However, regression estimates of k can lead to bias in the estimation of treatment effects. In Appendix B, we demonstrate that the bias from estimating k from within-sample data is,

$$\mathrm{E}\left[\frac{\widehat{Y_{1,R1}^T} - \widehat{Y_{0,R1}^T}}{N}\right] - \Delta = \frac{M}{N}\left(\mathrm{Cov}\left(\widehat{k}, \overline{n_{cj}}\right) - \mathrm{Cov}\left(\widehat{k}, \overline{n_{tj}}\right)\right),$$

where $\widehat{k}$ is an estimator of k, $\overline{n_{tj}}$ is the mean value of n_j for clusters in the treatment condition in a given randomization and $\overline{n_{cj}}$ is the mean value of n_j for units in the control condition in a given randomization.

Knowing the optimal value of k under the sharp null hypothesis of no treatment effect is nevertheless informative as we seek to construct principled prior estimates for k. By using the ordinary least squares estimator on auxiliary data with similar potential outcomes, we can approximate k_{optim*} with out-of-sample data, as we will demonstrate in section 6.1.

5.4 Des Raj difference estimator for cluster size and covariates

The Des Raj estimator may also be extended to include other covariates which may further reduce the sampling variability of the estimator. Consider that the researcher has access to A covariates for each individual i in cluster j, denoted by X_{aij}^T, $a \in 1, 2, ..., A$. If the researcher has an individual-level covariate for unit i in cluster j, X_{aij}, the researcher should use the cluster total of the covariate, so that $X_{aj}^T = \sum_{i=1}^{n_j} X_{aij}$. Define the sum of the X_{aij} across all individuals in all clusters $X_a^T = \sum_{j=1}^{M}\sum_{i=1}^{n_j} X_{aij}$. It is simple to adapt the Des Raj estimator to incorporate these additional covariates. Define constants k' and k_a $(\forall a)$ as prior estimates of the coefficients associated with a regression of Y_j on cluster size and cluster-level covariates, respectively. Again, k' and k_a do not have causal interpretations. It follows that we may define

$$\widehat{Y_{0,R2}^T} = \frac{M}{m_c}\sum_{j\in J_0}\left(Y_j^T - k'(n_j - N/M) - \sum_{a=1}^{A} k_a(X_{aj}^T - X_a^T/M)\right)$$

and

$$\widehat{Y_{1,R2}^T} = \frac{M}{m_t}\sum_{j\in J_1}\left(Y_j^T - k'(n_j - N/M) - \sum_{a=1}^{A} k_a(X_{aj}^T - X_a^T/M)\right).$$

By the logic of equation 20, $\widehat{Y^T_{0,R2}}$ and $\widehat{Y^T_{1,R2}}$ are unbiased estimators of Y^T_0 and Y^T_1, respectively. It follows that we may again construct an unbiased estimator of Δ,

$$\widehat{\Delta_{R2}} = \frac{1}{N}\left[\widehat{Y^T_{1,R2}} - \widehat{Y^T_{0,R2}}\right].$$

Following the same steps as in equation 22, it is trivial to show that as long as k' undergoes the same linear transformation as the original data and k_a ($\forall a \in A$) undergoes the same multiplicative scale shift, the Des Raj estimator with covariates will also be invariant. It will also be more efficient than the preceding estimators if the researcher's estimates for k' and k_a are reasonable; constructing variance estimators for $\widehat{\Delta_{R2}}$ is simple and follows directly from section 5.1. Note that the efficiency characteristics of this Des Raj estimator may be derived as in section 5.3, where the same intuitions about efficiency hold.

6 Application

We discuss the case of a randomized experiment designed to assess the causal effect of mobilization efforts on voter turnout. A mobilization organization targeted 10,592 registered voters in African American neighborhoods in Columbus, OH for the purpose of getting out the vote for Democratic presidential candidate John Kerry in the 2004 American presidential election. For exposition, this data set is taken from one stratum of a much larger multistate experiment. Analysis of the single stratum avoids an in-depth discussion of the study design, which was quite involved.

The voters were canvassed on Election Day by workers hired from the local community. The canvassers were instructed to knock on each door in an assigned area and deliver a standard voter mobilization message (without reference to a particular candidate). If there was no answer, the canvasser was instructed to leave a door hanger with a voter mobilization message emphasizing community empowerment. These canvassers were scheduled to work from 2 pm until 7 pm. In that time, if canvassers had finished canvassing their scheduled houses, they were instructed to return to homes where no contact was made initially until the end of the shift.

There were two important design constraints imposed on the experimenter: 1) very few of the experimental blocks could be assigned to the control condition (due to the competitive environment) and 2) only entire street blocks could be assigned to the treatment condition or the control condition. Note that pair matching, as recommended by Imai, King and Nall (2009), would not be feasible in this environment due to the necessary imbalance between treatment and control (without omitting the vast majority of blocks). While Imai, King and Nall's recommendation of pair matching is often a sound design decision, this experiment provides an example where such a design would be infeasible.

There are a total of 347 street blocks (i.e., clusters), of which 345 are assigned to treatment and 2 are assigned to control. These clusters range in size from 10 to 160 units. The outcome measure of interest, Y_{ij}, is whether or not the individual i on block j voted in the 2004 American presidential election (coded 1 if the individual voted, 0 if the individual did not vote). In addition, covariate information on the vote history of each individual in the 2004 primary election, the 2002 general election and the 2000 general election is available.

6.1 Selection of k, k', and k_a

For the Des Raj estimators, it is important that the researcher be able to have a principled method for the selection of k that does not utilize any within-sample data. We use a dataset consisting of 78,929 individuals in 3,246 street blocks located in 9 distinct regions in Columbus, OH to estimate k, k', and k_a. This dataset was originally from the other 9 randomization strata in Columbus, OH in the aforementioned larger experiment. Following from section 5.3, we use ordinary least squares (OLS) regression on this external dataset to estimate k, k', and k_a. To estimate k for the Des Raj estimator with only n_j, we use the following model:

$$Y_j^T = \alpha + kn_j + \sum_{f=1}^{F-1} \gamma_f \Gamma_f + e_j,$$

where α is a constant, F is the number of regions (in this case, 9), γ_f is the fixed effect coefficient for region f, Γ_f is an indicator variable indicating whether cluster j is in region f, and e_j is a random disturbance. Fixed effects are specified for region to reduce sampling variability. This estimation procedure yields a principled estimate for k. In column 1 of table 1, the estimate for k is listed.

Covariate	Des Raj (n_j)	Des Raj (n_j & X_{aj})
n_j	0.283	0.251
Total Voting in 2004 Primary		0.134
Total Voting in 2002 General		0.414
Total Voting in 2000 General		0.102

Table 1: Selection of k, k', and k_a

To estimate k' and k_a for the Des Raj estimator with both n_j and covariates, we use the following model:

$$Y_j^T = \alpha' + k'n_j + k_1 X_{1j}^T + k_2 X_{2j}^T + k_3 X_{3j}^T + \sum_{f=1}^{F-1} \gamma_f \Gamma_f + e_j,$$

where α' is a constant, X_{1j}^T is the total voting in the 2004 primary election on cluster j, X_{2j}^T is the total voting in the 2002 general election on cluster j, and X_{3j}^T is the total voting in the 2000 general election on cluster j. In column 2 of table 1, the estimates for k' and k_a are listed. Note that k is not generally equal to k', due to correlation between n_j and the other covariates. In the following section, we use these estimates for the Des Raj estimator.

6.2 Treatment effect estimates

In this section, we estimate the ATE using five common estimators, as well as the three posited in this paper. We begin by detailing each of these estimators. The first estimator under consideration is the difference-in-means estimator, $\widehat{\Delta}_S$. As detailed in section 3.2, the difference-in-means

estimator is prone to bias. And as Freedman (2008a) notes, the difference-in-means estimator is equivalent to regression with ordinary least squares.

We then consider multiple regression with OLS to reduce sampling variability (also known as regression adjustment). As Freedman (2008a) notes, even without clustering, regression adjustment may be biased if either treatment assignment is imbalanced (i.e., $n_t \neq n_c$) or there exists treatment effect heterogeneity (i.e., $\exists i, j$ s.t. $\tau_i \neq \tau_j$). First, we include a model where we simply adjust for n_j:

$$Y_{ij} = \beta_0 + \beta_1 D_j + \beta_2 n_j + e_{ij},$$

where β_0 is a constant, β_1 is an estimate of Δ, β_2 is an estimate of k and e_{ij} is an individual-level random disturbance. Second, we also use OLS with n_j and vote history:

$$Y_{ij} = \beta_0 + \beta_1 D_j + \beta_2 n_j + \beta_3 X_{1ij} + \beta_4 X_{2ij} + \beta_5 X_{3ij} + e_{ij},$$

where X_{1ij} indicates whether or not unit i in cluster j voted in the 2004 primary election, X_{2ij} indicates whether or not unit i in cluster j voted in the 2002 general election, X_{3ij} indicates whether or not unit i in cluster j voted in the 2000 general election, and $\beta_2 - \beta_5$ are corresponding coefficients. For all three OLS models (including the difference-in-means), Huber-White "robust" cluster standard errors are estimated. While often sufficient for inference, these standard errors may be unreliable in finite samples (Freedman, 2006; Angrist and Pischke, 2009).

Random effects estimation is often recommended for the analysis of cluster-randomized experiments (Green and Vavreck, 2008). However, this estimator is not guaranteed to be unbiased. We use the following specification:

$$Y_{ij} = \beta_0 + \beta_1 D_j + \beta_2 n_j + \beta_3 X_{1ij} + \beta_4 X_{2ij} + \beta_5 X_{3ij} + e_j + e_{ij},$$

where e_j is a normally distributed cluster-level disturbance (and e_{ij} is also distributed normally). This model is estimated using the `lmer()` function in the lme4 (Bates and Maechler, 2010) package in R (R Development Core Team, 2010) using the default settings. Standard errors are empirical Bayes estimates also produced by the `lmer()` function.

And, finally, we present treatment effect estimates for the Horvitz-Thompson estimator, the Des Raj difference estimator (with n_j) and the Des Raj difference estimator (with n_j and history). The standard error estimates are the square root of our estimated variances. As noted above, unlike our variance estimators, however, the standard error estimators are not guaranteed to be generally conservative (due to Jensen's inequality). This phenomenon is generally true; even an unbiased estimator of the variance rarely guarantees that the associated standard error estimator will be unbiased.

The ATE estimates and associated SE estimates are presented illustratively in table 2. These estimates may be of substantive interest, but do not speak to the properties of the estimator. We will more formally benchmark these estimators. In the following section, we will demonstrate how to calculate the bias and variance (and therefore standard error) of each estimator under different null hypotheses using randomization inference.

Estimator	$\widehat{\Delta}$	$\widehat{\text{SE}}$	$\widehat{\text{SE}}^N$
Difference-in-Means	4.6	(1.1)	
Ordinary Least Squares (n_j)	8.1	(1.4)	
Ordinary Least Squares (n_j, history)	5.7	(0.8)	
Random Effects	5.5	(7.2)	
Horvitz-Thompson	11.1	(3.5)†	(15.4)
Des Raj Difference (n_j)	7.3	(2.0)†	(8.3)
Des Raj Difference (n_j, X_{aj})	5.7	(1.0)†	(8.2)

Table 2: Treatment effect estimates in percentage points. † indicates SEs estimated from variance estimators derived in this paper. $\widehat{\text{SE}}^N$ refers to SEs derived under the sharp null hypothesis of no treatment effect. All other SE estimators described in section 6.2.

6.3 Randomization inference

Randomization inference (RI) will allow us to assess the bias and variance of any given estimator. In addition, RI allows the researcher to perform completely nonparametric significance testing (see, e.g., Rosenbaum, 2002). We refer to the estimate produced by a given estimator as the test statistic. RI assumes that a given sharp null hypothesis holds and evaluates the test statistic for every possible random assignment of units to treatment and control. By recalculating the test statistic for each possible treatment assignment, the reference distribution of the test statistic is constructed. Fisher's exact test is a well known form of RI for significance testing, but the method is much more general.

Because the total possible permutations increase rapidly with population size, RI may be computationally infeasible. We may use Monte Carlo simulations to approximate RI by repeatedly assigning units to treatment and control groups randomly and estimating the test statistic that would be observed for each repetition. The distribution of the test statistic across randomizations forms the reference distribution of the statistic. As the number of repetitions gets large, the distribution of the test statistic based on repeated randomizations converges to that of the the full RI. This method can achieve results arbitrarily close to RI by increasing the number of repetitions.

6.4 Randomization inference with the sharp null hypothesis of no treatment effect

The hypothesis most commonly associated with RI is the sharp null hypothesis of no treatment effect, outlined in section 4.2. We may compute sampling distributions (and thus bias and standard errors) for all posited estimators of the ATE under the assumption of no treatment effect whatsoever.

In Figure 1, we present the sampling distributions associated with each estimator. As expected, the HT estimator and the two Des Raj estimators are unbiased. Notably, the difference-in-means estimator has a bias of -2.1 percentage points (pp); in a study with no actual effect, this bias is substantively meaningful. Even controlling for n_j through regression adjustment, there still exists

a bias of -0.6 pp. OLS controlling for n_j and history performs better, reducing the bias further to -0.1 pp. Random effects estimation is biased by -0.1 pp as well.

We can also use the reference distributions in figure 1 to compute the variance of each estimator if the sharp null hypothesis of no effect holds. As expected, the HT estimator is the least efficient of all estimators; the SE for this estimator is 15.4 pp. However, the Des Raj estimators reduce the sampling variability associated with the estimator greatly: differencing out n_j and X_{aj}^T reduces the SE to 8.2 pp. This SE is superior to all other estimators except the OLS (with controls for n_j and history) and the random effects estimator, which have SEs of 7.1 pp and 7.0 pp respectively. At the very least, the Des Raj estimator is comparable in efficiency to all of the other estimators.

We need not limit ourselves to only benchmarking the ATE estimators, though. We may use the same RI procedure with the SE estimates as the statistics of interest. In figure 2, we present the distributions of estimated SEs associated with the sharp null hypothesis of no treatment effect. Aside from the SEs assuming no effect, only the random effects SE estimator is conservative and its SE has very low variance, suggesting that the estimated SE is very reliable. All other SE estimators are both biased downward and unreliable. The Huber-White SE estimators are anticonservative and skewed; e.g., the SE for OLS (controlling for n_j) is biased by -4.1 pp. And, although our variance estimators for the HT and Des Raj estimators are unbiased because the sharp null hypothesis of no treatment effect holds, the high variance of the estimators leads to considerable downward bias in SE estimation. The SE estimator for the HT estimator has the highest variance of any of the SE estimators and is therefore one of the most biased. And, as expected, the SEs assuming no treatment effect whatsoever are exactly correct and have no variance when there is indeed no treatment effect. When the SE estimators are as unreliable as they are in this application, RI calculations (assuming no effect) will be preferable.

As mentioned above, RI also facilitates significance testing. We may ask if the observed $\widehat{\Delta}$ is significantly distinguishable from the distribution of $\widehat{\Delta}$ that would be observed if H_0 were to hold. We refer to the null distribution (under H_0) of $\widehat{\Delta}$ as $\widehat{\mathbf{\Delta}^{\mathbf{N}}}$ and a particular draw from this distribution as $\widehat{\Delta^N}$. Roughly following Ho and Imai (2006), a one-tailed p-value,

$$p = \Pr(\widehat{\mathbf{\Delta}^{\mathbf{N}}} \leq \widehat{\Delta}) = |\{\widehat{\Delta^N} \in \widehat{\mathbf{\Delta}^{\mathbf{N}}} : \widehat{\Delta^N} \leq \widehat{\Delta}\}|/|\Omega|,$$

where Ω is the set of all possible randomizations (estimated via a large random sample from the full set of randomizations). For example, for the Des Raj estimator (with n_j and history), 55.2% of randomizations yield a $\widehat{\Delta^N} \geq 5.7$; the p-value is thus approximately 0.55. None of the ATE estimates have p-values smaller than 0.43 and we therefore do not reject the sharp null hypothesis of no treatment effect.

6.5 Randomization inference with a sharp null hypothesis of no average treatment effect with treatment effect heterogeneity

We may also wish to know the performance of the estimators if there exists treatment effect heterogeneity; in the following section, we posit a particular sharp null hypothesis of no *average* treatment effect. We will see that the assumption of this hypothesis leads to considerably different inferences about the performances of the estimators.

Figure 1: ATE estimator sampling distributions associated with the sharp null hypothesis of no treatment effect detailed in section 6.4. 250,000 randomizations were used to estimate the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team, 2010), with the default settings and a bandwidth of 2 percentage points. Each estimator is detailed in section 6.2. The vertical line indicates the expected value, and therefore bias, of the estimator. Bias and SE estimates in the upper-right of each plot are computed from each empirical distribution.

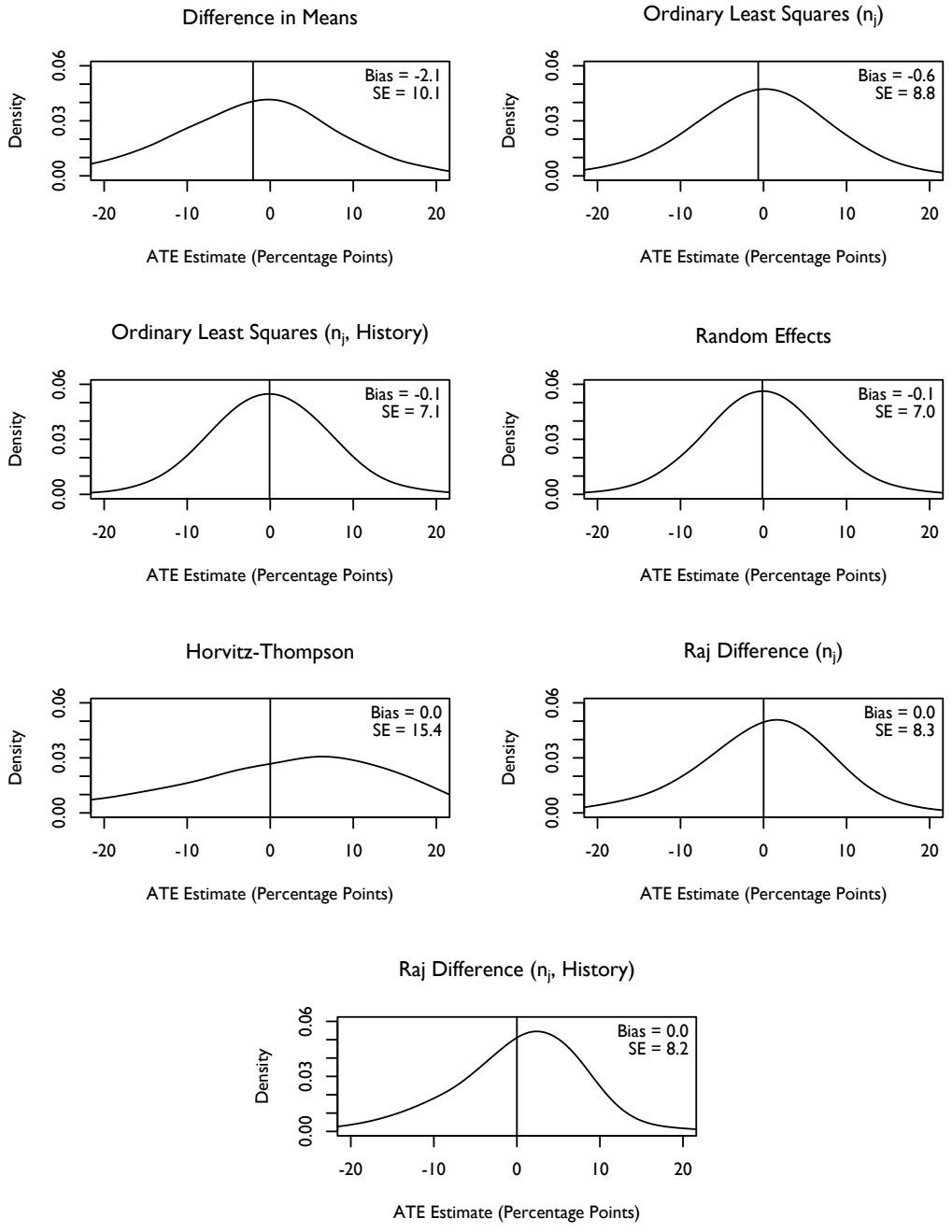
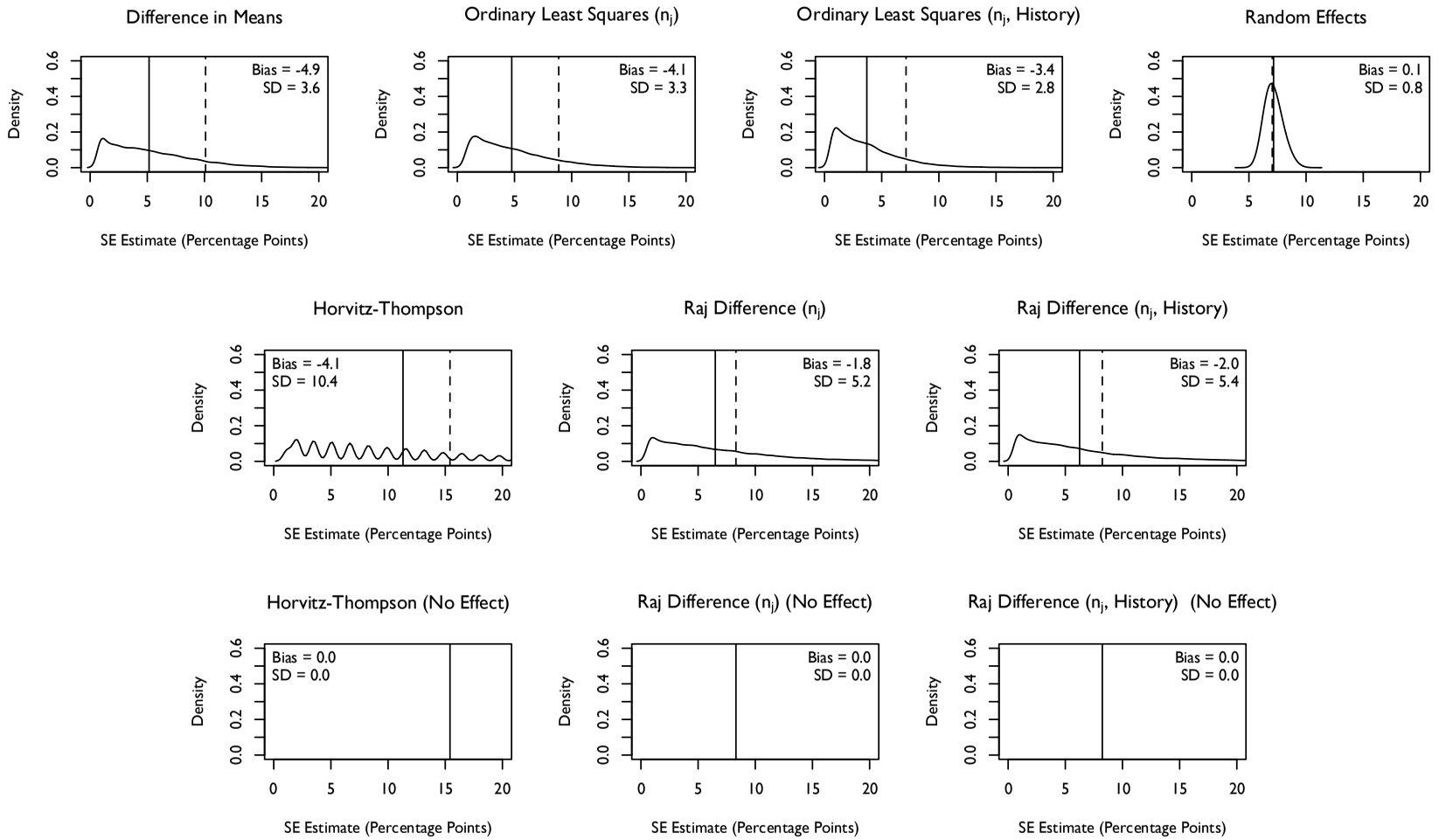


Figure 2: SE estimator sampling distributions associated with the sharp null hypothesis of no treatment effect detailed in section 6.5. 50,000 randomizations were used to estimate the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team, 2010), with the default settings and a bandwidth of 2 percentage points; this display method smoothes over discontinuities in the density plots, most noticably for the HT SE estimator. Each SE estimator is detailed in section 6.2. The solid vertical line indicates the expected value of the SE estimator. The dotted vertical line indicates the true standard error of the estimator. Bias and SD estimates are computed from each empirical distribution.



Using RI, we analyze the properties of each of the estimators with a sharp null hypothesis of no average treatment effect. Now, what if the average treatment effect were zero, but the treatment effect were negatively correlated with n_j? For illustration, we assume that

$$H_0^a : \tau_{ij} = \frac{{n_j}^{0.5} - \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{n_j}{n_j}^{0.5}}{4\max_j(n_j^{0.5})}.$$

Transforming n_j breaks τ_{ij}'s perfect correlation with n_j, and the denominator normalizes the magnitude of τ_{ij}. H_0^a implies that $\Delta = 0$. If H_0^a holds, we may compute both potential outcomes for each unit: $Y_{0ij}^{N'} = Y_{ij} - D_j\tau_{ij}$ and $Y_{1ij}^{N'} = Y_{ij} + (1 - D_j)\tau_{ij}$. In Figure 3, we present the sampling distributions of the ATE estimators under H_0^a. The bias associated with all of the biased estimators is now magnified greatly such that every non-design-based estimator has a bias < -0.9 pp. For example, the random effects estimator, previously with a bias of -0.1 pp under the sharp null of no effect, now has a bias of -1.1 pp. However, our design-based estimators remain unbiased with heterogeneous treatment effects.

For a given estimator, the true standard error depends on the distribution of treatment effects. In this case, the particular posited treatment effect reduces the variance in the control group cluster totals. Since control group totals contribute more to sampling variability in this experiment (from equation 17), the variance of the HT and Des Raj estimators are therefore reduced. All other estimators suffer from reduced precision with treatment effect heterogeneity and the Des Raj estimator (n_j, history) now has the lowest variance of all estimators. However, although not reported, when the sign of the treatment effect is reversed, this efficiency gain is lost. Even though its superiority does not generally hold, the Des Raj estimator, at the very least, has efficiency on par with existing, biased estimators of the ATE.

We may again assess the performance of the SE estimators, now under this sharp null hypothesis of no average treatment effect. While most of the estimators are substantively unchanged by the different null hypothesis, notably, the random effects estimator is now downwardly biased. It is also instructive to examine the performance of the SEs assuming the sharp null hypothesis of no treatment effect. While these SEs now have some variance and are no longer exactly correct, their variance is negligible (at most, the SE estimator has an SD of 0.1 pp) and they are all conservative. For example, the Des Raj estimator (n_j, history) is biased upwards by 0.8 pp and has almost no variance (SD $<$ 0.1 pp). Again, the RI-based SEs are preferable even though they are assuming an incorrect hypothesis about treatment effects.

7 Conclusion

The unbiased estimation of the ATE in cluster-randomized experiments has been elusive. In unpacking the source of the bias in the difference-in-means estimator, this paper has returned to the first principles of randomization and sampling theory. This paper shows that the fundamental statistical properties of randomization can be applied to modern causal inferential problems. Not only does the Des Raj estimator provide the basis for an unbiased and location-invariant estimator

Figure 3: ATE estimator sampling distributions associated with the sharp null hypothesis of no average treatment effect detailed in section 6.5. 250,000 randomizations were used to estimate the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team, 2010), with the default settings and a bandwidth of 2 percentage points. Each estimator is detailed in section 6.2. The vertical line indicates the expected value, and therefore bias, of the estimator. Bias and SE estimates in the upper-right of each plot are computed from each empirical distribution.

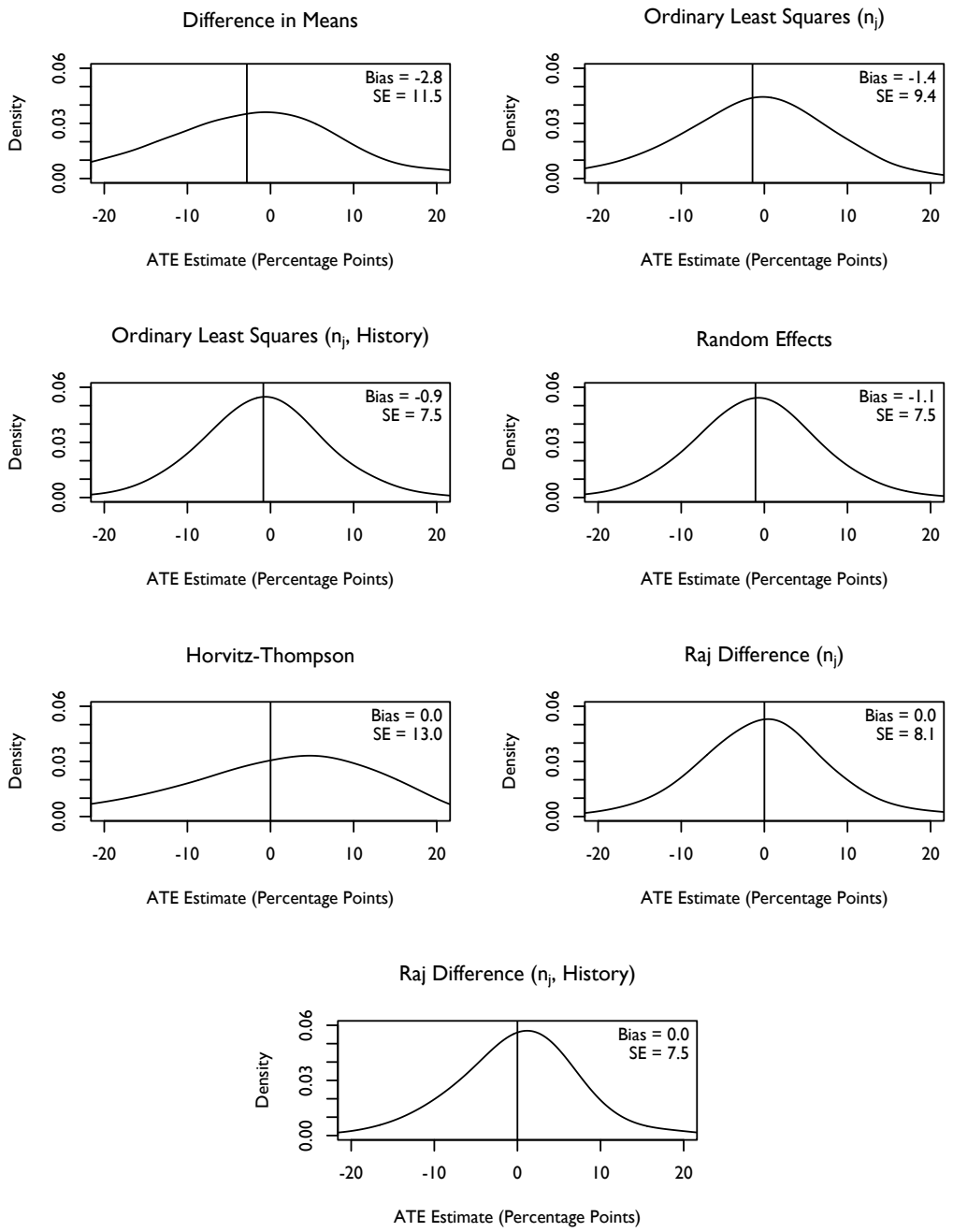
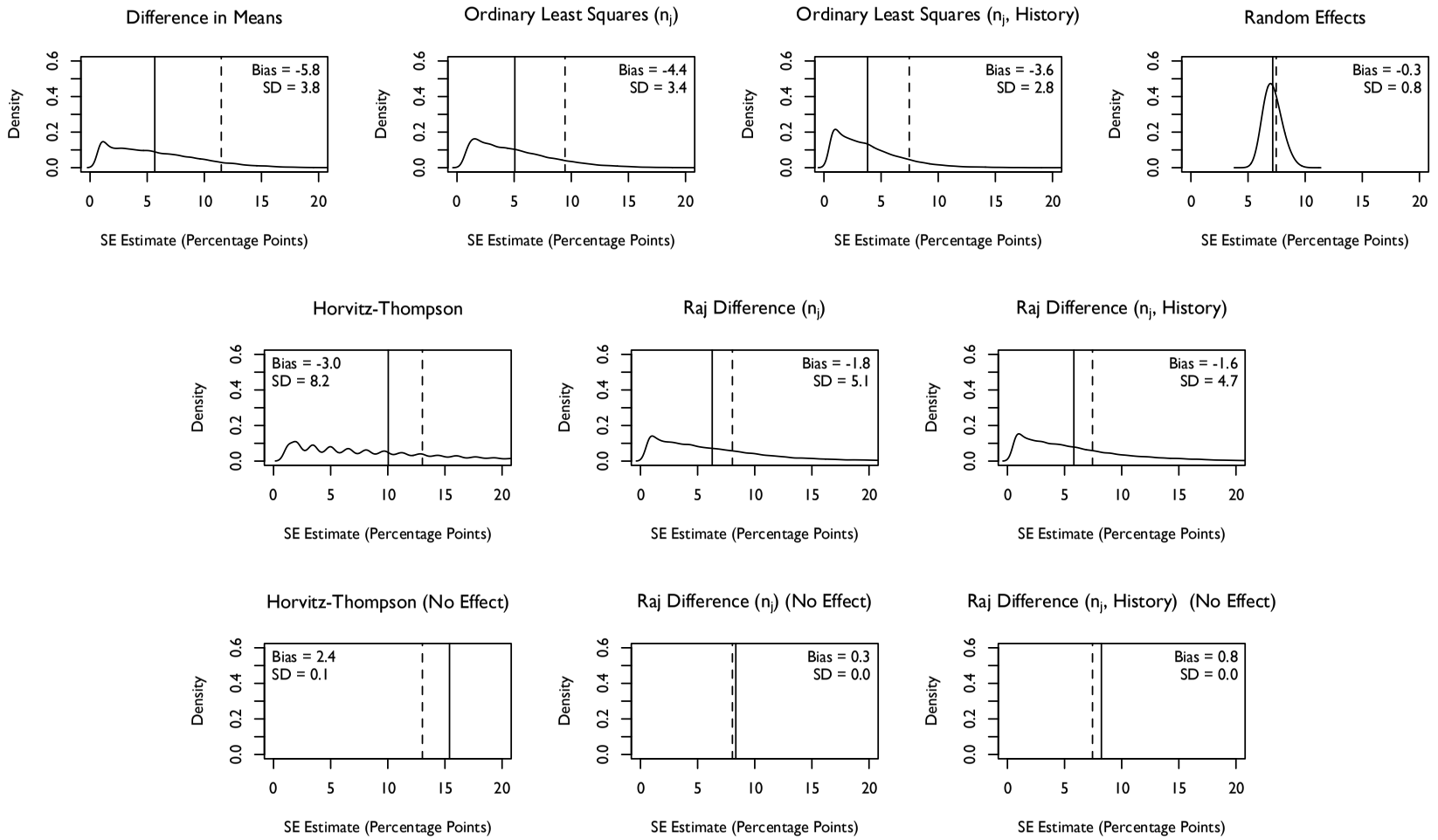


Figure 4: SE estimator sampling distributions associated with the sharp null hypothesis of no average treatment effect detailed in section 6.5. 50,000 randomizations were used to estimate the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team, 2010), with the default settings and a bandwidth of 0.33 percentage points; this display method smoothes over discontinuities in the density plots, most noticably for the HT SE estimator. Each SE estimator is detailed in section 6.2. The solid vertical line indicates the expected value of the SE estimator. The dotted vertical line indicates the true standard error of the estimator. Bias and SD estimates are computed from each empirical distribution. Distributions for the SEs under the sharp null hypothesis of no treatment effect were too narrow to display.



for the analysis of cluster-randomized experiments, it also achieves improved precision through covariate adjustment.

There are a number of theoretical implications of this return to the first principles of randomization. First, machinery based solely on sampling theoretic ideas can be sufficient for precise and unbiased estimation of causal parameters. Second, researchers need not feel that achieving precise and unbiased causal estimates requires an up-to-date knowledge of complex statistical models: we may easily derive estimators with good statistical properties using only fundamental concepts. Third, utilizing such estimators serves to remind us of the importance of this distinction between observational studies and randomized experiments. The importance of the logic of the experiment, with its reliance on randomization, may be lost when researchers rely on model-based estimators that may or may not reflect the experimental design.

Acknowledgements

The authors acknowledge support from the Yale University Faculty of Arts and Sciences High Performance Computing facility and staff. The authors would also like to thank Adam Dynes, Don Green, Mary McGrath, David Nickerson, Cyrus Samii and Allie Sovey for helpful comments. The usual caveat applies.

References

Angrist, J.D. and Pischke, J. (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

Bates, D. and Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package, version 0.999375-37.

Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74** 911–915.

Chaudhuri, A. and Stenger, H. (2005). *Survey Sampling*. Chapman and Hall, Boca Raton.

Cochran, W.G. (1977.) *Sampling Techniques*, 3rd Ed. John Wiley, New York.

Des Raj. (1965). On a method of using multi-auxiliary information in sample surveys. *J. Amer. Statist. Assoc.* **60** 270–277.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Oxford Univ. Press, New York.

Freedman, D.A. (2006). On the so-called 'Huber Sandwich Estimator' and 'robust' standard errors. *Amer. Statistician* **60** 299–302.

Freedman, D.A. (2008a). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193.

Freedman, D.A. (2008b). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196.

Freedman, D.A., Pisani R. and Purves, R.A. (1998). *Statistics*, 3rd edition. W. W. Norton, Inc., New York.

Green, D.P. and Vavreck, L. (2008). Analysis of cluster-randomized experiments: A comparison of alternative estimation approaches. *Political Analysis* **16** 138–152.

Hansen, B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Stat. Sci.* **23** 219–236.

Hansen, B. and Bowers, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885.

Hartley, H.O. and Ross, A. (1954). Unbiased ratio estimators. *Nature* **174** 270.

Ho, D.E. and Imai, K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *J. Amer. Statist. Assoc.* **101** 888–900.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663-684.

Humphreys, M. 2009. Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. Working paper. Available at: http://www.columbia.edu/˜mh2245/papers1/monotonicity4.pdf.

Imai, K., King, G. and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statist. Sci*. **24** 29–53.

Lin, W. In press. Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *Annals of Applied Statistics*.

Middleton, J.A. (2008). Bias of the regression estimator for experiments using clustered random assignment. *Stat. Probability Lett.* **78** 2654–2659.

Miratrix, L., Sekhon, J. and Yu, B. In press. Adjusting Treatment Effect Estimates by Post-Stratication in Randomized Experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statist. Sci.* **5** 465–480. (Translated in 1990.)

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, Vol. 97, No. 4: 558–625

R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. Version 2.12.0.

Rosenbaum, P.R. (2002). *Observational Studies*, second edition. Springer, New York.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

Rubin, D.B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331.

Samii, C. and Aronow, P.M. 2012. On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments. *Statistics and Probability Letters*. **82**: 365–370.

Sarndal, C-E. (1978). Design-Based and Model-Based Inference in Survey Sampling. *Scandinavian Journal of Statistics*. **5**, 1: 27–52.

Williams, W. H. (1961). Generating unbiased ratio and regression estimators. *Biometrics* **17** 267–274.

A Proof of non-invariance of the Horvitz-Thompson estimator

To prove that the HT estimator is not invariant to location shifts, we need only replace Y_j^T with its linear transformation:

$$
\begin{aligned}
\widehat{\Delta}^*_{HT} &= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} Y_j^{T*} - \frac{1}{m_c}\sum_{j\in J_0} Y_j^{T*}\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\sum_{i=1}^{n_j} Y_{ij}^{*} - \frac{1}{m_c}\sum_{j\in J_0}\sum_{i=1}^{n_j} Y_{ij}^{*}\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\sum_{i=1}^{n_j} (b_0 + b_1\cdot Y_{ij}) - \frac{1}{m_c}\sum_{j\in J_0}\sum_{i=1}^{n_j} (b_0 + b_1\cdot Y_{ij})\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\left(n_j\cdot b_0 + \sum_{i=1}^{n_j} b_1\cdot Y_{ij}\right)\right. \\
&\qquad \left. -\frac{1}{m_c}\sum_{j\in J_0}\left(n_j\cdot b_0 + \sum_{i=1}^{n_j} b_1\cdot Y_{ij}\right)\right] \\
&= b_0\cdot\frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} n_j - \frac{1}{m_c}\sum_{j\in J_0} n_j\right] \\
&\qquad + b_1\cdot\frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\sum_{i=1}^{n_j} Y_{ij} - \frac{1}{m_c}\sum_{j\in J_0}\sum_{i=1}^{n_j} Y_{ij}\right] \\
&= b_0\cdot\frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} n_j - \frac{1}{m_c}\sum_{j\in J_0} n_j\right] + b_1\cdot\widehat{\Delta}_{HT}.
\end{aligned}
$$

B Bias from estimating k from within-sample data

Consider the situation where one wishes to improve upon the HT estimator by adjusting for cluster size; in other words, one wishes to estimate k in equations 18 and 19 from the data to approximate the optimal value of k with an estimator $\widehat{k}$. In this scenario, the expected value of equation 18 yields

$$
\begin{aligned}
\mathrm{E}\left[\widehat{Y_{1,R1}^{T}}\right] &= \mathrm{E}\left[\frac{M}{m_t}\sum_{j\in J_1}\left(Y_j^T - \widehat{k}(n_j - N/M)\right)\right] \\
&= \frac{M}{m_t}\left(\mathrm{E}\left[\sum_{j\in J_1} Y_j^T\right] - \mathrm{E}\left[\sum_{j\in J_1}\widehat{k}n_j\right] + \mathrm{E}\left[\sum_{j\in J_1}\widehat{k}N/M\right]\right) \\
&= \frac{M}{m_t}\left(\mathrm{E}\left[m_t\overline{Y_{1j}^T}\right] - \mathrm{E}\left[\widehat{k}m_t\overline{n_{tj}}\right] + \mathrm{E}\left[\widehat{k}m_t N/M\right]\right) \\
&= Y_1^T - M\left(\mathrm{E}\left[\widehat{k}\overline{n_{tj}}\right] - \mathrm{E}\left[\widehat{k}\right]\mathrm{E}\left[\overline{n_{tj}}\right]\right) \\
&= Y_1^T - M\mathrm{Cov}\left(\widehat{k}, \overline{n_{tj}}\right), \qquad (24)
\end{aligned}
$$

where $\overline{n_{tj}}$ is the mean value of n_j for clusters in the treatment condition in a given randomization. In the third line of equation 24, $\widehat{k}$ moves outside the summation operator because it is a constant for a given randomization. Likewise,

$$
\mathrm{E}\left[\widehat{Y_{0,R1}^{T}}\right] = Y_0^T - M\mathrm{Cov}\left(\widehat{k}, \overline{n_{cj}}\right), \qquad (25)
$$

where $\overline{n_{cj}}$ is the mean value of n_j for units in the control condition in a given randomization. So the expected value of the estimator will be

$$
\mathrm{E}\left[\frac{\widehat{Y_{1,R1}^{T}} - \widehat{Y_{0,R1}^{T}}}{N}\right] = \Delta + \frac{M}{N}\left(\mathrm{Cov}\left(\widehat{k}, \overline{n_{cj}}\right) - \mathrm{Cov}\left(\widehat{k}, \overline{n_{tj}}\right)\right). \qquad (26)
$$

The term on the right of equation 26 represents the bias. A special case with no bias is when the sharp null hypothesis of no treatment effect holds and treatment and control groups have equal numbers of clusters. We refer the reader to Williams (1961), Freedman (2008a) and Freedman (2008b) for additional reading on the particular bias associated with the regression adjustment of random samples and experimental data.

C Derivation of the optimal value of k

To identify a single optimal value of k, k_{optim*}, we refer to the first line of equation 17,

$$
v\mathrm{V}\left(\widehat{\Delta_{R1}}\right) = c\sigma^2\left(U_{j0}^T\right) + t\sigma^2\left(U_{j1}^T\right) + 2\sigma\left(U_{j0}^T, U_{j1}^T\right) \qquad (27)
$$

where $v = \frac{(M-1)N^2}{M^2}$, $c = \frac{M-m_c}{m_c}$ and $t = \frac{M-m_t}{m_t}$. Now note that the terms $\sigma^2\left(U_{j0}^T\right)$, $\sigma^2\left(U_{j0}^T\right)$, and $\sigma\left(U_{j0}^T, U_{j1}^T\right)$ in equation 27 can be written as follows:

$$
\sigma^2\left(U_{j1}^T\right) = \sigma^2\left(Y_{j1}^T\right) + k^2\sigma^2\left(n_j\right) - 2k\sigma\left(Y_{j1}^T, n_j\right), \qquad (28)
$$

$$\sigma^2\left(U_{j0}^T\right) = \sigma^2\left(Y_{j0}^T\right) + k^2\sigma^2\left(n_j\right) - 2k\sigma\left(Y_{j0}^T, n_j\right), \tag{29}$$

and, defining $\delta_j = (n_j - N/M)$,

$$\begin{aligned}
\sigma\left(U_{j0}^T, U_{j1}^T\right) &= \mathrm{E}\left[U_{j0}^T U_{j1}^T\right] - \overline{U_0^T U_1^T} \\
&= \mathrm{E}\left[Y_{j0}^T - k\delta_j\right]\left(Y_{j1}^T - k\delta_j\right) - \overline{Y_0^T Y_1^T} \\
&= \mathrm{E}\left[Y_{j0}^T Y_{j1}^T - Y_{j0}^T k\delta_j - Y_{j1}^T k\delta_j + k^2\delta_j^2\right] - \overline{Y_0^T Y_1^T} \\
&= \mathrm{E}\left[Y_{j0}^T Y_{j1}^T\right] - \overline{Y_0^T Y_1^T} - \mathrm{E}\left[Y_{j0}^T k\delta_j\right] - \mathrm{E}\left[Y_{j1}^T k\delta_j\right] + \mathrm{E}\left[k^2\delta_j^2\right] \\
&= \sigma\left(Y_{j0}^T, Y_{j1}^T\right) - k\left[\sigma\left(Y_{j0}^T, n_j\right) + \mathrm{E}\left[Y_{j0}^T\right]\mathrm{E}\left[\delta_j\right]\right] \\
&- k\left[\sigma\left(Y_{j1}^T, n_j\right) + \mathrm{E}\left[Y_{j1}^T\right]\mathrm{E}\left[\delta_j\right]\right] + k^2\sigma^2\left(n_j\right) \\
&= \sigma\left(Y_{j0}^T, Y_{j1}^T\right) - k\left[\sigma\left(Y_{j0}^T, n_j\right) + \mathrm{E}\left[Y_{j0}^T\right]\cdot 0\right] \\
&\quad - k\left[\sigma\left(Y_{j1}^T, n_j\right) + \mathrm{E}\left[Y_{j1}^T\right]\cdot 0\right] + k^2\sigma^2\left(n_j\right) \\
&= \sigma\left(Y_{j0}^T, Y_{j1}^T\right) - k\sigma\left(Y_{j0}^T, n_j\right) - k\sigma\left(Y_{j1}^T, n_j\right) + k^2\sigma^2\left(n_j\right),
\end{aligned} \tag{30}$$

respectively. Substituting equations 28, 29, and 30 into equation 27,

$$\begin{aligned}
v\mathrm{V}\left(\widehat{\Delta_{R1}}\right) &= c\left[\sigma^2\left(Y_{j0}^T\right) + k^2\sigma^2\left(n_j\right) - 2k\sigma\left(Y_{j0}^T, n_j\right)\right] \\
&+ t\left[\sigma^2\left(Y_{j1}^T\right) + k^2\sigma^2\left(n_j\right) - 2k\sigma\left(Y_{j1}^T, n_j\right)\right] \\
&+ 2\left[\sigma\left(Y_{j0}^T, Y_{j1}^T\right) - k\sigma\left(Y_{j0}^T, n_j\right) - k\sigma\left(Y_{j1}^T, n_j\right) + k^2\sigma^2\left(n_j\right)\right].
\end{aligned}$$

Setting the first derivative with respect to k equal to zero,

$$\begin{aligned}
0 &= c\left[2k_{optim*}\sigma^2\left(n_j\right) - 2\sigma\left(Y_{j0}^T, n_j\right)\right] \\
&+ t\left[2k_{optim*}\sigma^2\left(n_j\right) - 2\sigma\left(Y_{j1}^T, n_j\right)\right] \\
&+ 2\left[-\sigma\left(Y_{j0}^T, n_j\right) - \sigma\left(Y_{j1}^T, n_j\right) + 2k_{optim*}\sigma^2\left(n_j\right)\right],
\end{aligned}$$

$$\begin{aligned}
ck_{optim*}\sigma^2\left(n_j\right) + tk_{optim*}\sigma^2\left(n_j\right) + 2k_{optim*}\sigma^2\left(n_j\right) &= c\sigma\left(Y_{j0}^T, n_j\right) + t\sigma\left(Y_{j1}^T, n_j\right) \\
&+ \sigma\left(Y_{j0}^T, n_j\right) + \sigma\left(Y_{j1}^T, n_j\right)
\end{aligned}$$

$$\begin{aligned}
\left(\frac{M-m_c}{m_c} + \frac{M-m_t}{m_t} + \frac{m_c}{m_c} + \frac{m_t}{m_t}\right) k_{optim*}\sigma^2\left(n_j\right) &= \left(\frac{M-m_c}{m_c} + \frac{m_c}{m_c}\right) \\
&\cdot \sigma\left(Y_{j0}^T, n_j\right) \\
&+ \left(\frac{M-m_t}{m_t} + \frac{m_t}{m_t}\right) \\
&\cdot \sigma\left(Y_{j1}^T, n_j\right)
\end{aligned}$$

$$\left(\frac{M}{m_c} + \frac{M}{m_t}\right) k_{optim*}\sigma^2\left(n_j\right) = \left(\frac{M}{m_c}\right)\sigma\left(Y_{j0}^T, n_j\right) + \left(\frac{M}{m_t}\right)\sigma\left(Y_{j1}^T, n_j\right)$$

$$k_{optim*} = \left(\frac{1}{m_c} + \frac{1}{m_t}\right)^{-1} \left[\left(\frac{1}{m_c}\right) \frac{\sigma\left(Y_{j0}^T, n_j\right)}{\sigma^2\left(n_j\right)} + \left(\frac{1}{m_t}\right) \frac{\sigma\left(Y_{j1}^T, n_j\right)}{\sigma^2\left(n_j\right)}\right]$$

$$k_{optim*} = \left(\frac{1}{m_c} + \frac{1}{m_t}\right)^{-1} \left[\left(\frac{1}{m_c}\right) k_{optim_c} + \left(\frac{1}{m_t}\right) k_{optim_t}\right]$$

$$k_{optim*} = \frac{m_t}{M} k_{optim_c} + \frac{m_c}{M} k_{optim_t}.$$

The Des Raj estimator will be more efficient than the HT estimator when

$$\begin{aligned}
\left(c+t+2\right) k^2 \sigma^2\left(n_j\right) &< 2k\left[\left(c+1\right)\sigma\left(Y_{j0}^T, n_j\right) + \left(t+1\right)\sigma\left(Y_{j1}^T, n_j\right)\right] \\
\left(c+t+2\right) k^2 &< 2k\left[\left(c+1\right)\frac{\sigma\left(Y_{j0}^T, n_j\right)}{\sigma^2\left(n_j\right)} + \left(t+1\right)\frac{\sigma\left(Y_{j1}^T, n_j\right)}{\sigma^2\left(n_j\right)}\right] \\
\left(c+t+2\right) k^2 &< 2k\left[\left(c+1\right) k_{optim_c} + \left(t+1\right) k_{optim_t}\right] \\
k^2 &< 2k\left[\frac{m_t}{M} k_{optim_c} + \frac{m_c}{M} k_{optim_t}\right] \\
k^2 &< 2k \cdot k_{optim*}.
\end{aligned}$$