DISSERTATION

RANDOMIZATION TESTS FOR EXPERIMENTS EMBEDDED IN COMPLEX SURVEYS

Submitted by

David A. Brown

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2022

Doctoral Committee:

Advisor: F. Jay Breidt

Julia Sharp
Tianjian Zhou
Stephen Ogle

ABSTRACT

RANDOMIZATION TESTS FOR EXPERIMENTS EMBEDDED IN COMPLEX SURVEYS

Embedding experiments in complex surveys has become increasingly important. For scientific questions, such embedding allows researchers to take advantage of both the internal validity of controlled experiments and the external validity of probability-based samples of a population. Within survey statistics, declining response rates have led to the development of new methods, known as adaptive and responsive survey designs, that try to increase or maintain response rates without negatively impacting survey quality. Such methodologies are assessed experimentally. Examples include a series of embedded experiments in the 2019 Triennial Community Health Survey (TCHS), conducted by the Health District of Northern Larimer County in collaboration with the Department of Statistics at Colorado State University, to determine the effects of monetary incentives, targeted mailing of reminders, and double-stuffed envelopes (including both English and Spanish versions of the survey) on response rates, cost, and representativeness of the sample.

This dissertation develops methodology and theory of randomization-based tests embedded in complex surveys, assesses the methodology via simulation, and applies the methods to data from the 2019 TCHS.

An important consideration in experiments to increase response rates is the overall balance of the sample, because higher overall response might still underrepresent important groups. There have been advances in recent years on methods to assess the representativeness of samples, including application of the dissimilarity index (DI) to help evaluate the representativeness of a sample under the different conditions in an incentive experiment (Biemer et al. [2018]). We develop theory and methodology for design-based inference for the DI when used in a complex survey. Simulation studies show that the linearization method has good properties, with good confidence interval

coverage even in cases when the true DI is close to zero, even though point estimates may be biased.

We then develop a class of randomization tests for evaluating experiments embedded in complex surveys. We consider a general parametric contrast, estimated using the design-weighted Narain-Horvitz-Thompson (NHT) approach, in either a completely randomized design or a randomized complete block design embedded in a complex survey. We derive asymptotic normal approximations for the randomization distribution of a general contrast, from which critical values can be derived for testing the null hypothesis that the contrast is zero. The asymptotic results are conditioned on the complex sample, but we include results showing that, under mild conditions, the inference extends to the finite population. Further, we develop asymptotic power properties of the tests under moderate conditions. Through simulation, we illustrate asymptotic properties of the randomization tests and compare the normal approximations of the randomization tests with corresponding Monte Carlo tests, with a design-based test developed by van den Brakel, and with randomization tests developed by Fisher-Pitman-Welch and Neyman. The randomization approach generalizes broadly to other kinds of embedded experimental designs and null hypothesis testing problems, for very general survey designs.

The randomization approach is then extended from NHT estimators to generalized regression estimators that incorporate auxiliary information, and from linear contrasts to comparisons of nonlinear functions.

ACKNOWLEDGEMENTS

I give major thanks F. Jay Breidt for his consistent advice and encouragement as I worked through my dissertation. I thank my committee, Steve Ogle, Julia Sharp, and Tianjian Zhou for their efforts reading this dissertation and offering helpful suggestions to make it better. Major thanks also go to the Health District of Northern Larimer County and Susan Hewitt, Suman Mathur, and Angela Castillo for allowing Jay and I to help design their survey and use their data in my dissertation. Lastly, thanks go to Elliott Forney for creating the Colorado State dissertation template in LaTeX.

DEDICATION

I would like to dedicate this thesis to my parents, Russell and Kathy, my brother, Michael, and all of the friends I made in Fort Collins that helped me grow into the person I am continually becoming.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Preliminaries

Now is a time of major growth and change in the field of survey statistics. A positive development has been the explosion of technologies that make it easier to gain auxiliary information. But there has also been a long trend of declining survey response rates (RRs) (Steeh [1981], Brick and Williams [2013], Peytchev [2013], National Research Council [2013]). This has led to increased focus on different strategies to try to improve response quality. These methods include adaptive and responsive survey designs (Wagner [2008], Groves and Heeringa [2006], Schouten et al. [2018]) and integration with non-probability samples (Elliott and Valliant [2017]). Concerns with data quality have also led to a renewed focus on the use of experiments embedded in surveys (Lavrakas et al. [2019]). In this dissertation, we develop a new method, based in randomization theory, to analyze experiments embedded in complex surveys accounting for both the survey design and the experimental design.

Before delving into the new material in this dissertation, we will start with a brief introduction to survey statistics, randomization inference, and experiments in surveys. We then finish the introduction with a review of the major contributions and an overview of the structure of this dissertation.

1.1 Survey Statistics

A major theme throughout statistics is trying to understand how and when you can generalize inference from a sample to a more general population, and additionally understanding how uncertain you are about the population values inferred. While in many other fields of statistics, models are used to approximate the data generation process, in survey statistics, the dominant mode of inference has been design-based inference (see Smith [1976] or Fienberg and Tanur [1987]) for more in depth discussion). In this section, we provide a brief review of some of the major concepts in design-based inference, starting with the Narain-Horvitz-Thompson (NHT) estimator (Narain

[1951], Horvitz and Thompson [1952]). Some recent books that explain the methodology of survey statistics include Heeringa et al. [2017] and Wu and Thompson [2020].

We will start out this discussion with a finite population $U = \{1, 2, \ldots, N\}$ consisting of N units. From here we consider a sampling design $p(S)$ which assigns probabilities for the selection of any possible subset $S \subset U$. In general when examining statistics with surveys, directly using the probabilities of drawing each sample is often not particularly useful. The standard framework of design-based inference involves expressing survey designs in terms of their first and second order inclusion probabilities.

For an arbitrary unit $i \in U$, the first-order inclusion probability (denoted π_i) is the probability that unit i is included in the selected sample, and satisifies $\pi_i = \sum_{\{S:i\in S\}} p(S)$. Similarly, for arbitrary units $i, j \in U$, the second-order inclusion probability (π_{ij}) is the probability that both units i and j are included in the selected sample, and satisfies $\pi_{ij} = \sum_{\{S:i,j\in S\}} P(S)$ (implying that $\pi_{ii} = \pi_i$). Inclusion probabilities reflect how representative the sample is of a finite population. One simple requirement of sampling is that each element of the population has a chance of appearing in the sample, or $\pi_i > 0$ for each $i \in U$. A design satisfying this property is called a probability sampling design, and admits unbiased estimation of finite population totals (Särndal et al. [1992] Section 1.3). Another valuable property related to inclusion probabilities is a measurable sampling design. A measurable sampling design is a design where any two elements in the population have a positive probability of both being selected, or $\pi_{ij} > 0$ for all i and j in the population. Measurable designs admit unbiased variance estimators (Särndal et al. [1992] p. 33).

To define the NHT estimator, we assume a sample $S \subset U$ is drawn via a probability sampling design, and the data Y_i for $i \in S$ are observed and measured without error. Define the sampling weights as $w_i = \pi_i^{-1}$. The NHT estimator for the sampling design is then

$$\hat{Y} = \sum_{i\in S} \frac{Y_i}{\pi_i} = \sum_{i\in S} w_i Y_i.$$

The NHT estimator is unbiased under the sampling design because, if E_S denotes expectation over all possible random selections under the sampling design,

$$\mathrm{E}_S\left[\sum_{i\in S}\frac{Y_i}{\pi_i}\right] = \sum_{i=1}^{N}\pi_i\frac{Y_i}{\pi_i} = \sum_{i=1}^{N} Y_i.$$

This follows because π_i is the probability that unit i will be included in the sample. If the sample is a measurable sampling design, then the variance of the NHT estimator has an unbiased estimator.

The NHT estimator can be extended to more general cases, including multiple stages of sampling and stratification. When considering multiple stages of sampling, the units selected at the first stage are called called primary sampling units (PSUs) are selected from the population. Subsampling within PSUs then proceeds in one or more additional stages of selection, until the sampling stops with the ultimate sampling units, from which data are obtained. Information from the earlier stages of sampling may be used in later stages.

To write the estimator with multiple stages of selection, we index the PSUs by $i = 1, \ldots, M$ in the population, and denote the sample of PSUs by S, where the sample size is $|S| = m$. We denote the estimate for the total within the i^{th} PSU as $\hat{Y}_{i+}$. For the overall estimator to be an NHT estimator, the estimate $\hat{Y}_{i+}$ should be an NHT estimator conditioned on i being a sampled PSU. In this case, the NHT estimator can be written as

$$\hat{Y} = \sum_{i\in S} w_i \hat{Y}_{i+}.$$

We also note that using an estimator other than the NHT for the PSU total estimates may still give an estimator with good statistical properties, but the resulting estimator will not be an NHT estimator. For example, estimators of PSU totals may incorporate auxiliary information, as is the case for generalized regression estimators (GREG; reference Chapters 6-8 of Särndal et al. [1992]) and more general calibration estimators (Deville and Sarndal [1992]).

Many surveys incorporate stratification, in which the population U is partitioned into disjoint (and, ideally, homogeneous) strata $U = \cup_{h=1}^{H} U_h$, and then samples are selected independently

from each stratum, $S_h \subset U_h$. Then the NHT estimator is

$$\hat{Y} = \sum_{h=1} \sum_{i \in S_h} w_{hi} Y_{hi}.$$

It should also be noted that clustered sampling can occur within each stratum. If we then let $\hat{Y}_{hi+}$ denote NHT estimators of totals within each PSU, the NHT estimator for the full sample can then be written as

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{hi+}.$$

One caveat about notation needs to be given now. Throughout the rest of this dissertation, the notation $h = 1, \ldots, H$ will be used to refer to blocks in a randomized complete block experiment embedded in a complex survey. It is possible that these blocks will correspond to strata used in the sampling, but this does not need to be true.

An important topic in complex surveys is variance estimation, in part because of the heavy reliance on normal approximations for the distribution of estimators under the sampling design. As noted above, the NHT estimator admits an unbiased variance estimator under a measurable sampling design, and accurate approximations are commonly used for multistage designs, alternative estimators like GREG, and nonlinear statistics. Related topics in variance estimation under randomized treatment assignment for experiments embedded in surveys will be addressed in Chapters 4 and 5, and 6 of this dissertation.

1.2 Randomization Inference

When trying to extract information from a sample, there are many philosophical frameworks that statisticians use to understand the data. The two most well known are frequentist inference and Bayesian inference. Another philosophy that has shown to be useful due to its lack of distributional assumptions is the philosophy of randomization inference. Randomization inference is useful in situations where data are in different groups, possibly assigned in an experimental design, and focuses on how the data are influenced by the group membership.

Within randomization inference, there is a distinction between permutation tests and randomization tests that is not consistently observed (see, for example, Ernst [2004] or Onghena [2017]). We will follow the language used in Onghena [2017] to describe the differences between these two procedures. Permutation tests view the data as coming from multiple populations, and the test is to make inferences about differences between the populations. With a randomization test, the data are viewed as coming from a single population, usually having been assigned treatments in an experiment, and inferences are made about the difference between the treatment effects on the population. One recent example of a permutation test in surveys is a test proposed by Toth [2020] to investigate differences in population domain parameters in a complex survey. This dissertation will discuss randomization tests for survey experiments.

Within randomization tests, there have also been two major approaches that have been taken: an approach originally attributed to Fisher [1971] (first edition 1935) and expanded by Pitman [1937a, 1937b, 1938] and Welch [1937], and an approach suggested by Neyman (Splawa-Neyman et al. [1990] (translation of work published in 1923), Neyman et al. [1935]). A good introduction to these approaches and a comparison between them is presented in Ding [2017]. We briefly introduce these tests, starting with the method proposed by Neyman.

The Neyman randomization method posits a model of potential outcomes for each experimental unit (EU) in the sample. Specifically, for each unit i, there are K potential outcomes, $Y_i^{(1)}, \ldots, Y_i^{(K)}$ corresponding to each of the possible K treatments that could be assigned. This model depends on the stable unit treatment value assumption (SUTVA; Rubin [1980]), which is that the observed values for a given unit depend only on the treatment that unit was assigned. Using this model, Neyman computes variances within each treatment, and uses these to compute an upper bound on the variability of the difference.

The Fisher-Pitman-Welch (FPW) randomization method uses a strong null hypothesis to derive a reference distribution for the test statistic when the treatment has no effect. This null hypothesis is that for each unit i and treatment k, $Y_i^{(k)} = Y_i$, no matter how the treatments are assigned, which is stronger than SUTVA. The reference distribution for the FPW method is the distribution

of the elements of the sample with the treatments randomly relabeled by the same distribution of the original treatment assignment. If the null hypothesis is true, and there really is no difference between the treatments, this reference distribution exactly corresponds to the distribution of the test statistic under different randomizations of the original experiment.

There are two major differences between the two methodologies. One difference is that because the FPW method involves only relabeling the treatment assignments, it generally admits a test based on Monte Carlo simulation of the reference distribution under its null hypothesis. The Neyman method depends on an assumption about the differences between the observations assigned each treatment, and thus only lends itself to statistics that can be well approximated by differences between the two treatments. Additionally, also because of the approximation of the variance components between the two treatments, the Neyman method usually has better power than the FPW method (Ding [2017]). This is because by relying on the relabeling of treatments, the FPW method includes the treatment effect in the estimate of the variability of the statistic, when such treatment effect exists. The Neyman method, on the other hand, only considers the variability within each of the treatments in estimating the variability of the estimated treatment effect.

1.3 Experiments in Surveys

There is a long history of experiments embedded in complex surveys (e.g. Mahalanobis [1958], Fienberg and Tanur [1987], and Fienberg and Tanur [1988]). More recently, since survey RRs have been falling for decades (Steeh [1981] National Research Council [2013]), there has been a push for adaptive and responsive designs for improving response and data quality, which can involve embedded experimentation. Additionally, there has been a push for "population-based experiments" in the social sciences (Mutz [2011]). These population-based experiments are experiments embedded in population-based surveys to try to leverage both the internal validity of experiments and the external validity of surveys.

With the development of experiments embedded in surveys, some researchers have thought about analyzing experiments in surveys, but the literature on such methods is not very large. Smith

[1983] used likelihood theory to argue that when the sample selection does not depend on the value of the response of inference, then the sampling procedure can be ignored in model-based inferences. Fienberg and Tanur [1988] (Section 6) discuss three different approaches for analyzing such experiments: standard model-based experimental methods, survey-based methods treating each treatment as a different sample, and methods viewing the experiment as a sample in the population of all survey experiments. Van den Brakel in a series of papers (van den Brakel and Renssen [1998], van den Brakel [2001], van den Brakel and van Berkel [2002], van den Brakel and Renssen [2005], van den Brakel [2008], van den Brakel [2013], and van den Brakel [2019]) has developed a method using a design-based procedure. Van den Brakel [2001] further argued in Section 2.5 that standard model-based methods may work for analyzing experimental impacts on certain survey features such as response rates, but that methods that take into account both the survey and experimental designs are valuable for analyzing differences in survey trends.

Van den Brakel's method is based on a contrast statistic $\mathbf{C}\hat{\boldsymbol{Y}}$, where $\mathbf{C} \neq \mathbf{0}$ is a matrix such that $\mathbf{C}\mathbf{1}_K = \mathbf{0}$ and $\hat{\boldsymbol{Y}}$ is a $K \times 1$ vector of estimated totals for the K treatments. In reality, van den Brakel discusses means rather than totals, but the difference is immaterial; we use means here for consistency with the discussion of the randomization test in this dissertation, which uses totals. Van den Brakel computed the variance of the contrast statistic using the design-based variance for the three sources of error present in experiments in surveys: sampling from the finite population, assigning the treatments, and measuring the responses. Using these computation, van den Brakel computes a design-based variance estimator that allows for inference to be conducted using a Wald test. Van den Brakel's methodology can be viewed as an extension of the Neyman style randomization tests, and additionally accounts for measurement error in the responses.

In this dissertation, we will provide theory for extending the FPW randomization methodology to experiments embedded in complex surveys. We additionally provide theoretical justification for the use of a Wald test under moderate conditions (Chapters 4–6).

We conclude with another note on language. In the rest of this dissertation, when we talk about a "randomization test" or "randomization procedures" we will be talking about the FPW

randomization test, or its extension to complex surveys described in this dissertation. We will always specify Neyman's name when referring to the Neyman randomization test, and we will refer to the test proposed by van den Brakel [2001] as a "design-based" method, consistent with how van den Brakel has referred to the test in his writings. When referring to design-based inference only applying to sample surveys, we use the descriptor "survey design-based."

1.4 Contribution and Chapter Outlines

The main contribution in this dissertation is the development of new methodologies for evaluating experiments embedded in complex surveys based on randomization procedures, and application of these methodologies to experiments embedded in the 2019 Triennial Community Health Survey conducted by the Health District of Northern Larimer County. Additionally, we provide theoretical justification of the Wald test in this randomization method and in the design-based method proposed by van den Brakel [2001]. We also propose a method for survey design-based inference of the dissimilarity index to evaluate the representativeness of survey samples.

In Chapter 2, we discuss the Health District survey, and apply our new methodologies in the analysis of that data. In Chapter 3, we describe our methodology for survey design-based infercncs for the dissimilarity index. In Chapter 4, we introduce the randomization test for experiments embedded in complex surveys in completely randomized experiments. In Chapter 5, we extend the randomization methodology for randomized complete block designs. In Chapter 6, we discuss some extensions of the randomization methodology to calibration estimators and nonlinear statistics and some practical issues of using the randomization methodology for experiments embedded in complex surveys. We conclude in Chapter 7 with discussion of contributions of this dissertation, and future directions of research to extend this methodology.

Chapter 2

An Incentive Experiment Embedded in a Community Health Survey

2.1 Introduction

Household survey response rates (RRs) have been falling for decades (see, for example, Steeh [1981], Brick and Williams [2013], Peytchev [2013] and National Research Council [2013]). Accordingly, survey practitioners have tried various methods to increase or maintain RRs (see Kanuk and Berenson [1975] or Edwards et al. [2009] for reviews). As described in Schouten [2018] (pp. 19-25), these methods can be studied with randomized experiments embedded within surveys, and experimental results may be used to identify effective protocols for future waves within the same survey (responsive designs; Groves and Heeringa [2006]) or for future surveys (adaptive designs; Wagner [2008]).

Methods that focus on increasing RRs may increase nonresponse bias or other survey errors (Groves [2006], Groves and Peytcheva [2008]). There is some evidence in the literature that providing incentives may reduce nonresponse bias, though perhaps not for all variables. Kanuk and Berenson [1975] suggest that incentives are effective at increasing response rates across income levels. Groves and Peytcheva [2008] (p. 176) say there is evidence people with lower incomes are more sensitive to incentives than people with higher incomes, which may induce bias on variables related to income. There is also evidence incentives may help increase RRs among those surveyed who may be less intrinsically interested in the survey topic (Groves et al. [2006], Peytchev [2013] p. 97).

Monetary incentives, especially when provided with the survey packet, have consistently been shown effective at increasing RRs (Bevis [1948], Kanuk and Berenson [1975], Church [1993], Avdeyeva and Matland [2013], Singer and Ye [2013], Zhang et al. [2018]). While there is a good

body of research on pre-incentives and post-incentives individually (Rao [2020], Church [1993]), very little research exists comparing the effects of combined pre- and post-incentives versus pre- or post- incentives alone (Beydoun et al. [2006], Avdeyeva and Matland [2013] and Coopersmith et al. [2016] are examples), or on push-to-web incentives (Biemer et al. [2018] is an example). The previous literature on pre- and post- incentives is sparse and inconclusive. Church [1993] found in a meta-analysis that there is an effect on response from pre-incentives, but not an effect from conditional incentives. Avdeyeva and Matland [2013] found that a post-incentive had no improvement in response after a pre-incentive had been given. Rao [2020] found a higher RR with a pre-incentive only than higher post-incentives alone. Beydoun et al. [2006] and Coopersmith et al. [2016] both tested pre- and post-incentives vs. post-incentives alone and do not find much evidence of a difference between the incentive groups in terms of contact (in the case of Beydoun et al. [2006]) or response (in the case of Coopersmith et al. [2016]). Beydoun et al. [2006] did find some evidence that contact rates were improved with the early incentive when they did not have a telephone number on file. Biemer et al. [2018] found some evidence that providing a $20 post-incentive can increase RR compared to a $10 post-incentive, and that an incentive to complete the survey online can greatly increase web response over no such incentive when both mail and web options are provided concurrently. Biemer et al. [2018] also found the highest overall RRs with the push-to-web incentive, though not enough to provide solid statistical evidence on its own, and found no evidence that the different incentive structures they tried had any impact on the representativeness of the sample.

Previous literature on mixed-mode surveys shows that higher RRs are obtained when surveys offer a mail option first, and then a web option, and that higher web RRs are obtained from having a web option first (Messer and Dillman [2011], Millar and Dillman [2011], Patrick et al. [2018]). Research into whether a concurrent design has higher RRs than a sequential design has mixed results, with Millar and Dillman [2011] finding the sequential design led to higher RRs, and Biemer et al. [2018] and Mauz et al. [2018] finding higher RRs in the concurrent design. Patrick et al. [2018] found that sequential mixed-mode surveys had lower cost-per-response than mail-only sur-

veys. These studies have not shown evidence that sequential mixed-mode surveys are more or less representative than other designs.

Since 1995, the Health District of Northern Larimer County (Colorado) has surveyed the adult population of the county every three years about health status, needs, and behaviors. The ninth iteration of Larimer County's Triennial Community Health Survey (TCHS) was fielded in the fall of 2019, using an address-based sample of 12,000 randomly selected addresses and a six-wave, push-to-web approach. The TCHS design has been modified over time to maintain validity and control costs, from random-digit dialing with mail follow-up to a push-to-web design in 2019. We assessed the effectiveness of the push-to-web incentive with a randomized experiment, comparing a $1 prepaid incentive with a promise of $5 for online survey completion ($1+$5web) to the $2 prepaid incentive only ($2pre) given in previous waves of the TCHS. Additional experiments embedded in the survey assessed the effectiveness of a targeted postcard reminder and of "double-stuffed" mailings of paper questionnaires printed in English and Spanish. These adaptive strategies seek to increase (or at least maintain) survey RRs and assure that those who complete the survey represent Larimer County's adult population, while reducing costs by pushing respondents online.

Based on American Community Survey (ACS) 5-year estimates from 2018, the population of Larimer County (excluding census tract 6) is predominantly White (92.7%), and relatively highly educated (41.7% with Bachelor's degrees; Table 2.1). Respondents from the survey skew older, female, and more educated than the general population (Table 2.1). The demographics provided for survey respondents in Table 2.1 are not adjusted for nonresponse, though nonresponse adjustments, including a propensity model based on auxiliary variables and calibration to census data, were performed for reporting the official statistics from the 2019 TCHS.

In this chapter, we focus on the incentives study and compare the effects of combined pre- and post-incentives versus pre-incentives alone on RRs, on the representativeness of the respondents, and on cost-effectiveness of the survey. Specifically, we focus in three main questions: 1) does the $1+$5web incentive treatment lead to a higher RR than the $2pre incentive treatment, 2) how do

the incentive treatments compare in terms of cost-effectiveness, and 3) does the incentive treatment affect how well the sample represents the adult population of Larimer County.

Table 2.1: Demographics of the adult population of Larimer County, Colorado excluding census tract 6, along with design-weighted estimates from the survey and from each incentive treatment ($2pre and $1+$5web) and each response mode (paper and online). Design-weighted estimates are not corrected for survey nonresponse. Standard errors are computed from the survey design, where the experimental treatments are treated as separate surveys, and the paper and online responses are treated as domain estimates.

Category	Pop	Survey	$2pre	$1+$5web	Paper	Online
Age						
18-24	16.3	3.0 (0.4)	2.7 (0.6)	3.4 (0.7)	1.1 (0.5)	3.6 (0.6)
25-34	18.3	8.8 (0.6)	8.7 (0.8)	8.9 (1.0)	8.2 (1.2)	9.0 (0.7)
35-44	15.6	11.7 (0.7)	11.8 (0.9)	11.6 (1.1)	9.7 (1.2)	12.4 (0.8)
45-54	14.7	13.7 (0.7)	13.3 (0.9)	14.1 (1.2)	10.4 (1.3)	14.8 (0.9)
55-64	16.3	23.8 (0.9)	24.3 (1.2)	22.8 (1.5)	23.4 (1.9)	23.9 (1.1)
65-74	11.5	25.4 (0.9)	24.9 (1.2)	26.4 (1.5)	27.5 (1.9)	24.8 (1.0)
75-84	5.1	10.9 (0.7)	11.6 (0.9)	9.7 (1.0)	14.3 (1.5)	9.7 (0.7)
85+	2.1	2.8 (0.3)	2.6 (0.4)	3.2 (0.6)	5.4 (0.9)	1.9 (0.3)
Sex						
Female	50.4	62.3 (1.0)	62.2 (1.3)	62.5 (1.7)	69.2 (2.0)	60.0 (1.2)
Male	49.6	37.7 (1.0)	37.8 (1.3)	37.5 (1.7)	30.8 (2.0)	40.0 (1.2)
Race						
White	92.7	93.9 (0.5)	94.1 (0.7)	93.6 (0.9)	95.7 (0.8)	93.4 (0.6)
Black	1.0	0.4 (0.1)	0.3 (0.2)	0.5 (0.2)	0.2 (0.2)	0.5 (0.2)
AIAN	0.7	0.6 (0.2)	0.7 (0.2)	0.5 (0.2)	0.4 (0.2)	0.6 (0.2)
API	2.2	1.3 (0.3)	1.4 (0.3)	1.1 (0.4)	1.6 (0.5)	1.3 (0.3)
Other	1.3	2.3 (0.3)	2.2 (0.4)	2.4 (0.6)	2.1 (0.6)	2.3 (0.4)
More than one	2.1	1.5 (0.3)	1.2 (0.3)	2.0 (0.5)	0.0 (0.0)	2.0 (0.4)
Hispanic						
Yes	9.4	5.4 (0.5)	5.1 (0.6)	5.9 (0.8)	3.8 (0.9)	5.9 (0.6)
No	90.6	94.6 (0.5)	94.9 (0.6)	94.6 (0.8)	96.2 (0.9)	94.1 (0.6)
Education						
Less than High School	4.6	1.1 (0.2)	1.1 (0.2)	1.1 (0.4)	2.0 (0.5)	0.9 (0.2)
High School Graduate	19.8	8.0 (0.6)	7.3 (0.7)	9.0 (1.0)	9.3 (1.2)	7.6 (0.7)
Some Coll./Assoc.	33.8	25.1 (0.9)	26.4 (1.2)	23.1 (1.5)	28.6 (1.9)	24.0 (1.0)
Bachelor's or more	41.7	65.8 (1.0)	65.1 (1.3)	66.7 (1.6)	60.2 (2.1)	67.6 (1.1)
Health Insurance						
Yes	93.3	96.6 (0.4)	96.3 (0.5)	97.1 (0.6)	96.7 (0.8)	96.6 (0.5)
No	6.7	3.4 (0.4)	3.7 (0.5)	2.9 (0.6)	3.3 (0.8)	3.4 (0.5)
Income						
<$25,000	16.8	12.4 (0.6)	11.9 (0.8)	13.4 (1.1)	16.1 (1.5)	11.3 (0.7)
$25,000-49,999	21.4	19.3 (0.8)	19.8 (1.0)	18.5 (1.3)	21.1 (1.7)	18.8 (0.9)
$50,000-99,999	31.1	34.3 (1.0)	33.9 (1.2)	34.9 (1.6)	29.5 (1.9)	35.7 (1.1)
≥$100,000	30.7	34.0 (1.0)	34.4 (1.2)	33.3 (1.6)	33.3 (2.0)	34.2 (1.1)

2.2 Methods

2.2.1 Survey Methods

In the 2019 TCHS, 12000 addresses were sampled using an address-based sample from 72 of the 73 census tracts comprising Larimer County, Colorado (Census tract 6, which corresponds to Colorado State University, was excluded; additionally there was only one address sampled from census tract 28.03, which is a National Park). Three census tracts (13.04, 13.06, and 17.04) were oversampled at 2-3 times the rate of the other census tracts to increase representation of Hispanics and younger people, who were known to have lower RRs from previous surveys. To allow for this oversampling and ensure representation of multi-family addresses from the sample, these census tracts were stratified out, and strata were also created by dwelling type (single-family, multi-family, and only way to get mail (OWTGM) PO Box). As OWTGM PO boxes were not present in the over-sampled census tracts, this led to a total of 9 strata. Additionally, some addresses in Fort Collins were removed from consideration as they were sampled by another organization conducted a survey simultaneously. In addition to the list of addresses, the auxiliary variables Hispanic surname, dwelling type, housing tenure, and income were purchased for the sampled addresses. This was a web-mail mixed mode survey, meaning addresses were initially only able to respond online and paper surveys were mailed in wave 4 of data collection (with the double-stuff experiment).

The 2019 TCHS included embedded experiments to assess survey design features aimed at improving survey outcomes, especially response. These adaptive features included differential treatments for different addresses in the sample. In the first embedded experiment, addresses in one-third of the census tracts were randomly assigned to receive the $1+$5web incentive, while addresses in the other two-thirds of the census tracts received the $2pre incentive, which had been standard practice for the TCHS in the past. To improve balance in the treatment assignments, the census tracts were divided into blocks of about 3 census tracts each by geography, education, and census tract population. The differential incentives were assigned at the tract level to minimize the chances of sampled neighbors comparing their incentives.

The other experiments in the survey were designed to increase RRs for lower-income groups. A targeted mailing was sent to approximately 5000 addresses thought to be less likely to respond based on the auxiliary variables. Addresses flagged as having a Hispanic surname were thought to be less likely to respond, and a categorical variable with three levels was created to identify addresses not flagged as having a Hispanic surname thought to be less likely to respond. The level 0 addresses were considered the most likely to respond and the level 2 addresses were considered the least likely to respond (see Figure 2.1 for details). The targeted mailing was sent to approximately one-half of the addresses flagged as having a Hispanic surname, one-half of the level 1 addresses and three-quarters of the level 2 addresses. The level 0 addresses did not receive the targeted mailing. This does mean that the treatment was not assigned to the entire sample. This was done because the Health District was more interested in understanding how the targeted mailing worked for addresses thought to be less likely to respond and improving balance in response than on understanding how the targeted mailing worked for the entire population. Approximately 1000 of the targeted addresses that had not responded by wave 6 received another targeted mailing close to the end of survey collection.

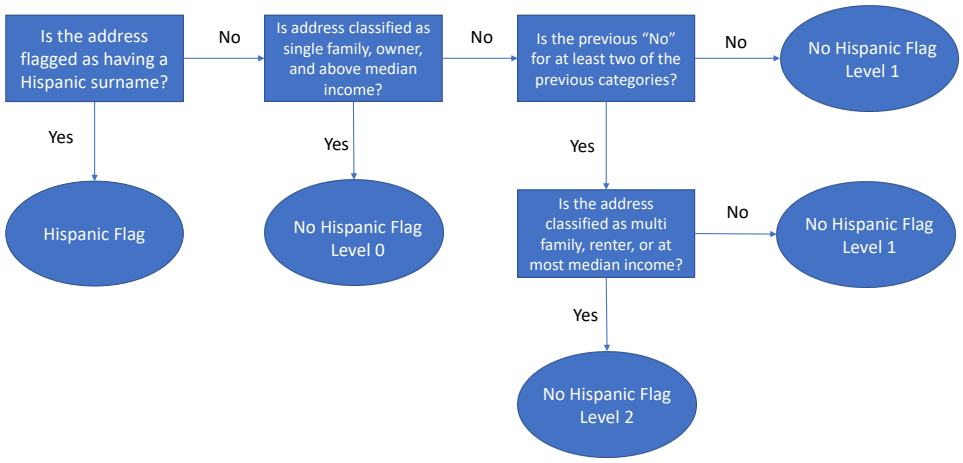

Figure 2.1: Flowchart showing the determination of addresses judged less likely to respond based on available vendor data.

To improve response of Hispanics in the TCHS, a "double-stuffed" mailing packet was mailed to approximately half the addresses flagged as having a Hispanic surname, and approximately 300 additional addresses as a control. The double-stuffed mailing was to include one questionnaire printed in English and one in Spanish. Due to a printing error, the addresses assigned to the double-stuff treatment received two mailings: the first mailing contained only a Spanish-language questionnaire with some pages missing, and the second mailing contained the correctly printed packet with both questionnaires.

To avoid issues of confounding between the factors, the incentive, targeted mailing, and double-stuff treatments were assigned in an approximately balanced manner (Table 2.2). Taken as a whole the experimental design of the experiments embedded in the 2019 TCHS form a split-plot design, with incentive, being assigned at the census tract level, as the "whole-plot" factor, and the targeted mailing and double-stuff, being assigned at the address level, as the "split-plot" factors. In the remainder of this chapter we will focus solely on the incentive experiment. Since the census tracts were divided into blocks before assigning treatments, the design for the incentive experiment alone is a randomized complete block design.

2.2.2 Statistical Methods

To reflect the sampling design used in the survey, addresses were assigned weights inverse to the probability of being sampled. Analyses in this chapter using these weights will be referred to as "design-weighted" and do not use additional corrections for nonresponse. Weights were computed by dividing the number of addresses in each stratum by the number of responses in each stratum. For weighting purposes the strata were separated by inside/outside of Fort Collins to account for avoiding addresses sampled for another survey being conducted simultaneously by another organization. Additional weighting to adjust for nonresponse was conducted for analyzing the responses to this survey. As we focus on response behavior as opposed to responses to the survey, the additional weighting was not used in the analyses in this paper and is not further discussed here. All statistical analyses were conducted in R version 3.6.3 (R Core Team [2020]).

Table 2.2: Treatments assignment of eligible addresses in the 2019 Larimer County TCHS overall and by the auxiliary variables Hispanic Surname Flag, Tenure, Dwelling Type, and Income. NG means "not given", S means "single-family", M means "multi-family", and PO means "POBOX only way to get mail."

Treatment Group	Incentive	Target Mailing	Double Stuff	Number Assigned	Hispanic Surname Flag		Tenure			Dwelling Type			Income	
					Yes	No	Rent	Own	NG	S	M	PO	Given	NG
1	\$2pre	No	No	4061	138	3923	351	3253	457	3460	601	0	3863	198
2	\$2pre	No	Yes	240	125	115	35	178	27	168	72	0	237	3
3	\$2pre	Yes	No	3145	156	2989	880	1218	1047	1726	1419	0	2661	484
4	\$2pre	Yes	Yes	234	148	86	62	109	63	122	112	0	217	17
Subtotal	\$2pre			7680	567	7113	1328	4758	1594	5476	2204	0	6978	702
5	\$1+\$5web	No	No	2181	58	2123	155	1738	288	1879	289	13	2031	150
6	\$1+\$5web	No	Yes	99	46	53	16	66	17	78	21	0	94	5
7	\$1+\$5web	Yes	No	1544	54	1490	427	543	574	808	718	18	1253	291
8	\$1+\$5web	Yes	Yes	94	50	44	25	45	24	46	47	1	87	7
Subtotal	\$1+\$5web			3918	208	3710	623	2392	903	2811	1075	32	3465	453
Total				11598	775	10823	1951	7150	2497	8287	3279	32	10443	1155

Survey-weighted analyses were conducted to examine differences in RRs, response modes, survey costs, and representativeness between the groups receiving each of the survey incentives.

To study RRs, subgroup design-weighted RRs were calculated for categories defined by the auxiliary variables purchased for this study. Hypothesis testing was performed using a randomization distribution of the treatment assignments conditioned on the sample (Chapters 4 through 6). To examine trends of RRs throughout the survey fielding period, design-weighted cumulative response curves were used to visualize the differences in response behavior for all protocols used in this survey. A design-weighted curve showing the difference in cumulative RRs between the incentive treatments is shown with curves in the background drawn from the randomization distribution of the treatment assignment to provide a sense of scale for the differences. A survey-weighted log-rank test (Rader [2014]) was also used to analyze the difference between these response curves, with inference based on the randomization distribution of the treatment assignment conditioned on the sample though a Monte Carlo test (Section 6.2).

To account for potential covariates, a design-weighted logistic regression model was used as well. The effects of the incentive structure on survey response (encoded as the binary variable 1 = responded, 0 = did not respond) and on mode of response (encoded as 1 = web response, 0 = mail, given that the address responded) were studied with design-weighted mixed-effects logistic regression models. These models included additive fixed effects to account for incentive type and to control for demographic characteristics, using the auxiliary variables housing tenure (own/rent/not given), dwelling type (single-family, multi-family, or OWTGM PO box: post office box as only way to get mail), income (given/not given) and Hispanic surname. Income was coded by given or not given because the auxiliary income variable did not correspond well to the reported income from the survey among the respondents. Random effects were used to account for the tract-level treatment assignment. Further, logistic regression models including interactions between incentive structure and each of the demographic variables were also fit for both response and web response. Inference was based on the randomization distribution conditioned on the sample. Confidence

intervals were calculated using a Robins-Munro algorithm as described in Garthwaite [1996]. The tests for alternative hypotheses were modified from Gail et al. [1988].

To analyze cost-effectiveness, design-weighted estimates of incentive costs per response associated with both incentive treatments were computed overall and within demographic groups identified using the auxiliary variables. This analysis involved comparing design-weighted costs of the incentive and associated mailing costs divided by the design-weighted estimate of the number of responses in each incentive treatment, which can be analyzed using the linearization discussed in Section 6.4 applied to a ratio. The analysis for cost was assuming that the cost per response in each census tract will be the same under both treatments, and thus only the cost under the assigned treatment was considered. The randomization test, as suggested by the theory (Chapter 5), only altered the labels under different randomizations, using the cost estimates under the original treatment. This analysis of costs excludes costs associated with later waves of the survey. See Section 2.4 for further discussion of technical details in analyzing cost per response.

To analyze the representativeness of the responses, the RRs and the proportion of online responses were examined for each treatment within several demographic categories identified from the auxiliary variables. Analyses on interaction effects in the logistic regression were also used to investigate this question. Lastly, we compared the survey-weighted differences in demographic variable between the responding addresses of the 2019 TCHS and the 2018 American Communities Survey (ACS) 5-year estimates by each of incentive and response mode using the dissimilarity index (DI, see also Biemer et al. [2018]). The DI is calculated for a categorical variable with G categories by $\hat{D} = 1/2 \sum_{g=1}^{G} |\hat{p}_g - \alpha_g|$, where $\hat{p}_g$ is the estimated proportion within category g and α_g is the proportion of the category from the 2018 ACS 5-year estimates. The DI can take values in the range [0,1), and is described further in Chapter 3). A linearization was applied to estimate standard errors for differences in the DI by incentive treatment using the randomization distribution of the treatment assignment conditioned on the sample (Section 6.4). Variation in the DI by response mode was quantified using a linearization approach for sampling distribution of the

survey (Chapter 3). This is because the only randomization in the response mode analysis was the sampling mechanism.

2.3 Results

2.3.1 Response Rate and Mode

To consider differences in response rate between the treatments, we looked at overall response rate, differences in response rate across groups, and used a design-based regression model to determine differences accounting for subgroups. Additionally, we examined response curves to determine differences over time between treatments.

We compared RRs between addresses in census tracts receiving the $2pre and $1+$5web incentives according to auxiliary demographic variables purchased for the addresses: Hispanic surname flag, housing tenure (coded as rent, own, not given), dwelling type (coded as single-family, multi-family, PO Box), and income (coded as given, not given). In each of these comparisons, the empirical RR is higher for the $1+$5web group than the $2pre group (Table 2.3).

Overall, there was a slightly higher RR from the $1+$5web incentive (24.8%) than the $2pre incentive ($22.2\%$, standard error of difference [SE(diff)] $= 1.4\%$, $p = 0.029$). Within each of the subgroups, there was a slightly higher RR to the $1+$5web incentive versus the $2pre incentive. The differences appeared to be larger among groups from which less information was known in the auxiliary variables. When tenure was not given, the RR was 3.3 percentage points higher for addresses receiving the $1+$5web incentive than the $2pre incentive ($2pre RR $= 14.2\%$, $1+$5pre RR $= 17.5\%$, SE(diff) = 1.5%, $p = 0.016$). When income was not given, the RR was 3.5 percentage points higher for addresses receiving the $1+$5web incentive than the $2pre incentive ($2pre RR $= 10.8\%$, $1+$5pre RR $= 14.4\%$, SE(diff) $= 2.0\%$, $p = 0.037$).

Table 2.3: Total numbers of eligible addresses, responses, response rates, and web response rates (conditional on response) by incentive treatment overall and by categories of auxiliary variables. Response rates and web response rates are weighted to reflect sampling design of the TCHS. Inference was based on the asymptotic randomization distribution of the treatment assignment conditioned on the sample. Significance tests were one-sided to test if the $1+$5web incentive improved response or web response compared to the $2pre incentive.

Group	Incentive	Tot	Resp	Resp Rate	Diff *SE*	z p	Web Resp	Web Rate	Diff *SE*	z p
Overall										
	$2pre	7680	1666	0.222	0.026	1.913	1234	0.743	0.025	1.400
	$1+$5web	3918	966	0.248	*0.014*	*0.028*	741	0.768	*0.018*	*0.081*
Hispanic Surname Flag										
Yes	$2pre	567	69	0.130	0.010	0.383	58	0.863	-0.106	-1.204
	$1+$5web	208	29	0.140	*0.025*	*0.351*	22	0.757	*0.088*	*0.886*
No	$2pre	7113	1597	0.228	0.027	1.853	1176	0.739	0.030	1.729
	$1+$5web	3710	937	0.254	*0.014*	*0.032*	719	0.769	*0.017*	*0.042*
Housing Tenure										
Own	$2pre	4758	1146	0.267	0.033	2.116	922	0.741	0.009	0.437
	$1+$5web	2392	713	0.300	*0.016*	*0.017*	534	0.750	*0.020*	*0.331*
Rent	$2pre	1328	197	0.152	0.001	0.071	150	0.768	0.054	1.142
	$1+$5web	623	96	0.154	*0.019*	*0.472*	78	0.822	*0.048*	*0.127*
Not Given	$2pre	1594	223	0.142	0.033	2.137	162	0.732	0.087	1.660
	$1+$5web	903	157	0.175	*0.015*	*0.016*	129	0.819	*0.052*	*0.048*
Dwelling Type										
Multi Family	$2pre	2204	310	0.142	0.031	1.859	227	0.739	0.096	2.465
	$1+$5web	1075	185	0.173	*0.017*	*0.032*	153	0.835	*0.039*	*0.007*
Single Family	$2pre	5476	1356	0.251	0.026	1.715	1007	0.744	0.009	0.451
	$1+$5web	2811	773	0.277	*0.015*	*0.043*	582	0.753	*0.020*	*0.326*
POBOX	$2pre	0	0				0			
	$1+$5web	32	8	0.25			6	0.75		
Income										
Given	$2pre	6978	1593	0.233	0.029	2.024	1180	0.743	0.019	1.026
	$1+$5web	3465	901	0.262	*0.014*	*0.021*	685	0.762	*0.018*	*0.152*
Not Given	$2pre	702	73	0.108	0.035	1.787	54	0.746	0.114	1.658
	$1+$5web	453	65	0.144	*0.020*	*0.037*	56	0.860	*0.069*	*0.049*

Among addresses not receiving the double-stuff mailing, those receiving the $1+$5web incentive responded at a higher rate than addresses receiving the $2pre incentive. Among addresses that did receive the double-stuff mailing, however, there was a higher RR among the $2pre incentive group than the $1+$5web group (Figure 2.2). Formal testing reveals evidence at the

$\alpha = 0.05$ level of higher response in the \$1+\$5web group (RR $= 28.1\%$) than the \$2pre group (RR $= 24.9\%$) among addresses not receiving either the double-stuff or target mailing treatments (SE(diff) $= 1.6\%$, $p = 0.025$). There is also evidence at the $\alpha = 0.1$ level of higher response in the \$1+\$5web group (RR $= 20.2\%$) than the \$2pre group (RR $= 18.2\%$) among addresses receiving the targeted mailing but not the double-stuff (SE(diff) $= 1.3\%$, $p = 0.057$). There was not evidence of differences in response rates between incentives in groups that received the double-stuff mailing (received targeted mailing $p = 0.535$, did not receive targeted mailing $p = 0.759$).

When only observing the incentive group, there appears to be evidence of higher RRs overall among addresses receiving the \$1+\$5web treatment than the \$2pre treatment (Figure 2.3). Between days 30 and 40 the \$2pre treatment response seems to surpass the RR in the \$1+\$5web treatment for a few days, but this is an artifact of the data collection that should be disregarded by viewers of the graph. This artifact appears because the mailing of the \$2pre treatment was delayed by four days due to issues obtaining two-dollar bills for the \$2pre treatment. When adjusted so that the timelines for both treatments start with day 0 being the first mailing, the treatment effects appear four days earlier in the \$2pre treatment.

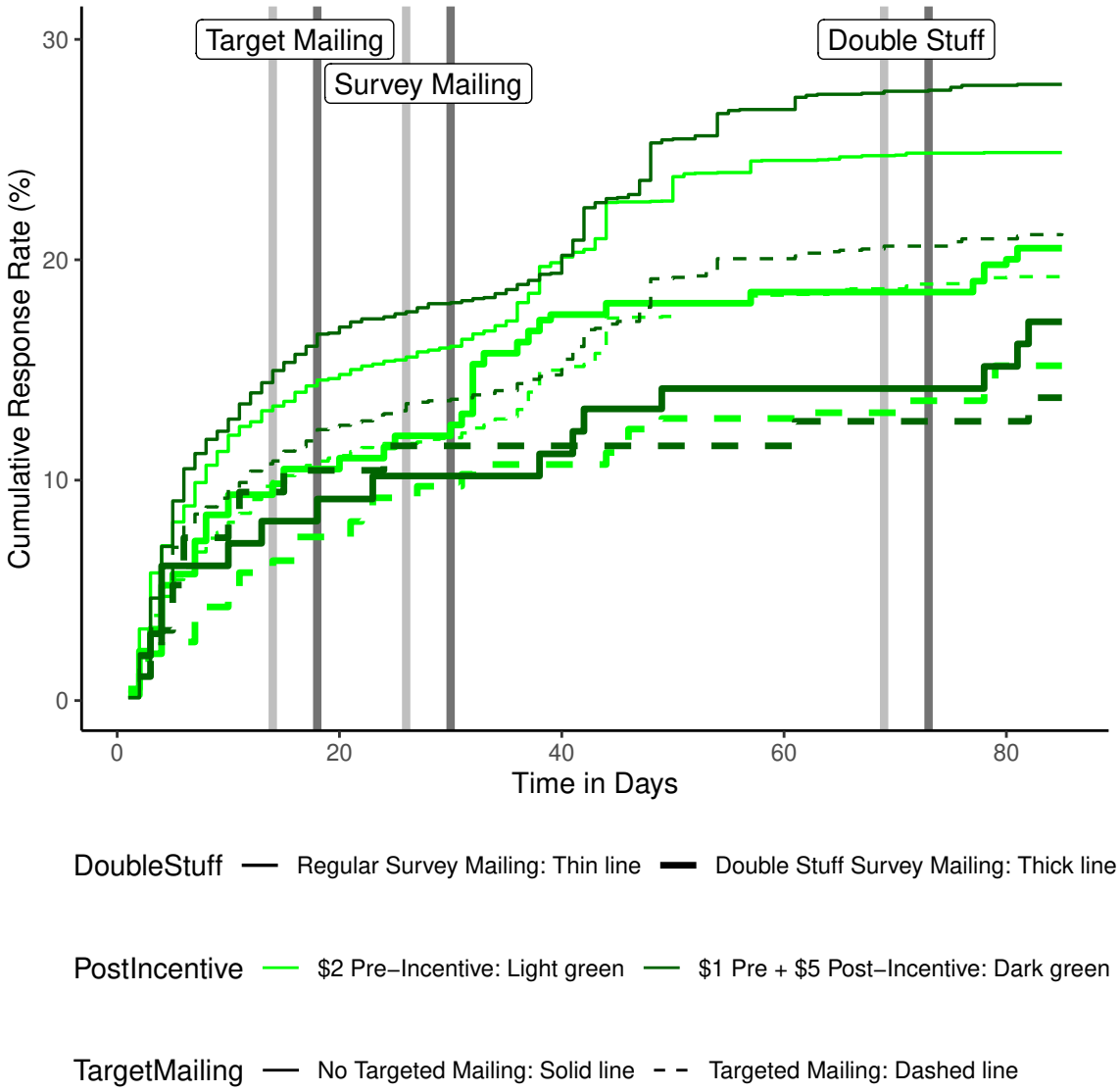

Figure 2.2: Response curves for all different treatments in the survey experiment. The lighter grey vertical lines indicate when the $2pre group received additional intervention, and the darker grey vertical lines indicate when the $1+$5web group received additional intervention, as noted in the text box (target mailing, survey mailing or double stuff).

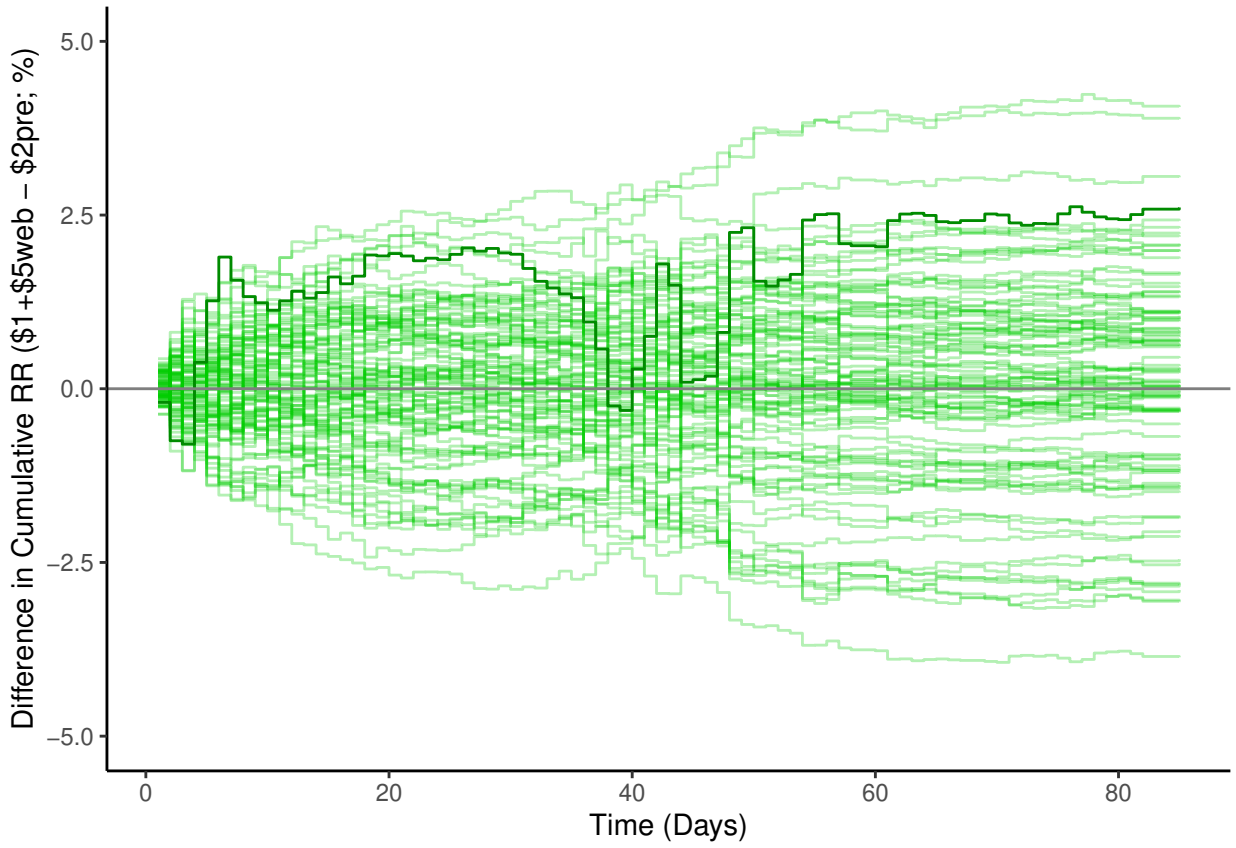

Figure 2.3: Response curves for the difference between treatment and control for the incentive experiment. The actual curve is in darker green, and curves from other simulations are in lighter green. The apparent dips in the actual curve are an artifact of the $2 incentives being mailed 4 days late.

The randomization-based survey-weighted log-rank test provided evidence at the $\alpha = 0.1$ level that the $1+$5web incentive may have more and/or faster response to the survey than the $2pre incentive ($\chi^2 = 7.854$, $p = 0.068$).

The randomization-based logistic regression analysis did not show evidence of interactions between the demographic variables and the incentive treatment ($F_{5,15} = 1.168$, $p = 0.369$; Table 2.4, Model 2), therefore we proceeded with inference using the model without interactions. The logistic regression analysis for incentive structure provided evidence at the $\alpha = 0.05$ level that addresses that received the $1+$5web incentive responded at a higher rate than the addresses that received the $2pre incentive (Odds ratio (OR) $= 1.183$, 95% confidence interval (CI) $= (1.011, 1.384)$, $p = 0.037$; Table 2.4, Model 1).

There was evidence at the $\alpha = 0.1$ level that there may be interaction between the auxiliary variables and the incentive treatment ($F_{5,15} = 2.820$, $p = 0.054$; Table 2.5, Model 2), suggesting that odds ratios of response mode by incentive may be different for members of some demographic groups represented in the auxiliary variables than for members of other demographic groups. When examining interactions individually, there was evidence at the $\alpha = 0.1$ level that the push-to-web impact of receiving the \$1+\$5web incentive is less for Hispanics than for non-Hispanics (OR = 0.365, 95% CI $= (0.115, 1.155)$, $p = 0.083$). There was no evidence found for interactions among other groups identified with the auxiliary variables (min $p = 0.326$). From the model without interactions (Model 1), there was not evidence of differences in response mode by incentive treatment (OR $= 1.295$, 95% CI $= (0.927, 1.810)$, $p = 0.122$; Table 2.5, Model 1). There was evidence at the $\alpha = 0.1$ level that addresses flagged as Hispanic are more likely to respond by web than addresses not flagged as Hispanic (OR $= 1.639$, 95% CI $= (0.974, 2.758)$, $p = 0.062$), but there is not evidence of differences in response mode by any of the demographic variables (smallest $p = 0.445$).

Table 2.4: Estimated model coefficients, standard errors (SE), and p-values (for testing the hypothesis that a single model coefficient is zero) for the mixed-effect logistic regression models to examine the effect of the incentive on survey response (1 = responded, 0 = did not respond). Coefficients are estimated using weighting to reflect the true design of the survey and the standard error are estimated using the delete-a-group jackknife allowing for small strata [Kott, 2001] with 20 jackknife replicates, and the tests for each of the model coefficients was a t-test with 19 degrees of freedom.

	Model 1			Model 2		
Coefficient	B	SE	p	B	SE	p
Intercept	-1.286	0.044	<0.001	-1.355	0.064	<0.001
Incentive - $1+$5web	0.168	0.075	0.037	0.350	0.148	0.029
Tenure - Rent	-0.620	0.081	<0.001	-0.568	0.116	<0.001
Tenure - Not Given	-0.430	0.075	<0.001	-0.440	0.071	<0.001
Dwelling Type - Single	0.329	0.043	<0.001	0.407	0.073	<0.001
Dwelling Type - PO Box	0.582	0.507	0.266	0.458	0.531	0.399
Income - Not Given	-0.391	0.116	0.003	-0.413	0.107	0.001
Hispanic Surname Flag	-0.606	0.135	<0.001	-0.541	0.160	0.003
($1+$5web) x Tenure Rent				-0.154	0.216	0.484
($1+$5web) x Tenure Not Given				0.019	0.161	0.907
($1+$5web) x Single				-0.204	0.140	0.162
($1+$5web) x PO Box				NA*	NA*	NA*
($1+$5web) x Income Not Given				0.026	0.203	0.899
($1+$5web) x Hispanic Flag				-0.181	0.289	0.538
Tract Standard Deviation	0.280	0.025		0.280	0.025	
Interactions Test	$F_{5,15} = 1.168, p = 0.369$					

In summary, there was evidence of a slightly higher response rate for addresses assigned the $1+$5web incentive than for addresses assigned the $2 incentive. We observed higher response rates in all groups based on auxiliary variables, but the differences may have been slightly higher in groups from which less auxiliary information was known. Looking at response rates over time, we see that the difference in response between the incentive treatments started early, and was maintained throughout the fielding period. There was some evidence from analyzing response mode that Hispanics may be more likely to respond by web and are less affected (or perhaps negatively affected) by push-to-web incentives, but more research would need to be done to confirm these results.

Table 2.5: Estimated model coefficients, standard errors (SE), and p-values (for testing the hypothesis that a single model coefficient is zero) for the mixed-effect logistic regression models to examine the effect of the incentive on mode of response (1 = web, 0 = mail), conditioned on responding. Coefficients are estimated using weighting to reflect the true design of the survey and the standard error are estimated using the delete-a-group jackknife allowing for small strata [Kott, 2001] with 20 jackknife replicates, and the tests for each of the model coefficients was a t-test with 19 degrees of freedom.

Coefficient	Model 1			Model 2		
	B	SE	p	B	SE	p
Intercept	0.934	0.188	<0.001	0.852	0.195	<0.001
Incentive - $1+$5web	0.259	0.160	0.122	0.452	0.325	0.181
Tenure - Rent	0.120	0.172	0.493	0.067	0.160	0.682
Tenure - Not Given	0.052	0.134	0.703	-0.056	0.173	0.749
Dwelling Type - Single	0.142	0.183	0.445	0.270	0.216	0.226
Dwelling Type - PO Box	-0.390	1.000	0.701	-0.765	1.092	0.492
Income - Not Given	0.210	0.332	0.535	0.040	0.354	0.912
Hispanic Surname Flag	0.494	0.249	0.062	0.838	0.358	0.030
($1+$5web) x Tenure Rent				0.170	0.262	0.523
($1+$5web) x Tenure Not Given				0.298	0.414	0.481
($1+$5web) x Single				-0.312	0.309	0.326
($1+$5web) x PO Box				NA*	NA*	NA*
($1+$5web) x Income Not Given				0.357	0.591	0.553
($1+$5web) x Hispanic Flag				-1.009	0.551	0.083
Tract Standard Deviation	0.485	0.094		0.484	0.093	
Interactions Test	$F_{5,15} = 2.820, p = 0.054$					

2.3.2 Survey Costs

Overall, estimated incentive cost per response were similar for the $1+$5web incentive than for the $2pre incentive ($2pre: $\$9.01$, $1+$5web: $\$8.80$; SE(diff) $= 0.45$, $p = 0.644$; Table 2.6). Breaking down the costs by the auxiliary variables, we find $1+$5web incentive requires a similar incentive cost per response to the $2pre incentive for higher-response groups but cheaper for some lower-response groups, namely addresses classified as multi-family housing units, without income given, or without tenure given. For example, giving the incentive to addresses listed as owner-occupied cost $\$7.49$ per response when given the $2pre incentive and $\$7.99$ per response with the $1+$5web incentive (SE(diff) $= 0.35$, $p = 0.151$), providing the incentive to addresses listed as renter-occupied cost $\$13.13$ per response with the $2pre incentive and $\$11.61$ per response with the $1+$5web incentive (SE(diff) $= 1.48$, $p = 0.305$), and providing the incentive to addresses with-

out a tenure given cost $\$14.09$ with the \$2pre incentive and $\$10.80$ with the \$1+\$5web incentive (SE(diff) $= 1.23$, $p = 0.007$).

In summary, the incentive cost per response from the \$1+\$5web incentive is similar to the \$2pre incentive only for addresses that are more likely to respond, but lower than the \$2pre incentive for some addresses that are less likely to respond, such as addresses classified as multi-family housing units, without income given, or without tenure given.

Table 2.6: Incentive cost per response (in dollars) of different incentives overall and for subgroups identified by the auxiliary variables. The costs only include the cost of the incentive and mailing the \$5 incentive to each of the web-respondents. Costs not included include downstream cost differences due to earlier response times and potential cost savings of not needing to process as many paper surveys. Significance tests are based on the randomization distribution of treatment assignment and are two-sided to identify any difference in costs.

Group	\$2pre cost	\$1+\$5web cost	Difference cost (SE)	z	p
Overall	9.01	8.80	-0.21 (0.45)	-0.811	0.417
Hispanic Flag					
Yes	15.30	11.82	-3.48 (2.46)	-1.413	0.158
No	8.79	8.70	-0.09 (0.45)	-0.189	0.850
Income					
Given	8.58	8.55	-0.03 (0.41)	-0.067	0.946
Not Given	18.50	12.31	-6.19 (2.44)	-2.537	0.011
Dwelling Type					
Multi-Family	14.04	10.96	-3.08 (1.39)	-2.207	0.027
Single-Family	7.97	8.29	0.32 (0.40)	0.795	0.426
POBOX					
Tenure					
Own	7.49	7.99	0.50 (0.35)	1.437	0.151
Rent	13.13	11.61	-1.52 (1.48)	-1.026	0.305
Not Given	14.09	10.80	-3.29 (1.23)	-2.680	0.007

2.3.3 Representativeness of Response

In this section, we review and quantify differences in what types of people were represented in the respondents when given a \$2pre incentive compared to when given a \$1+\$5web incentive. We first review the results from the logistic regression analyses discussed in Section 2.3.1 from

the lens of representativeness. Additionally, we consider evidence from comparing differences in deviation from American Communities Survey (ACS) estimates, as measured by the dissimilarity index (DI), between the treatments.

From the logistic regression analyses, we found evidence that minorities and people of lower socioeconomic status are underrepresented in the respondents to the TCHS. Renter-occupied addresses (OR $= 0.538$, 95% CI $= (0.454, 0.638)$, $p < 0.001$; Table 2.4, Model 1) or addresses whose tenure was not given (OR $= 0.651$, 95% CI $= (0.556, 0.761)$, $p < 0.001$) were less likely to respond than owner-occupied addresses. Single-family addresses (OR $= 1.389$, 95% CI $= (1.269, 1.521)$, $p < 0.001$) were more likely to respond than multi-family addresses. Residents of addresses whose income was not provided (OR $= 0.676$, 95% CI $= (0.530, 0.862)$, $p = 0.003$) were less likely to respond than residents whose income was provided. Residents of addresses flagged as Hispanic (OR $= 0.546$, 95% CI $= (0.411, 0.724)$, $p < 0.001$) were less likely to respond than residents of addresses not flagged as Hispanic.

For both treatment groups, the proportion of survey respondents who responded by web was over 70% for any subgroup examined based on auxiliary variables. Somewhat surprisingly, among addresses flagged as having a Hispanic surname, addresses receiving the $2pre incentive were more likely to respond online (86.3%) than addresses receiving the $1+$5web incentive (75.7%; SE(diff) $= 8.8\%$, $p = 0.886$; Table 2.3), though there is not evidence to conclude a difference for the subpopulation of flagged Hispanic addresses. Among addresses not flagged as having a Hispanic surname, there was evidence that addresses receiving the $2pre incentive were less likely to respond online (73.9%) than addresses receiving the $1+$5web incentive (76.9%, SE(diff) = 1.7%, $p = 0.042$). For the other auxiliary variables, we found higher web RRs for the $1+$5web incentive than the $2pre incentive, and the differences appear to be greater in the categories with lower RRs.

The interactions between the incentive type and the auxiliary variables in the logistic regression models, as previously noted, were not found to be significant (Table 2.4).

Between the two incentive treatments, we did not find evidence of differences in deviation from proportions from the ACS the DIs in any of the variables that we measured (min $p = 0.404$; Table 2.7). This is consistent with Table 2.1, where the demographics look similar between the two treatment columns. However, we did see differences when we examined responses by response mode. The dissimilarity index is lower for online respondents than paper respondents for age (Paper $= 0.35$, Online $= 0.25$, SE(diff) $= 0.02$, $p < 0.001$), sex (Paper $= 0.19$, Online $= 0.10$, SE(diff) $= 0.02$, $p < 0.001$), race (Paper $= 0.04$, Online $= 0.02$, SE(diff) $= 0.01$, $p = 0.007$) and Hispanic ethnicity (Paper $= 0.35$, Online $= 0.25$, SE(diff) $= 0.02$, p = 0.035). We found evidence that the dissimilarity index was higher for online respondents than paper respondents for education (Paper $= 0.18$, Online $= 0.26$, SE(diff) $= 0.02$, p = 0.002; Table 2.8) and income (Paper $= 0.03$, Online $= 0.08$, SE(diff) $= 0.02$, $p = 0.014$). Closer examination into the demographics reveal that for age, online respondents tend to be younger, more male, less White and more Hispanic than paper respondents, making them closer to the general population in those categories (Table 2.1). We also see that paper respondents include more people without college experience, fewer people with Bachelor's degrees, and more people with incomes less than $25,000 than online respondents, making paper respondents more representative of the population in those categories.

Table 2.7: Dissimilarity indices by treatment and differences of the dissimilarity index between treatments with standard errors (SE(diff)), test statistics (z) and p-values (p) for a test of a difference between treatments for age, sex, race, Hispanic ethnicity, education, health insurance, and income. Hypothesis test is two-sided and standard errors are calculated using the randomization distribution of the treatment assignment.

Variable	$2pre	$1+$5web	Difference	SE(diff)	z	p
Age	0.28	0.27	-0.01	0.03	-0.350	0.727
Sex	0.12	0.12	0.00	0.02	0.121	0.904
Race	0.02	0.02	-0.00	0.01	-0.381	0.703
Hispanic	0.04	0.04	-0.01	0.01	-0.834	0.404
Education	0.24	0.25	0.01	0.03	0.448	0.654
Health Insurance	0.03	0.04	0.01	0.01	0.829	0.407
Income	0.07	0.06	-0.01	0.03	-0.183	0.855

Table 2.8: Dissimilarity indices by response mode and differences of the dissimilarity index between response modes with standard errors (SE(diff)), test statistics (z) and p-values (p) for a test of a difference between response mode for age, sex, race, Hispanic ethnicity, education, health insurance, and income. Hypothesis test is two-sided and standard errors are calculated using the sampling distribution for the survey design.

Variable	Paper	Online	Difference	SE(diff)	z	p
Age	0.35	0.25	-0.10	0.02	-4.547	<0.001
Sex	0.19	0.10	-0.09	0.02	-4.002	<0.001
Race	0.04	0.02	-0.02	0.01	-2.699	0.007
Hispanic	0.06	0.04	-0.02	0.01	-2.105	0.035
Education	0.18	0.26	0.07	0.02	3.108	0.002
Health Insurance	0.03	0.03	-0.00	0.01	-0.085	0.933
Income	0.03	0.08	0.06	0.02	2.454	0.014

2.4 Discussion

In this chapter, we discussed the analysis of the incentive experiment in a web-mail mixed-mode survey with six waves of data collection and treatments assigned before the start of data collection. Embedded in the survey were three experiments: an incentive treatment ($2pre or $1+$5web) assigned at the census tract level at the beginning of the survey, a targeted mailing sent to some of the addresses thought to be less likely to respond at wave 3, and a double-stuffed survey packet (including English-language and Spanish-language surveys) sent out with the survey mailing at wave 4. We focused our analysis on the incentive experiment from the beginning of the survey with the goal of evaluating whether the $1+$5pre incentive improved RRs and changed the cost-per-response while maintaining the representativeness of the $2pre design.

We found that in the 2019 Larimer County TCHS, addresses receiving the $1+$5web incentive responded at higher rates relative to addresses receiving the $2pre incentive. The increased response is seen in all subgroups of the population, as identified by the auxiliary variables. We found no evidence of different incentive costs per response overall for the $1+$5web incentive compared to the $2pre incentive, but found some evidence that the $1+$5web incentive provided more responses per dollar of incentive given than the $2pre incentive for addresses that were multi-family

housing units, did not have an income listed, or did not have the tenure listed. There was also no evidence that the representativeness of response, measured as a deviation from ACS statistics using the DI, was different between the incentive structures.

This suggests that overall the $1+5web incentive is more effective than the $2pre incentive for the adult population in Larimer County, Colorado, because it resulted in slightly higher response rates, and did not substantially change the demographic makeup of who responded. However, there were some changes in the mode by which people responded on this setup. Specifically, online respondents were more representative of the population in terms of age, sex, race and ethnicity, and paper respondents were more representative of the population in terms of income and education. We also note that measures of response quality were not taken into account for this study, but would be useful for decided what type of incentive to use.

For the analysis of incentive costs per response, we did not include costs related to later waves of the survey and processing mail responses. This is because we did not have precise cost information on how early response affected costs of later waves of the survey, including information on how quickly respondents were removed from the mailing lists and the costs of processing paper surveys vs. online surveys.

We had initially considered an analysis of incentive costs that examined the costs by changing the label to reflect the cost for the treatment it was randomly relabeled as in the randomization test. This method, however, does not lead to a randomization distribution with mean zero. It is possible to adjust the distribution to have mean zero, but the form of the test in that case has a more complicated interpretation, and does not seem to analyze the question of cost per response. For this reason, we decided to only change the labels and not change costs. If the costs per response are not different between treatments, then a randomization that only changes the labels will not change the mean by a large amount when the labels are changed. This method is also consistent with how other variables are measured.

As mentioned above, the analysis did not consider the downstream costs from the incentives. However, we can make some guesses based on online responses, and when people who received

each of the incentive structures responded. The rate of online response was similar for the two incentive treatments but was slightly higher in the $1+$5web incentive group than the $2pre group. There were also more early responses in the $1+$5web group than in the $2pre group. Both observations suggest that the $1+$5web treatment may be even more cost-effective relative to the $2pre treatment than reported.

The similarity of the dissimilarity indices between the two survey treatments means that we found no evidence that the respondents differed on measurable demographic characteristics based on the incentives they received. This confirms the result of not having significant interaction effects in the model evaluating response by treatment. This information, combined with the information that there may be cost savings (per response) for at least some subgroups and that the RRs were higher for the $1+$5web incentive than with the $2pre incentive, suggests that, at least for subgroups thought to be less likely to respond, the $1+$5web incentive may be better than the $2pre incentive, at least for Larimer County.

Comparing our results with previous literature, we find that our work aligns with some of the previous literature. Agreeing most closely with Biemer et al. [2018], our results suggest that in Larimer County, Colorado, an incentive to respond online can result in slightly higher RRs. We did not find evidence of higher web RRs with the $1+$5web incentive; however, in this survey, the only way to respond at the start of fielding was by web. Web RRs have consistently been found to be higher with a web-first design than a mail-first or concurrent design (Messer and Dillman [2011], Millar and Dillman [2011], Patrick et al. [2018], Biemer et al. [2018]), so it is possible that the web first design may have been more successful at pushing people to web than the incentive.

When thinking about the differences in the dissimilarity index on web response, we noticed more balance in web response on the variables of sex and age. One interesting finding is that we found better representation of Hispanic ethnicity and race, but less representation among education level and income among web respondents. This could be an incidental finding, but it would be interesting to see if similar patterns hold in other studies.

A key issue with analyzing these data were that not many methodologies had been developed for analyzing embedded experiments in surveys. We developed new methodologies to analyze these experiments. One possible approach to analyze these data that was available in the literature was the approach proposed by van den Brakel [2001]. Van den Brakel's approach will work for many of the analyses used in the 2019 TCHs, including comparison of response rates, incentive cost per response, and the dissimilarity index. Van den Brakel's approach, however, does not allow for studying statistics, such as the log-rank test, that do not have an asymptotic normal distribution. It also does not allow for plots like those in Figure 2.3, which show the distribution of different samples. In order to allow researchers to make comparisons based on arbitrarily complicated statistics derived from experiments embedded in surveys, new methodology was needed, which is presented in Chapters 4–6 of this dissertation.

Chapter 3

Survey Design-Based Inference for the Dissimilarity Index

3.1 Introduction

The dissimilarity index (DI) was originally proposed by Jahn et al. [1947], and brought into the literature with major early contributions by Williams [1948] and Duncan and Duncan [1955]. The original problem was to measure segregation between Black and White Americans in cities by totaling up differences between the Black and White populations within each city, but it has been generalized to measure the difference between the two row conditional distributions in any $2 \times G$ table, one at each level of a binary variable. The DI has been applied to surveys by Biemer et al. [2018] to evaluate the representativeness of response in the national pilot of the Energy Information Administration's Renewable Energy Consumption Survey.

In conjunction with the rise of adaptive and responsive survey designs, there has been a push to propose estimators to evaluate how well the respondents of surveys using such designs represent the target population of the survey ("representativeness"; Groves et al. [2008]). These and help researchers understand who is not responding to surveys to try to see where to target efforts to improve response. This development led to the R indicator (Schouten et al. [2009]), the fraction of missing information (Wagner [2010]), and several imbalance indicators proposed by Särndal [2011]. These have been discussed both as measures of response quality at the end of the survey and as ways to monitor and target responses (for example Lundquist and Särndal [2013], Wagner and Hubbard [2014]). These indicators are fairly diverse, and some of these indicators are based on propensity models (R indicator), or are specific to response variables (fraction of missing information), meaning that they can serve different purposes. The DI is a good addition to this toolkit as a simple indicator of representativeness for a categorical auxiliary variable.

When using a statistical measure, one natural question is how to describe the variability in that measure. Cortese et al. [1976] develop some theory for the distribution of the dissimilarity index under the case of no segregation based on the hypergeometric distribution. Some properties of the central limit theory for the DI were studied in more generality by Ransom [2000] and Allen et al. [2015]. Ransom derived asymptotics using a multinomial distribution for the number of persons with each characteristic in each group of interest (Ransom [2000]). Allen et al. also derived central limit theory for the dissimilarity index under a conditional multinomial model (Allen et al. [2015]). While Ransom considered the population distribution of all combinations of the binary variable and the categorical variable, Allen et al. considered the distributions of the categorical variable conditioned on the levels of the binary variable. While Ransom's model treated the binary variable as varying, Allen et al. treated the binary variable as a fixed characteristic.

While progress has been made on the distribution of the dissimilarity index in general, the survey context requires different derivations to obtain relevant reference distributions. We will consider this problem for examining the difference between the available sample and the general population in surveys. We will use the term "available sample" to indicate the subset sampled from the frame and/or for whom responses were collected. Using this term allows for discussion of problems of coverage, nonresponse, or a combination of both. We will also use the term "available population" to refer to the part of the population that is in the frame and/or will provide a response, considering a fixed response model. In other words, the available population is the part of the population from which information can be obtained from researchers, and the available sample is the part of the sample that provides such information. We will derive an expression for the DI allowing for linearization variance estimation for complex survey designs, under the hypothesis that the group membership probabilities for the sample differ from those in the population in all categories. To do this, we will first focus on the DI for the population, and then generalize to create measures for a DI for subgroups that one might want to compare. We begin by discussing the DI as it has been traditionally used and in the context of surveys in Section 3.2. The derivation and

assumptions for the DI variance calculations are discussed in Section 3.3. A simulation study is discussed in Section 3.4.

3.2 The Dissimilarity Index in the Survey Context

The DI was originally formulated to analyze segregation in American cities Jahn et al. [1947], and has historically been used to measure the degree of economic and social segregation in various contexts (see Ransom [2000] for some examples). In surveys, the DI is used to determine how representative a survey sample is compared to the general population. Specifically, in surveys, the DI examines the dissimilarity between the available population and a known reference to the finite population of interest. Because of some differences in these problems, the DI is expressed differently in the context of surveys than it is in the traditional social science context.

We now define notation and describe the dissimilarity index as it has been used in the traditional social science context, and how it has been used in surveys. Let $g = 1, \ldots, G$ be levels of a categorical variable (traditionally jobs or census tracts, in the survey context race or level of education, for example), and let $h = 0, 1$ be a state (traditionally white/non-white or male/female, in surveys respondents/nonrespondents or in-frame/out-of-frame). Further, let p_g^h be the proportion of individuals in category g out of all individuals having state h (or the conditional proportion of having category g given state h).

$$D = \frac{1}{2} \sum_{g=1}^{G} |p_g^1 - p_g^0|. \tag{3.1}$$

This formulation works when good information is available about individuals in both state $h = 0$ and $h = 1$. In the problem of survey representativeness, we do not have information on one of the states because they are unavailable, either due to not being in the frame or not responding. There are, however, often good references for the complete population of interest in survey problems. Therefore, for surveys, we use reference proportions $\alpha_1, \ldots, \alpha_g$ in calculations. If we let $h = 1$ denote the available population and $h = 0$ denote the unavailable population, then,

following Biemer et al. [2018], we write the version of the DI used in surveys as

$$D = \frac{1}{2}\sum_{g=1}^{G} |p_g^1 - \alpha_g|. \tag{3.2}$$

We note that these formulations (3.1) and (3.2) are not mathematically equivalent. However, if we use the fact that α_g is the proportion of the overall population of size N, it follows that if N_0 is the number of individuals with $h = 0$ and N_1 is the number of individuals with $h = 1$, then

$$\begin{aligned}\frac{1}{2}\sum_{g=1}^{G} |p_g^1 - \alpha_g| &= \frac{1}{2}\sum_{g=1}^{G} |p_g^1 - \frac{N_0}{N}p_g^0 - \frac{N_1}{N}p_g^1| \\ &= \frac{N_0}{N}\frac{1}{2}\sum_{g=1}^{G} |p_g^1 - p_g^0|.\end{aligned}$$

3.3 Survey Design-Based Variance Estimation

Let $U = \{1, 2, \ldots, N\}$ denote the finite population, $S \subset U$ denote the set of respondents, and $A_g \subset U$ denote the set of units having level g of a categorical variable of interest, where $U = \cup_{g=1}^{G} A_g$. Let the (known) distribution of the categorical variable in the population of interest be denoted by $\boldsymbol{\alpha}_N = (\alpha_{N,1}, \ldots, \alpha_{N,G})$, the (unknown) distribution in the available population as $\boldsymbol{p}_N = (p_{N,1}, \ldots, p_{N,G})$. The available population may be the population in the frame, or the population among respondents in the population, depending on the context. We denote the estimates of the distribution of the finite population from a sample of size $n = n(N)$ by $\hat{\boldsymbol{p}}_N = (\hat{p}_{N,1}, \ldots, \hat{p}_{N,G})$. We also let w_i denote sampling weights. We will drop the subscript N from the sample distributions in the sequel, unless it is relevant in context to view these as a sequence.

The DI is $D = \frac{1}{2}\sum_{g=1}^{G} |p_g - \alpha_g|$, which can be estimated by

$$\hat{D} = \frac{1}{2}\sum_{g=1}^{G} |\hat{p}_g - \alpha_g|.$$

To derive an expression that can be used for variance estimation, we rewrite the DI to remove the absolute value. First, $\hat{p}_g - \alpha_g = (\hat{p}_g - p_g) + (p_g - \alpha_g)$. From equation (2.9) in Breidt et al.

[2006], we see

$$\frac{1}{2}\sum_{g=1}^{G}|(\hat{p}_g - p_g) + (p_g - \alpha_g)| = \frac{1}{2}\sum_{g=1}^{G}\Big[\ |p_g - \alpha_g| + (\hat{p}_g - p_g)\text{sign}(p_g - \alpha_g) + 2(\hat{p}_g - \alpha_g)(I(0 < \alpha_g - p_g < \hat{p}_g - p_g) - I(\hat{p}_g - p_g < \alpha_g - p_g < 0))\ \Big],$$

where $I(B) = 1$ if the event B is true, and 0 otherwise.

In the expression above, let $e_g = (\hat{p}_g - \alpha_g)(I(0 < \alpha_g - p_g < \hat{p}_g - p_g) - I(\hat{p}_g - p_g < \alpha_g - p_g < 0))$, $e = \sum_{g=1}^{G} e_g$ and $R = \frac{1}{2}\sum_{g=1}^{G}(\hat{p}_g - p_g)\text{sign}(p_g - \alpha_g)$. This gives

$$\hat{D} = D + R + e.$$

where D is the population DI as defined above, e is a random expression with negligible variance contribution that converges in probability to 0, and R can be written as

$$\begin{aligned} R =& \sum_{g=1}^{G}\frac{1}{2}(\hat{p}_g - p_g)\text{sign}(p_g - \alpha_g) \\ =& \sum_{g=1}^{G}\frac{1}{2}\left(\frac{\sum_{i\in U} w_i I(i \in A_g)I(i \in S)}{\sum_{i\in U} w_i I(i \in S)} - p_g\right)\text{sign}(p_g - \alpha_g) \\ =& \frac{\sum_{i\in U} w_i I(i \in S)\left[\frac{1}{2}\sum_{g=1}^{G} I(i \in A_g)\text{sign}(p_g - \alpha_g)\right]}{\sum_{i\in U} w_i I(i \in S)} - \sum_{g=1}^{G}\frac{1}{2}p_g\text{sign}(p_g - \alpha_g), \end{aligned}$$

a ratio of two estimated population totals plus a constant. Hence, the variance can be approximated and estimated using standard procedures for an estimated ratio. To use this variance estimator in practice, we replace the term in square brackets in the numerator sum,

$$m_i = \frac{1}{2}\sum_{g=1}^{G} I(i \in A_g)\text{sign}(p_g - \alpha_g),$$

by its plug-in estimator,

$$\hat{m}_i = \frac{1}{2}\sum_{g=1}^{G} I(i \in A_g)\text{sign}(\hat{p}_g - \alpha_g),$$

giving the expression

$$\hat{R} = \frac{\sum_{i \in U} w_i I(i \in S) \hat{m}_i}{\sum_{i \in U} w_i I(i \in S)}.$$

The variance can now be estimated by `svymean` or `svyratio` in R (or similar commands in other survey software) by the estimated variance of the finite population mean of $\hat{m}_i$.

We now introduce assumptions to formalize the arguments about the asymptotic variance of the DI. Suppose that $\hat{\mathbf{\Gamma}}_N$ is the estimate of the design-based covariance matrix of $\hat{\boldsymbol{p}}_N$. Assume there is a sequence of populations and designs, indexed by N, such that the following assumptions hold as $N \to \infty$:

D1 $nVar(\hat{\boldsymbol{p}}_N) \to \mathbf{\Gamma}$, where $\mathbf{\Gamma}$ is a positive definite $G \times G$ matrix.

D2 $n^{1/2}(\hat{\boldsymbol{p}}_N - \boldsymbol{p}_N) \xrightarrow{d} N_G(\mathbf{0}, \mathbf{\Gamma})$.

D3 $\hat{\mathbf{\Gamma}}_N \xrightarrow{p} \mathbf{\Gamma}$.

D4 For each $g = 1, \ldots, G$, there is an $\epsilon > 0$ such that either $p_{N,g} - \alpha_{N,g} > \epsilon$ for all N or $\alpha_{N,g} - p_{N,g} > \epsilon$ for all N.

Assumption *D1* means that the normalized covariance matrix of the point estimates of the proportions approaches an asymptotic limit. Assumption *D2* says the point estimate for the proportions is asymptotically normal, and that the asymptotic variance is the limit of the covariance matrix. Assumption *D3* states that the estimated covariance matrix of the estimated proportions is consistent. Assumption *D4* states that the expected proportions from the available sample are bounded away from the true population proportions in every level of the categorical variables. This assumes away issues raised in Allen et al. [2015] that the asymptotic distribution of the DI will not be normal when $p_g - \alpha_g = O(n^{-1/2})$. We will consider this case later.

Note that $\hat{\boldsymbol{p}}_N - \boldsymbol{p}_N \xrightarrow{p} \mathbf{0}$, by Assumption *D2* above.

We will now show some results about the asymptotic distribution of R_N.

Theorem 3.3.1. *Let* $\boldsymbol{a} = sign(\boldsymbol{p}_N - \boldsymbol{\alpha}_N)$, $\hat{\boldsymbol{a}}_N = sign(\hat{\boldsymbol{p}}_N - \boldsymbol{\alpha}_N)$, $V_R = \boldsymbol{a}'\boldsymbol{\Gamma}\boldsymbol{a}$ *and* $\hat{V}_{R,N} = \hat{\boldsymbol{a}}_N'\hat{\boldsymbol{\Gamma}}_N\hat{\boldsymbol{a}}_N$ *(note* $V_R > 0$ *as* $\boldsymbol{\Gamma}$ *is positive definite and* $\boldsymbol{a} \neq \boldsymbol{0}$*). Then under Assumptions* D1–D4 *above,* $nVar(R_N) \to V_R$, $n^{1/2}R_N \xrightarrow{d} N(0, V_R)$, *and* $\hat{V}_{R,N} \xrightarrow{p} V_R$ *as* $N \to \infty$.

Proof. Since $R_N = \boldsymbol{a}'(\hat{\boldsymbol{p}}_N - \boldsymbol{p}_N)$, we have $nVar(R_N) = \boldsymbol{a}'nVar(\hat{\boldsymbol{p}}_N)\boldsymbol{a} \to \boldsymbol{a}'\boldsymbol{\Gamma}\boldsymbol{a} = V_R$ by Assumption *D1* and $n^{1/2}R_N \xrightarrow{d} N(0, V_R)$ by Assumption *D2*.

To show $\hat{V}_{R,N} \xrightarrow{p} V_R$ we need to show that $\hat{\boldsymbol{a}} \xrightarrow{p} \boldsymbol{a}$. This involves showing that $P(\text{sign}(\hat{p}_{N,g} - \alpha_{N,g}) \neq \text{sign}(p_{N,g} - \alpha_{N,g})) \to 0$ for each $g \in \{1, \ldots, G\}$.

To prove this, suppose we are given a g and without loss of generality $p_{N,g} - \alpha_{N,g} > \epsilon$ under Assumption *D4*. Then the events $\{\text{sign}(\hat{p}_{N,g} - \alpha_{N,g}) \neq \text{sign}(p_{N,g} - \alpha_{N,g})\}$ and $\{\hat{p}_{N,g} < \alpha_{N,g}\}$ are equivalent. Using Chebyshev's inequality, we find

$$P(\hat{p}_{N,g} < \alpha_{N,g}) \leq P(|\hat{p}_{N,g} - p_{N,g}| > \alpha_{N,g} - p_{N,g}) \leq \frac{Var(\hat{p}_{N,g})}{(\alpha_{N,g} - p_{N,g})^2}.$$

By Assumption *D1*, we know that $Var(\hat{p}_g) = O(n^{-1})$. As $(\alpha_{N,g} - p_{N,g})^2$ is bounded from below, we find $P(\hat{p}_{N,g} < \alpha_{N,g}) \to 0$ as desired.

Having shown $\hat{\boldsymbol{a}} \xrightarrow{p} \boldsymbol{a}$, an application of Slutsky's theorem and Assumption *D3* gives that $\hat{V}_R = \hat{\boldsymbol{a}}_N'\hat{\boldsymbol{\Gamma}}_N\hat{\boldsymbol{a}}_N \xrightarrow{p} \boldsymbol{a}'\boldsymbol{\Gamma}\boldsymbol{a} = V_R$.

□

Since $R = \boldsymbol{a}'\boldsymbol{p}$, the above theorem shows that the calculated variance of $\hat{R} = \hat{\boldsymbol{a}}'\hat{\boldsymbol{p}}$ from the `survey` package, $\frac{1}{4}\hat{\boldsymbol{a}}'\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{a}}$, is a consistent estimator for the sampling variance of $\hat{R}$ under Assumptions *D1–D4*.

We next show that the remainder term e_N is negligible under our regularity conditions.

Theorem 3.3.2. *Under Assumptions* D1–D4, $n^{1/2}e_g \xrightarrow{ms} 0$ *for each* g, *and thus* $n^{1/2}e_N \xrightarrow{ms} 0$ *as* $N \to \infty$.

Proof. Assume without loss of generality and using Assumption *D4* that $\tilde{\alpha}_g = \alpha_g - p_g > 0$ and define $Y_g = |Y_g| = n(\hat{p}_g - p_g)^2$, which is uniformly integrable by Assumptions *D1* and *D2* and

Lemma 1.4B (p. 15) of Serfling [1980]. Hence,

$$\begin{aligned}
\mathrm{E}[ne_g^2] &= \mathrm{E}\left[n(\hat{p}_g - \alpha_g)^2 I(\hat{p}_g > \alpha_g)\right] \\
&\leq \mathrm{E}\left[n(\hat{p}_g - p_g)^2 I(\hat{p}_g > \alpha_g)\right] \\
&= \mathrm{E}\left[Y_g I\left(n^{1/2}(\hat{p}_g - p_g) > n^{1/2}(\alpha_g - p_g)\right)\right] \\
&\leq \mathrm{E}\left[Y_g I\left(Y_g > n\tilde{\alpha}_g^2\right)\right] \\
&\leq \sup_m \mathrm{E}\left[Y_{g,m} I\left(Y_{g,m} > n\tilde{\alpha}_g^2\right)\right] \to 0 \text{ as } n \to \infty.
\end{aligned}$$

□

Theorems 3.3.1 and 3.3.2 show that the DI $\hat{D}_N$ has an asymptotic normal distribution. Formally, if

$$n^{1/2}(R - R) \xrightarrow{d} (0, V_R),$$

then

$$\begin{aligned}
n^{1/2}(\hat{D} - D) =& n^{1/2}(R + e) \\
&\xrightarrow{d} N(0, V_R)
\end{aligned}$$

as $n^{1/2}R_N \xrightarrow{d} N(0, V_R)$ and $n^{1/2}e_N \xrightarrow{p} 0$.

The above discussion supposes Assumption *D4*, that the proportions from the available population are bounded away from the true population proportions in every level of the categorical variables. There may be many cases in real surveys where the categories match for some levels, either by happenstance or because the survey has good representation for that variable. In these cases the asymptotic contribution to the DI for this is not normal but folded normal. If $X \sim N(\mu, \sigma^2)$, then we say $Y = |X|$ has a folded normal distribution, denoted $Y = |X| \sim FN(\mu, \sigma^2)$ (Leone et al. [1961]). The folded normal asymptotics were stated without proof in Allen et al. [2015]; a proof is given below.

Theorem 3.3.3. *Suppose that samples of size n are drawn from a sequence of populations of size N where the vector of proportions of the categorical variable in the available population is denoted $\boldsymbol{p}_N$, Assumptions* D1–D3 *above hold and Assumption* D4 *is replaced with the following assumption for some level g:*

D4' $n^{1/2}(p_{N,g} - \alpha_{N,g}) \to c_g$, *with* $|c_g| < \infty$.

Then $n^{1/2}|\hat{p}_g - \alpha_g|$ converges in distribution to a folded normal distribution with parameters $\mu = |c_g|$ and $\sigma^2 = \gamma_{gg}$, the g^{th} diagonal element of $\boldsymbol{\Gamma}$.

Proof. Note that

$$n^{1/2}(\hat{p}_g - \alpha_g) = n^{1/2}(\hat{p}_g - p_g) + n^{1/2}(p_g - \alpha_g)$$

In the right hand side of the above expression, the first piece is asymptotically normal from Assumption *D2*, and the second piece is a nonrandom sequence converging to c_g by Assumption *D4'*. Thus we see that

$$n^{1/2}(\hat{p}_g - \alpha_g) \xrightarrow{d} N(c_g, \gamma_{gg})$$

Since the absolute value is a continuous transformation, and the absolute value of a normal is defined as a folded normal distribution, the continuous mapping theorem implies that

$$n^{1/2}|\hat{p}_g - \alpha_g| = |n^{1/2}(\hat{p}_g - \alpha_g)| \xrightarrow{d} FN(|c_g|, \gamma_{gg}).$$

□

The distribution theory for the sum of these components is beyond the scope of this chapter. However, there are some theoretical results on the folded normal distribution that are useful to this discussion. Consider a random variable $X \sim N(\mu, \sigma^2)$ and $Y = |X| \sim FN(\mu, \sigma^2)$. From Leone

et al. [1961],

$$E[Y] = \sqrt{\frac{2}{\pi}}\sigma \exp\{-\frac{\mu^2}{2\sigma^2}\} + \mu[1 - 2\Phi(-\frac{\mu}{\sigma})] \tag{3.3}$$

$$Var(Y) = \mu^2 + \sigma^2 - (E[Y])^2. \tag{3.4}$$

Since $E[|X|] > E[X]$, we know that $E[Y] > \mu$, which also implies that $Var(Y) < \sigma^2$. This means that when $p_g - \alpha_g$ is small for some g, we will expect that the plug-in estimator will be positively biased. Specifically, we see that for the half-normal distribution ($\mu = 0, \sigma^2 = 1$), $E[Y] = \sqrt{2/\pi}$. A linearization based estimator assumes a normal distribution when $p_g - \alpha_g$ is small, rather than the folded normal. Thus we expect that in this case a linearization variance estimator will overestimate the variance of the dissimilarity index.

We will compare point and interval estimates for the DI in a simulation study. This study includes robustness to Assumption *D4*.

3.4 Simulation Study

While the dissimilarity index can be used to quantify coverage bias, nonresponse bias, or a combination of the two, we will consider coverage bias in our simulation to avoid discussion of nonresponse adjustments. In the setting of coverage bias, the above theory shows that a normal approximation to the DI can work very well if frame proportions for a categorical variable differ at all levels from the corresponding population proportions. There may be cases in practice where frame proportions match population proportions at some levels. To understand the implications for such cases, we conduct a simulation study below.

3.4.1 Simulated Population and Frames

We created an artificial population and artificial frames with coverage error to study properties of the DI. The population was based on data downloaded from the US Census Bureau's Public Use Microdata Sample (PUMS) (U.S. Census Bureau [2021]), including Public Use Microdata Areas

(PUMAs) 100, 103, 300, and 400 in Northern Colorado. These PUMAs correspond respectively to Northeast Colorado (Eastern Plains); Northern Larimer County; South Central Weld County (Greeley, Evans, and Windsor); and Eagle, Summit, Grand and Jackson Counties. Available variables are age, sex, race, Hispanic origin, marital status, personal income, level of education, and access to high-speed internet, along with individual weights and household weights. Records were deleted for persons under the age of 18 and for households with household weight of 0, to simulate the population of interest as the adult, noninstitutionalized population, which is a common target population for surveys. Person weights were treated as the number of individuals in the population having a given set of characteristics.

Table 3.1: Demographics for each of the sampling frames used in the simulation study of different estimators of the DI

Variable	Level	Population	$\lambda = 0.0$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$
Age	18-34	33.4	33.4	31.3	24.1	18.9
	35-49	24.6	24.6	24.4	24.3	24.6
	50-64	23.7	23.6	25.0	29.2	32.0
	65+	18.3	18.3	19.3	22.3	24.5
Sex	Male	50.8	50.9	49.2	44.1	40.8
	Female	49.2	49.1	50.8	55.9	59.2
Race	White	91.8	91.9	92.1	93.0	93.5
	Black	1.0	1.0	0.9	0.7	0.6
	AIAN	1.1	1.1	1.0	0.9	0.8
	API	1.7	1.6	1.7	1.6	1.6
	Other	2.3	2.3	2.2	1.8	1.5
	Multi	2.2	2.1	2.1	2.0	1.9
Hispanic	No	81.4	81.3	82.5	86.0	88.5
	Yes	18.6	18.7	17.5	14.0	11.5
Marital	Married	54.8	54.8	56.3	61.4	65.2
	Widowed	4.3	4.3	4.5	5.2	5.7
	Divorced	10.7	10.7	10.8	11.4	11.9
	Separated	0.9	0.9	0.9	0.9	0.9
	Single	29.3	29.4	27.5	21.1	16.4
Income	Negative	7.7	7.7	7.6	7.2	6.8
	Under $15K	19.1	19.2	18.7	17.8	16.7
	$15-50K	39.9	39.9	39.7	38.9	38.5
	$50-100K	23.5	23.5	23.8	24.8	25.6
	$100K or More	9.8	9.7	10.1	11.4	12.5
Education	$<$ High School	9.3	9.2	8.6	6.5	5.0
	High School/Eq	21.6	21.6	21.1	19.2	18.2
	Some Coll/Assoc	32.3	32.4	32.3	32.5	32.3
	Bachelors	23.5	23.4	24.0	25.3	26.4
	Advanced	13.3	13.4	14.1	16.4	18.1
Internet	NA	7.0	7.0	6.4	4.8	3.8
	Yes	77.0	77.0	78.1	81.3	83.2
	No	16.0	16.0	15.5	13.9	12.9
PUMA	100	15.9	15.9	15.5	14.3	13.5
	103	34.2	34.2	34.7	36.2	36.8
	300	30.0	30.1	30.2	30.7	31.1
	400	19.8	19.9	19.6	18.9	18.5

To simulate frames with varying degrees of coverage bias, we constructed a stochastic model in which the probability of frame inclusion for individual i is

$$
\begin{aligned}
p_i(\lambda) = \text{expit}\{&4.8 + (1-\lambda)(-2.9) + \lambda[-2.8(\text{Age18-34}) - 1.6(\text{Age35-49}) - 1.6(\text{SexMale}) - \\
&0.4(\text{RaceBlack}) - 0.4(\text{RaceAIAN}) - 0.1(\text{RaceAPI}) - 0.1(\text{RaceOther}) - \\
&0.1(\text{RaceMulti}) - 1.0(\text{HispanicYes}) - 0.1(\text{MaritalWidowed}) - \\
&0.6(\text{MaritalDivorced}) - 0.6(\text{MaritalSeparated}) - 1.0(\text{MaritalSingle}) - \\
&0.2(\text{IncomeNegative}) - 0.1(\text{Income0-15K}) - 0.1(\text{Income>100K}) - \\
&1.5(\text{EducNoHS}) - 1.0(\text{EducHS/Equiv}) - 0.3(\text{EducSomeCollege/Assoc}) - \\
&0.2(\text{EducBachelor}) - 2.5(\text{HispeedNA}) - 1.0(\text{HispeedNo}) - 0.9(\text{PUMA100}) - \\
&0.9(\text{PUMA400})]\}.
\end{aligned}
$$

Representativeness of the frame for the target population is determined by the parameter λ. Four frames were generated, with λ equal to 0, 0.1, 0.5, or 1. The frames with λ close to zero represent frames that are very representative of the population, and the frames with high levels of λ have more dissimilarity from the population (Table 3.1). Because the derivation of the variance estimator for the DI depends on the proportions in the frame not being equal to the population totals, it is expected that the variance estimator will perform better in frames with high λ.

In creating these frames, we did intentionally keep some categories close to their true population values. These categories were the age category 35-49 and the education category some college or Associate's degree. This was done to ensure that the simulation study includes cases where Assumption *D4* is not satisfied.

3.4.2 Estimators Considered

We considered four estimators of DI, along with their respective variance estimators and confidence intervals. The first DI estimator was the plug-in estimator,

$$\frac{1}{2}\sum_{g=1}^{G}|\hat{p}_g - \alpha_g|.$$

As shown theoretically above and noted in many sources (e.g. Cortese et al. [1976], Ransom [2000], Allen et al. [2015]), this estimator is consistent and asymptotically normal when $|p_g - \alpha_g|$ is "large" for all g (Assumption *D4*), but can be positively biased and non-normal when $|p_g - \alpha_g|$ is "small" for some g (Assumption *D4'*). Theory also suggests heuristically that the variance estimator will be positively biased in the latter case. To attempt to correct for these biases, we consider two additional DI estimators that set the gth term of DI equal to zero and the corresponding variance contribution equal to zero when the gth sample estimate is close to the corresponding population proportion. The two estimators differ in the definition of "close": the "cut-SE" estimator modifies the gth contribution if $|\hat{p}_g - \alpha_g| < \text{SE}(\hat{p}_g)$, and the "cut-Z" estimator modifies those contributions if $|\hat{p}_g - \alpha_g| < z_{0.975}\text{SE}(\hat{p}_g)$. Additionally, we consider a bias-corrected estimator proposed by Allen et al. [2015]. This estimator starts, like the previous estimators, by computing $|\hat{p}_g - \alpha_g|$ for each g. Then, as noted in Allen et al. [2015] and proved above, $|\hat{p}_g - \alpha_g|$ asymptotically follows a folded normal distribution when $|p_g - \alpha_g| = O(n^{-1/2})$. The DI is then computed as the sum over levels g of the values where this folded normal distribution is maximized. That is, if we let f_g denote the pdf of each of these folded normal distributions, the estimate of the DI is $\sum_{g=1}^{G} \text{argmax}_x f_g(x)$. For this fourth estimator, we do not use normal theory but instead use a parametric bootstrap to find the confidence intervals. For the parametric bootstrap confidence intervals, the bootstrap replicate probability vectors are drawn from the multivariate normal distribution with mean vector $\hat{\boldsymbol{p}}$ and covariance matrix calculated using standard survey arguments. The variance estimator is then the variance of the DIs calculated from the bootstrap values. The confidence interval is computed with

a percentile bootstrap taking the 0.025 and the 0.975 quantiles of the bootstrap distribution as the confidence limits.

3.4.3 Simulation Methods

As a comparison and to check confidence interval (CI) coverage, the population DI was calculated. The simulation study consisted of 1000 replicated samples each of size $n = 200$ and $n = 1000$, drawn from each of the four frames. Each sample was selected via stratified simple random sampling, with PUMAs as strata, with stratum allocations given in Table 3.2. From each sample the DI and its sampling variance were estimated using each of the four estimators as described above. From these replicates, we computed relative bias (relbias) of the point estimators, relative root mean squared error (relRMSE) of the variance estimators, and coverage rates and mean widths of confidence intervals.

Table 3.2: Sample allocations by PUMA for stratified simple random sampling in the simulated replicate samples.

PUMA	100	103	300	400
$n = 200$	78	45	26	51
$n = 1000$	390	225	130	255

3.4.4 Results

In the study, the cut-Z estimator overcorrected for the positive bias of the plug-in estimator when the DI is somewhat small. This led to worse coverage properties than the other estimators, so we do not include the cut-Z estimator in the tables or discuss it further.

For DI point estimation, we notice that the cut-SE and Allen estimators, as expected, had less bias than the plug-in estimator, especially in the harder cases (lower sample size, lower DI), consistent with Allen et al. [2015]. For variance estimation, we see that the plug-in linearization estimator, as expected, had a noticeable positive bias when the true DI was lower. We also notice

that the cut-SE linearization variance estimator is negatively biased particularly when the true DI is low. The Monte Carlo variance is higher for the Allen and cut-SE estimators than for the plug-in estimator.

For confidence intervals, we see that all estimators had less than nominal coverage when the DI is close to zero. The best coverage properties are from the plug-in linearization confidence interval, and the Allen-bootstrap confidence interval, with noticeably worse performance in the cut-SE linearization confidence interval. This may be explained by noting that the linearization variance overestimates the true variance when the estimate is positively biased. The cut-SE linearization confidence interval uses a less biased point estimator, but with a negatively biased variance estimator in many cases. We see that the dissimilarity index needs to be relatively high (>10%) before the variance estimators have low bias (<10%) when the sample size is 200. The variance estimators have low bias in the sample size of 1000 when the true DI is greater than 5.

This simulation suggests that although the plug-in estimator is biased when the DI is small, the linearization confidence interval from the plug-in point estimator has reasonable coverage properties most of the time, and is no wider than the other confidence intervals considered.

Table 3.3: Simulation results for sample size $n = 200$ with 1000 replicated samples at each setting. For DI point estimates, "truth" refers to the true DI of the frame. For variance, "MC" refers to the Monte-Carlo variances of the estimates. Also, "est" is the mean of the estimates, and "rbias" is the mean % relative bias (bias/truth) of the estimates compared to "truth", and "rRMSE" is the relative mean-squared error of the variance estimates. For 95% CIs, "cover" is empirical coverage of the true DI by the interval, and "width" is the mean width of the interval. To save space, marital status and internet access are not shown in this table, and we restrict the focus to $\lambda = 0.1, 1.0$.

			DI Point Estimate			Variance of DI Est.				95% CI	
Variable	λ	Estimator	truth	est	rbias	MC	est	rbias	rRMSE	cover	width
Age	0.1	Plug-in	2.30	6.28	173.3	7.58	16.34	115.7	120.5	0.898	13.62
		Cut-SE	2.30	4.05	76.2	11.91	5.90	-50.4	65.0	0.647	7.42
		Allen	2.30	3.69	60.4	11.30	14.48	28.1	37.2	0.934	13.93
	1.0	Plug-in	14.55	16.18	11.2	11.58	15.34	32.5	46.7	0.932	15.20
		Cut-SE	14.55	15.24	4.7	13.91	12.25	-11.9	24.3	0.912	13.62
		Allen	14.55	15.00	3.1	14.36	16.15	12.5	25.9	0.948	15.63
Sex	0.1	Plug-in	1.59	3.61	127.0	7.27	18.10	149.2	149.3	0.969	11.82
		Cut-SE	1.59	2.30	44.5	11.47	6.14	-46.4	87.1	0.315	5.03
		Allen	1.59	2.07	30.3	10.75	15.19	41.4	59.5	0.993	12.05
	1.0	Plug-in	9.99	9.80	-2.0	16.81	18.53	10.2	11.5	0.956	15.75
		Cut-SE	9.99	9.55	-4.4	20.73	16.74	-19.3	32.5	0.884	14.71
		Allen	9.99	9.43	-5.7	22.03	20.87	-5.3	16.0	0.966	16.62
Race	0.1	Plug-in	0.24	3.12	1195.7	1.37	3.75	173.6	262.6	0.677	6.09
		Cut-SE	0.24	1.90	689.9	2.09	1.04	-50.1	94.8	0.397	2.86
		Allen	0.24	1.74	625.4	1.88	3.29	75.4	130.9	0.241	6.48
	1.0	Plug-in	1.66	3.47	108.9	1.40	2.74	95.5	194.3	0.691	5.23
		Cut-SE	1.66	2.42	45.6	2.45	0.80	-67.5	78.9	0.546	2.81
		Allen	1.66	2.28	37.3	2.40	2.62	9.3	60.7	0.733	5.90
Hispanic	0.1	Plug-in	1.11	2.80	152.5	4.35	10.74	146.9	152.8	0.952	9.06
		Cut-SE	1.11	1.84	65.4	6.83	3.46	-49.4	86.5	0.310	3.85
		Allen	1.11	1.70	53.5	6.47	8.90	37.5	57.6	0.984	9.27
	1.0	Plug-in	7.16	7.07	-1.1	8.10	8.58	5.9	26.4	0.951	10.53
		Cut-SE	7.16	6.85	-4.3	10.70	7.26	-32.2	43.2	0.843	9.55
		Allen	7.16	6.78	-5.3	11.17	9.34	-16.4	28.2	0.955	10.99
Income	0.1	Plug-in	0.66	6.35	868.8	6.28	15.70	149.8	157.3	0.793	13.55
		Cut-SE	0.66	3.88	491.7	9.81	5.10	-48.1	66.1	0.561	7.05
		Allen	0.66	3.49	431.9	9.05	13.61	50.4	58.6	0.737	13.81
	1.0	Plug-in	4.78	7.74	61.7	8.33	16.56	98.7	107.2	0.948	14.57
		Cut-SE	4.78	5.41	13.1	12.60	6.62	-47.4	61.1	0.760	8.78
		Allen	4.78	5.01	4.7	12.02	15.09	25.5	36.3	0.963	14.78
Education	0.1	Plug-in	1.21	6.45	434.3	5.96	15.81	165.2	171.2	0.835	13.66
		Cut-SE	1.21	3.88	221.3	9.34	4.92	-47.3	64.4	0.631	7.02
		Allen	1.21	3.51	190.6	8.61	13.61	58.1	64.2	0.797	13.89
	1.0	Plug-in	7.68	10.22	33.0	9.51	15.96	67.8	75.6	0.924	15.19
		Cut-SE	7.68	8.35	8.6	14.23	8.25	-42.1	51.0	0.836	10.71
		Allen	7.68	7.94	3.4	14.25	15.37	7.8	20.5	0.956	15.11

Table 3.4: Simulation results for sample size $n = 1000$ with 1000 replicated samples at each setting. For DI point estimates, "truth" refers to the true DI of the frame. For variance, "MC" refers to the Monte-Carlo variances of the estimates. Also, "est" is the mean of the estimates, and "rbias" is the mean % relative bias (bias/truth) of the estimates compared to "truth", and "rRMSE" is the relative mean-squared error of the variance estimates. For 95% CIs, "cover" is empirical coverage of the true DI by the interval, and "width" is the mean width of the interval. To save space, marital status and internet access are not shown in this table, and we restrict the focus to $\lambda = 0.1, 1.0$.

			DI Point Estimates			Variance of DI Est.				95% CI	
Variable	λ	Estimator	truth	est	rbias	MC	est	rbias	rRMSE	cover	width
Age	0.1	Plug-in	2.30	3.40	47.8	1.94	3.31	70.5	73.3	0.956	6.48
		Cut-SE	2.30	2.54	10.4	2.99	1.59	-46.7	57.2	0.778	4.26
		Allen	2.30	2.37	3.0	2.95	3.12	5.6	18.8	0.955	6.64
	1.0	Plug-in	14.55	15.14	4.0	2.76	3.02	9.2	26.4	0.946	6.77
		Cut-SE	14.55	14.88	2.3	2.91	2.60	-10.8	20.6	0.938	6.29
		Allen	14.55	14.83	1.9	2.89	2.98	3.1	18.5	0.949	6.71
Sex	0.1	Plug-in	1.59	2.03	27.5	1.93	3.61	87.1	87.1	0.982	5.64
		Cut-SE	1.59	1.56	-2.1	3.03	1.75	-42.2	72.8	0.469	3.26
		Allen	1.59	1.45	-8.8	2.98	3.34	12.3	35.9	0.992	5.78
	1.0	Plug-in	9.99	10.12	1.3	3.49	3.69	5.8	6.2	0.953	7.53
		Cut-SE	9.99	10.12	1.3	3.49	3.69	5.8	6.2	0.953	7.53
		Allen	9.99	10.12	1.3	3.49	3.70	6.2	8.9	0.957	7.50
Race	0.1	Plug-in	0.24	1.43	493.3	0.28	0.77	170.0	197.2	0.802	2.98
		Cut-SE	0.24	0.89	270.0	0.45	0.25	-44.9	74.5	0.685	1.57
		Allen	0.24	0.80	234.7	0.42	0.67	58.8	80.3	0.656	3.05
	1.0	Plug-in	1.66	2.11	27.4	0.50	0.70	39.0	58.2	0.917	3.11
		Cut-SE	1.66	1.72	3.7	0.79	0.39	-50.7	57.1	0.736	2.27
		Allen	1.66	1.65	-0.5	0.81	0.69	-14.9	22.8	0.918	3.17
Hispanic	0.1	Plug-in	1.11	1.48	33.6	1.09	2.15	96.5	97.5	0.977	4.27
		Cut-SE	1.11	1.11	0.3	1.71	0.94	-44.9	75.0	0.437	2.31
		Allen	1.11	1.04	-6.7	1.66	1.91	15.2	35.1	0.989	4.38
	1.0	Plug-in	7.16	7.16	0.1	1.67	1.70	1.6	11.0	0.942	5.10
		Cut-SE	7.16	7.16	0.1	1.67	1.70	1.6	11.0	0.942	5.10
		Allen	7.16	7.16	0.1	1.67	1.72	2.6	14.3	0.942	5.11
Income	0.1	Plug-in	0.66	2.90	342.4	1.15	3.16	175.5	180.5	0.860	6.16
		Cut-SE	0.66	1.79	173.0	1.77	1.06	-40.2	62.7	0.691	3.30
		Allen	0.66	1.61	145.8	1.64	2.76	68.5	75.4	0.872	6.26
	1.0	Plug-in	4.78	5.29	10.6	2.49	3.37	35.3	42.0	0.963	7.10
		Cut-SE	4.78	4.62	-3.5	3.56	2.10	-41.1	49.7	0.819	5.47
		Allen	4.78	4.47	-6.7	3.57	3.39	-5.1	18.9	0.977	7.11
Education	0.1	Plug-in	1.21	3.09	155.7	1.21	3.19	164.1	168.6	0.912	6.28
		Cut-SE	1.21	1.99	64.8	1.92	1.12	-41.6	59.3	0.769	3.52
		Allen	1.21	1.81	50.2	1.80	2.80	55.2	60.7	0.920	6.35
	1.0	Plug-in	7.68	8.45	10.0	2.49	3.11	24.7	32.7	0.946	6.88
		Cut-SE	7.68	8.00	4.2	3.09	2.30	-25.7	32.8	0.892	5.88
		Allen	7.68	7.90	2.9	3.18	3.19	0.3	15.2	0.954	6.92

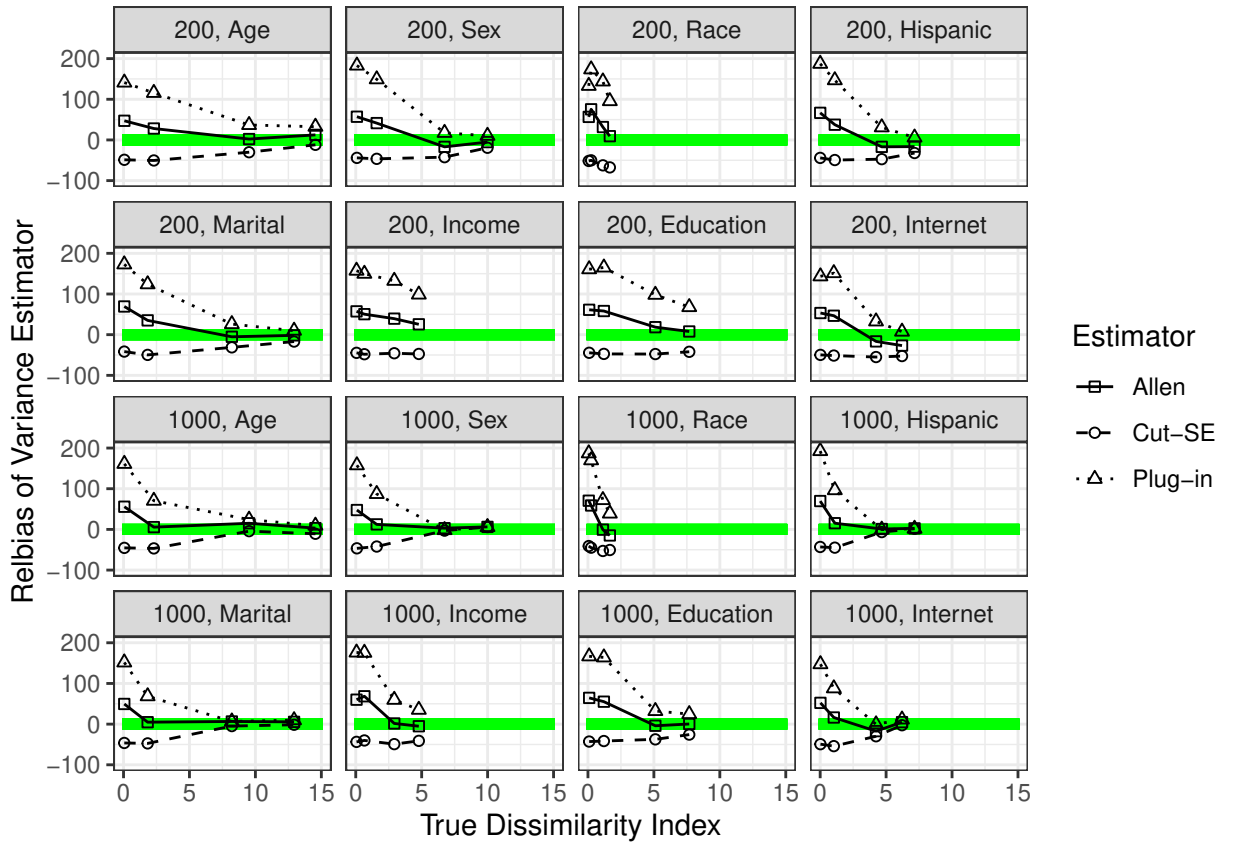

Figure 3.1: Relative bias squared error (relbias) for the linearization variance estimator plotted against population level DI for each variable at different levels of the parameter λ, with larger values of λ corresponding to higher DI. Points in the green band have less than 10% relbias.

3.4.5 Discussion of Simulation Study

The simulation results show that although the linearization variance estimator is biased when the DI is close to zero, confidence intervals formed using the linearization variance and the plug-in DI estimator performed well except in the toughest of cases. Additionally, attempts to improve on the performance of the linearization confidence interval were not successful.

It is worth considering how the DI estimators performed across the different variables. In interpreting this, it is relevant to note that the impact of estimating the DI depends more on the values within each category than what the categories represent. In this sample, we see the worst coverage in the variables Race and Income. This makes sense because these variables also have

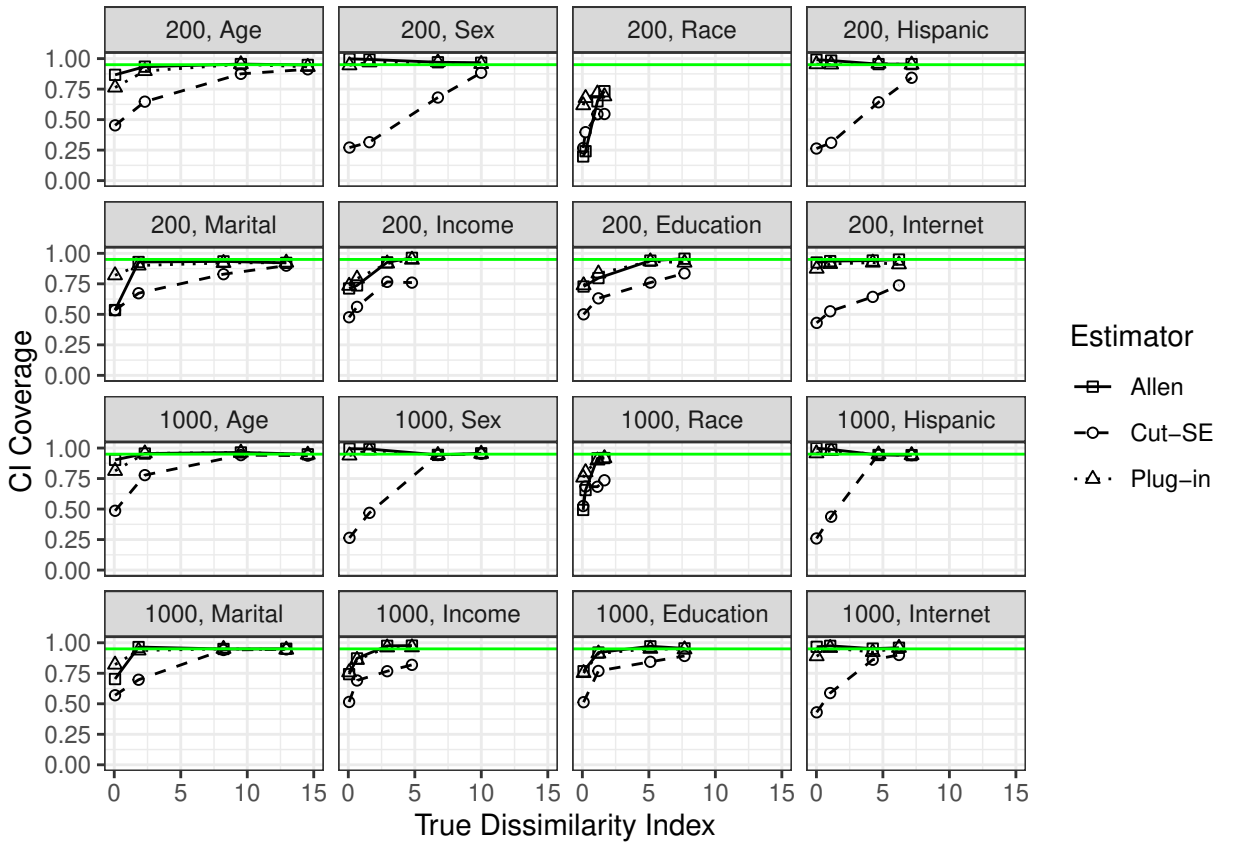

Figure 3.2: Coverage rates for 95% confidence intervals plotted against the true DI. The green line represents 95% coverage.

the lowest DIs. We see that the coverage properties and biases of the variance estimators for all variables follow a similar pattern based on the value of the true DI (Figures 3.2 and 3.1). This even holds true for the variables Age and Education, for which one category was intentionally held close to the truth.

When dealing with categorical variables, it is always important to consider the coding of the variables. The results above suggest that the true value of the DI is the main determinant of coverage properties and variance estimation accuracy. Since interpretation is important, the categories should be coded in a way that is highly relevant to the researcher, not necessarily in a way to give the best statistical properties. That said, we also note that the race variable had several small categories, and also had among the worst bias and coverage properties of the variables. Given that estimating small proportions is often difficult, and estimation is less good when the true value is

near to the truth, categories that are as coarse as practical may be a good way to go. We will also note that there are characteristics of our population that may not generalize to other populations (for example, our population was 91.8% white).

It is valuable to consider ways to improve interval estimation for the DI. With the linearization estimators, it is noted that the true distribution is not asymptotically normal when proportions of categories in the available population approximate categories in the target population, but the good coverage properties of the uncorrected linearization estimator in this case suggest that the normal approximation may be adequate. As the variances of the corrected linearization estimators are underestimates of the true variance when the variable is close to zero, setting the variance equal to zero for each proportion when the point estimate is close to the truth appears to be an overcorrection. It could be worth investigating if there is a multiplicative factor in (0,1) that can result in as good coverage as and smaller interval width than the uncorrected estimator when the point estimates for several categories are close to the truth.

There is also potential for improvement of the estimator developed by Allen et al. [2015]. The DI is a sum over levels of a categorical variable between estimates in a sample and the truth. In the approach from Allen et al. [2015], each of the categories is treated separately in the maximum likelihood estimation. One promising possibility is to apply the Allen et al. [2015] method to the sum of the components of the multivariate folded normal distributions (Chakraborty and Chatterjee [2013]) of the deviations of each level of the categorical variable. Investigation into the distribution of the sum of the folded multivariate normal components may also improve inference for the linearization estimators.

The bootstrap confidence intervals may also have room for improvement. When computing confidence intervals using the bootstrap, we used the 0.025 and 0.975 quantiles of the bootstrap distribution (percentile bootstrap). Another common method for bootstrap confidence intervals involves pivoting a t-style statistic around the true values. One issue with the t-style bootstrap CIs for this problem is that if the point estimate for the DI is 0, then the CI is $\{0\}$.

3.5 Summary and Future Work

In this chapter, we further developed theory for the dissimilarity index (DI). We established a mathematical connection between the DI as used in the social sciences, and how it is now being used in surveys. We also derived survey design-based linearization variance expressions for the DI, and showed their usefulness for real data in a simulation. We also pointed out some potential ways to improve interval estimation of the DI.

We derived a survey design-based expression for the variance of the DI. This expression allows the DI to be discussed more fully as a measure of representativeness from a survey sample. Expressing the degree of uncertainty is a key contribution of statistics to the practice of science, so having a design-based expression for the variance of the DI keeps it on the same footing as other statistics, like the R indicator (Shlomo et al. [2012]).

Properties of point and interval estimates for the DI were tested in a simulation study. Results from this study showed, as expected, that the plug-in estimator for the DI was found to be positively biased when the proportion of some levels of the categorical variable in the available population and target population were similar. This bias could be greatly reduced, though not eliminated, by either using the maximum-likelihood estimator DI described by Allen et al. [2015], or by the similar cut-Z estimator, setting the category-level difference equal to zero when the difference is within one standard error of zero. For interval estimation, the plug-in estimator and associated linearization estimator performed as well as any other estimator in terms of coverage and interval width. This is partially because setting the variance contribution to zero for of categories when the cut-Z estimator was set to zero underestimated the variance. Additionally, more research could be done to improve bootstrap methods for estimating the distribution of the estimator proposed by Allen et al. [2015].

We also recognize that there is room for improvement on interval estimation for the DI. One key idea would be learning more about the distribution of the multivariate folded normal to have a better established CI. Another idea involves improvements to the bootstrap CI and connecting with the literature about statistics near a boundary (for example Self and Liang [1987]).

Chapter 4

Randomization Test for Completely Randomized Experiments Embedded in Complex Surveys

4.1 Introduction

Experimental design and survey inference have a long history. Randomization tests were first developed in the context of agricultural statistics in the 1920s and 30s (Fisher [1971] (first edition 1935), Pitman [1937a], Welch [1937]). Randomization inference has been promoted by scholars like Kempthorne (Kempthorne [1955], Kempthorne and Doerfler [1969]) in the mid 20th century, and has had a resurgence recently in evaluating clinical trials because of its adaptability to complex designs and lack of distributional assumptions (see, for example Proschan and Dodd [2019], Wang and Rosenberger [2020]).

In this chapter, we will review randomization inference and extend it to experiments embedded in complex surveys. This chapter will focus on developing theory for randomization tests that can be used for experiments using a completely randomized design (CRD) conditioned on the sample, with inferential targets estimated using the Narain-Horvitz-Thompson (NHT) approach. We will develop theory for a normal approximation for the randomization tests and show how such inference is valid for not just the sample, but also the finite population from which the sample was drawn. In Chapter 5, we develop parallel results for randomized complete block experiments. In Chapter 6, we consider extensions including use of the generalized regression (GREG) estimator and inferential targets that are differentiable functions of finite population totals.

In this chapter, we will first define notation and discuss the sources of randomness in Section 4.2. We will then derive expressions for the mean and variance of the treatment assignment in Section 4.3. We explain the randomization test in Section 4.4. We describe central limit theory for a normal approximation in Section 4.5. We consider the sampling distribution and extend

the results to the finite population in Section 4.7. We describe the distribution of the test statistic under both procedures and power properties of the randomization test in Section 4.6. We discuss simulation studies in Section 4.8, and we conclude with a summary and discussion of future work in Section 4.9.

4.2 Notation and Sources of Randomness

Randomization tests are a relatively simple way to ensure valid inference from a finite population. These tests are based on rerandomizing the experimental units (EUs) to treatments based on the distribution of the original randomization. If the treatment has no effect, then the test statistic will be the same under this rerandomization distribution as it was under the original randomization of treatments. This rerandomization provides a reference distribution that can be used in a simple test of whether the treatment had an effect. In order for this inference to make sense, however, one needs to ensure that there is enough variability in the treatment randomization (Basu [1980]).

In the context of randomization tests for complex surveys, there are three sources of randomness that need to be considered: randomness due to drawing a sample, which will be denoted with a subscript S; randomness due to assigning the treatments for each experimental unit in the experiment, denoted with a subscript T; and randomness due to relabelling the treatments in a null-hypothesis randomization test, denoted with a subscript R. In this chapter, we ignore issues of nonresponse and assume that all units sampled for the survey respond. There will be discussion of nonresponse and other issues with practical surveys in Section 6.5. The randomization test will condition on the S and T distributions, and focus on the R distribution, which is used to simulate the T distribution. We will discuss the S distribution to explain how inference is extended to the finite population from which the original sample was drawn.

To investigate a CRD embedded in a survey, we assume that we have m experimental units (EUs) sampled from a finite population of M elements (EUs) indexed $1, \ldots, M$, that have been assigned treatments $k = 1, \ldots, K$ by a CRD. We assume that the M elements consist of all possible

sampling units at some stage of a multi-stage sampling design. We first consider the test statistic, and then calculate the distribution under the randomization distribution.

We consider an experimental design embedded in a survey where the EUs assigned each treatment are sampling units at some stage in the sampling design. The observational units (OUs) are the ultimate sampling units of the survey. The treatment assignment, through a CRD in this case, can be viewed as the second phase of a two-phase sampling design. In a two-phase sampling design, a subsample is drawn in the second phase from a preliminary sample (drawn in the first phase). In this example, the first phase will be drawing the EUs from the population, and the second phase, for each treatment, will be assigning a random subsample of the EUs from the original sample to be exposed to the specified treatment. To denote this, we denote the treatments using superscripts $k = 1, \ldots, K$. We further let $T_i^{(k)}$ be the indicator that treatment k was assigned to experimental unit i, and we let $m^{(k)}$ be the number of experimental units assigned treatment k. Therefore the NHT estimator for the subsample exposed to experimental treatment k is

$$\hat{Y}^{(k)} = \frac{m}{m^{(k)}} \sum_{i \in S} w_i \hat{Y}_{i+} T_i^{(k)} = \sum_{i \in S} w_i \hat{Y}_{i+} \check{T}_i^{(k)},$$

where $\check{T}_i^{(k)} = \frac{m}{m^{(k)}} T_i^{(k)}$. The treatment total estimate $\hat{Y}^{(k)}$ is a unbiased estimator of the NHT estimator of the finite population Y-total conditioned on the sample, if the entire sample were exposed to treatment k. Therefore it is an unbiased estimator of finite population Y-total, if the entire population were exposed to treatment k. To allow for general contrasts of treatment totals, we let $\hat{\boldsymbol{Y}}$ be the K-vector of the treatment total estimates for the finite population, that is $\hat{\boldsymbol{Y}} = (\hat{Y}^{(1)}, \hat{Y}^{(2)}, \ldots, \hat{Y}^{(K)})'$.

In survey practice, all the weights are typically accrued to the observational unit, but we are using EU-level weights for generality to the case when EUs are OUs. This randomization test can still be performed with all of the weights on the observational unit level. The survey weighted totals within each experimental unit can be computed as the sum of the survey weighted experimental unit values.

4.3 Treatment Assignment

Randomization procedures are based on simulating the distribution that the response variable would have under different treatment assignments. To introduce how this works, we will begin by computing how the test statistic could vary with different treatment assignments (henceforth the "treatment assignment distribution"). Design-based approaches to inference estimate the distribution analytically. The randomization approach uses the data more directly to simulate what would happen if the treatment had no effect (henceforth the "randomization distribution"). Though we will be using a randomization approach in this chapter, we consider differences in the treatment effects across units in this section for more generality and to allow for power calculations. In the next section, we will specialize this approach to the randomization test.

In the previous section, we derived the formula for the test statistic under a random treatment assignment. We will now begin a statistical examination of this statistic, starting with computing the mean and the variance under the treatment assignment distribution.

We begin the development of the theory with general contrasts of treatment totals. Let $\mathbf{C}$ be a $L \times K$ matrix with the property that $\mathbf{C1} = \mathbf{0}$, where $\mathbf{1}$ is a $K \times 1$ vector of ones, so that

$$\mathbf{C}\hat{\boldsymbol{Y}} = \sum_{k=1}^{K} \boldsymbol{c}^{(k)} \hat{Y}^{(k)} = \sum_{k=1}^{K} \boldsymbol{c}^{(k)} \sum_{i \in S} w_i \check{T}_i^{(k)} \hat{Y}_{i+}^{(k)}$$

(where $\boldsymbol{c}^{(k)}$ is a row of $\mathbf{C}$) is a $L \times 1$ vector of contrasts. Another way to write this expression is

$$\mathbf{C}\hat{\boldsymbol{Y}} = \mathbf{C} \sum_{i \in S} w_i \hat{\boldsymbol{Y}}_{i+} \circ \check{\boldsymbol{T}}_i,$$

where $\circ$ signifies the Hadamard (element-wise) product, and $\hat{\boldsymbol{Y}}_{i+}$ is the column-vector with the k^{th} entry of $\hat{Y}_{i+}^{(k)}$.

4.3.1 Hypergeometric Argument for Mean and Variance Computations

We will now use hypergeometric arguments to compute some expectations and covariances of treatment assignment indicators, leading up to the vector versions. We will also define some addi-

tional notation. Let $d^{(k)}$ be the treatment assignment weights, that is $d^{(k)} = m/m^{(k)}$. Additionally let $\boldsymbol{d} = \mathrm{vec}(d^{(k)})$ be the K-vector with elements $d^{(k)}$ and $\mathbf{D} = \mathrm{diag}(\boldsymbol{d})$ be the diagonal matrix with diagonal entries $d^{(k)}$.

The expectation averaging over all possible treatment assignments, holding other stochastic elements fixed, is $\mathrm{E}_T[T_i^{(k)}] = m^{(k)}/m_k$, and thus $\mathrm{E}_T[\check{T}_i^{(k)}] = 1$. If we allow $\boldsymbol{m} = [m^{(1)}, \ldots, m^{(K)}]'$, we get $\mathrm{E}_T[\boldsymbol{T}_i] = \boldsymbol{m}/m$ and $\mathrm{E}_T[\check{\boldsymbol{T}}_i] = \mathbf{1}$. The expectation of a product is

$$\mathrm{E}_T[T_i^{(k)} T_{i'}^{(k')}] = \begin{cases} \frac{m^{(k)}}{m} & k = k', i = i' \\ \frac{m^{(k)}(m^{(k)}-1)}{m_h(m_h-1)} & k = k', i \neq i' \\ 0 & k \neq k', i = i' \\ \frac{m^{(k)}m^{(k')}}{m(m-1)} & k \neq k', i \neq i' \end{cases}. \tag{4.1}$$

Thus the covariances are

$$\mathrm{Cov}_T(T_i^{(k)}, T_{i'}^{(k')}) = \begin{cases} \frac{m^{(k)}(m-m^{(k)})}{m^2} & k = k', i = i' \\ -\frac{m^{(k)}(m-m^{(k)})}{m^2(m-1)} & k = k', i \neq i' \\ -\frac{m^{(k)}m^{(k')}}{m^2} & k \neq k', i = i' \\ \frac{m^{(k)}m^{(k')}}{m^2(m-1)} & k \neq k', i \neq i' \end{cases}. \tag{4.2}$$

If we let $\check{T}_{hi}^{(k)} = \frac{m_h}{m_h^{(k)}} T_{hi}^{(k)}$, then we see that the covariances are

$$\mathrm{Cov}_T(\check{T}_i^{(k)}, \check{T}_{i'}^{(k')}) = \begin{cases} \frac{(m-m^{(k)})}{m^{(k)}} & k = k', i = i' \\ -\frac{(m-m^{(k)})}{m^{(k)}(m-1)} & k = k', i \neq i' \\ -1 & k \neq k', i = i' \\ \frac{1}{m-1} & k \neq k', i \neq i' \end{cases}. \tag{4.3}$$

We next allow $\mathbf{D}$ to be the diagonal matrix with entries $d^{(k)} = m/m^{(k)}$. Taking these results and applying them to vectors gives

$$\mathrm{Cov}_T(\boldsymbol{T}_i, \boldsymbol{T}_{i'}) = \begin{cases} \mathbf{D}^{-1} - \frac{\boldsymbol{m}\boldsymbol{m}'}{m^2} & : \quad i = i' \\ \frac{-1}{m-1}\left(\mathbf{D}^{-1} - \frac{\boldsymbol{m}\boldsymbol{m}'}{m^2}\right) & : \quad i \neq i' \end{cases} \tag{4.4}$$

and

$$\mathrm{Cov}_T(\check{\boldsymbol{T}}_i, \check{\boldsymbol{T}}_{i'}) = \begin{cases} \mathbf{D} - \mathbf{J} & : \quad i = i' \\ \frac{-1}{m-1}\left(\mathbf{D} - \mathbf{J}\right) & : \quad i \neq i' \end{cases}, \tag{4.5}$$

where $\mathbf{J}$ is the $K \times K$ matrix where each element is 1.

We also use a result about Hadamard products. It can be easily shown that if $\boldsymbol{a}, \boldsymbol{b}$ are fixed vectors and $\boldsymbol{X}, \boldsymbol{Y}$ are random vectors, then $\mathrm{Cov}(\boldsymbol{a} \circ \boldsymbol{X}, \boldsymbol{b} \circ \boldsymbol{Y}) = \boldsymbol{a}\boldsymbol{b}' \circ \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})$.

Using these results, we can compute the expectation and variance of the contrast under the treatment assignment distribution.

Let

$$\dot{Y}^{(k)} = \sum_{i \in S} w_i \hat{Y}_{i+}^{(k)}$$

be the vector of estimated finite-population totals, as if all elements of the sample had been assigned treatment k. Further let $\dot{\boldsymbol{Y}} = [\dot{Y}^{(1)}, \ldots, \dot{Y}^{(K)}]$ be the K-vector of these estimated totals. Let

$$\hat{\mathbf{V}}_{w\hat{Y}} = \frac{1}{2m(m-1)} \sum_{i \in S} \sum_{i' \in S} (w_i \hat{\boldsymbol{Y}}_{i+} - w_{i'} \hat{\boldsymbol{Y}}_{i'+})(w_i \hat{\boldsymbol{Y}}_{i+} - w_{i'} \hat{\boldsymbol{Y}}_{i'+})'$$

be the sample covariance matrix of the weighted EU, and let the diagonal elements of $\hat{\mathbf{V}}_{w\hat{Y}}$ be

$$s^2_{w\hat{Y}^{(k)}} = \frac{1}{2m(m-1)} \sum_{i \in S} \sum_{i' \in S} (w_i \hat{Y}_{i+}^{(k)} - w_{i'} \hat{Y}_{i'+}^{(k)})^2,$$

the sample variances of the EU totals as if they had received treatment k.

Lemma 4.3.1. *If* $\mathbf{C}$ *is a contrast matrix, then, assuming additivity 4.6, the mean and variance under the treatment assignment distribution in a CRD are given by*

$$\begin{aligned}\mathrm{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}] &= \mathbf{C}\dot{\boldsymbol{Y}} \\ \mathrm{Var}_T(\mathbf{C}\hat{\boldsymbol{Y}}) &= m\mathbf{C}\left[\hat{\mathbf{V}}_{w\hat{Y}} \circ (\mathbf{D}-\mathbf{J})\right]\mathbf{C}'.\end{aligned}$$

Proof. We will use the vector versions of the above results to prove this theorem.

Recall that

$$\hat{\boldsymbol{Y}} = \sum_{i\in S} w_i \hat{\boldsymbol{Y}}_{i+} \circ \check{\boldsymbol{T}}_i.$$

From the fact that the vector $\mathbf{1}$ is the identity element under Hadamard multiplication, we find that

$$\mathrm{E}_T[\hat{\boldsymbol{Y}}] = \mathrm{E}_T\left[\sum_{i\in S} w_i \hat{\boldsymbol{Y}}_{i+} \circ \check{\boldsymbol{T}}_i\right] = \sum_{i\in S} w_i \hat{\boldsymbol{Y}}_{i+} \circ \mathbf{1} = \sum_{i\in S} w_i \hat{\boldsymbol{Y}}_{i+} = \dot{\boldsymbol{Y}}.$$

To compute the variance, observe

$$\begin{aligned}\mathrm{Var}_T(\boldsymbol{Y}) &= \mathrm{Var}_T\left(\sum_{i\in S} w_i \hat{\boldsymbol{Y}}_i \circ \check{\boldsymbol{T}}_i\right) \\ &= \sum_{i\in S}\sum_{i'\in S} w_i w_{i'} \mathrm{Cov}_T(\hat{\boldsymbol{Y}}_{i+} \circ \check{\boldsymbol{T}}_i, \hat{\boldsymbol{Y}}_{i'+} \circ \check{\boldsymbol{T}}_{i'}) \\ &= \sum_{i\in S}\sum_{i'\in S} w_i w_{i'} \hat{\boldsymbol{Y}}_{i+}\hat{\boldsymbol{Y}}'_{i'+} \circ \mathrm{Cov}_T(\check{\boldsymbol{T}}_i, \check{\boldsymbol{T}}_{i'}).\end{aligned}$$

From equation (4.5) above, it follows that

$$\begin{aligned}\mathrm{Var}_T(\boldsymbol{Y}) &= \sum_{i\in S} w_i^2 \hat{\boldsymbol{Y}}_{i+}\hat{\boldsymbol{Y}}'_{i+} \circ (\mathbf{D}-\mathbf{J}) - \frac{1}{m-1}\sum_{i\in S}\sum_{i'\in S\setminus\{i\}} w_i w_{i'} \hat{\boldsymbol{Y}}_i \hat{\boldsymbol{Y}}'_{i'} \circ (\mathbf{D}-\mathbf{J}) \\ &= \frac{1}{m-1}\sum_{i\in S}\sum_{i'\in S} w_i \hat{\boldsymbol{Y}}_{i+}(w_i \hat{\boldsymbol{Y}}_{i+} - w_{i'}\hat{\boldsymbol{Y}}_{i'+})' \circ (\mathbf{D}-\mathbf{J}) \\ &= m\hat{\mathbf{V}}_{w\hat{Y}} \circ (\mathbf{D}-\mathbf{J}),\end{aligned}$$

as desired. □

In this dissertation, we will derive further results under the alternative hypothesis using the simplifying assumption that the treatments are additive at the observational unit level. This means that each treatment results in an equal additive shift for all observational units. When weighted up to the EU level, this can be written as

$$\hat{Y}_{i+}^{(k)} = \hat{Y}_{i+}^{*} + \beta^{(k)}\hat{N}_i. \tag{4.6}$$

In this expression, $\hat{Y}_{i+}^{*}$ is the intrinsic (estimated) value of the response on EU i, $\hat{N}_i$ is the weighted sum of the number of OUs in EU i, used to estimate the total number of OUs within EU i.

We introduce new notation to explain these results. Let $\mathbf{D}_{w\hat{Y}}$ be the diagonal matrix with entries $d_{w\hat{Y}}^{(k)} = d^{(k)} s^2_{w\hat{Y}^{(k)}}$.

Under this assumption, we modify the above result in the following lemma.

Lemma 4.3.2. *If* $\mathbf{C}$ *is a contrast matrix and we assume additivity 4.6, the mean and variance under the randomness due to treatment assignment in a completely randomized experiment embedded in a complex survey are given by*

$$\mathrm{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}] = \hat{N}\mathbf{C}\boldsymbol{\beta} \tag{4.7}$$

$$\mathrm{Var}_T(\mathbf{C}\hat{\boldsymbol{Y}}) = m\mathbf{C}\left[\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}'\right]\mathbf{C}'. \tag{4.8}$$

Proof. For the expectation, we have

$$\dot{\boldsymbol{Y}} = \sum_{i\in S} w_i \hat{\boldsymbol{Y}}_{i+} = \sum_{i\in S} w_i(\hat{Y}_{i+}^{*}\mathbf{1} + \hat{N}_i\boldsymbol{\beta}) = \left(\sum_{i\in S} w_i \hat{Y}_{i+}^{*}\right)\mathbf{1} + \hat{N}\boldsymbol{\beta}.$$

From this it follows that

$$\mathrm{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}] = \mathbf{C}\left[\left(\sum_{i\in S} w_i \boldsymbol{Y}_{i+}^{*}\right)\mathbf{1} + \hat{N}\boldsymbol{\beta}\right] = \hat{N}\mathbf{C}\boldsymbol{\beta}.$$

To simplify this expression, we will consider the form of $\boldsymbol{V}_{w\hat{Y}}$ assuming (4.6):

$$\begin{aligned}
\hat{\mathbf{V}}_{w\hat{Y}} &= \frac{1}{2m(m-1)} \sum_{i\in S}(w_i\hat{\boldsymbol{Y}}_{i+} - w_{i'}\hat{\boldsymbol{Y}}_{i'+})(w_i\hat{\boldsymbol{Y}}_{i+} - w_{i'}\hat{\boldsymbol{Y}}_{i'+})' \\
&= \frac{1}{2m(m-1)} \sum_{i\in S}\sum_{i\in S}[(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})\mathbf{1} + (w_i\hat{N}_i - w_{i'}\hat{N}_{i'})\boldsymbol{\beta}] \\
&\qquad\qquad [(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})\mathbf{1} + (w_i\hat{N}_i - w_{i'}\hat{N}_{i'})\boldsymbol{\beta}]' \\
&= \frac{1}{2m(m-1)} \sum_{i\in S}\sum_{i\in S}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})^2\mathbf{J}_K + \\
&\qquad\qquad (w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(w_i\hat{N}_i - w_{i'}\hat{N}_{i'})(\mathbf{1}\boldsymbol{\beta}' + \boldsymbol{\beta}\mathbf{1}') + \\
&\qquad\qquad (w_i\hat{N}_i - w_{i'}\hat{N}_{i'})^2\boldsymbol{\beta}\boldsymbol{\beta}'.
\end{aligned}$$

Examining the term $\mathbf{C}\hat{\mathbf{V}}_{h,w\hat{Y}}\mathbf{C}'$ using this form, we note that $\mathbf{C}\mathbf{J}_K\mathbf{C}'$, $\mathbf{C}\mathbf{1}\boldsymbol{\beta}'\mathbf{C}'$, and $\mathbf{C}\boldsymbol{\beta}\mathbf{1}'\mathbf{C}'$ are all $\mathbf{0}$ because $\mathbf{C}$ is a contrast matrix. Thus

$$\mathbf{C}\hat{\mathbf{V}}_{h,w\hat{Y}}\mathbf{C}' = \frac{1}{2m(m-1)} \sum_{i\in S}\sum_{i'\in S}(w_i\hat{N}_i - w_{i'}\hat{N}_{i'})^2\mathbf{C}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{C}' = s^2_{w\hat{N}}\mathbf{C}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{C}'.$$

Therefore, we have that assuming additivity 4.6

$$\mathrm{Var}_T\left(\mathbf{C}\hat{\boldsymbol{Y}}\right) = m\mathbf{C}(\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\mathbf{C}',$$

as desired. □

For the randomization test, we will be interested in examining the distribution of the observed data under the null hypothesis (H_0) of no treatment effect. When we set $\boldsymbol{\beta} = \mathbf{0}$, we find that under H_0,

$$\begin{aligned}
\mathrm{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}] &= \mathbf{0} \\
\mathrm{Var}_T(\mathbf{C}\hat{\boldsymbol{Y}}) &= m\mathbf{C}\mathbf{D}_{w\hat{Y}}\mathbf{C}'.
\end{aligned}$$

4.4 Randomization Distribution and Test

The key idea of randomization procedures is to compare the test statistic obtained to the distribution of test statistics that could have been obtained under different labels of the treatments. This tests the null hypothesis

$$H_0 : \hat{Y}_{i+}^{(k)} = \hat{Y}_{i+}^{(k')} \; \forall i, k, k'. \tag{4.9}$$

Randomization procedures are valuable for making inference when there is not a lot of previous information on the phenomenon you are studying, the experimental design is quite complex and hard to model using other means, or when a researcher does not want to use outside knowledge to provide a more unbiased analysis due to standards in their field (Lane [1980], Kempthorne [1980], Proschan and Dodd [2019], for example). Embedding an experiment in a complex survey with its stratification, clustering and unequal probabilities of selection inherently makes experimental designs within surveys complex and hard to model, and survey practitioners are often in a setting where analysis is expected to be as model free as possible, due to the multipurpose nature of the surveys. Hence, randomization inference fits naturally in the context of complex surveys.

In this section, we will explain how the randomization procedure works in the context of a CRD embedded in a survey. In a general randomized experiment, one can reassign the treatment labels and recompute the statistic several times (using the same observed data) to get an idea of the range of plausible values that the statistic may take if the treatment had no effect. With small sample sizes, it may be feasible to enumerate the complete randomization distribution of all possible treatment assignments. With moderate or large sample sizes, this randomization distribution is infeasible to compute exactly but can be approximated via Monte Carlo or an asymptotic normal distribution.

To describe the randomization distribution for a linear statistic in a survey, we use the notation $R_i^{(k)}$ (and $\check{R}_i^{(k)} = \frac{m}{m^{(k)}} R_i^{(k)}$) to represent the randomized treatment labels. We use different, but parallel, notation than the treatment assignment to emphasize that the randomized labels are independent of the original treatment assignments $T_i^{(k)}$, but follow the same distribution. In fact, the randomization test is conducted conditioned on the original treatment assignment. We denote the

randomization total for each treatment as

$$\tilde{Y}^{(k)} = \frac{m}{m^{(k)}} \sum_{i \in S} w_i \hat{Y}_{i+} R_i^{(k)} = \sum_{i \in S} w_i \hat{Y}_{i+} \check{R}_i^{(k)}.$$

Using this notation, the original contrast is denoted $\mathbf{C}\hat{\boldsymbol{Y}}$ and the randomization version is denoted $\mathbf{C}\tilde{\boldsymbol{Y}}$.

The rest of this section will be devoted to explaining the logic of the randomization test and finding the mean and variance of the randomization distribution that we have described. The latter computations are used in subsequent sections for our normal-theory approximations.

4.4.1 Randomization Test

We now describe the randomization procedure for experiments embedded in surveys. The null hypothesis 4.9 discussed above is equivalent to additivity (4.6) with $\boldsymbol{\beta} = \beta \mathbf{1}$, for some constant β. Since we could change the intrinsic values of the response on EUs to match an additive shift in $\boldsymbol{\beta}$ by a multiple of $\mathbf{1}$, we can assume without loss of generality that under H_0, $\boldsymbol{\beta} = \mathbf{0}$. In this case, we have that for any i, k, $\hat{Y}_{i+}^{(k)} = \hat{Y}_{i+}^{*} = \hat{Y}_{i+}$.

Thus, under H_0, the distribution of the response vector

$$\hat{\boldsymbol{Y}} = \sum_{i \in S} w_i \hat{Y}_{i+} \check{\boldsymbol{T}}_i$$

under random treatment assignment (conditioned on the sample) is the same as the distribution of

$$\tilde{\boldsymbol{Y}} = \sum_{i \in S} w_i \hat{Y}_{i+} \check{\boldsymbol{R}}_i$$

under randomization of treatment labels (conditioned on the sample and the original treatment assignment), since the $\check{\boldsymbol{T}}_i$ have the same distribution as the $\check{\boldsymbol{R}}_i$. Hence, the randomization distribution of the statistic recreates the treatment assignment distribution under H_0, providing a basis for randomization inference.

4.4.2 Mean and Variance

To use a randomization test, one can always simulate results under the randomization distribution. However, it is often computationally and practically convenient to have a normal approximation for the null distribution. We derive the mean and variance for the test statistic here. Because the distribution of the treatment labels is the same as the original distribution of the treatments, we can reuse much of the work from the previous section.

Theorem 4.4.1. *The randomization contrast*

$$\mathbf{C}\tilde{\boldsymbol{Y}} = \mathbf{C}\sum_{i\in S} w_i \hat{Y}_{i+} \check{\boldsymbol{R}}_i$$

has mean and variance under the randomization distribution given by

$$\mathrm{E}_R[\mathbf{C}\tilde{\boldsymbol{Y}}] = \mathbf{0}$$
$$\mathrm{Var}_R(\mathbf{C}\tilde{\boldsymbol{Y}}) = m s^2_{w\hat{Y}} \mathbf{C}\mathbf{D}\mathbf{C}',$$

where

$$s^2_{w\hat{Y}} = \frac{1}{2m(m-1)} \sum_{i\in S} \sum_{i'\in S\setminus\{i\}} (w_i \hat{Y}_{i+} - w_{i'} \hat{Y}_{i'+})^2$$

is the variance of the observed responses at the EU level.

Proof. From the same arguments in Lemma 4.3.1, we find that

$$\mathrm{E}_R[\tilde{\boldsymbol{Y}}] = \sum_{i\in S} w_i \hat{Y}_{i+} \mathbf{1}.$$

Thus we have

$$\mathrm{E}_R[\mathbf{C}\tilde{\boldsymbol{Y}}] = \mathbf{C}\mathbf{1}\sum_{i\in S} w_i \hat{Y}_{i+} = \mathbf{0}$$

as $\mathbf{C}$ is a contrast matrix.

Similarly, we see that

$$\mathrm{Var}_R(\tilde{\boldsymbol{Y}}) = m s^2_{w\hat{Y}} (\mathbf{D} - \mathbf{J}),$$

yielding

$$\mathrm{Var}_R(\mathbf{C}\tilde{\boldsymbol{Y}}) = m s^2_{w\hat{Y}} \mathbf{C}(\mathbf{D} - \mathbf{J}_K)\mathbf{C}' = m s^2_{w\hat{Y}} \mathbf{C}\mathbf{D}\mathbf{C}',$$

again using the fact that $\mathbf{C}$ is a contrast matrix. □

4.5 Central Limit Theory

In this section, we will provide conditions under which the randomization distribution approximately follows a normal distribution. This proof is based heavily on results from Hoeffding [1951]. We will first prove some lemmas in Section 4.5.1, and then show the main result in Section 4.5.2.

4.5.1 Introductory Lemmas

Before proving the central limit theorem, there are lemmas that need to be established. The first is a lemma that is adapted from Theorem 1 in Hoeffding [1951]

Lemma 4.5.1. *Let $a_{n,i}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$ be a double sequence such that for some $\delta > 0$, $n^{-1} \sum_{i=1}^n |a_{n,i}|^{2+\delta} = O(1)$. Then*

1. $\lim_{n\to\infty} n^{-1/2} \max\{|a_{n,i}|, i = 1, \ldots, n\} = 0$
2. $\lim_{n\to\infty} n^{-(2+\gamma)/2} \sum_{i=1}^n |a_{n,i}|^{2+\gamma} = 0$ *for all* $\gamma > 0$

Proof. Let $b_{n,i} = n^{-1/2} a_{n,i}$ and $B_n = \max\{|b_{n,i}| : i = 1, \ldots, n\}$.

By the hypothesis, there is some $\delta > 0$ such that

$$B_n^{2+\delta} \leq \sum_{i=1}^n |b_{n,i}|^{2+\delta} \to 0$$

as $n \to \infty$. This proves conclusion 1.

To prove conclusion 2, notice by the hypothesis, we have $n^{-1} \sum_{i=1}^n a_{n,i}^2 = \sum_{i=1}^n b_{n,i}^2 = O(1)$. Thus there is some $F > 0$ such that $\sum_{i=1}^n b_{n,i}^2 \leq F$ for all $n > 0$. Now for any $\gamma > 0$,

$$\sum_{i=1}^n |b_{n,i}|^{2+\gamma} \leq B_n^\gamma \sum_{i=1}^n b_{n,i}^2 \leq F B_n^\gamma \to 0$$

as $n \to \infty$. □

The next two lemmas work to provide a lower bound to variance expressions in the denominators when the additivity assumption (4.6) is used.

Lemma 4.5.2. *Let $\boldsymbol{v} \in \mathbb{R}^d$, and let $a_i > 0$, $i = 1, \ldots, d$ be constants such that $\sum_{i=1}^d a_i = 1$. Let $\boldsymbol{w} \in \mathbb{R}^d$ be the vector with elements $w_i = a_i^{-1} v_i^2$. Then it follows that the matrix*

$$\mathrm{diag}(\boldsymbol{w}) - \boldsymbol{v}\boldsymbol{v}'$$

is nonnegative definite, and the vector with elements $a_i v_i^{-1}$ is in its null space.

Proof. Let $\boldsymbol{X} \in \mathbb{R}^d$ be the random vector that has $a_i^{-1} v_i$ in the i^{th} position (and all other elements 0) with probability a_i, giving

$$\begin{aligned} E[\boldsymbol{X}] &= \boldsymbol{v} \\ E[\boldsymbol{X}\boldsymbol{X}'] &= \mathrm{diag}(\boldsymbol{w}), \end{aligned}$$

and therefore

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{diag}(\boldsymbol{w}) - \boldsymbol{v}\boldsymbol{v}'.$$

Since we have expressed $\mathrm{diag}(\boldsymbol{w}) - \boldsymbol{v}\boldsymbol{v}'$ as a covariance matrix of a random variable, it must be nonnegative definite.

To show the vector with elements $a_i v_i^{-1}$ is in the null space of $\mathrm{diag}(\boldsymbol{w}) - \boldsymbol{v}\boldsymbol{v}'$ we evaluate the k^{th} element in the product of this matrix and vector is

$$\begin{aligned} \sum_{i=1}^d \left[w_i I(i=k) - v_i v_k \right] a_i v_i^{-1} &= \sum_{i=1}^d (v_i^2 a_i^{-1})(a_i v_i^{-1}) I(i=k) - a_i v_k \\ &= v_k - v_k \sum_{i=1}^d a_i. \end{aligned}$$

This is zero by the assumption that $\sum_{i=1}^d a_i = 1$. □

We note that the vectors $\check{\boldsymbol{R}}_{hi}$ and $\check{\boldsymbol{T}}_{hi}$ have the same properties of the vector $\boldsymbol{X}$ in the proof of Lemma 4.5.2, with $\boldsymbol{a} \in \mathbb{R}^K$ satisfying $a^{(k)} = m^{(k)}/m$ and $\boldsymbol{v} = \mathbf{1}_K$.

Lemma 4.5.3. *Let $\boldsymbol{c} \in \mathbb{R}^K$ be an arbitrary contrast vector, then*

$$\boldsymbol{c}'(\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{c} = \left(\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}s^2_{w\hat{Y}^{(k)}}\right) - (\boldsymbol{c}'\boldsymbol{\beta})^2 s^2_{w\hat{N}} \geq \sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}s^2_{w\hat{Y}^*}(1-r^2),$$

where r, the correlation between the $w_i\hat{Y}^*_{i+}$ and the $w_i\hat{N}_i$, is

$$r = \frac{\sum_{i\in S}\sum_{i'\in S}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}_{i'+})(w_i\hat{N}_i - w_{i'}\hat{N}_{i'})}{2m(m-1)s_{w\hat{Y}^*}s_{w\hat{N}}},$$

if $s_{w\hat{Y}^*}s_{w\hat{N}} > 0$, and $r = 0$ otherwise ($s^2_{w\hat{Y}^{(k)}}$ was defined in Lemma 4.3.1).

Proof. The left-hand side of the above expression can be written using covariances as

$$\left(\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}\left(s^2_{w\hat{Y}^*} + 2\beta^{(k)}rs_{w\hat{Y}^*}s_{w\hat{N}} + (\beta^{(k)})^2 s^2_{w\hat{N}}\right)\right) - \left(\sum_{k=1}^{K}c^{(k)}\beta^{(k)}\right)^2 s^2_{w\hat{N}}.$$

If $s_{w\hat{Y}^*}s_{w\hat{N}} = 0$, then either $s_{w\hat{Y}^*} = 0$ or $s_{w\hat{N}} = 0$. In the former case, the right hand side of the above is 0 and the left hand side is

$$\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}(\beta^{(k)})^2 s^2_{w\hat{N}} - \left(\sum_{k=1}^{K}c^{(k)}\beta^{(k)}\right)^2 s^2_{w\hat{N}} = s^2_{w\hat{N}}\left[\sum_{k=1}^{K}\frac{m}{m^{(k)}}(c^{(k)}\beta^{(k)})^2 - \left(\sum_{k=1}^{K}c^{(k)}\beta^{(k)}\right)^2\right].$$

This is nonnegative by Lemma 4.5.2. In the latter case, we see that the both the left and right hand sides of the above expression equal

$$\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}s^2_{w\hat{Y}^*},$$

by the definition that $r = 0$.

Now we consider the cases where $s_{w\hat{Y}^*}s_{w\hat{N}} > 0$. Taking the partial derivative of the left hand side with respect to $\beta^{(k)}$ (for arbitrary k), we find that

$$\frac{\partial}{\partial\beta^{(k)}}\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}_h) = 2m_h(c^{(k)})^2\frac{m_h}{m_h^{(k)}}(r_h s_{h,w\hat{Y}^*}s_{h,w\hat{N}} + \beta^{(k)}s_{h,w\hat{N}}^2) - 2m_h s_{h,w\hat{N}}^2\sum_{k'=1}^{K} c^{(k)}c^{(k')}\beta^{(k')}$$

Now we will check that the Hessian matrix is everywhere non-negative definite, ensuring that the solution obtained by setting this derivative equal to zero will be a global minimum.

Note

$$\frac{\partial^2}{\partial(\beta^{(k)})^2}\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}) = 2m(c^{(k)})^2\frac{m}{m^{(k)}}s_{h,w\hat{N}}^2 - 2s_{w\hat{N}}^2c^{(k)}c^{(k')}$$
$$\frac{\partial^2}{\partial\beta^{(k)}\partial\beta^{(k')}}\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}) = -2ms_{w\hat{N}}c^{(k)}c^{(k')}.$$

Expressing the above as a matrix, we find that the Hessian is

$$\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}) = 2ms_{w\hat{N}}^2\left[\mathrm{diag}\left((c^{(k)})^2\frac{m}{m^{(k)}}\right) - \boldsymbol{c}\boldsymbol{c}'\right].$$

Since $\sum_{k=1}^{K}\frac{m^{(k)}}{m} = 1$,

$$\mathrm{diag}\left((c^{(k)})^2\frac{m}{m^{(k)}}\right) - \boldsymbol{c}\boldsymbol{c}'$$

is everywhere non-negative definite by Lemma 4.5.2. Since $2ms_{w\hat{N}}^2 > 0$, we have verified that the Hessian is everywhere non-negative definite.

We now set the gradient with respect to $\boldsymbol{\beta}$ equal to zero to find a minimum. Recall that

$$\frac{\partial}{\partial\beta^{(k)}}\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}) = 2ms_{w\hat{N}}\left[(c^{(k)})^2\frac{m_h}{m_h^{(k)}}(rs_{w\hat{Y}^*} + \beta^{(k)}s_{w\hat{N}}) - s_{w\hat{N}}\sum_{k'=1}^{K}c^{(k)}c^{(k')}\beta^{(k')}.\right]$$

Taking all terms involving $\boldsymbol{\beta}$ to one side of the equation, and combining these equations into a system, we see that

$$\left[\mathrm{diag}\left((c^{(k)})^2\frac{m}{m^{(k)}}\right) - \boldsymbol{c}\boldsymbol{c}'\right]\boldsymbol{\beta} = -\frac{rs_{w\hat{Y}^*}}{s_{w\hat{N}}}\mathrm{vec}\left((c^{(k)})^2\frac{m}{m^{(k)}}\right).$$

Since $\boldsymbol{c}$ is a contrast vector, $\boldsymbol{\beta} = -rs_{w\hat{Y}^*}/s_{w\hat{N}}\mathbf{1}$ is a solution. From Lemma 4.5.2, we know that there are other solutions of the form

$$\boldsymbol{\beta} = -\frac{rs_{w\hat{Y}^*}}{s_{w\hat{N}}}\mathbf{1} + a\text{vec}\left(\frac{m^{(k)}c^{(k)}}{m}\right),$$

for any constant $a \in \mathbb{R}$. Since the gradient of the function is $\mathbf{0}$ on this subspace, the value of the variance of the numerator of the test statistic will be the same for all treatment effects in this subspace.

Substituting this solution into the original equation, we see that

$$\begin{aligned}
\text{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}) &= m\left[\left(\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}\left(s^2_{w\hat{Y}^*} + 2\beta^{(k)}rs_{w\hat{Y}^*}s_{w\hat{N}} + (\beta^{(k)})^2 s^2_{w\hat{N}}\right)\right) - \right. \\
&\qquad \left.\left(\sum_{k=1}^{K} c^{(k)}\beta^{(k)}\right)^2 s_{w\hat{N}}\right] \\
&\geq m\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}(s^2_{w\hat{Y}^*} - 2r^2 s^2_{w\hat{Y}^*} + r^2 s^2_{w\hat{Y}^*}) \\
&= m\sum_{k=1}^{K}(c^{(k)})^2\frac{m}{m^{(k)}}s^2_{w\hat{Y}^*}(1-r^2)
\end{aligned}$$

completing the proof. □

Lastly, here is a lemma bounding central moments.

Lemma 4.5.4. *For all $q \geq 1$ we have that for any $r \in \mathbb{R}$,*

$$\sum_{i=1}^{n}|y_i - \bar{y}|^q \leq 2^q\sum_{i=1}^{n}|y_i - r|^q.$$

Proof. First notice

$$\sum_{i=1}^{n}|y_i - \bar{y}|^q = \sum_{i=1}^{n}|y_i - r + r - \bar{y}|^q.$$

By Jensen's inequality, we have that

$$\left|\frac{1}{2}[(y_i - r) + (r - \bar{y})]\right|^q \leq \frac{1}{2}[|y_i - r|^q + |r - \bar{y}|^q],$$

or

$$|y_i - r + r - \bar{y}|^q \leq 2^{q-1}\left[|y_i - r|^q + |r - \bar{y}|^q\right].$$

Substituting in the original form, and applying Jensen's inequality again, we obtain

$$\begin{aligned}
\sum_{i=1}^{n} |y_i - \bar{y}|^q &\leq 2^{q-1}\left[\left(\sum_{i=1}^{n} |y_i - r|^q\right) + n|r - \bar{y}|^q\right] \\
&= 2^{q-1}\left[\left(\sum_{i=1}^{n} |y_i - r|^q\right) + n\left|\frac{1}{n}\sum_{i=1}^{n}(y_i - r)\right|^q\right] \\
&\leq 2^{q-1}\left[\left(\sum_{i=1}^{n} |y_i - r|^q\right) + \frac{n}{n}\sum_{i=1}^{n} |y_i - r|^q\right] \\
&= 2^q \sum_{i=1}^{n} |y_i - r|^q,
\end{aligned}$$

as desired. □

4.5.2 Main Theorem

In Chapter 5, we will consider the case of randomized complete block designs. Two asymptotic formulations are of potential interest in block designs: a fixed number of blocks with an increasing number of EUs within each block, or an increasing number of blocks with a fixed number of EUs within each block. In Chapter 5, we consider the second asymptotic setting of an increasing number of blocks. Here, we consider a completely randomized design with an increasing number of EUs, noting that the extension to the first asymptotic setting of a fixed number of blocks is immediate, because the CRD is a block design with one block.

To show asymptotic normality in the increasing block size case ($m \to \infty$), sufficient conditions are

A1 There are $A, B < \infty$ such that $\frac{\max_k\{m^{(k)}\}}{\min_k\{m^{(k)}\}} < B$ when $m > A$.

A2 There are $A, b, B \in \mathbb{R}^+$ such that for all i $b < w_i < B$ when $m > A$.

A3 For some $\delta > 0$, there are $A, B < \infty$ such that $m^{-1} \sum_{i \in S} |\hat{N}_i|^{2+\delta} \leq B$ when $m > A$.

A4 For some $\delta > 0$, there are $A, B < \infty$ such that $m^{-1} \sum_{i \in S} |\hat{Y}^*_{i+}|^{2+\delta} \leq B$ when $m > A$.

A5 Let r be as defined in Lemma 4.5.3. There are $A < \infty$ and $b > 0$ such that if $m \geq A$, then $s^2_{w\hat{Y}^*}(1 - r^2) \geq b$.

Notes on Assumptions: Assumption *A1* guarantees that the treatment assignment is relatively balanced, with the bound between treatment assignment proportions bounded. This is a desirable property of CRDs, and would be expected to hold. Assumption *A2* ensures that the weights are bounded and not terribly imbalanced. This should be a goal of survey designs. However, if the weights are adjusted for nonresponse, differential nonresponse can result in large variations between weights. Assumption *A3* ensures that the estimated EU sizes are not too variable, which is desirable and should hold in practice with good survey and experimental design. Assumption *A4* states that the distribution of the EU totals has some moment beyond a variance, which is a standard assumption in survey analysis. Assumption *A5* is a technical assumption that rules out cases where the variance of the treatment assignment distribution becomes arbitrarily small. Overall, these assumptions are fairly standard, and should hold in many cases for real surveys, barring excessive nonresponse.

This theorem shows the asymptotic normality of the treatment assignment procedure under additivity (4.6). This implies the asymptotic normality of the randomization procedure, because this is a special case under the null hypothesis.

Theorem 4.5.1. *Under Assumptions* A1–A5 *above, we have that as* $m \to \infty$

$$\left[m\mathbf{C}(\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\mathbf{C}'\right]^{-1/2} \left[\mathbf{C}\hat{\boldsymbol{Y}} - \mathbf{C}\boldsymbol{\beta}\hat{N}\right] \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}).$$

Proof. First, note that by the Cramér-Wold Device, it suffices to show that for any contrast vector $\boldsymbol{v}$,

$$\left[m\boldsymbol{v}'\mathbf{C}(\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\mathbf{C}'\boldsymbol{v}\right]^{-1/2}\left[\boldsymbol{v}'\mathbf{C}\hat{\boldsymbol{Y}} - \boldsymbol{v}'\mathbf{C}\boldsymbol{\beta}\hat{N}\right] \xrightarrow{d} \mathcal{N}(0,1).$$

Since $\boldsymbol{v}'\mathbf{C}$ is always a contrast vector when $\mathbf{C}$ is a contrast matrix, it suffices to show that for any contrast vector $\boldsymbol{c}$ we have

$$\left[m\boldsymbol{c}'(\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{c}\right]^{-1/2}\left[\boldsymbol{c}'\hat{\boldsymbol{Y}} - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}\right] \xrightarrow{d} \mathcal{N}(0,1).$$

Now we will complete the proof by considering Theorem 3 from Hoeffding [1951], and applying it to $\boldsymbol{c}'\hat{\boldsymbol{Y}}$.

To simplify notation of this proof, we will reindex the $\hat{\boldsymbol{Y}}_{i+}$ so that i ranges from 1 to m, ignoring the original indexing from the finite population.

We will begin by defining, using the additivity assumption (4.6),

$$c_m(i,j) = c^{(k)}\frac{m}{m^{(k)}}w_i\hat{Y}^{(k)}_{i+} = c^{(k)}\frac{m}{m^{(k)}}w_i(\hat{Y}^*_{i+} + \hat{N}_i\beta^{(k)})$$

where $k = k(j) = \min\{k^* : \sum_{k=1}^{k^*} m^{(k)} > j\}$. Then we see that if $\pi_m(\cdot)$ is a random function permuting the numbers $1,\ldots,m$, then the treatment assignment distribution of $\boldsymbol{c}'\hat{\boldsymbol{Y}}$ is the same as the distribution under the randomness defined by π_m of

$$\sum_{i=1}^{m} c_m(i,\pi_m(i)),$$

by construction.

Let

$$d_m(i,j) = c_m(i,j) - m^{-1}\sum_{i'=1}^{m} c_m(i',j) - m^{-1}\sum_{j'=1}^{m} c_m(i,j') + m^{-2}\sum_{i'=1}^{m}\sum_{j'=1}^{m} c_m(i',j').$$

Substituting the values for $c_m(i,j)$ in the above and simplifying gives that

$$d_m(i,j) = \frac{m}{m^{(k)}} c^{(k)} \left(w_i \hat{Y}^*_{i+} - \frac{1}{m} \sum_{i' \in S} w_{i'} \hat{Y}^*_{i'+} \right) + \left(\frac{m}{m^{(k)}} c^{(k)} \beta^{(k)} - \boldsymbol{c}'\boldsymbol{\beta} \right) \left(w_i \hat{N}_i - \frac{1}{m} \sum_{i' \in S} w_{i'} \hat{N}_{i'} \right).$$

Now to complete the proof, by Theorem 3 in Hoeffding [1951], it suffices to show that for any $\gamma > 0$, we have that as $m \to \infty$

$$\frac{m^{-1} \sum_{i=1}^m \sum_{j=1}^m |d_m(i,j)|^{2+\gamma}}{\left[m^{-1} \sum_{i=1}^m \sum_{j=1}^m \{d_m(i,j)\}^2 \right]^{(2+\gamma)/2}} \to 0.$$

At this step, we examine the expression $m^{-1} \sum_{j=1}^m |d_m(i,j)|^{2+\gamma}$:

$$\frac{1}{m} \sum_{j=1}^m |d_h(i,j)|^{2+\gamma} \leq \sum_{k=1}^K \frac{m^{(k)}}{m} 2^{1+\gamma} \left[\left(\frac{m}{m^{(k)}} \right)^{2+\gamma} |c^{(k)}|^{2+\gamma} \left| w_i \hat{Y}^*_{i+} - \frac{1}{m} \sum_{i' \in S} w_{i'} \hat{Y}^*_{i'+} \right|^{2+\gamma} + \left| \frac{m}{m^{(k)}} c^k \beta^{(k)} - \boldsymbol{c}'\boldsymbol{\beta} \right|^{2+\gamma} \left| w_i \hat{N}_i - \frac{1}{m} \sum_{i' \in S} w_{i'} \hat{N}_{i'} \right|^{2+\gamma} \right].$$

From Assumptions *A1* and *A2* and Lemma 4.5.4, there are constants B_Y and B_N (which may depend on γ, but not m) such that $m^{-1} \sum_{i=1}^m \sum_{j=1}^m |d_m(i,j)|^{2+\gamma} \leq \sum_{i=1}^m B_Y |\hat{Y}_{i+}|^{2+\gamma} + B_N |\hat{N}_i|^{2+\gamma}$.

Additionally, Theorem 2 of Hoeffding [1951] gives that

$$\text{Var} \left(\sum_{i=1}^m c_m(i, \pi_h(i)) \right) = \frac{1}{m-1} \sum_{i=1}^m \sum_{j=1}^m \{d_m(i,j)\}^2.$$

This implies, using Lemma 4.5.3 that

$$\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m \{d_m(i,j)\}^2 = (m-1) \boldsymbol{c}' (\mathbf{D}_{w\hat{Y}} - s^2_{w\hat{N}} \boldsymbol{\beta}\boldsymbol{\beta}') \boldsymbol{c} \geq (m-1) \sum_{k=1}^K (c^{(k)})^2 \frac{m}{m^{(k)}} s^2_{w\hat{Y}^*} (1 - r^2).$$

By Assumptions *A1* and *A5*, there is some $\epsilon > 0$ such that $m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{m}\{d_m(i,j)\}^2 \geq \epsilon m$. Using the bounds derived above, we obtain

$$\frac{m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{m}|d_m(i,j)|^{2+\gamma}}{\left[m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{m}\{d_m(i,j)\}^2\right]^{(2+\gamma)/2}} \leq \frac{\sum_{i=1}^{m} B_Y|\hat{Y}^*_{i+}|^{2+\gamma} + B_N|\hat{N}_i|^{2+\gamma}}{(\epsilon m)^{(2+\gamma)/2}}.$$

Assumptions *A3* and *A4* and Lemma 4.5.1 give that the right hand side of the above goes to zero for all $\gamma > 0$, completing the proof.

□

4.6 Treatment Assignment and Randomization: Power

Previously, we considered the variance computations for contrasts under the randomization distribution, conditioned on the sample and treatment assignment. In this subsection, we will consider the variance computations under both treatment assignment and randomization. Understanding the distribution under both of these sources of variability allows us estimate power, conditioned on the sample.

To begin, we will partition the variance to understand the variation of the randomization distribution averaged over all possible treatment assignments. This yields

$$\mathrm{Var}_{TR}(\mathbf{C}\tilde{\boldsymbol{Y}}) = \mathrm{Var}_T(\mathrm{E}_R[\mathbf{C}\tilde{\boldsymbol{Y}}]) + \mathrm{E}_T[\mathrm{Var}_R(\mathbf{C}\tilde{\boldsymbol{Y}})].$$

We have already shown that $\mathrm{E}_R[\mathbf{C}\tilde{\boldsymbol{Y}}] = \mathbf{0}$, so the first term is $\mathbf{0}$ and we need only to focus on the second term.

From Theorem 4.4.1, we know that

$$\mathrm{Var}_R(\mathbf{C}\tilde{\boldsymbol{Y}}) = m s^2_{w\hat{Y}}\mathbf{C}\mathbf{D}\mathbf{C}'$$

where $m\mathbf{CDC}'$ is fixed. Therefore we need only focus on $s^2_{w\hat{Y}}$, for which

$$\mathrm{E}_T[s^2_{w\hat{Y}}] = \mathrm{E}_T\left[\frac{1}{2m(m-1)}\sum_{i\in S}\sum_{i'\in S}(w_i\hat{Y}_{i+} - w_{i'}\hat{Y}_{i'+})^2\right].$$

When $i = i'$, $w_i\hat{Y}_{i+} - w_{i'}\hat{Y}_{i'+} = 0$, so we restrict our focus to deriving $\mathrm{E}_T[(w_i\hat{Y}_{i+} - w_{i'}\hat{Y}_{i'+})^2]$ when $i \neq i'$. Note

$$\begin{aligned}\mathrm{E}_T\left[(w_i\hat{Y}_{i+} - w_{i'}\hat{Y}_{i'+})^2\right] &= \mathrm{E}_T\left[\left(w_i(\hat{Y}^*_{i+} + \hat{N}_i\boldsymbol{T}'_i\boldsymbol{\beta}) - w_{i'}(\hat{Y}^*_{i'+} + \hat{N}_{i'}\boldsymbol{T}'_{i'}\boldsymbol{\beta})\right)^2\right] \\ &= (w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})^2 + 2(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(w_i\hat{N}_i\mathrm{E}_T[\boldsymbol{T}'_i\boldsymbol{\beta}] - w_{i'}\hat{N}_{i'}\mathrm{E}_T[\boldsymbol{T}'_{i'}\boldsymbol{\beta}]) + \\ &\quad \mathrm{E}_T[(w_i\hat{N}_i\boldsymbol{T}'_i\boldsymbol{\beta} - w_{i'}\hat{N}_{i'}\boldsymbol{T}'_{i'}\boldsymbol{\beta})^2].\end{aligned}$$

Now we focus on the third term,

$$\begin{aligned}\mathrm{E}_T[(w_i\hat{N}_i\boldsymbol{T}'_i\boldsymbol{\beta} - w_{i'}\hat{N}_{i'}\boldsymbol{T}'_{i'}\boldsymbol{\beta})^2] = w_i^2\hat{N}_i^2\boldsymbol{\beta}'\mathrm{E}_T[\boldsymbol{T}_i\boldsymbol{T}'_i]\boldsymbol{\beta} - w_iw_{i'}\hat{N}_i\hat{N}_{i'}\boldsymbol{\beta}'\mathrm{E}_T[\boldsymbol{T}_i\boldsymbol{T}'_{i'} + \boldsymbol{T}_{i'}\boldsymbol{T}'_i]\boldsymbol{\beta} + \\ w_{i'}^2\hat{N}_{i'}^2\boldsymbol{\beta}'\mathrm{E}_T[\boldsymbol{T}_{i'}\boldsymbol{T}'_{i'}]\boldsymbol{\beta}.\end{aligned}$$

We define $\bar{\beta} = \mathrm{E}_T[\boldsymbol{T}'_i\boldsymbol{\beta}] = m^{-1}\sum_{k=1}^k m^{(k)}\beta^{(k)}$. Recall that $\boldsymbol{d} = (m/m^{(1)}, \ldots, m/m^{(K)})'$ and $\mathbf{D} = \mathrm{diag}(\boldsymbol{d}_h)$. Also, let $\boldsymbol{m} = (m^1, \ldots, m^{(K)})'$. Using results from (4.1), we obtain

$$\begin{aligned}\mathrm{E}_T[\boldsymbol{T}_i\boldsymbol{T}'_i] &= \mathbf{D}^{-1} \\ \mathrm{E}_T[\boldsymbol{T}_i\boldsymbol{T}'_{i'}] &= \frac{m}{m-1}\left(\frac{\boldsymbol{m}\boldsymbol{m}'}{m^2} - \frac{\mathbf{D}^{-1}}{m}\right).\end{aligned}$$

If we let

$$\overline{\beta^2} = \boldsymbol{\beta}'\mathbf{D}^{-1}\boldsymbol{\beta} = m^{-1}\sum_{k=1}^K m^{(k)}(\beta^{(k)})^2$$

then

$$\begin{aligned}\mathrm{E}_T[(w_i\hat{N}_i\boldsymbol{T}'_i\boldsymbol{\beta} - w_{i'}\hat{N}_{i'}\boldsymbol{T}'_{i'}\boldsymbol{\beta})^2] &= (w_i^2\hat{N}_i^2 + w_{i'}^2\hat{N}_{i'}^2)\overline{\beta^2} - \\ &\quad 2w_iw_{i'}\hat{N}_i\hat{N}_{i'}\frac{m}{m-1}(\overline{\beta}^2 - m^{-1}\overline{\beta^2}) \\ &= (w_i\hat{N}_i - w_{i'}\hat{N}_{i'})^2\overline{\beta^2} + 2w_iw_{i'}\hat{N}_i\hat{N}_{i'}\frac{m}{m-1}(\overline{\beta^2} - \overline{\beta}^2).\end{aligned}$$

Putting this all together, we have

$$\begin{aligned}\mathrm{E}_T\left[(w_i\hat{Y}_{i+} - w_{i'}\hat{Y}_{i'+})^2\right] &= (w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})^2 + 2\overline{\beta}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(w_i\hat{N}_i - w_{i'}\hat{N}_{i'}) + \\ &\quad \overline{\beta^2_h}(w_i\hat{N}_i - w_{i'}\hat{N}_{i'})^2 + 2w_iw_{i'}\hat{N}_i\hat{N}_{i'}\frac{m}{m-1}(\overline{\beta^2} - \overline{\beta}^2).\end{aligned}$$

Thus

$$\begin{aligned}\mathrm{E}_T[s^2_{w\hat{Y}}] &= \frac{1}{2m(m-1)}\sum_{i\in S}\sum_{i'\in S\setminus\{i\}}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})^2 + 2\overline{\beta}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(w_i\hat{N}_i - w_{i'}\hat{N}_{i'}) + \\ &\quad \overline{\beta^2}(w_i\hat{N}_i - w_{i'}\hat{N}_{i'})^2 + 2\frac{w_iw_{i'}\hat{N}_i\hat{N}_{i'}}{m-1}(\overline{\beta^2_h} - \overline{\beta}^2_h) \\ &= \frac{1}{2m(m-1)}\sum_{i\in S}\sum_{i'\in S\setminus\{i\}}\left[\left(w_i(\hat{Y}^*_{i+} + \hat{N}_i\overline{\beta}) - w_{i'}(\hat{Y}^*_{i'+} + \hat{N}_{i'}\overline{\beta})\right)^2 + \right. \\ &\quad \left.\left((w_i\hat{N}_i - w_{i'}\hat{N}_{i'})^2 + 2w_iw_{i'}\hat{N}_i\hat{N}_{i'}\frac{m}{m-1}\right)(\overline{\beta^2} - \overline{\beta}^2)\right].\end{aligned}$$

This is the variance of the weighted $\hat{Y}^*_{i+}$, including the average treatment effect, plus the variance of the weighted estimated EU sizes multiplied by the variance of the treatment effects, plus an additional term involving the weighted estimated EU sizes multiplied by the variance of the treatment effects.

Notice that we could define $\overline{\beta} = 0$ here, but we will not impose this constraint to preserve generality for cases with multiple blocks where the treatment may not be balanced. However, if

we define

$$s^2_{w\hat{Y}^*_{\text{adj}}} = \frac{1}{2m(m-1)} \sum_{i \in S} \sum_{i' \in S} \left(w_i(\hat{Y}^*_{i+} + \hat{N}_i \overline{\beta}) - w_{i'}(\hat{Y}^*_{i+} + \hat{N}_{i'} \overline{\beta}) \right)^2,$$

we observe that the above can be rewritten as

$$\mathrm{E}_T[s^2_{w\hat{Y}}] = s^2_{w\hat{Y}^*_{\text{adj}}} + (\overline{\beta^2} - \overline{\beta}^2) s^2_{w\hat{N}} + \frac{\overline{\beta^2} - \overline{\beta}^2}{(m-1)^2} \sum_{i \in S} \sum_{i' \in S \setminus \{i\}} w_i w_{i'} \hat{N}_i \hat{N}_{i'}.$$

Therefore we have

$$\mathrm{E}_T[\mathrm{Var}_R(\mathbf{C}\tilde{\boldsymbol{Y}})] = m \left[s^2_{w\hat{Y}^*_{\text{adj}}} + (\overline{\beta^2} - \overline{\beta}^2) s^2_{w\hat{N}} + \frac{\overline{\beta^2} - \overline{\beta}^2}{(m-1)^2} \sum_{i \in S} \sum_{i' \in S \setminus \{i\}} w_i w_{i'} \hat{N}_i \hat{N}_{i'} \right] \mathbf{C}\mathbf{D}\mathbf{C}'. \tag{4.10}$$

4.6.1 Power Conditioned on the Sample

Conditioned on the sample, we will follow methods in Robinson [1973] which followed Hoeffding [1952]. This approach will involve showing that the variance (when scaled appropriately) converges in probability to a constant.

Conditioned on the sample, there are still two sources of randomness: the randomness due to experimental treatment assignment and variance due to randomizing treatment labels for the randomization test.

To simplify computation, we consider the case where experimental units are sampled with probability proportional to the size of the experimental unit so that the product of the experimental unit weight and the estimated sample size is constant. We write this as

$$w_i \hat{N}_i = W \tag{4.11}$$

for all $i \in S_h$ (we assume also that w_{hi} does not depend on the sample that was selected). One example of this is a common self-weighted design scheme for two-stage samples, where the first

stage is selected with probability proportional to the number of experimental units, and the second stage is selected as a stratified simple random sample with constant size. In this case, the second stage selection probabilities will be such that $\hat{N}_i = N_i$ for each i, but in some other cases there may be calibration to ensure this equality.

Lemma 4.6.1. *Assuming (4.6) and (4.11), we have that under the treatment assignment distribution, the randomization distribution of an arbitrary entry of the randomization covariance matrix satisfies*

$$\begin{aligned}
\mathrm{E}_T[\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})] &= m\left[s^2_{w\hat{Y}^*} + \frac{W^2 m}{m-1}(\overline{\beta^2} - \overline{\beta}^2)\right]\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2 \\
\mathrm{Var}_T(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})) &= \frac{4W^2 m^3}{(m-1)^2} s^2_{w\hat{Y}^*}(\overline{\beta^2} - \overline{\beta}^2)(\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2)^2.
\end{aligned}$$

Proof. We know from earlier calculations that

$$\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}}) = m s^2_{w\hat{Y}}\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2,$$

and the only randomness in this expression is from the term $s^2_{w\hat{Y}}$.

From (4.10), we know that

$$\mathrm{E}_T[\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})] = m\left[s^2_{w\hat{Y}^*_{\mathrm{adj}}} + (\overline{\beta^2} - \overline{\beta}^2)s^2_{w\hat{N}} + \frac{\overline{\beta^2} - \overline{\beta}^2}{(m-1)^2}\sum_{i\in S}\sum_{i'\in S\setminus\{i\}} w_i w_{i'}\hat{N}_i\hat{N}_{i'}\right]\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2.$$

Assuming equation (4.11), it can be shown that $s^2_{w\hat{Y}^*_{\mathrm{adj}}} = s^2_{w\hat{Y}^*}$ and $s^2_{w\hat{N}} = 0$, and thus the above expression simplifies to

$$\mathrm{E}_T[\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})] = m\left[s^2_{w\hat{Y}^*} + \frac{W^2 m}{m-1}(\overline{\beta^2} - \overline{\beta}^2)\right]\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2.$$

To discuss the treatment assignment variance of $\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})$, we again focus on the $s^2_{w\hat{Y}}$ term. Specifically, we note that we can rewrite $2m(m-1)s^2_{w\hat{Y}}$ as

$$\sum_{i\in S}\sum_{i'\in S}(w_i\hat{Y}_{i+} - w_{i'}\hat{Y}_{i'+})^2 = \sum_{i\in S}\sum_{i'\in S}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})^2 + 2(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(w_i\hat{N}_i\boldsymbol{T}_i'\boldsymbol{\beta} - w_{i'}\hat{N}_{i'}\boldsymbol{T}_{i'}'\boldsymbol{\beta}) + (w_i\hat{N}_i\boldsymbol{T}_i'\boldsymbol{\beta} - w_{i'}\hat{N}_{i'}\boldsymbol{T}_{i'}'\boldsymbol{\beta})^2.$$

In this expression, the first term is not random. Additionally, if we assume (4.11), we see by counting terms that the last term is

$$\sum_{i\in S}\sum_{i'\in S}W^2(\boldsymbol{T}_i'\boldsymbol{\beta} - \boldsymbol{T}_{i'}'\boldsymbol{\beta})^2 = W^2\sum_{k=1}^{K}\sum_{k'=1}^{K}m^{(k)}m^{(k')}(\beta^{(k)} - \beta^{(k')})^2,$$

which is also not random. Therefore we will focus on the middle term for the rest of this argument. Assuming (4.11), this is

$$2W\sum_{i\in S}\sum_{i'\in S}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(\boldsymbol{T}_i'\boldsymbol{\beta} - \boldsymbol{T}_{i'}\boldsymbol{\beta}).$$

Taking the covariances term-by-term (and ignoring for now the constant of $2W$), we find

$$\begin{aligned}
&\mathrm{Var}_T\left(\sum_{i\in S}\sum_{i'\in S}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(\boldsymbol{T}_i'\boldsymbol{\beta} - \boldsymbol{T}_{i'}'\boldsymbol{\beta})\right)\\
&= \sum_{i\in S}\sum_{i'\in S}\sum_{j\in S}\sum_{j'\in S}(w_i\hat{Y}^*_{i+} - w_{i'}\hat{Y}^*_{i'+})(w_j\hat{Y}^*_{j+} - w_{j'}\hat{Y}^*_{j'+})\mathrm{Cov}_T(\boldsymbol{T}_i'\boldsymbol{\beta} - \boldsymbol{T}_{i'}'\boldsymbol{\beta}, \boldsymbol{T}_j'\boldsymbol{\beta} - \boldsymbol{T}_{j'}'\boldsymbol{\beta})\\
&= \sum_{i\in S}\sum_{i'\in S}\sum_{j\in S}\sum_{j'\in S}[w_i\hat{Y}^*_{i+}w_j\hat{Y}^*_{j+} - w_i\hat{Y}^*_{i+}w_{j'}\hat{Y}^*_{j'+} - w_{i'}\hat{Y}^*_{i'+}w_j\hat{Y}^*_{j+} + w_{i'}\hat{Y}^*_{i'+}w_{j'}\hat{Y}^*_{j'+}]\bullet\\
&\qquad [\boldsymbol{\beta}'\mathrm{Cov}_T(\boldsymbol{T}_i, \boldsymbol{T}_j)\boldsymbol{\beta} - \boldsymbol{\beta}'\mathrm{Cov}_T(\boldsymbol{T}_i, \boldsymbol{T}_{j'})\boldsymbol{\beta} - \boldsymbol{\beta}'\mathrm{Cov}_T(\boldsymbol{T}_{i'}, \boldsymbol{T}_j)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathrm{Cov}_T(\boldsymbol{T}_{i'}, \boldsymbol{T}_{j'})\boldsymbol{\beta}],
\end{aligned}$$

where $\bullet$ denotes multiplication across lines.

Using the covariance expressions in (4.4), we find that, if $\overline{\beta^2} = m^{-1}\sum_{k=1}^{K} m^{(k)}(\beta^{(k)})^2$,

$$\boldsymbol{\beta}'\mathrm{Cov}_T(\boldsymbol{T}_i, \boldsymbol{T}_{i'})\boldsymbol{\beta} = \begin{cases} \overline{\beta^2} - \overline{\beta}^2 & : \quad i = i' \\ \frac{-1}{m-1}(\overline{\beta^2} - \overline{\beta}^2) & : \quad i \neq i'. \end{cases}$$

To understand this notation, notice that the signs in the first term correspond to the signs in the second term, so that the sum just obtained is the sum of four copies of the same expression. We will proceed by evaluating the first of these copies, and will later multiply the result by four:

$$\begin{aligned}
&\sum_{i\in S}\sum_{i'\in S}\sum_{j\in S}\sum_{j'\in S}[w_i\hat{Y}^*_{i+}w_j\hat{Y}^*_{j+} - w_i\hat{Y}^*_{i+}w_{j'}\hat{Y}^*_{j'+} - w_{i'}\hat{Y}^*_{i'+}w_j\hat{Y}^*_{j+} + w_{i'}\hat{Y}^*_{i'+}w_{j'}\hat{Y}^*_{j'+}] \bullet \\
&\qquad\qquad \boldsymbol{\beta}'\mathrm{Cov}_T(\boldsymbol{T}_i, \boldsymbol{T}_j)\boldsymbol{\beta} \\
&= (\overline{\beta^2} - \overline{\beta}^2)\sum_{i\in S}\left[m^2w_i^2(\hat{Y}^*_{i+})^2 - 2mw_i\hat{Y}^*_{i+}\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+} + \left(\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+}\right)^2\right] - \frac{(\overline{\beta^2} - \overline{\beta}^2)}{m-1} \\
&\quad \sum_{i\in S}\sum_{j\in S\setminus\{i\}}\left[m^2w_i\hat{Y}^*_{i+}w_j\hat{Y}^*_{j+} - m(w_i\hat{Y}^*_{i+} + w_j\hat{Y}^*_{j+})\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+} + \left(\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+}\right)^2\right] \\
&= \frac{m^2(\overline{\beta^2} - \overline{\beta}^2)}{m-1}\left[(m-1)^2 s^2_{w\hat{Y}^*} - \right. \\
&\qquad\qquad \left. \sum_{i\in S}\sum_{j\in S\setminus\{i\}}\left(w_i\hat{Y}^*_{i+} - \frac{1}{m}\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+}\right)\left(w_j\hat{Y}^*_{j+} - \frac{1}{m}\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+}\right)\right],
\end{aligned}$$

where we have used

$$s^2_{w\hat{Y}^*} = \frac{1}{m-1}\sum_{i\in S}\left(w_i\hat{Y}^*_{i+} - \frac{1}{m}\sum_{i'\in S}w_{i'}\hat{Y}^*_{i'+}\right)^2.$$

Additionally, since

$$\sum_{i\in S}\sum_{j\in S}(y_i - \bar{y})(y_j - \bar{y}) = \left[\sum_{i\in S}(y_i - \bar{y})\right]^2 = 0,$$

we have that the expression above equals

$$\frac{m^2(\overline{\beta^2}-\overline{\beta}^2)}{m-1}[(m-1)^2 s^2_{w\hat{Y}*}+(m-1)s^2_{w\hat{Y}*}] = m^3 s^2_{w\hat{Y}*}(\overline{\beta^2}-\overline{\beta}^2).$$

Replacing the constants that have been ignored, we see

$$\mathrm{Var}_T(s^2_{w\hat{Y}}) = 4\left(\frac{2W}{2m(m-1)}\right)^2 m^3 s^2_{w\hat{Y}*}(\overline{\beta^2}-\overline{\beta}^2) = \frac{4W^2 m}{(m-1)^2} s^2_{w\hat{Y}*}(\overline{\beta^2}-\overline{\beta}^2).$$

Therefore we conclude that

$$\mathrm{Var}_T(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})) = (m\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2)^2 \mathrm{Var}_T(s^2_{w\hat{Y}}) = \frac{4W^2m^3}{(m-1)^2} s^2_{w\hat{Y}*}(\overline{\beta^2}-\overline{\beta}^2)(\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2)^2.$$

□

Using the above results, we state the following theorem providing the following conclusions when m is large, under the moderate assumptions we have been using: the expected value of the randomization variance for any given treatment assignment is close to the expected randomization variance averaging over all possible treatment assignments, and the variation in the estimated correlations are negligible. This suggests that for large samples, the computed correlation matrix for the randomization distribution will have asymptotically negligible variation under the treatment assignment distribution.

Theorem 4.6.1. *Assume (4.6) and (4.11), and Assumptions* A1–A5. *Then for any nonzero contrast vectors* $\boldsymbol{c}$, $\boldsymbol{c}_1$, *and* $\boldsymbol{c}_2$, *we have*

1.

$$\frac{\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})}{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})\right]} \xrightarrow{p} 1.$$

2.

$$\frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right)}}{\sqrt{\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}})\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}})}} \xrightarrow{p} 0.$$

Proof. From the previous lemma,

$$\mathrm{E}_T[\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})] = m\left[s^2_{w\hat{Y}^*} + \frac{W^2 m}{m-1}(\overline{\beta^2} - \overline{\beta}^2)\right]\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2,$$
$$\mathrm{Var}_T(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})) = \frac{4W^2m^3}{(m-1)^2}s^2_{w\hat{Y}^*}(\overline{\beta^2} - \overline{\beta}^2)(\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2)^2.$$

Before the asymptotic analysis, we will show that $\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})\right] \neq 0$. From the expression, we see that the expectation can only be zero if $\boldsymbol{c}'\mathbf{D}\boldsymbol{c} = 0$ or if both $s^2_{w\hat{Y}^*} = 0$ and $\boldsymbol{\beta} = \beta\mathbf{1}$ for some $\beta \in \mathbb{R}$. Since $\boldsymbol{c}$ is nonzero and $\mathbf{D}$ is positive definite, we have that $\boldsymbol{c}'\mathbf{D}\boldsymbol{c} > 0$. Also, Assumption *A5* implies that $s^2_{w\hat{Y}^*}$ is bounded away from zero. Thus we conclude the denominator is nonzero.

Based on the formulas above, we obtain

$$\begin{aligned}\frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})\right)}}{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})\right]} &= \frac{2W\frac{\sqrt{m}}{m-1}s_{w\hat{Y}^*}\sqrt{\overline{\beta^2} - \overline{\beta}^2}}{s^2_{w\hat{Y}^*} + W^2\frac{m}{m-1}(\overline{\beta^2} - \overline{\beta}^2)} \\ &= \frac{1}{\sqrt{m-1}}\left(\frac{2s_{w\hat{Y}^*}W\sqrt{\frac{m}{m-1}(\overline{\beta^2} - \overline{\beta}^2)}}{s^2_{w\hat{Y}^*} + W^2\frac{m}{m-1}(\overline{\beta^2} - \overline{\beta}^2)}\right).\end{aligned}$$

If we let $a = s^2_{w\hat{Y}^*} > 0$ and $b = W^2\frac{m}{m-1}(\overline{\beta^2} - \overline{\beta}^2) > 0$, the expression in parantheses is $2\sqrt{ab}/(a+b)$. Since for positive real numbers the arithmetic mean is always greater than or equal to the geometric mean, as $m \to \infty$

$$\frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Var}_R(\mathbf{c}'\tilde{\boldsymbol{Y}})\right)}}{\mathrm{E}_T\left[\mathrm{Var}_R(\mathbf{c}'\tilde{\boldsymbol{Y}})\right]} \leq \frac{1}{\sqrt{m-1}} \to 0.$$

It follows from Chebyshev's inequality that this is a sufficient condition for

$$\frac{\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})}{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})\right]} \xrightarrow{p} 1.$$

For the second part,

$$\frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right)}}{\sqrt{\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}})\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}})}} = \frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right)}}{\sqrt{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}})\right]\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right]}} \frac{\sqrt{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}})\right]\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right]}}{\sqrt{\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}})\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}})}}.$$

By the result just proven, we know that the second part converges in probability to 1. Therefore, we focus on the first part.

For the standard error,

$$\begin{aligned}\frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right)}}{\sqrt{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}})\right]\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}})\right]}} &= \frac{2s_{w\hat{Y}^*}W\frac{\sqrt{m}}{m-1}\sqrt{\overline{\beta^2}-\overline{\beta}^2}}{s^2_{w\hat{Y}^*}+W^2\frac{m}{m-1}(\overline{\beta^2}-\overline{\beta}^2)}\frac{\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_2}{\sqrt{(\boldsymbol{c}_1'\mathbf{D}\boldsymbol{c}_1)(\boldsymbol{c}_2'\mathbf{D}\boldsymbol{c}_2)}} \\ &\leq \frac{1}{\sqrt{m-1}} \to 0\end{aligned}$$

by the arguments above and properties of inner products.

□

If we consider a single contrast and a one sample test, the above result implies

$$P_T\left(\frac{\boldsymbol{c}'\hat{\boldsymbol{Y}}-0}{\sqrt{\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})}} \geq \Phi^{-1}(1-\alpha)\right).$$

Under the treatment assignment distribution, $\boldsymbol{c}'\hat{\boldsymbol{Y}}$ approximately follows a normal distribution with mean $\boldsymbol{c}'\boldsymbol{\beta}\hat{N}$ and variance $\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}})$, defined earlier. Thus we find the power is

$$P_T\left(\frac{\boldsymbol{c}'\hat{\boldsymbol{Y}}-\boldsymbol{c}'\boldsymbol{\beta}\hat{N}}{\sqrt{\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}})}} \geq \frac{\Phi^{-1}(1-\alpha)\sqrt{\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})}-\boldsymbol{c}'\boldsymbol{\beta}\hat{N}}{\sqrt{\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}})}}\right).$$

From the above result, we can derive the following approximate formula for power, which can be known from the sample and the true treatment effect size, and is fixed conditioned on the sample:

$$\text{Power} \approx 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)\sqrt{\text{E}_T[\text{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}})]} - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}}{\sqrt{\text{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}})}}\right).$$

For a two sided test, one can replace α with $\alpha/2$ and add the two pieces corresponding to both sides of the test.

4.7 Extending Inference to the Finite Population

In the previous chapters, we showed that the test conditioned on the sample admits a central limit theorem conditioned on the sample. In this section, we will argue that this inference can be extended to the finite population as well. To do this, we will demonstrate the distribution of the estimators given a finite population.

The finite population limit theorem corresponds to the following infeasible test on the finite population: 1) take a sample of size m from a population of size M using a probability sampling design, 2) assign treatments to the sampled units in a completely randomized design, 3) randomize the treatment labels independently of and with the same distribution as the true treatment assignment (the randomized treatment labels will not be independent of each other, just independent of the true treatment assignments). We will also assume in this discussion that the sampling design is a fixed-size design, at least at the experimental unit level.

4.7.1 Notation and Lemmas

To simplify notation, we will denote N as the size of the finite population and n as the size of the sample drawn. As shown above, the expected value of the contrast test statistic under the treatment assignment distribution is

$$\text{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}] = \mathbf{C}\boldsymbol{\beta}\sum_{i\in S} w_i\hat{N}_i = \mathbf{C}\boldsymbol{\beta}\hat{N},$$

or the contrast of treatment effects multiplied by the estimated number of observational units in the population. The contrast of treatment effects is a constant, and under many sampling designs, the estimated total number of observational units in the sample is asymptotically normal. This leads to the reasonable assumption that $mM^{-2}\mathrm{Var}_S(\sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\hat{N}_{hi}) = \sigma_N^2$. The full list of assumptions to prove asymptotic normality is below.

*A1** The ratio of treatment assignments is not too unbalanced, *i.e.* there is a $B < \infty$ such that $\max_k m^{(k)} / \min_k m^{(k)} < B$ for all possible sequences of samples.

*A2** The sampling design is such that there is a $b > 0$ and $A, B < \infty$ for which $b < w_i < B$ for all experimental units i in the finite populations, uniformly for all $m > A$.

*A3** There is some $B < \infty$ such that the sequence of finite populations and sampling designs satisfies $\sum_{m=1}^{\infty} P\left(m^{-1}\sum_{i\in S}|\hat{N}_i|^{2+\delta} \geq B\right) < \infty$.

*A4** There is some $B < \infty$ such that the sequence of finite populations and sampling designs satisfies $\sum_{m=1}^{\infty} P\left(m^{-1}\sum_{i\in S}|\hat{Y}^*_{i+}|^{2+\delta} \geq B\right) < \infty$.

*A5** For some $\epsilon > 0$, the sequence of finite populations and sampling designs satisfies

$$P\left(s^2_{w\hat{Y}^*}(1-r^2) < \epsilon\right) < \infty.$$

*A6** The sampling design is such that the estimator of the number of observational units (OUs) in the population is asymptotically normal.

$$\frac{m^{1/2}}{M}(\hat{N} - N) \xrightarrow{d} \mathcal{N}(0, \sigma_N^2).$$

The variance in the limiting normal distribution is allowed to be zero.

The condition on the weights is already a condition on the sampling design, and the assumptions on independence of blocks and balance of treatments only apply to the experimental design. Therefore, if we have that for almost all sequences of samples drawn from the sampling design

the assumptions on the distributions of $\hat{Y}^*_{hi+}$, $\hat{N}_{hi}$, and the assumption ensuring that the variance is positive hold for all "large" m, that will suffice to show that the conditional CLT will hold for almost all sequences of samples as well.

Given that we have shown above that the test statistic (conditioned on the sample) converges to a normal distribution, we can show that under sampling designs satisfying Assumption *A6**, the distribution of the test statistic (conditioned on the finite population) is asymptotically normal as well. This result is stated in the following lemma.

Lemma 4.7.1. *From Theorem 1.3.6 in Fuller [2009].*
Let $\{\mathcal{F}_N\}$ be a sequence of finite populations and $\{\mathcal{S}_N\}$ be a sequence of samples, and let $\hat{\theta}_N$ be a function of the elements of the sample, $\tilde{\theta}_N$ be $\hat{\theta}_N$ calculated on a permuted version of the sample, and θ_N be a sequence of functions of the elements of $\mathcal{F}_N$. Then suppose that we have, conditionally on the sequence $\{\mathcal{F}_N\}$

$$s_{1,N}^{-1}(\tilde{\theta}_N - \hat{\theta}_N)|\mathcal{S}_N \xrightarrow{d} \mathcal{N}(0,1) \text{ almost surely}$$
$$s_{2,N}^{-1}(\hat{\theta}_N - \theta_N)|\mathcal{F}_N \xrightarrow{d} \mathcal{N}(0,1)$$

where in the first line, the randomness is induced by the possible sequences of samples drawn from the sequences of finite population.

It then follows that

$$(s_{1,N}^2 + s_{2,N}^2)^{-1/2}(\tilde{\theta}_N - \theta_N)|\mathcal{F}_N \xrightarrow{d} \mathcal{N}(0,1).$$

Now that we have proven the preliminary results, we can state and prove the main theorem.

Theorem 4.7.1. *Under Assumptions* A1*–A6* *above, the distribution of $\mathbf{C}\hat{\boldsymbol{Y}}$ is asymptotically normal with respect to both the sampling distribution and the treatment assignment distribution.*

Proof. The proof is an application of Theorem 1.3.6 in Fuller [2009].

It is clear that Assumptions *A1** and *A2** state that Assumptions *A1* and *A2* hold for almost all sequences of samples. Assumptions *A3*–A5** verify through the first Borel-Cantelli lemma that almost surely, under the randomness due to the sampling design, there will be some $A > 0$ such that Assumptions *A3–A5* will hold for all $m > A$. Assumption *A6** shows that the asymptotic distribution of the estimated population size is normal. This implies that the estimate of the treatment effect $\mathbf{C}\boldsymbol{\beta}\hat{N}$ is asymptotically normal as well, completing the required assumptions for Lemma 4.7.1. □

Since $\boldsymbol{\beta} = \mathbf{0}$ under the null hypothesis, the additional variance added due to sampling under the null hypothesis is exactly 0. Thus the above theorem shows that when the null hypothesis is true, the asymptotic distribution of the randomization test conditioned on the sample is the same as the asymptotic distribution of the test statistic for an infeasible experiment where a sample is repeatedly drawn from a population and experiments are assigned for each sample. This result shows that inference conditioned on the sample does actually apply to the finite population when a probability sample is selected.

One interesting consequence of this lemma is that if $\mathrm{Var}(\hat{N}) = o(M^2 m^{-1})$, then the variance due to sampling is negligible. This implies that, when the normality assumptions are met, the procedure of randomizing the treatment assignments is asymptotically equivalent in distribution to a procedure that involves taking a random sample from the population several times, and then assigning treatments each time.

4.8 Simulation Experiments

We investigate the theory above, for linear contrasts in completely randomized designs, via simulation. For simplicity of exposition, we include in this section additional simulation results for nonlinear test statistics, the theory for which is discussed in Chapter 6.

4.8.1 Simulated Population

As in Section 3.4, we created an artificial population based on data downloaded from the public use microdata sample (PUMS) dataset (U.S. Census Bureau [2021]) for four PUMAs in northeastern Colorado. Data were recoded using the same methods as in Section 3.4. Most importantly, the Hispanic ethnicities were recoded into a binary flag for Hispanics. A propensity to respond was then computed for each record, using record-level data in the following model:

$$\begin{aligned} p_i = \text{expit}\{&2.5 - 1.0(\text{Age18-34})_i - 0.5(\text{Age35-49})_i - 0.5(\text{RaceBlack})_i - 0.5(\text{RaceAIAN})_i - \\ &0.2(\text{RaceAPI})_i - 0.2(\text{RaceOther})_i - 0.2(\text{RaceMulti})_i - 0.5(\text{HispanicYes})_i - \\ &0.5(\text{MaritalDivorced}) - 0.5(\text{MaritalSeparated})_i - 1.0(\text{MaritalSingle})_i - \\ &0.3(\text{IncomeNegative})_i - 1.0(\text{Income0-15K})_i - 0.6(\text{Income15-50K})_i - \\ &0.2(\text{Income50-100K})_i - 1.0(\text{EducNoHS})_i - 0.8(\text{EducHS/Equiv})_i - \\ &0.3(\text{EducSomeCollege/Assoc})_i - 0.1(\text{EducBachelor})_i - 2.0(\text{HispeedNA})_i - \\ &1.2(\text{HispeedNo})_i - 0.9(\text{PUMA100})_i - 0.9(\text{PUMA400})_i + 2Z_i]\}, \end{aligned}$$

where Z_i is a standard normal random variable. After the propensities were calculated, data were cut using a procedure to generate 615 artificial "districts" of at least 500 people each. Out of the 615 districts, 105 were in PUMA 100, 215 were in PUMA 103, 180 were in PUMA 300, and 115 were in PUMA 400. The study was designed so that there was no correlation between district size and any of the variables in the district.

4.8.2 Variables and Settings

We considered both weighted and unweighted analysis of the experiment. Weighted analysis is expected to be important when weights are related to the study variables of interest in the experiment, and not already accounted otherwise (e.g., via blocking in the design or analysis of covariance). Accordingly, we created study variables with different amounts of correlation with

the weights, via the equation

$$y_i = 300 + 5a + 2ab\frac{w_i - \bar{w}}{\bar{w}} + 35z_i, \tag{4.12}$$

where $\bar{w}$ is the mean of the weights, s_w is the standard deviation of the weights, z_i is a standard normal random variable, and a and b are constants that determine the treatment effect and the variation of the treatment effect based on the weights. While the construction of these variables is artificial, it is common in practice to have variables that are correlated with the weights.

This simulation included 4 settings each for a and b: $a = 0, 1, 2, 3$ and $b = 0, 1, 2, 3$. The setting where $a = 0$ corresponds to no treatment effect for all values of b. The setting $a = 3$ gives an average treatment effect of 12 across all elements. The setting $b = 0$ corresponds to the case where the treatment is constant for all units. The setting $b = 3$ corresponds to a treatment effect that varies greatly with weights, with larger treatment effects with larger weights. Setting a or b equal to 1 or 2 represent intermediate cases.

Binary variables were also created for this study and were defined as 1 if $y_i > 350$ and 0 otherwise.

4.8.3 Overview of Studies

We used the simulated population in two studies. In the first study, we compared two approximations for the randomization distribution (the distribution of the the test statistic under randomized treatment labels): the Monte Carlo approximation based on 1000 randomizations and the asymptotic normal approximation given by Theorem 4.5.1 above. In the second study, we compared the size and power properties of the survey-weighted randomization test proposed here to other competing methods.

For both studies, we considered four different test statistics: difference in means, difference in log odds, odds ratio, and difference in the dissimilarity index. For the power curves, we compared the survey weighted methods to their non-weighted equivalents. The methods included in this study are Van den Brakel's design-based method (VdB), the randomization method included in this paper

(both the normal approximation, Rdmz, and the Monte Carlo version, Rdmz MC), and versions of both of these methods ignoring the survey weights (VdB uw and Rdmz uw). Additionally, for the difference in means tests, we also included differences of estimated totals using both van den Brakel's design-based method (VdB tot) and the randomization test discussed in this dissertation (Rdmz tot).

4.8.4 Sampling Plan and Treatment Assignment

For the simulation study evaluating asymptotic normality, we started with three different samples with treatments assigned. The samples were a two-stage samples. The first-stage sample of 20, 40, or 60 districts with probability proportional to the proportion of Hispanics in the district. The second stage was a sample of 10, 20, or 30 individuals, respectively, within each district by simple random sampling. Treatments were assigned to districts by assigned half (10, 20, or 30, respectively) of the districts to the treatment, and the other half to the control.

The asymptotic normality was tested by repeatedly randomizing treatment labels and computing the test statistic, fora total of 1000 statistics. This Monte Carlo distribution was compared to the theoretical distribution derived in Theorems 4.4.1 and 4.5.1. The variables used in this study were from generated from equation (4.12) with $a = b = 2$.

For the simulation study evaluating power, we evaluated the methods by repeatedly drawing a sample from a population, assigning treatments, and conducting a hypothesis test. In each repetition we again selected a two-stage sample. The first-stage sample of 60 districts with probability proportional to the proportion of Hispanics in the district. The second stage was a sample of 30 individuals within each district by simple random sampling.

Treatments were assigned to districts by a completely randomized design, randomly assigning 30 districts to the treatment and 30 districts to the control. Experimental units (EUs) are then districts and observational units (OUs) are individuals.

The simulation study for power consisted of 1000 replications of the full process of drawing a sample, assigning treatments, and analyzing the results with each of the methods discussed. Power

curves were generated by using different variables generated by equation (4.12), varying a from 0 to 3 with a fixed value of b.

4.8.5 Results

Asymptotic Normality

To assess asymptotic normality, we included three settings with sample sizes of 20, 40, and 60 EUs total for each of the four statistics. Figure 4.1 shows the normal approximation (solid curve) from Theorem 4.4.1 and a kernel density estimate (KDE) for the Monte Carlo approximation (dashed curve). For all four statistics (rows), the KDE is a good approximation to the normal distribution for the largest number of EUs, shown in the last column. The approximation is not as good at the smallest number of EUs, shown in the first column.

One complication in this study is that in the middle case, $m = 40$, there was a district that was an outlier that had the largest value and a weight that was about two and a half times the size of the next smallest weight (Figure 4.2). Including this outlier resulted in a bimodal distribution for all of the statistics (Figure 4.3 is a representative example). Therefore, the outlier was removed in the final figures. The removal of the outlier is justified in this case because Assumption *A2* for asymptotic normality requires that the weights are bounded. Therefore, the inclusion of unusually large weights would not be theoretically guaranteed to provide asymptotic normality.

Additionally, we notice a spike in the kernel density estimate at the maximum and minimum of the distributions for the difference in DI at size $m = 39$. While it is surprising that this effect is not also seen at size $m = 20$, one possible explanation for this behavior has to do with a property of the DI that we now discuss.

Borrowing notation from Chapter 3, the difference in the DI for a variable with two categories and an experiment with two treatments between groups that were randomly relabeled is

$$\tilde{D}^{(1)} - \tilde{D}^{(2)} = \frac{1}{2}\left(|\tilde{p}_1^{(1)} - \alpha_1| + |\tilde{p}_2^{(1)} - \alpha_2| - |\tilde{p}_1^{(2)} - \alpha_1| - |\tilde{p}_2^{(2)} - \alpha_2|\right).$$

Since there are only two proportions, we have $\tilde{p}_1^{(k)} + \tilde{p}_2^{(k)} = \alpha_1 + \alpha_2 = 1$, for $k = 1, 2$. Thus this expression simplifies to

$$\tilde{D}^{(1)} - \tilde{D}^{(2)} = |\tilde{p}_1^{(1)} - \alpha_1| - |\tilde{p}_1^{(2)} - \alpha_1|.$$

Now if $\tilde{p}_1^{(1)}$ and $\tilde{p}_1^{(2)}$ are both on the same side of α_1, or $\text{sign}(\tilde{p}_1^{(1)} - \alpha_1) = \text{sign}(\tilde{p}_1^{(2)} - \alpha_1)$, the absolute value has the same effect on both terms, and the above expression becomes $\tilde{D}^{(1)} - \tilde{D}^{(2)} = \text{sign}(\tilde{p}_1^{(1)} - \alpha_1)(|\tilde{p}_1^{(1)}| - |\tilde{p}_1^{(2)}|)$. However, when $\text{sign}(\tilde{p}_1^{(1)} - \alpha_1) \neq \text{sign}(\tilde{p}_1^{(2)} - \alpha_1)$, the absolute value has a different effect on both terms, and thus we have

$$\tilde{D}^{(1)} - \tilde{D}^{(2)} = \text{sign}(\tilde{p}_1^{(1)} - \alpha_1)(\tilde{p}_1^{(1)} + \tilde{p}_1^{(2)} - 2\alpha_1).$$

To see how these calculations affect the graphical results, we recall that $\tilde{p}_1^{(1)}$ and $\tilde{p}_1^{(2)}$ are both proportions of randomly assigned groups from the same sample. Thus if the sample did not align with the target population, the means of the distributions of $\tilde{p}_1^{(k)}$, which would approximately be the sample mean $\hat{p}_1$, would be on the same side of α_1. However, if the sample mean $\hat{p}_1$ were close to α_1, then the sampling variability would push to a situation where α_1 is between $\tilde{p}_1^{(1)}$ and $\tilde{p}_1^{(2)}$. If the sample were approximately balanced, then $\tilde{p}_1^{(1)} + \tilde{p}_1^{(2)} \approx 2\hat{p}_1$ for all samples. This would set the extreme values of this distribution at approximately $\pm|2\hat{p}_1 - 2\alpha_1|$, which would give a distribution like that seen in the case $m = 39$.

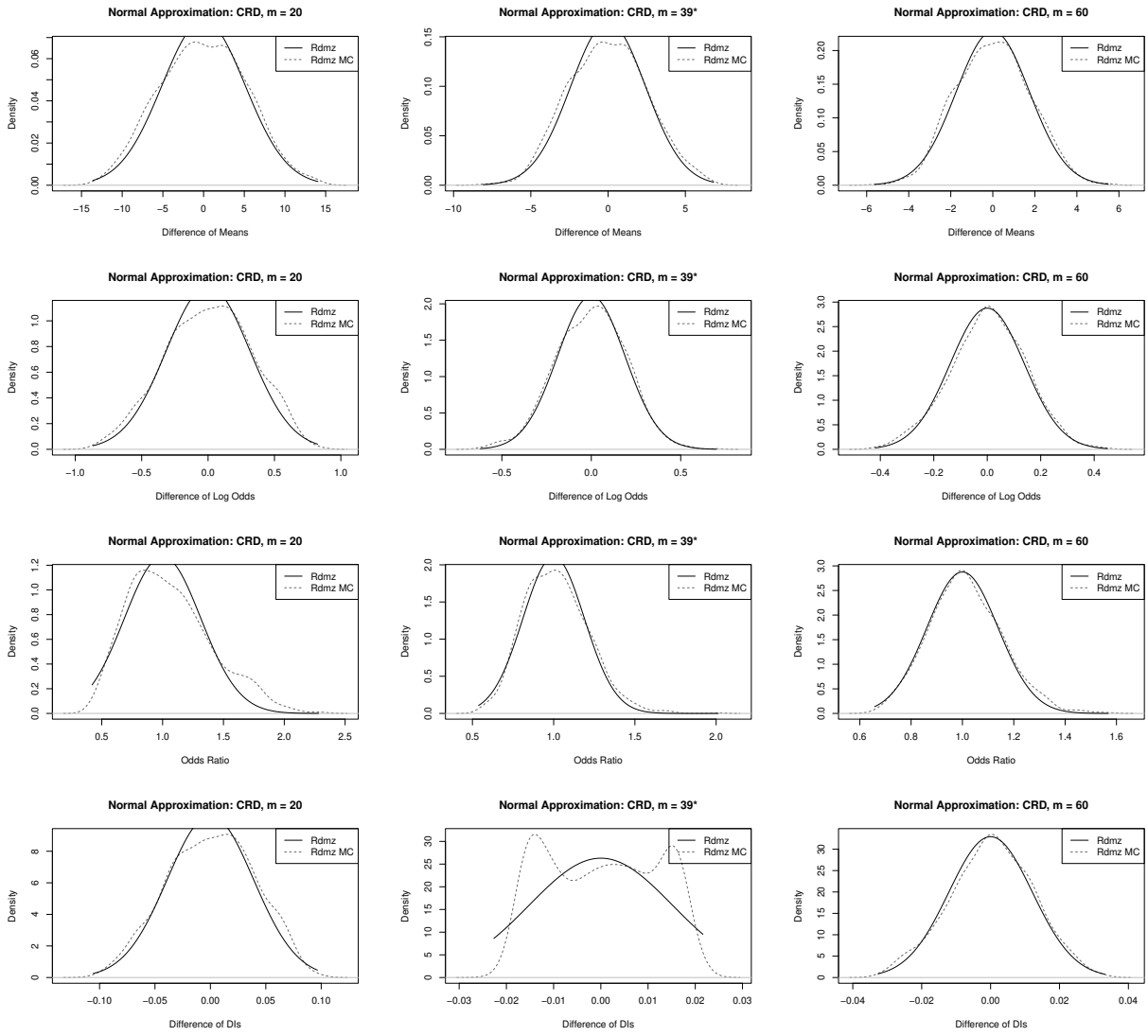

Figure 4.1: Curves showing the normal approximation to the statistic, and a kernel density estimate of the Monte Carlo distribution. The Monte Carlo distribution is simulated with 1000 draws. The rows are different statistics (top-bottom: difference of means, difference of log odds, odds ratio, and difference of DIs), and the columns are different sample sizes (left-right: 20, 39, and 60). The second sample size is 39 rather than 40 due to an outlier.

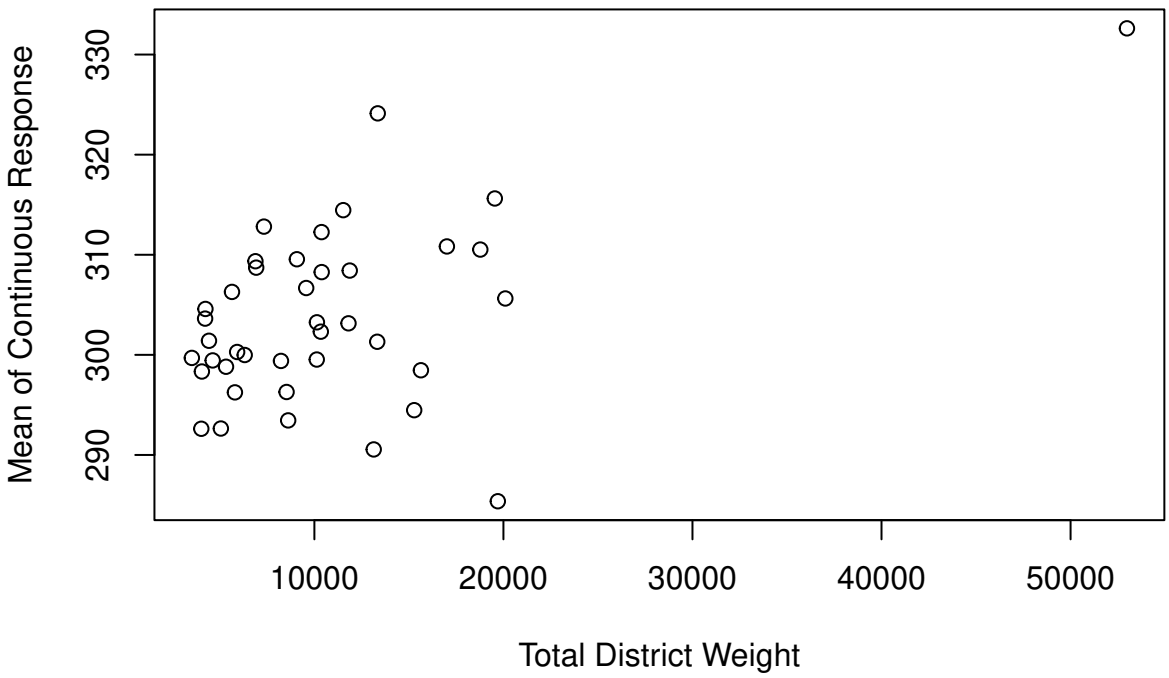

Figure 4.2: Scatterplot of the means of each of the EUs versus the total weights of the EUs in the sample of 40 EUs for the CRD normality simulation.

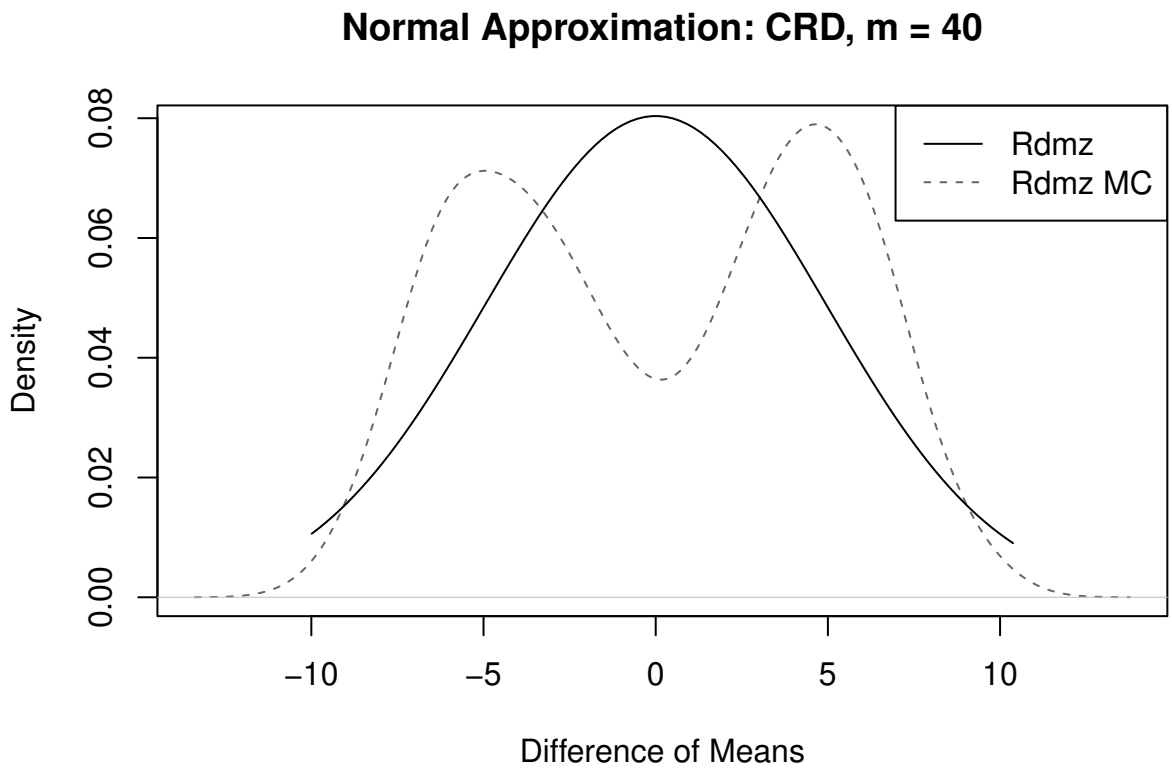

Figure 4.3: Distribution of the difference of means for $m = 40$, including the outlier.

Power Curves

For all of the statistics, the unweighted methods outperform the weighted methods on each of the statistics in the case where the weights are unrelated to the variables of interest ($b = 0$; Figure 4.4). By contrast, when there is a large positive relationship between the effect and the weights in the second column ($b = 3$), the effect is reversed. We additionally see that the weighted methods maintain size in all of the statistics except for the difference of DIs.

While the methods held size for most of the statistics, only the Monte Carlo test held size for the difference of DIs. One reason of this is that the null case was a situation where the population dissimilarity index was close to 0 for both treatment and control, which is a difficult case to linearize the DI (see Chapter 3 for more discussion). The Monte Carlo test does not rely on

linearization, and thus can do fine. As one of the estimates gets further from 0, the linearization works better, and we thus see the power increase.

It may be surprising to some that the unweighted tests held their size when there was no treatment effect for all statistics except for the difference in DIs. One explanation for this has to do with the nature of experimental designs embedded in surveys. Since the treatments are assigned randomly, on a dataset, if there is a bias due to weighting in estimating totals, such bias will be equally present in all treatments. This means that if the null hypothesis (4.9) holds that there is no difference between the treatment and the control groups, the bias in estimating the totals for each group will be removed by multiplying by a contrast.

The observation of weighted randomization tests having better power when there is a strong positive relationship between the weights and the treatment effect can be explained by looking at the correlations. When there is a strong positive correlation between the treatment effect and the weights, then units with a stronger treatment effect are less likely to be sampled, so the unweighted analysis will underestimate the treatment effect. This underestimation leads to smaller contrasts, which makes it harder to reject the null hypothesis. The power curves are similar at different levels of relationship between weights and treatment effect for the weighted tests, but there is lower power when the positive relationship between the weights and the treatment effect is stronger for the unweighted test.

In this simulation study, all cases of no average treatment effect were the null hypothesis (4.9) of no difference between treatments. It is interesting to consider, however, what happens when there is no average treatment effect in the population, but the effect is correlated with the survey weights. This correlation analysis above suggests that if there is no average treatment effect, but the treatment effect for individual elements varies with weights, the unweighted tests may not get the correct size. This is because the estimation of the treatment effect would be biased, suggesting that there is a treatment effect when, overall, there is not one.

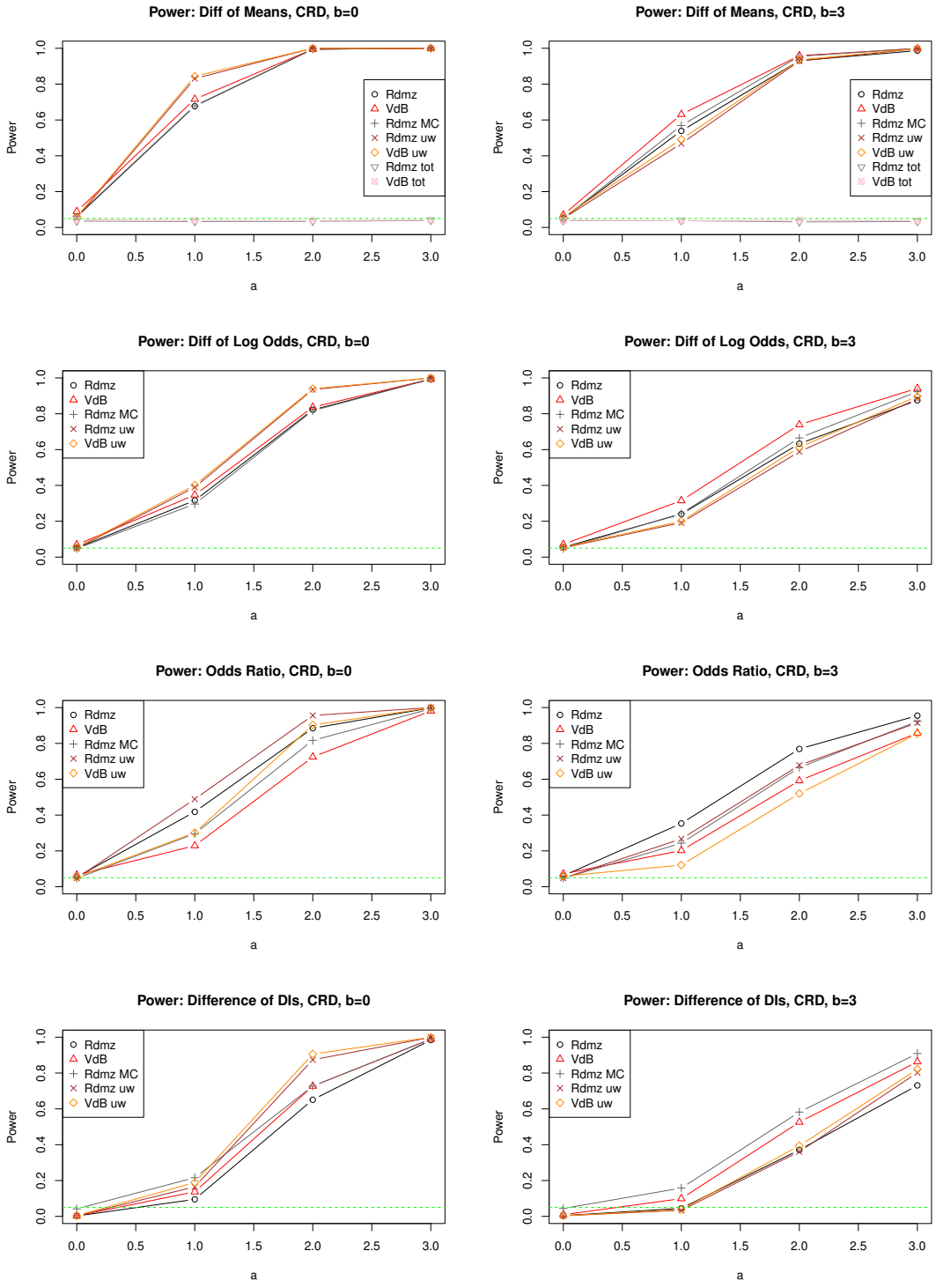

Figure 4.4: Curves showing power of all the methods for two-sided hypothesis tests at the $\alpha = 0.05$ level. The green dashed lines represents a rejection rate of 0.05. Each power calculation used 1000 replicates. The statistics are in rows (top-bottom: difference of means, difference of log odds, odds ratio, and difference of DIs). The columns are two different settings of the effect of the treatment on the response (left is no relationship, right is a large relationship). The a and b in the axis labels refer to equation (4.12)

4.9 Summary and Future Work

In this section, we derived a randomization test conditioned on the sample for a CRD embedded in a complex survey. We derived expressions for the mean and variance of the components under randomization of treatment labels. We showed that under mild conditions, contrasts of totals have an asymptotic normal distribution under the distribution of treatment assignments, and therefore also under randomization of labels. We showed that under moderate conditions, the randomization variance does not vary too much under the treatment assignment distribution, making power calculations feasible. We also showed that under mild conditions, the asymptotic normality extends to the finite population as well. Through simulations, we confirmed these results, and showed that when there is positive correlation between weights and the treatment effect, then the weighted methods have more power to detect the differences than unweighted methods do.

In the simulation studies presented in this chapter, the only cases tested where there was no average treatment effect followed the null hypothesis (4.9) of zero treatment effect for each unit. Simulation studies would be useful to further examine the case where there is no average treatment effect, but the treatment effect varies with the weights. While the randomization test invokes the null hypothesis (4.9) that there is no difference in the treatment effect, there is no reason to believe that randomization tests will have power to detect deviations on average. I would suspect that unweighted methods would have biased estimated for the average treatment effect due to the correlation with weights, and have incorrect size, while the weighted methods would have appropriate size in this case, providing stronger rationale for the use of weighted analyses.

Chapter 5

Randomization Test for Randomized Complete Block Experiments Embedded in Complex Surveys

5.1 Introduction

While CRDs are a good place to start for examining experiments, due to their simplicity, experiments embedded in surveys are frequently more complex. Often, the survey design will have clustering and/or stratification. These features can make for natural blocks for experiments using a randomized block design. In this chapter, we consider randomized complete block designs (RCBDs), in which every block contains every level of the treatment.

In this chapter, we will generalize the theory developed in the last chapter to extend to randomized complete block experiments embedded in complex surveys. We introduce the notation for the block structure, and generalize the results for the treatment assignment distribution in Section 5.2. In Section 5.3, we extend the mean and variance results for the randomization distribution. Asymptotic normality is established for the case of an increasing number of small blocks in Section 5.4. Power properties conditioned on the sample are discussed in Section 5.5. In Section 5.6, results are extended to the finite population. We discuss simulation studies in Section 5.7, and conclude with a summary and directions for future work in Section 5.8.

5.2 Notation and Treatment Assignment

Before calculating the variances, we describe the adjustments to the notation for randomized complete block designs.

For a randomized complete block design, we assume the sample is divided into H subsamples $S_1, \ldots, S_H$, which will serve as blocks for the experiment. These blocks could be strata used in sampling, clusters used at an early stage in a multi-stage sampling design, or something else. For

purposes of simplifying notation in this manuscript, we will assume additionally that the partition $1, \ldots, H$ can also be used to represent the entire population. It is possible that blocks could be sampled from a larger population of blocks in a randomized complete block design, but that would make no difference for inference on the sample, and would still extend to the population by the arguments to be discussed in Section 5.6.

Using the partitioned structure, the NHT estimator can be written as

$$\hat{Y}_+ = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{hi+}.$$

Now we consider the sampling design. We denote the number of EUs assigned each treatment as $|S_h| = m_h$. The number of these m_h experimental units sampled in block h that were assigned treatment k is $m_h^{(k)}$. Analogous to the completely randomized design, we include treatment assignment indicators $T_{hi}^{(k)}$, and define $\check{T}_{hi}^{(k)} = \frac{m_h}{m_h^{(k)}} T_{hi}^{(k)}$. This gives the NHT estimator for the total as if everyone had been assigned treatment k as

$$\hat{Y}_+^{(k)} = \sum_{h=1}^{H} \frac{m_h}{m_h^{(k)}} \sum_{i \in S_h} w_{hi} \hat{Y}_{hi+} T_{hi}^{(k)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{hi+} \check{T}_{hi}^{(k)}.$$

We write the vector of these results as

$$\hat{\boldsymbol{Y}}_+ = \begin{bmatrix} \hat{Y}_+^{(1)} \\ \vdots \\ \hat{Y}_+^{(K)} \end{bmatrix}.$$

In the rest of this section we will study the distribution of the contrast $\mathbf{C}\hat{\boldsymbol{Y}}_+$.

5.2.1 Treatment Assignment and Variance

As in the CRD, we consider the distribution of the vector $\hat{\boldsymbol{Y}}_+$ under the random treatment assignment. For the RCBD, the experimental units are assigned to all of the treatments by a CRD within each block. This assignment is done in each block without regard to the assignments

in other blocks, meaning that the treatment assignment indicators $T_{hi}^{(k)}$ and $T_{h'i'}^{(k)}$ are independent when $h \neq h'$.

Because the assignment of treatments in a RCBD is a CRD within each block, and the assignment is independent across blocks, we have that the distribution of the statistic $\hat{\boldsymbol{Y}}_h$ from a single block of an RCBD can be calculated using the distribution of the CRD derived in Chapter 4.

Before stating these results, we introduce additional notation. Let the sample covariance matrix of the weighted estimates under all treatment assignments be denoted

$$\hat{\mathbf{V}}_{h,w\hat{Y}} = \frac{1}{2m_h(m_h-1)} \sum_{i \in S_h} \sum_{i' \in S_h} (w_{hi}\hat{\boldsymbol{Y}}_{hi+} - w_{hi'}\hat{\boldsymbol{Y}}_{hi'+})(w_{hi}\hat{\boldsymbol{Y}}_{hi+} - w_{hi'}\hat{\boldsymbol{Y}}_{hi'+})',$$

and let the diagonal elements of this matrix be denoted as

$$s^2_{h,w\hat{Y}^{(k)}} = \frac{1}{2m_h(m_h-1)} \sum_{i \in S_h} \sum_{i' \in S_h} (w_{hi}\hat{Y}^{(k)}_{hi+} - w_{hi'}\hat{Y}^{(k)}_{hi'+}).$$

Additionally, and also similar to the notation in Chapter 4, we define treatment assignment weights as $\mathbf{d}_h = (d_h^{(1)}, \ldots, d_h^{(K)}) = (m_h/m_h^{(1)}, \ldots, m_h/m_h^{(K)})$.

Now we state the following corollaries, which provide formulas for the mean and variance of the contrast test statistic under the treatment assignment distribution with and without assuming additivity of treatment effects at the observational unit level. This can be expressed in blocks as

$$\hat{Y}_{hi}^{(k)} = \hat{Y}_{hi}^{*} + \hat{N}_{hi}\beta^{(k)}. \tag{5.1}$$

Corollary 5.2.1. *If* $\mathbf{C}$ *is a contrast matrix, then the mean and variance under the randomness due to treatment assignment in a RCBD are given by*

$$\begin{aligned} \mathrm{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}_+] &= \mathbf{C}\dot{\boldsymbol{Y}}_+ \\ \mathrm{Var}_T(\mathbf{C}\hat{\boldsymbol{Y}}_+) &= \sum_{h=1}^{H} m_h \mathbf{C}[\hat{\mathbf{V}}_{h,w\hat{Y}} \circ (\mathbf{D}_h - \mathbf{J})]\mathbf{C}' \end{aligned}$$

where $\dot{\boldsymbol{Y}}$ is the NHT estimator for the sample vector without treatments being assigned, and $\mathbf{D}_h = \mathrm{diag}(\boldsymbol{d}_h)$ is the diagonal matrix with the k^{th} diagonal element $d_h^{(k)}$.

Corollary 5.2.2. *If $\mathbf{C}$ is a contrast matrix, then, assuming additivity (5.1), the mean and variance under the randomness due to treatment assignment in a RCBD are given by*

$$\begin{aligned}\mathrm{E}_T[\mathbf{C}\hat{\boldsymbol{Y}}_+] &= \mathbf{C}\boldsymbol{\beta}\hat{N}_+ \\ \mathrm{Var}_T(\mathbf{C}\hat{\boldsymbol{Y}}_+) &= \sum_{h=1}^{H} m_h \mathbf{C}(\mathbf{D}_{h,w\hat{Y}} - s^2_{h,w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\mathbf{C}'\end{aligned}$$

where $\hat{N}_+ = \sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\hat{N}_{hi}$ is the estimated number of OUs in block h and $\mathbf{D}_{h,w\hat{Y}}$ is the diagonal $K \times K$ matrix with the k^{th} diagonal element $d_h^{(k)} s^2_{h,w\hat{Y}^{(k)}}$.

5.3 Randomization Distribution and Test

Similar to the previous chapter, we explain a hypothesis test comparing the test statistic $\mathbf{C}\hat{\boldsymbol{Y}}_+$ to the distribution obtained from randomizing the treatment labels on the observed dataset. To do this comparison, we examine the distribution of the randomization contrast $\mathbf{C}\tilde{\boldsymbol{Y}}_+$, where $\tilde{\boldsymbol{Y}}_+$ is the K-vector defined by

$$\tilde{Y}_+^{(k)} = \sum_{h=1}^{H} \frac{m_h}{m_h^{(k)}} \sum_{i\in S_h} w_{hi} R_{hi}^{(k)} \hat{Y}_{hi+} = \sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\check{R}_{hi}\hat{Y}_{hi+}.$$

As in the previous section, we can directly apply results from the last chapter because the treatment assignments in different blocks do not depend on each other.

Corollary 5.3.1. *Conditioned on the sample, the randomization contrast $\mathbf{C}\tilde{\boldsymbol{Y}}_+$ has mean and variance given by*

$$\begin{aligned}\mathrm{E}[\mathbf{C}\tilde{\boldsymbol{Y}}_+] &= \mathbf{0} \\ \mathrm{Var}(\mathbf{C}\tilde{\boldsymbol{Y}}_+) &= \sum_{h=1}^{H} m_h s^2_{h,w\hat{Y}} \mathbf{C}\mathbf{D}_h\mathbf{C}',\end{aligned}$$

where

$$s^2_{h,w\hat{Y}} = \frac{1}{2m(m-1)} \sum_{i \in S_h} \sum_{i' \in S_h \setminus \{i\}} (w_{hi}\hat{Y}_{hi+} - w_{hi'}\hat{Y}_{hi'+})^2$$

is the variance of the observed responses from the experimental units level in block h*.*

5.4 Central Limit Theory

In this section, we will prove central limit theory for the treatment assignment contrast in the situation where there are many blocks, each with few experimental units. In the previous chapter, we proved asymptotic normality where there was one block with an increasing number of experimental units. Because of the independence of blocks, this argument could also work if there are a small (fixed) number of blocks, and increasing numbers of experimental units within each block.

The central limit theory from the previous chapter provides a theoretical basis for a normal approximation in an experiment in a survey where there are a few large regions that could be used as blocks. The central limit theory in the this section provides such justification in a case where there are many regions that can be divided into blocks. An example of this might be the Health District survey (Chapter 2), where census tracts were divided into 24 blocks of approximately 3 census tracts each, and the census tracts are considered the experimental units for this study.

A third situation that could arise in surveys is an RCBD where both the block size and number of blocks are increasing. A central limit theorem could be established for this case in the future.

5.4.1 The Central Limit Theorem

Since the treatment assignments are independent between the blocks, we can develop central limit theory based on the Lyapunov central limit theorem for independent but not identically distributed random variables. We now discuss conditions on the contrast vectors, the treatment assignment, the weights, the block size, and the response. We condition on a sequence of samples in which the Hth sample of the sequence is divided into H blocks and all levels of treatment are randomly assigned within each block. The following assumptions hold as $H \to \infty$.

B1 For any blocks h, h', the treatment assignment in block h is independent of the treatment assignment in block h'.

B2 There exist $A, B < \infty$ such that $k \leq m_h$ for all h when $H > A$.

B3 There are $b > 0$ and $A, B < \infty$ such that $b < w_{hi} < B$ for all h, i when $H > A$.

B4 There exist $\delta > 0$ and $A, B < \infty$ such that

$$\frac{\sum_{h=1}^{H} \sum_{i \in S_h} |\hat{N}_{hi}|^{2+\delta}}{\sum_{h=1}^{H} m_h} < B.$$

for all $H > A$

B5 There exist $A, B > \infty$ such that for some $\delta > 0$,

$$\frac{\sum_{h=1}^{H} \sum_{i \in S_h} |\hat{Y}^*_{hi+}|^{2+\delta}}{\sum_{h=1}^{H} m_h} < B)$$

for all $H > A$.

B6 Let

$$r_h = \frac{\sum_{i \in S_h} (w_{hi} \hat{Y}^*_{hi+} - w_{hi'} \hat{Y}^*_{hi'})(w_{hi} \hat{N}_{hi} - w_{hi'} \hat{N}_{hi'})}{2m_h(m_h - 1) s_{h,wY^*} s_{h,wN}}$$

For some $A < \infty$, $b, \epsilon > 0$, we have

$$H^{-1} \sum_{h=1}^{H} I\left(s^2_{h,w\hat{Y}^*}(1 - r_h^2) > \epsilon\right) > b.$$

for all $H > A$.

Notes on Assumptions: Assumption *B1* states that the treatment assignment is independent across blocks, which is standard for RCBDs. The next two assumptions are about weights and block size. Assumption *B2* ensures that the blocks are not too variable in size. Assumption *B3* ensures that the weights are bounded and not too unbalanced, like was assumed in the previous

chapter. This ensures the setting of many small blocks, and is useful for the Lyapunov assumption to ensure that one block cannot be too big and dominate the variance term. The next two assumptions guarantee the Lyapunov condition for the response variable under any of the treatments. Assumption *B4* establishes that the estimated number of observational units in each block has a finite moment beyond the second moment. This is important because this is what is added to the base response value to get responses under a given treatment under additivity (5.1). Assumption *B5* establishes the same bound on the intrinsic response values. The final assumption, Assumption *B6*, ensures that there is no situation where a few blocks dominate because the block variance goes to zero for reasons not addressed in previous assumptions. While this could happen in theory with probability proportional to size sampling, this is not a situation likely to occur in practice, but needs to be ruled out for formality.

Theorem 5.4.1. *Assuming the additivity condition 5.1 and Assumptions* B1–B6 *above, we have, conditioned on the sample, that*

$$\left[\sum_{h=1}^{H} m_h \mathbf{C}(\mathbf{D}_{h,w\hat{Y}} - s^2_{h,w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\mathbf{C}'\right]^{-1/2} \sum_{h=1}^{H}(\mathbf{C}\hat{\boldsymbol{Y}}_h - \mathbf{C}\boldsymbol{\beta}\hat{N}_h) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$$

as $H \to \infty$.

Proof. For any column-vector $\boldsymbol{v}$ and contrast matrix $\mathbf{C}$, $\boldsymbol{v}'\mathbf{C}$ is a contrast row-vector. Therefore, by the Cramér-Wold Device it suffices to show that for any contrast vector $\boldsymbol{c}$, we have

$$\left[\sum_{h=1}^{H} m_h \boldsymbol{c}'(\mathbf{D}_{h,w\hat{Y}} - s^2_{h,w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{c}\right]^{-1/2} \sum_{h=1}^{H}(\boldsymbol{c}'\hat{\boldsymbol{Y}}_h - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h) \xrightarrow{d} \mathcal{N}(0, 1).$$

Because treatments are assigned independently across blocks (Assumption *B1*), we can apply the Lyapunov CLT using assumptions on each block, that is show that for some $\delta > 0$

$$\left[\mathrm{Var}_T(\boldsymbol{c}\hat{\boldsymbol{Y}}_+)\right]^{-(2+\delta)/2} \sum_{h=1}^{H} \mathrm{E}_T\left[|\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|^{2+\delta}\right] \to 0.$$

We first focus on the overall variance term of the Lyapunov condition. Recall

$$\mathrm{Var}_T(\boldsymbol{c}\hat{\boldsymbol{Y}}_+) = \sum_{h=1}^{H} m_h \boldsymbol{c}'(\mathbf{D}_{h,w\hat{Y}} - s^2_{h,w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{c}.$$

From Lemma 4.5.3 and $m_h/m_h^{(k)} \geq 1$ for each k, we have

$$\begin{aligned}
\mathrm{Var}_T(\boldsymbol{c}\hat{\boldsymbol{Y}}_+) &= \sum_{h=1}^{H} m_h \boldsymbol{c}'(\mathbf{D}_{h,w\hat{Y}} - s^2_{h,w\hat{N}}\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{c} \\
&\geq \sum_{h=1}^{H} m_h \sum_{k=1}^{K} (c^{(k)})^2 \frac{m_h}{m_h^{(k)}} s^2_{h,w\hat{Y}^*}(1 - r_h^2) \\
&\geq \|\boldsymbol{c}\|_2^2 \sum_{h=1}^{H} m_h s^2_{h,w\hat{Y}^*}(1 - r_h^2).
\end{aligned}$$

From Assumptions *B2*, *B6*, we have that for some $b, \epsilon > 0$,

$$\sum_{h=1}^{H} m_h s^2_{h,w\hat{Y}}(1 - r_h^2) \geq Hb\epsilon.$$

We now focus on the sum $\sum_{h=1}^{H} \mathrm{E}_T[|\boldsymbol{c}'\hat{\boldsymbol{Y}}_h - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|^{2+\delta}]$. Applying Jensen's inequality gives

$$\begin{aligned}
|\boldsymbol{c}'\hat{\boldsymbol{Y}}_h - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|^{2+\delta} &\leq (|\boldsymbol{c}'\hat{\boldsymbol{Y}}_h| + |\boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|)^{2+\delta} \\
&\leq 2^{1+\delta}(|\boldsymbol{c}'\hat{\boldsymbol{Y}}_h|^{2+\delta} + |\boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|^{2+\delta}) \\
&= 2^{1+\delta}\left|\sum_{i\in S_h} m_h w_{hi}(\hat{Y}^*_{hi+} + \hat{N}_{hi}\boldsymbol{T}'_{hi}\boldsymbol{\beta})\right|^{2+\delta} + 2^{1+\delta}|\boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|^{2+\delta}.
\end{aligned}$$

Applying Jensen's inequality again, we learn that

$$\begin{aligned}
\left|\frac{\sum'_{i\in S_h} m_h w_{hi}(\hat{Y}^*_{hi+} + \hat{N}_{hi}\boldsymbol{T}'_{hi}\boldsymbol{\beta})}{\sum_{i\in S_h} m_h w_{hi}}\right|^{2+\delta} &\leq \frac{\sum_{i\in S_h} m_h w_{hi}|\hat{Y}^*_{hi+} + \hat{N}_{hi}\boldsymbol{T}'_{hi}\boldsymbol{\beta}|^{2+\delta}}{\sum_{i\in S_h} m_h w_{hi}} \\
&\leq 2^{1+\delta}\frac{\sum_{i\in S_h} m_h w_{hi}(|\hat{Y}^*_{hi+}|^{2+\delta} + |\hat{N}_{hi}\max_k \beta^{(k)}|^{2+\delta})}{\sum_{i\in S_h} m_h w_{hi}}.
\end{aligned}$$

We note that, conditioned on the sample, there is nothing random in the final expression, so the expectation can be dropped.

Using the Assumptions *B3* and *B2* to bound the weights and the block sizes and noting that the treatment effects are uniformly bounded, we find that there are constants $B_Y, B_N < \infty$ such that

$$\sum_{h=1}^{H} |\boldsymbol{c}^{(k)} \hat{\boldsymbol{Y}}_h|^{2+\delta} \leq B_Y \sum_{h=1}^{H} \sum_{i \in S_h} |Y^*_{hi+}|^{2+\delta} + B_N \sum_{h=1}^{H} \sum_{i \in S_h} |\hat{N}_{hi}|^{2+\delta}.$$

Putting everything together, we have that under the conditions stated above

$$\begin{aligned}
&\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}_+)^{-(2+\delta)/2} \sum_{h=1}^{H} \mathrm{E}_T[|\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_h|^{2+\delta}] \\
&\quad \leq \frac{B_Y \sum_{h=1}^{H} \sum_{i \in S_h} |Y^*_{hi+}|^{2+\delta} + B_N \sum_{h=1}^{H} \sum_{i \in S_h} |\hat{N}_{hi}|^{2+\delta}}{(Hb\epsilon)^{(2+\delta)/2}} \\
&\quad \leq H^{-\delta/2} \frac{B_Y \sum_{h=1}^{H} \sum_{i \in S_h} |Y^*_{hi+}|^{2+\delta} + B_N \sum_{h=1}^{H} \sum_{i \in S_h} |\hat{N}_{hi}|^{2+\delta}}{H(b\epsilon)^{(2+\delta)/2}}.
\end{aligned}$$

Now since $\delta > 0$ and the expression after $H^{-\delta/2}$ is $O(1)$, we have that this expression converges to 0, verifying the Lyapunov condition.

To complete the proof, we note that the centering and scaling of each block were verified in Lemma 4.3.1. □

We will later revisit asymptotic normality with respect to the general population.

5.5 Treatment Assignment and Randomization: Power

In this section, we examine both the treatment assignment and randomization distributions to understand how to estimate the power of a randomization test for a randomized complete block design experiment. As in Section 4.6, we will show that the variation of the randomization distribution under different treatment assignments is asymptotically negligible. However, like in Section 5.4, the asymptotics here will depend on the number of blocks increasing, rather than the

number of EUs within blocks getting larger. We will also explain how these results and the results from Section 4.6 can be used to calculate power for many RCBDs embedded in surveys.

As in Chapter 4, we add an assumption that experimental units are sampled with probability proportional to size. Using the new notation to allow for blocks, this can be written as

$$\text{For each block } h,\ w_{hi}\hat{N}_{hi} = W_h. \tag{5.2}$$

Theorem 5.5.1. *Assume (5.1), (5.2), and Assumptions* B1–B6*. Then for any nonzero contrast vectors* $\boldsymbol{c}$*,* $\boldsymbol{c}_1$*, and* $\boldsymbol{c}_2$*,*

$$\frac{\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+)}{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+)\right]} = \frac{\sum_{h=1}^{H}\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h)}{\sum_{h=1}^{H}\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h)\right]} \xrightarrow{p} 1,$$

and

$$\frac{\sqrt{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)\right)}}{\sqrt{\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+)\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)}} \xrightarrow{p} 0.$$

Proof. Recall, assuming (5.1), (5.2), and the independence of treatment assignment across blocks (*B1*) that

$$\mathrm{E}_T[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+)] = \sum_{h=1}^{H} m_h\left[s^2_{h,w\hat{Y}^*} + \frac{W_h^2 m_h}{m_h-1}(\overline{\beta_h^2} - \overline{\beta}_h^2)\right]\boldsymbol{c}'\mathbf{D}_h\boldsymbol{c}$$

$$\mathrm{Var}_T\left(\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+)\right) = \sum_{h=1}^{H}\frac{4W_h^2 m_h^3}{(m_h-1)^2}s^2_{h,w\hat{Y}^*}(\overline{\beta_h^2} - \overline{\beta}_h^2)(\boldsymbol{c}'\mathbf{D}_h\boldsymbol{c})^2.$$

Under the assumptions above, we have established that the terms are such that uniformly for all $H > A$, there are constants B, b, and ϵ such that

$$m_h\left[s^2_{h,w\hat{Y}^*} + \frac{W_h^2 m_h}{m_h-1}(\overline{\beta_h^2} - \overline{\beta}_h^2)\right]\boldsymbol{c}'\mathbf{D}_h\boldsymbol{c} \le B$$

and that

$$\frac{1}{H}\sum_{h=1}^{H} I\left(m_h\left[s^2_{h,w\hat{Y}^*}+\frac{W_h^2 m_h}{m_h-1}(\overline{\beta_h^2}-\overline{\beta}_h^2)\right]\boldsymbol{c}'\mathbf{D}_h\boldsymbol{c}\geq b\right)\geq\epsilon.$$

By the fact that arithmetic means are larger than geometric means, we have that

$$\begin{aligned}\mathrm{Var}_T\left(\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h)\right) &= 2m_h s_{h,w\hat{Y}^*}\frac{W_h\sqrt{m_h}}{m_h-1}\sqrt{\overline{\beta^2}-\overline{\beta}^2}(\boldsymbol{c}'\mathbf{D}_h\boldsymbol{c})\\ &\leq m_h\left[s^2_{h,w\hat{Y}^*}+\frac{W_h^2 m_h}{m_h-1}(\overline{\beta_h^2}-\overline{\beta}_h^2)\right]\boldsymbol{c}'\mathbf{D}_h\boldsymbol{c}.\end{aligned}$$

Thus we find

$$\sqrt{\sum_{h=1}^{H}\mathrm{Var}_T\left(\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h)\right)}\leq\sqrt{\sum_{h=1}^{H}B^2}=B\sqrt{H}.$$

Using Assumption *B3*, we obtain

$$\frac{\sqrt{\sum_{h=1}^{H}\mathrm{Var}_T\left(\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h)\right)}}{\sum_{h=1}^{H}\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_h)\right]}\leq\frac{B\sqrt{H}}{b\epsilon H}=O(H^{-1/2})\to 0.$$

Now we will show the result about the correlation matrix. As in the proof of Theorem 4.6.1, the result just proven means that it suffices to show that

$$\frac{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+,\boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)\right)}{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+)\right]\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)\right]}\xrightarrow{p}0.$$

Here,

$$\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_h,\boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_h)\right)=2m_h s^2_{h,w\hat{Y}^*}\frac{W_h^2 m_h^2}{(m_h-1)^2}(\overline{\beta^2}-\overline{\beta}^2)\boldsymbol{c}_1'\mathbf{D}_h\boldsymbol{c}_2$$

and

$$\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+)\right]=\sum_{h=1}^{H}m_h\left[s^2_{h,w\hat{Y}^*}+\frac{W_h^2 m_h}{m_h-1}(\overline{\beta_h^2}-\overline{\beta}_h^2)\right]\boldsymbol{c}_1'\mathbf{D}_h\boldsymbol{c}_1.$$

Since $\mathbf{D}_h$ is a diagonal matrix with all entries greater than 1, we have that for all h, $\boldsymbol{c}_1'\mathbf{D}_h\boldsymbol{c}_1 \geq \boldsymbol{c}_1'\boldsymbol{c}_1 \geq \min\{\boldsymbol{c}_1'\boldsymbol{c}_1, \boldsymbol{c}_2'\boldsymbol{c}_2\}$. Therefore, we have

$$\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+)\right]\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)\right] \geq \left(\sum_{h=1}^{H} m_h\left[s^2_{h,w\hat{Y}^*} + \frac{W_h^2 m_h}{m_h - 1}(\overline{\beta_h^2} - \overline{\beta}_h^2)\right]\right)^2 \min\{\boldsymbol{c}_1'\boldsymbol{c}_1, \boldsymbol{c}_2'\boldsymbol{c}_2\}.$$

Now from similar arguments to above, there are constants B, b, and ϵ such that, for $H > A$,

$$\frac{\mathrm{Var}_T\left(\mathrm{Cov}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+, \boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)\right)}{\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_1'\tilde{\boldsymbol{Y}}_+)\right]\mathrm{E}_T\left[\mathrm{Var}_R(\boldsymbol{c}_2'\tilde{\boldsymbol{Y}}_+)\right]} \leq \frac{HB}{H^2 b\epsilon} = O(H^{-1}) \xrightarrow{p} 0,$$

completing the proof.

□

The above results show that the variance of any single randomization contrast converges the expectation under treatment assignment of that variance in the sense that the ratio between the two statistics converges to 1. We also showed that the variance of the correlations is asymptotically negligible. We will now use these results to discuss power computations for many RCBDs.

In Section 4.6, we derived the asymptotic power of the one-sided test for a single contrast for a CRD, under the stated conditions. Those computations apply analogously to show that under the conditions in the hypothesis of Theorem 5.5.1,

$$P_T\left(\frac{\boldsymbol{c}'\hat{\boldsymbol{Y}}_+ - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_+}{\sqrt{\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}_+)}} \geq \frac{\Phi^{-1}(1-\alpha)\sqrt{\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+)} - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_+}{\sqrt{\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}_+)}}\right).$$

Applying the above theorem, we obtain an approximate formula for power in RCBDs with many small blocks (or few large blocks by the results from Section 4.6). Assuming (5.1), (5.2),

and Assumptions *B1–B6*, the power is approximately

$$1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)\sqrt{\mathrm{E}_T[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+)]} - \boldsymbol{c}'\boldsymbol{\beta}\hat{N}_+}{\sqrt{\mathrm{Var}_T(\boldsymbol{c}'\hat{\boldsymbol{Y}}_+)}}\right).$$

These terms can be approximated by investigator knowledge of the variability of the phenomenon being investigated, the expected effect size, and the formula provided above for $\mathrm{E}_T[\mathrm{Var}_R(\boldsymbol{c}'\tilde{\boldsymbol{Y}}_+]$.

5.6 Extending to Finite Population

In this section we show how the randomization methodology can be extended to the finite population. As in Chapter 4, we apply Theorem 1.3.6 from Fuller [2009]. To apply this theorem, we need to restate the assumptions to guarantee that the conditional asymptotic normality occurs in almost all sequences of samples from the sequence of finite populations.

*B1** For all possible sampling designs, the block sizes are bounded, meaning that there is a $B < \infty$ such that for all H, $\max_h m_h < B$.

*B2** The sampling design admits a $b > 0$ and $A, B < \infty$ such that uniformly for all $H > A$,

$$b < w_{hi} < B$$

for all experimental units hi in the finite population.

*B3** There exist $B < \infty$, $\delta > 0$ and a sequence $\epsilon_H > 0$ where $\sum_{H=1}^{\infty} \epsilon_H < \infty$ such that for all H

$$P\left(\frac{\sum_{h=1}^{H}\sum_{i=1}^{M_h}|\hat{N}_{hi}|^{2+\delta}}{\sum_{h=1}^{H} M_h} < B\right) > 1 - \epsilon_H.$$

*B4** There exist $B < \infty$, $\delta > 0$ and a sequence $\epsilon_H > 0$ where $\sum_{H=1}^{\infty} \epsilon_H < \infty$ such that for all H

$$P\left(\frac{\sum_{h=1}^{H}\sum_{i=1}^{M_h}|\hat{Y}^*_{hi+}|^{2+\delta}}{\sum_{h=1}^{H} M_h} < B\right) > 1 - \epsilon_H.$$

*B5** For some $\epsilon > 0$ and $b > 0$, we have that the sequence of finite populations and sampling designs satisfies

$$P_S(\lim_{H\to\infty} H^{-1}\sum_{h=1}^{H} I(s^2_{h,w\hat{Y}^*}(1-r_h^2) \geq \epsilon) \geq b) = 1,$$

where P_S is the probability with respect to the randomness induced by the sampling designs.

*B6** The sampling design is such that the estimator of the number of observational units (OUs) in the population is asymptotically normal, that is

$$\frac{m^{1/2}}{M}(\sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\hat{N}_{hi} - \sum_{h=1}^{H}\sum_{i=1}^{M_h} N_{hi}) \xrightarrow{d} \mathcal{N}(0,\sigma_N^2),$$

as $H \to \infty$. The variance in the limit is allowed to be zero.

Now that the assumptions have been appropriately modified, we state the theorem.

Theorem 5.6.1. *Under the Assumptions* B1*–B6* *above, the distribution of* $\mathbf{C}\hat{\boldsymbol{Y}}_+$ *is asymptotically normal with respect to both the sampling distribution and the treatment assignment distribution.*

Proof. The proof is an application of Theorem 1.3.6 in Fuller [2009].

Assumptions *B1** and *B2** show that Assumptions *B2* and *B3* hold for all samples. Assumptions *B3** and *B4** show via the first Borel-Cantelli Lemma that Assumptions *B4* and *B5* hold almost surely. Lastly Assumption *B6** above guarantees that the the expected value of the contrast under treatment assignment at the sample level is a consistent and asymptotically normal estimate for the value of the contrast as if it were applied to the whole population. This completes the necessary conditions for Theorem 1.3.6 in Fuller [2009], from which the result follows. □

We additionally note here, as in the previous chapter, that because we consider the variation of both treatment assignment and sampling, Theorem 5.6.1 does apply to show asymptotic normality for the design-based test statistic proposed by van den Brakel [2001], with respect to all stochastic elements except for the measurement error model.

5.7 Simulation Experiments

5.7.1 Population and Variables

The population used for this simulation study was the same as that used in the simulation study for the CRD case in Section 4.8 based on US Census Bureau data from four PUMAs in northeastern Colorado, grouped into artificial districts within the PUMAs by a propensity score method. The study variables for this experiment were created using the same methods as in the previous chapter, but with different weights, reflecting a different sampling design. The continuous response variable was created to be positively correlated with weights, following equation (4.12). The binary variables were again created by an indicator of whether the continuous variable was greater than 350. For more details, see Section 4.8.

5.7.2 Overview of Studies

We used the simulated population in two studies. In the first study, we compared two approximations for the randomization distribution (the distribution of the the test statistic under randomized treatment labels): the Monte Carlo approximation based on 1000 randomizations and the asymptotic normal approximation given by Theorem 5.4.1 above. In the second study, we compared the size and power properties of the survey-weighted randomization test proposed here to several alternative methods.

For both studies, we considered four different test statistics: difference in means, difference in log odds, odds ratio, and difference in the dissimilarity index. For the power curves, we compared the survey weighted methods to their non-weighted equivalents. The methods included in this study are van den Brakel's design-based method (VdB), the randomization method included in this paper (both the normal approximation, Rdmz, and the Monte Carlo version, Rdmz MC), and versions of both of these methods ignoring the survey weights (VdB uw and Rdmz uw). Additionally, for the difference in means tests, we also included differences of estimated totals using both van den Brakel's design-based method (VdB tot) and the randomization test discussed in this dissertation (Rdmz tot).

5.7.3 Sampling Plan and Treatment Assignment

For both of the simulation studies, the sample was a two-stage sample with stratified sampling in both stages. The first stage is sample of districts, stratified by PUMA. Six districts were selected within each PUMA with probability proportional to the number of individuals in the district. Within each district, individuals were stratified based on a noisy indicator of Hispanic origin. A noised Hispanic origin variable was created to mimic the noise found in auxiliary variables purchased in real surveys. This noised variable was created so that people were Hispanic had an 80% chance of being labeled as Hispanic and people who were not Hispanic had a 4% chance of being labeled as Hispanic. The second-stage sample within each district was then a stratified sample of 30 individuals labeled Hispanic and 30 individuals labeled non-Hispanic. For the size and power study, the sample was redrawn each time. For the normality study, only one sample was drawn and the test involved testing the normality of the randomization distribution.

For the assessment of normal approximation of the randomization test distribution, we used the same general sampling design described for the size and power experiment, but considered three different sample sizes. As with the CRD experiment, variables are considered with setting $a = b = 2$. The sample sizes considered were 2, 4 and 6 districts in each of the four PUMAs at the first stage, corresponding to 16, 32 or 48 blocks (predicted Hispanic/non-Hispanic strata within district). Within these blocks, there were, respectively, 10, 20, or 30 individuals per block.

For the size and power study, treatments were assigned to districts by a randomized complete block design where the experimental units were individuals and the blocks were strata within districts, divided by whether the sampled individuals were labeled Hispanic. Within each block, 15 EUs were assigned the treatment and 15 EUs were assigned the control. Power curves were generated by evaluating power for variables as a increased from 0 to 3 for each fixed value of b, as in the CRD case.

5.7.4 Results

Normal Approximation

Figure 5.1 shows the normal approximation from Theorem 5.6.1 and a kernel density estimate based on 1000 Monte Carlo draws from the randomization distribution. The normal approximation appears to work well for all statistics at the largest number of blocks tested in the third column (48 blocks), though a small amount of skewness may be present for testing odds ratios. The figure provides empirical support for Theorem 5.6.1.

In the smaller sample sizes, we see that the difference of means is very close to the normal distribution even with only 16 blocks. The difference of log odds has some abnormal deviation from normality at 16 blocks, but appears normal at 32 and 48 blocks. The deviation at 16 blocks may be due to the discrete nature of binary data. The odds ratio is very skewed at 16 blocks, and becomes more normal, with only slight skew present at 48 blocks. This is unsurprising as the odds ratio is the exponent of the difference in log odds (see discussion in Section 6.4), which is symmetric at all sample sizes. As there is more variability in the statistics at smaller sample sizes than larger sample sizes (all else held constant), the smaller range makes the skew disappear.

The randomization distribution of difference in DI is poorly approximated by the normal distribution at the smallest sample size, as was the case at the middle sample size in the CRD case described in Section 4.8.5. As the sample size gets larger, the precision of the estimates improves so the estimates of the difference are no longer impacted by the bounding effect explained in Section 4.8.5.

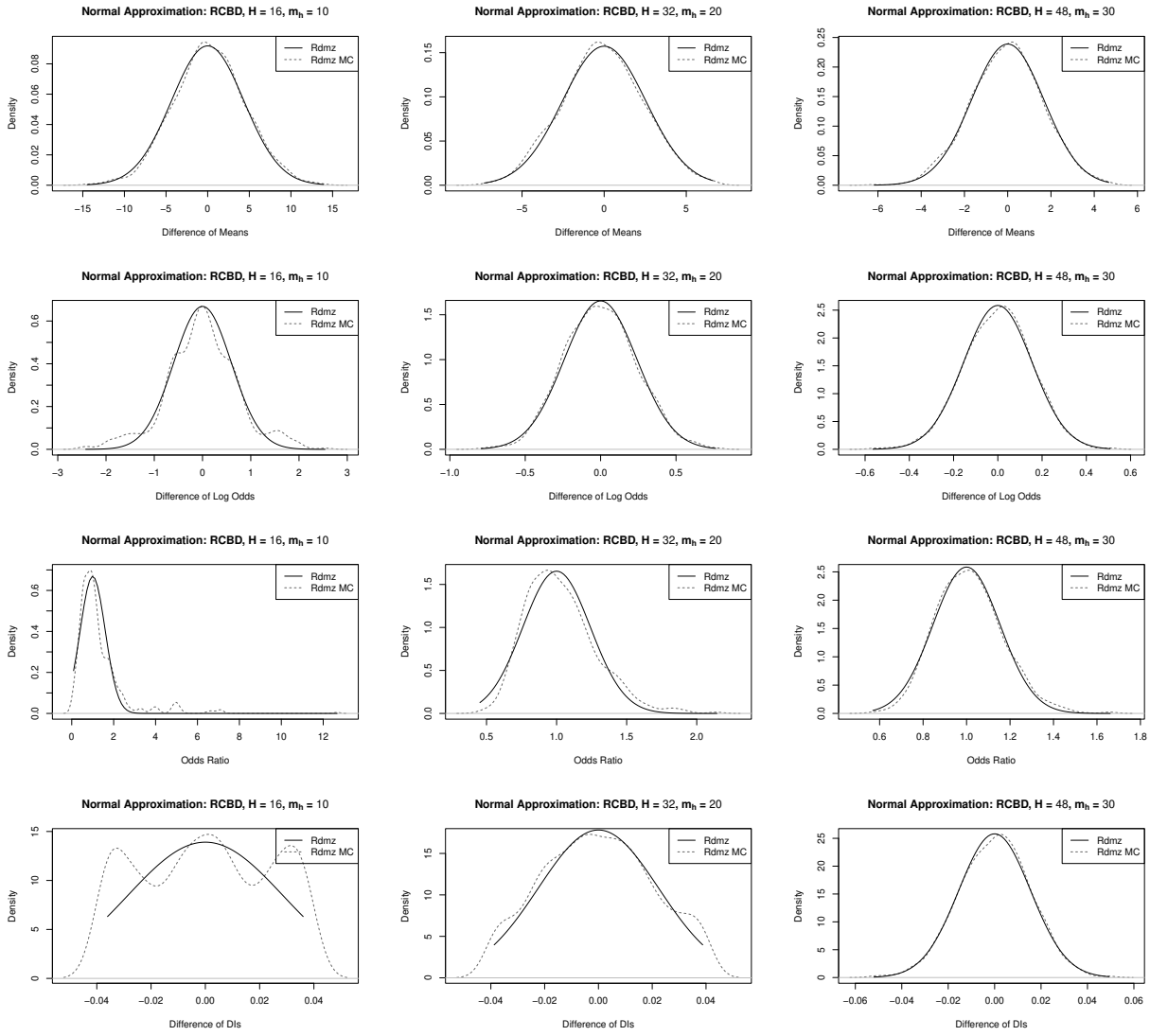

Figure 5.1: Curves showing the normal approximation to the statistic, and a kernel density estimate of the Monte Carlo distribution in the RCBD. The Monte Carlo distribution is simulated with 1000 draws. The rows are different statistics (top-bottom: difference of means, difference of log odds, odds ratio, and difference of DIs), and the columns are different sample sizes (left-right: 16, 32, and 48 blocks).

Power Curves

From examining the power curves in Figure 5.2, all of the methods maintain good size when there is actually no difference between the treatments, except for the difference of DIs. All of the tests have increasing power with greater sample size, which confirms that the methodology works. These confirm in the setting of an RCBD the results obtained in the CRD setting in Section 4.8.5.

For the randomization tests, we see the same story as in the CRD case, where we see lower power with little relationship between the weights and the treatment effects, but much more power when there is a large positive relationship between weights and treatment effects.

In the RCBD case, we also see that there is very little difference in power between the randomization test and the test proposed by van den Brakel (except in the case of the odds ratio). This is different from what was seen in a CRD. This may be because blocking reduced some of the variation. More investigation would be needed to determine what factors affect differences in power between these two methodologies.

For the odds ratio, the difference in power has to do with properties of the exponential function. The linearization is based on exponentiating the difference of log odds, and in this study, the difference in log-odds was positive. The slope of the exponential function at the true odds ratio (the center of linearization for the van den Brakel method) is greater than at 0 (the center of linearization for the randomization method). This means the estimated variance is higher for van den Brakel's method, which explains why the randomization method in this dissertation had more power in this situation. This suggests that if the odds ratio were less than 1, van den Brakel's method would have more power detecting such a difference. This would be worth checking in future simulations.

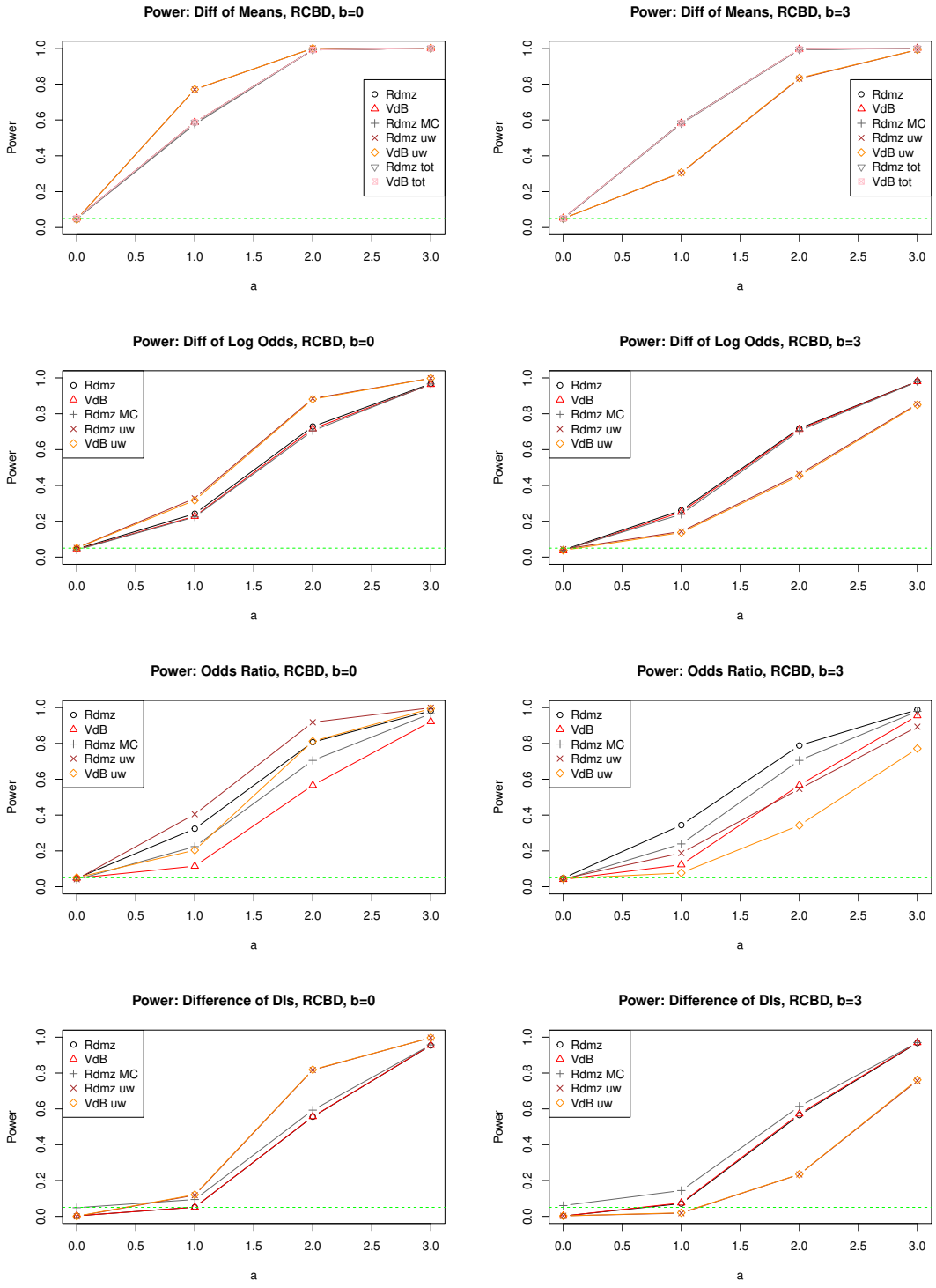

Figure 5.2: Curves showing power of all the methods for two-sided hypothesis tests at the $\alpha = 0.05$ level in the RCBD. The green dashed lines represents a rejection rate of 0.05. Each power calculation used 1000 replicates. The statistics are in rows (top-bottom: difference of means, difference of log odds, odds ratio, and difference of DIs). The columns are two different settings of the effect of the treatment on the response (left is no relationship, right is a large relationship). The a and b in the axis labels refer to equation (4.12)

5.8 Summary and Future Work

In this chapter, we extended the theory on the FPW style randomization test, conditioned on the sample, for randomization tests embedded in complex surveys to apply for randomized complete block designs. We derived expressions for the mean and variance of the treatment assignment distribution of the contrast and of the randomization contrast, and showed theory justifying a normal approximation in cases of a few large blocks, and of many small blocks. We further showed power properties of this method, and showed that the inference extended to the finite population. We concluded with simulation studies confirming asymptotic normality and comparing this method to other methods that had been proposed using power curves.

We have established asymptotic normality in the case of a few blocks of large size by adapting results from Chapter 4. We also developed more theory in this chapter to provide central limit theory to cases with many blocks of small size. More work could be done to fill the gap in the case of many blocks with large size. Additionally, providing a central limit theory for other experimental designs, including split-plot designs and incomplete block designs remains an open problem.

Simulation studies showed that the power loss compared to the design-based method proposed by van den Brakel [2001] is minimal in the case with many blocks. More investigation should be taken to understand why this may be. It would also be useful to have future simulation studies comparing these methods in experiments in surveys, analogous to the study undertaken by Ding [2017].

Chapter 6

Extensions of Randomization Tests for Experiments Embedded in Complex Surveys

6.1 Introduction

In the previous two chapters, we introduced randomization tests conditioned on the sample, and proved that the distribution of contrasts of totals satisfies a normal distribution both conditioned on the sample and unconditionally. Theory was developed for completely randomized designs (Chapter 4) and randomized complete block designs (Chapter 5), while using the NHT estimator for surveys.

In this chapter, we extend some of the theory and discuss some issues that can be present when using a randomization test in practice. We first discuss the Monte Carlo randomization test in more detail in Section 6.2. We then discuss how these randomization tests can be used with the generalized regression (GREG) estimator in Section 6.3, and how linearization techniques can work with the randomization methodology in Section 6.4. We discuss potential practical hurdles to randomization procedures with special focus on the presence of nonresponse in Section 6.5. We conclude in Section 6.6 with a summary of the contributions in this chapter and some directions for future research.

6.2 The Monte Carlo Test

While the previous chapters focused primarily on developing theory for a normal approximation to the randomization test for linear statistics, there are some cases in which the estimators may be nonlinear, and may not have asymptotic normal distributions. While many commonly used statistics can be approximated by a linear expression (cases like this are discussed in Sec-

tion 6.4), there are some statistics that cannot. It could also be that the experimental design is very complicated, or that a linear approximation may not be good enough for the investigators' needs.

To review, a Monte Carlo randomization test is a test that computes a test statistic over the randomization of treatment assignments for a large number of the possible treatment assignments that are possible for the given finite population. If there were no difference between treatments, then one would expect the statistic calculated from the original treatment assignment to be within the range of statistics calculated when the treatment assignments were randomized. Therefore if there are very few randomizations that give a statistic more "extreme" than obtained initially, there is evidence that the treatment had an effect. This theory also is the basis of the normal approximations discussed in Chapters 4 and 5.

The Monte Carlo test is a flexible test that can be used for any type of statistic that an investigator may be interested in. In some modern surveys, experiments can be designed in complicated ways. For example, in the targeted mailing for the Health District survey (discussed in Chapter 2), addresses that were assigned to receive the targeted mailing were disproportionately from groups that were thought to be less likely to respond. Without accounting for the complicated designs, the effect on the results would be difficult to measure.

In some cases, a statistic may not be able to be linearized. For example, if an investigator is interested in comparing log-rank tests under different survey treatments, then a linearization is impossible. It was also seen in the simulation studies (Sections 4.8 and 5.7) that linearization of the difference in DIs fails when the DIs are within a few standard errors of zero.

The simulations for both a completely randomized design and a randomized complete block design involved comparing inferences from the normal approximation to the randomization test to a Monte Carlo test. As a review, these results showed similarity between the randomization test and MC for the Hájek-type ratio estimator and the estimator of the difference of log-odds. We saw bigger differences in power for the odds ratio and the dissimilarity index.

This discussion shows that a Monte Carlo test can be a useful way to evaluate a wide variety of statistics for many types of randomized experiments. This is a major advantage of the

randomization method described in this dissertation over the design-based test of van den Brakel [2001]. Because van den Brakel's methodology requires the approximation the distribution of the test statistic with normal theory, it may not work for very complicated statistics or experimental design. Because the randomization test described in this dissertation relies on the treatment randomization conditioned on the sample, it allows for a Monte Carlo test that can be used in many situations.

6.3 The Generalized Regression Estimator

The generalized regression (GREG) estimator is one example of a calibration estimator (Deville and Sarndal [1992]). A calibration estimator of a survey total that allows for use of auxiliary information in the form of population totals of variables collected in the sampling. This often improves efficiency when estimating survey totals.

The GREG estimator for the full sample, ignoring the treatments, is

$$\hat{Y}_{+\mathrm{GR}} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi} - w_{hi} I(i \in S_h) \boldsymbol{x}_{hi})' \hat{\boldsymbol{b}},$$

where $\hat{\boldsymbol{b}}$ is the estimated value of the regression coefficient for the sample, specifically

$$\hat{\boldsymbol{b}} = \left(\sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \boldsymbol{x}_{hi} \boldsymbol{x}'_{hi} \right)^{-1} \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \boldsymbol{x}_{hi} \hat{Y}_{hi+}.$$

Here, only the sample weights are included in estimation of the coefficient vector, for ease of notation. It is straightforward to include an additional term to account for heteroscedasticity (Chapter 6 of Särndal et al. [1992], for example).

Considering a specific treatment, k, the GREG estimator for the population total from the subsample assigned treatment k is

$$\hat{Y}^{(k)}_{+\mathrm{GR}} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}^{(k)}_{hi} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi} - w_{hi} I(i \in S_h) \check{T}^{(k)}_{hi} \boldsymbol{x}_{hi})' \hat{\boldsymbol{b}}^{(k)},$$

where $\hat{\boldsymbol{b}}^{(k)}$ is the estimated regression coefficient on the subsample assigned treatment k, meaning

$$\hat{\boldsymbol{b}}^{(k)} = \left(\sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}_{hi}^{(k)} \boldsymbol{x}_{hi} \boldsymbol{x}'_{hi} \right)^{-1} \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}_{hi}^{(k)} \boldsymbol{x}_{hi} \hat{Y}_{hi+}^{(k)}.$$

As with the NHT estimator, we can consider a randomization version of this estimator to use in the randomization test. This is

$$\tilde{Y}_{+\mathrm{GR}}^{(k)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi} - w_{hi} I(i \in S_h) \check{R}_{hi}^{(k)} \boldsymbol{x}_{hi})' \tilde{\boldsymbol{b}}^{(k)},$$

where $\tilde{\boldsymbol{b}}^{(k)}$ is the estimated value of the regression coefficient for the subsample randomly labeled with treatment k, denoted

$$\tilde{\boldsymbol{b}}^{(k)} = \left(\sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \boldsymbol{x}_{hi} \boldsymbol{x}'_{hi} \right)^{-1} \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \boldsymbol{x}_{hi} \hat{Y}_{hi+}^{(k)}.$$

This will have the same distribution of test statistic under the null hypothesis (4.9) that the treatment had no effect.

Also like the NHT estimator, we can write the estimated totals under all treatments as a vector and consider contrasts. That is, we can write $\hat{\boldsymbol{Y}}_{+\mathrm{GR}} = (\hat{Y}_{+\mathrm{GR}}^{(1)}, \ldots, \hat{Y}_{+\mathrm{GR}}^{(1)})'$ and $\tilde{\boldsymbol{Y}}_{+\mathrm{GR}} = (\tilde{Y}_{+\mathrm{GR}}^{(1)}, \ldots, \tilde{Y}_{+\mathrm{GR}}^{(1)})'$, which leads to investigation of the test statistic $\mathbf{C}\hat{\boldsymbol{Y}}_{+\mathrm{GR}}$ with the randomization distribution of $\mathbf{C}\tilde{\boldsymbol{Y}}_{+\mathrm{GR}}$ via a randomization test.

6.3.1 GREG and the Treatment Assignment Distribution

From the above, the estimated total using GREG for treatment k is

$$\hat{Y}_{+\mathrm{GR}}^{(k)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}_{hi}^{(k)} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi} - w_{hi} I(i \in S_h) \check{T}_{hi}^{(k)} \boldsymbol{x}_{hi})' \hat{\boldsymbol{b}}^{(k)}.$$

In this section, we investigate the variance of this term under the treatment assignment distribution. The expected value of the regression coefficients under the treatment assignment distribution is

approximately

$$\dot{\boldsymbol{b}}^{(k)} = \left(\sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \boldsymbol{x}_{hi} \boldsymbol{x}'_{hi} \right)^{-1} \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \boldsymbol{x}_{hi} \hat{Y}^{(k)}_{hi+},$$

which is the estimated regression coefficient had everyone in the sample received treatment k. Therefore we can rewrite $\hat{Y}^{(k)}_{+\mathrm{GR}}$ as

$$\begin{aligned} \hat{Y}^{(k)}_{+\mathrm{GR}} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}^{(k)}_{hi} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi+} - w_{hi} I(i \in S_h) \check{T}^{(k)}_{hi} \hat{\boldsymbol{x}}_{hi+})' \dot{\boldsymbol{b}}^{(k)} + \\ \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi+} - w_{hi} I(i \in S_h) \check{T}^{(k)}_{hi} \hat{\boldsymbol{x}}_{hi+})' (\hat{\boldsymbol{b}}^{(k)} - \dot{\boldsymbol{b}}^{(k)}). \end{aligned}$$

Due to random treatment assignment, $\hat{\boldsymbol{b}}^{(k)}$ will often be close to $\dot{\boldsymbol{b}}^{(k)}$ in large samples. Therefore, we approximate

$$\begin{aligned} \hat{Y}^{(k)}_{+\mathrm{GR}} &\approx \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}^{(k)}_{hi} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\boldsymbol{x}_{hi+} - w_{hi} I(i \in S_h) \check{T}^{(k)}_{hi} \hat{\boldsymbol{x}}_{hi+})' \dot{\boldsymbol{b}}^{(k)} \\ &= \sum_{h=1}^{H} \sum_{i=1}^{M_h} \boldsymbol{x}_{hi+} \dot{\boldsymbol{b}}^{(k)} + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}^{(k)}_{hi} (\hat{Y}_{hi+} - \hat{\boldsymbol{x}}'_{hi+} \dot{\boldsymbol{b}}^{(k)}). \end{aligned}$$

In this statement, the first term is nonrandom, and the second term looks just like a total estimate under treatment k for the random variable $\hat{e}^{(k)}_{hi} = \hat{Y}^{(k)}_{hi+} - \hat{\boldsymbol{x}}'_{hi+} \dot{\boldsymbol{b}}^{(k)}$.

6.3.2 Variance Estimation with GREG

We begin with a brief review of the randomization distribution of the GREG estimator. Recall the following expression for the randomized version of the regression estimator for a single treatment label k:

$$\tilde{Y}^{(k)}_{+\mathrm{GR}} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}^{(k)}_{hi} \hat{Y}_{hi+} + \sum_{h=1}^{H} \sum_{i=1}^{M_h} (\hat{\boldsymbol{x}}_{hi+} - w_{hi} I(i \in S_h) \check{R}^{(k)}_{hi} \hat{\boldsymbol{x}}_{hi+})' \tilde{\boldsymbol{b}}^{(k)},$$

where $\tilde{\boldsymbol{b}}^{(k)}$ is the regression coefficient obtained by treating the subsample randomly labelled with treatment k as a probability sample from the finite population.

We can rewrite this as

$$\tilde{Y}^{(k)}_{+\mathrm{GR}} = \sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\check{R}^{(k)}_{hi}\hat{Y}_{hi+} + \sum_{h=1}^{H}\sum_{i=1}^{M_h}(\boldsymbol{x}_{hi+} - w_{hi}I(i\in S_h)\check{R}^{(k)}_{hi}\hat{\boldsymbol{x}}_{hi+})'\hat{\boldsymbol{b}} + \sum_{h=1}^{H}\sum_{i=1}^{M_h}(\boldsymbol{x}_{hi+} - w_{hi}I(i\in S_h)\check{R}^{(k)}_{hi}\hat{\boldsymbol{x}}_{hi+})'(\tilde{\boldsymbol{b}}^{(k)} - \hat{\boldsymbol{b}}),$$

where $\hat{\boldsymbol{b}}$ is the regression coefficient obtained from the full sample, ignoring treatment labels, as defined above. In large samples, $\tilde{\boldsymbol{b}}^{(k)}$ will be approximately equal to $\hat{\boldsymbol{b}}$ due to random treatment assignment. Therefore, in large samples, $\sum_{h=1}^{H}\sum_{i=1}^{M_h}(\boldsymbol{x}_{hi+} - w_{hi}I(i\in S_h)\check{R}^{(k)}_{hi}\hat{\boldsymbol{x}}_{hi+})'(\tilde{\boldsymbol{b}}^{(k)} - \hat{\boldsymbol{b}})$ will often be negligible. Therefore, one can approximate the regression estimator as

$$\begin{aligned}\tilde{Y}^{(k)}_{+\mathrm{GR}} &\approx \sum_{h=1}^{H}\sum_{i\in S_h} w_i\check{R}^{(k)}_{i}\hat{Y}_{hi+} + \sum_{h=1}^{H}\sum_{i=1}^{M_h}(\boldsymbol{x}_{hi+} - w_{hi}I(i\in S_h)\check{R}^{(k)}_{hi}\hat{\boldsymbol{x}}_{hi+})'\hat{\boldsymbol{b}} \\ &= \sum_{h=1}^{H}\sum_{i=1}^{M_h}\boldsymbol{x}'_{hi+}\hat{\boldsymbol{b}} + \sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\check{R}^{(k)}_{hi}(\hat{Y}_{hi+} - \hat{\boldsymbol{x}}'_{hi+}\hat{\boldsymbol{b}})\end{aligned}$$

An argument like the one presented in Särndal et al. [1992] sec. 6.6 shows that this is actually the Taylor linearization. This approximation can be used to derive a variance estimator.

Under the randomization distribution of treatment labels, the first term is not random. Standard sampling theory reveals that the variance of $\tilde{Y}^{(k)}_{+\mathrm{GR}}$ is approximately

$$\sum_{h=1}^{H}\frac{m_h(m_h - m_h^{(k)})}{m_h^{(k)}}\frac{1}{2m_h(m_h-1)}\sum_{i\in S_h}(w_{hi}\hat{e}_{hi+} - w_{hi'}\hat{e}_{hi'+})^2,$$

where $\hat{e}_{hi+} = \hat{Y}_{hi+} - \hat{\boldsymbol{x}}'_{hi+}\hat{\boldsymbol{b}}$.

We approximate the covariances involved in the linearization randomization test with the covariances of this linear approximation. Knowing that the blocks are independent, we will compute

the covariances within a block. When $k \neq k'$, we have

$$\begin{aligned}
\mathrm{Cov}_R\left(\sum_{i\in S_h} w_{hi}\check{R}_{hi}^{(k)}\hat{e}_{hi+}, \sum_{i\in S_h} w_{hi}\check{R}_{hi}^{(k')}\hat{e}_{hi+}\right) &= \sum_{i\in S_h}\sum_{i'\in S_h} w_{hi}w_{hi'}\hat{e}_{hi+}\hat{e}_{hi'+}\mathrm{Cov}_R(\check{R}_{hi}^{(k)}, \check{R}_{hi'}^{(k')}) \\
&= \sum_{i\in S_h} w_{hi}^2\hat{e}_{hi+}^2\mathrm{Cov}_R(\check{R}_{hi}^{(k)}, \check{R}_{hi}^{(k')}) + \sum_{i\in S_h}\sum_{i'\neq i} w_{hi}w_{hi'}\hat{e}_{hi+}\hat{e}_{hi'+}\mathrm{Cov}_R(\check{R}_{hi}^{(k)}, \check{R}_{hi'}^{(k')}) \\
&= \sum_{i\in S_h} w_{hi}^2\hat{e}_{hi+}^2(-1) + \sum_{i\in S_h}\sum_{i'\neq i} w_{hi}w_{hi'}\hat{e}_{hi+}\hat{e}_{hi'+}(\frac{1}{m_h-1}) \\
&= \frac{-1}{m_h-1}\sum_{i\in S_h}\sum_{i'\in S_h}(w_{hi}^2\hat{e}_{hi+}^2 - w_{hi}w_{hi'}\hat{e}_{hi+}\hat{e}_{hi'+}) \\
&= -m_h\frac{1}{2m_h(m_h-1)}\sum_{i\in S_h}\sum_{i'\in S_h}(w_{hi}\hat{e}_{hi} - w_{hi'}\hat{e}_{hi'+})^2
\end{aligned}$$

This derives an estimated covariance matrix of $\tilde{\boldsymbol{Y}}_{+\mathrm{GR}}$ as

$$\hat{\mathrm{Var}}_R(\tilde{\boldsymbol{Y}}_{+\mathrm{GR}}) = \sum_{h=1}^{H} m_h s_{h,w\hat{e}}^2(\mathbf{D}_h - \mathbf{J}),$$

which is the same as the covariance matrix under the NHT estimator with $\hat{e}_{hi+}$ replacing $\hat{Y}_{hi+}$.

6.4 Linearization of the Randomization Test

Many finite population parameters of interest that are not totals can be expressed as an explicit function of population totals. When a parameter of interest is an explicit function of finite population totals, a natural estimator is obtained by plugging in NHT-estimated totals to the function of totals. One common approach to estimate the design variance of this NHT plug-in estimator is to linearize the function. The linearization yields a "Taylor deviate", which has the property that the variance estimate for the "total" of the Taylor deviate approximates the variance of the original total. In this section, we explain how this linearization can be applied to estimate a contrast of nonlinear functions of experimental totals in the context of a randomization test for an experiment embedded in a complex survey.

We start the discussion with the difference of log odds, which is also known as the log odds ratio. If $\hat{N}^{(k)}$ is the estimated total of OUs in the population based on the subsample assigned treatment k and $\hat{Y}_+^{(k)}$ is the number of EUs receiving treatment k that had some outcome, then the odds ratio for that outcome between treatments k and k' is

$$\frac{\hat{Y}_+^{(k)}(\hat{N}^{(k')} - \hat{Y}^{(k')})}{(\hat{N}^{(k)} - \hat{Y}^{(k)})\hat{Y}_+^{(k')}}.$$

If we take the logarithm of this expression, we get the difference of log odds of

$$\log\left(\frac{\hat{Y}_+^{(k)}(\hat{N}^{(k')} - \hat{Y}^{(k')})}{(\hat{N}^{(k)} - \hat{Y}^{(k)})\hat{Y}_+^{(k')}}\right) = \log\left(\frac{\hat{Y}_+^{(k)}}{\hat{N}^{(k)} - \hat{Y}_+^{(k)}}\right) - \log\left(\frac{\hat{Y}_+^{(k')}}{\hat{N}^{(k')} - \hat{Y}_+^{(k')}}\right).$$

Thus, we see a contrast (a difference of two treatments in this case) of the same function of two survey totals, $f(y_1, y_2) = \log(y_1/(y_2 - y_1))$, evaluated first at $(y_1, y_2) = (\hat{Y}_+^{(k)}, \hat{N}_+^{(k)})$, and second at $(y_1, y_2) = (\hat{Y}_+^{(k')}, \hat{N}_+^{(k')})$.

Now that we have seen an example of how to identify the function, we will examine the theory of linearization for the case of general contrasts of functions.

6.4.1 Notation and Treatment Assignment Distribution

For notation, suppose there is a differentiable function $f(y_1, \ldots, y_P)$, and that

$$\dot{Y}_p^{(k)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{p,hi}^{(k)}$$
$$\hat{Y}_p^{(k)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{p,hi+} \check{T}_{hi}^{(k)}.$$

In this notation, $\dot{Y}_p^{(k)}$ is the estimator of the total from the entire sample as if each element of the sample had been assigned treatment k. These quantities is not observable from the sample data for any k or p because not all of the experimental units were assigned the same treatment. Following previous notation, $\hat{Y}_p^{(k)}$ is the estimator for the total from the part of the sample that was assigned

treatment k in the experimental design. These quantities are observable after the treatment has been assigned for all k and p. Both of these quantities can be expressed as vectors for each variable p as $\dot{\boldsymbol{Y}}_p = (\dot{Y}_p^{(1)}, \ldots, \dot{Y}_p^{(K)})$ and $\hat{\boldsymbol{Y}}_p = (\hat{Y}_p^{(1)}, \ldots, \hat{Y}_p^{(K)})$

In order to define a contrast, we will define a vector-valued function with P vector arguments. We define $f^{(k)}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P) = f(y_1^{(k)}, \ldots, y_P^{(k)})$, and we define $\boldsymbol{f}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)$ to be the K vector with the k^{th} element $f^{(k)}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)$. Expressed in context, the k^{th} element of $\boldsymbol{f}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)$ is the original function evaluated at the estimated totals under the k^{th} treatment in the survey experiment.

We now investigate the treatment assignment distribution of $\mathbf{C}\boldsymbol{f}(\hat{\boldsymbol{Y}}_1, \ldots, \hat{\boldsymbol{Y}}_P)$. Let $\boldsymbol{c}^{(k)}$ be the k^{th} column of $\mathbf{C}$. Then we have

$$\mathbf{C}\boldsymbol{f}(\hat{\boldsymbol{Y}}_1, \ldots, \hat{\boldsymbol{Y}}_P) = \sum_{k=1}^{K} \boldsymbol{c}^{(k)} f(\hat{Y}_1^{(k)}, \ldots, \hat{Y}_P^{(k)}).$$

Linearizing about the full-sample totals for each treatment, we have

$$\begin{aligned} f(\hat{Y}_1^{(k)}, \ldots, \hat{Y}_P^{(k)}) &\approx f(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)}) + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)}) \right] (\hat{Y}_p^{(k)} - \dot{Y}_p^{(k)}) \\ &= a^{(k)} + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi} \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)}) \right] \hat{Y}_{p,hi+}^{(k)}, \end{aligned}$$

where $a^{(k)}$ is a basket for all additive terms that are non-random conditioned on the sample. Let $z_{hi}^{(k)} = \sum_{p=1}^{P} [\frac{\partial f}{\partial y_p}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)})] \hat{Y}_{p,hi+}^{(k)}$. Then this can be expressed as

$$f(\hat{Y}_1^{(k)}, \ldots, \hat{Y}_P^{(k)}) \approx a^{(k)} + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} z_{hi}^{(k)}.$$

At this point, we can analyze this like a standard treatment assignment distribution. To estimate the variance from the sample, we would replace $z_{hi}^{(k)}$ with $\hat{z}_{hi}^{(k)} = \sum_{p=1}^{P} [\frac{\partial f}{\partial y_p}(\hat{Y}_1^{(k)}, \ldots, \hat{Y}_P^{(k)})] \hat{Y}_{p,hi+}^{(k)}$.

6.4.2 Randomization Test

For the randomization test, the test statistic is the function of the totals from the initial treatment assignment, as above. To compute the variance, we compute the Taylor deviate, and treat it like a study variable. The randomization distribution is derived under the null hypothesis of no treatment effect. Therefore, as in the randomization test for linear contrasts discussed in Chapters 4 and 5, the linearization totals will be based on randomized treatment labels ignoring the original treatment assignment and any treatment effects. That is, the randomization statistic for a total is

$$\tilde{Y}_p^{(k)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \hat{Y}_{p,hi+}.$$

We center the linearization of the function evaluated at the expected values of all totals under the randomization distribution, conditioned on the sample and treatment assignment, which is

$$\tilde{Y}_p^{(\bullet)} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{Y}_{p,hi+}.$$

This gives

$$\begin{aligned} f(\tilde{Y}_1^{(k)}, \ldots, \tilde{Y}_P^{(k)}) &\approx f(\tilde{Y}_1^{(\bullet)}, \ldots, \tilde{Y}_P^{(\bullet)}) + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p} (\tilde{Y}_1^{(\bullet)}, \ldots, \tilde{Y}_P^{(\bullet)}) \right] (\tilde{Y}_p^{(k)} - \tilde{Y}_p^{(\bullet)}) \\ &= a + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p} (\tilde{Y}_1^{(\bullet)}, \ldots, \tilde{Y}_P^{(\bullet)}) \right] \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \hat{Y}_{p,hi+} \\ &= a + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p} (\tilde{Y}_1^{(\bullet)}, \ldots, \tilde{Y}_P^{(\bullet)}) \right] \hat{Y}_{p,hi+}, \end{aligned}$$

where a is a catch-all term for everything nonrandom. Note there is no superscript k as this term does not depend on k here. The Taylor deviate is thus $\tilde{z}_{hi} = \sum_{p=1}^{P} [\frac{\partial f}{\partial y_p} (\tilde{Y}_1^{(\bullet)}, \ldots, \tilde{Y}_P^{(\bullet)})] \hat{Y}_{p,hi+}$. Then we have

$$f(\tilde{Y}_1^{(k)}, \ldots, \tilde{Y}_P^{(k)}) \approx a + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi}^{(k)} \tilde{z}_{hi}.$$

From here, we see that the standard theory works to get a variance estimate.

To do this, we note that (since a does not depend on k)

$$\mathbf{C}\boldsymbol{f}(\tilde{\boldsymbol{Y}}_1, \ldots, \tilde{\boldsymbol{Y}}_P) \approx \sum_{h=1}^{H} \sum_{i \in S_h} \mathbf{C}\boldsymbol{R}_{hi} w_{hi} \tilde{z}_{hi}.$$

Then, from the randomization distribution, we see that the mean is approximately 0, and the variance can be approximated by

$$\mathrm{Var}_R(\mathbf{C}\boldsymbol{f}(\tilde{\boldsymbol{Y}}_1, \ldots, \tilde{\boldsymbol{Y}}_P)) \approx \sum_{h=1}^{H} m_h s^2_{h,w\tilde{z}} \mathbf{C}\mathbf{D}_h\mathbf{C}',$$

where $s^2_{h,w\tilde{z}}$ is the variance of the weighted Taylor deviates, following previous notation.

6.4.3 Extension to Functions of Contrasts

Sometimes a researcher may be interested in functions of contrasts of functions. Returning to our earlier example, this may be relevant if a researcher was studying the odds ratio, rather than the difference of log odds. Using the notation in the introduction we have that the odds ratio can be expressed as

$$\mathrm{OR}_{k,k'} = \frac{\hat{Y}_+^{(k)}(\hat{N}^{(k')} - \hat{Y}_+^{(k')})}{(\hat{N}^{(k)} - \hat{Y}_+^{(k)})\hat{Y}_+^{(k')}} = \exp\left\{ \log\left(\frac{\hat{Y}_+^{(k)}}{\hat{N}^{(k)} - \hat{Y}_+^{(k)}} \right) - \log\left(\frac{\hat{Y}_+^{(k')}}{\hat{N}^{(k')} - \hat{Y}_+^{(k')}} \right) \right\}.$$

Thus

$$\mathrm{OR}_{k,k'} = g\left(f(\hat{Y}_+^{(k)}, \hat{N}^{(k)}) - f(\hat{Y}_+^{(k')}, \hat{N}^{(k')}) \right),$$

where $g(\gamma) = \exp(\gamma)$ and $f(y_1, y_2) = \log(y_1/(y_2 - y_1))$, and the argument of g is clearly a contrast of two functions.

We now consider the linearization of functions of contrasts in general. For ease of notation, we restrict ourselves to a contrast vector in this case; notes on how the notation can be extended to a contrast matrix are included at the end of this section. Extending notation from the previous sections, let

$$\hat{\theta}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P) = g\left(\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)\right),$$

where $g : \mathbb{R} \to \mathbb{R}$ is a function with argument γ and each $\boldsymbol{y}_p$ is a K-vector of terms $y_p^{(k)}$ corresponding to estimates of the value of variable p for the observations receiving treatment k. The chain rule yields

$$\begin{aligned}\frac{\partial \hat{\theta}}{\partial y_p^{(k)}}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P) &= \frac{\partial}{\partial y_p^{(k)}} g\left(\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)\right) \\ &= \left[\frac{dg}{d\gamma}\left(\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)\right)\right] c^{(k)} \frac{\partial f}{\partial y_p^{(k)}}(y_1^{(k)}, \ldots, y_P^{(k)}).\end{aligned}$$

In this expression, note that the first term (the derivative of g evaluated at the contrast) does not depend on k. This means that if we write out the linearization for the treatment assignment distribution, we have

$$\begin{aligned}\hat{\theta}(\hat{\boldsymbol{Y}}_1, \ldots, \hat{\boldsymbol{Y}}_P) &\approx \hat{\theta}(\dot{\boldsymbol{Y}}_1, \ldots, \dot{\boldsymbol{Y}}_P) + \sum_{k=1}^{K}\sum_{p=1}^{P}\left[\frac{\partial \hat{\theta}}{\partial y_p^{(k)}}(\dot{\boldsymbol{Y}}_1, \ldots, \dot{\boldsymbol{Y}}_P)\right](\hat{Y}_p^{(k)} - \dot{Y}_p^{(k)}) \\ &= a + \sum_{k=1}^{K}\sum_{p=1}^{P}\left[\frac{dg}{d\gamma}\left(\boldsymbol{c}'\boldsymbol{f}(\dot{\boldsymbol{Y}}_1, \ldots, \dot{\boldsymbol{Y}}_P)\right)\right] c^{(k)} \frac{\partial f}{\partial y_p^{(k)}}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)})\hat{Y}_p^{(k)} \\ &= a + \left[\frac{dg}{d\gamma}\left(\boldsymbol{c}'\boldsymbol{f}(\dot{\boldsymbol{Y}}_1, \ldots, \dot{\boldsymbol{Y}}_P)\right)\right] \sum_{k=1}^{K} c^{(k)} \sum_{p=1}^{P} \frac{\partial f}{\partial y_p^{(k)}}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)})\hat{Y}_p^{(k)}.\end{aligned}$$

Here, we see that the expression is a nonrandom constant plus the derivative evaluated at the value of the contrast as if measurements had been taken on the entire population. In an a priori power calculation, one could use the desired difference, or an estimate of what the difference is as a position to evaluate the derivative.

Similarly, we find that for the randomization distribution:

$$\begin{aligned}\hat{\theta}(\tilde{\boldsymbol{Y}}_1, \ldots, \tilde{\boldsymbol{Y}}_P) &\approx a + \left[\frac{dg}{d\gamma}\left(\boldsymbol{c}'\boldsymbol{f}(\tilde{Y}_1^{(\bullet)}\mathbf{1}, \ldots, \tilde{Y}_P^{(\bullet)}\mathbf{1}\right)\right] \sum_{k=1}^{K} c^{(k)} \sum_{p=1}^{P} \frac{\partial f}{\partial y_p^{(k)}}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)})\hat{Y}_p^{(k)} \\ &= a + \left[\frac{dg}{d\gamma}(0)\right] \sum_{k=1}^{K} c^{(k)} \sum_{p=1}^{P} \frac{\partial f}{\partial y_p^{(k)}}(\dot{Y}_1^{(k)}, \ldots, \dot{Y}_P^{(k)})\hat{Y}_p^{(k)},\end{aligned}$$

by the definition of $\boldsymbol{f}$ and the fact that $\boldsymbol{c}'\mathbf{1} = 0$. Thus for the randomization test, the constant multiplied by the linearization of the function is just $g'(0)$!

As noted, we have thus far only discussed functions a single contrast. For functions of a contrast matrix, one can define a vector-valued function $\boldsymbol{g}(\boldsymbol{\gamma})$, in a way similar to how $\boldsymbol{f}$ was defined from f, and evaluate the multiplier for each contrast individually. For power calculations, you can use the estimated difference in each contrast. For the randomization test, one would simply use a vector version of $\boldsymbol{g}(\mathbf{0})$. The multiplier would end up being $g'(0)\mathbf{1}$.

6.4.4 Linearization Using GREG

Many modern surveys use calibration estimators to improve precision. We now consider problems of linearization where the point estimate of the treatment totals is estimated using the GREG estimator rather than the NHT estimator. Considering the treatment assignment distribution, this leads to the expression

$$f(\hat{Y}^{(k)}_{\mathrm{GR},1}, \ldots, \hat{Y}^{(k)}_{\mathrm{GR},P}) \approx f(\dot{Y}^{(k)}_{\mathrm{GR},1}, \ldots, \dot{Y}^{(k)}_{\mathrm{GR},P}) + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p}(\dot{Y}^{(k)}_{\mathrm{GR},1}, \ldots, \dot{Y}^{(k)}_{\mathrm{GR},P}) \right] (\hat{Y}^{(k)}_{\mathrm{GR},p} - \dot{Y}^{(k)}_{\mathrm{GR},p})$$

where

$$\dot{Y}^{(k)}_{\mathrm{GR},p} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \boldsymbol{x}'_{hi+} \dot{\boldsymbol{b}}^{(k)} + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} (\hat{Y}^{(k)}_{hi+} - \hat{\boldsymbol{x}}'_{hi+} \dot{\boldsymbol{b}}^{(k)})$$

is the estimated total using the generalized regression estimator had all experimental units been assigned treatment k. Note that $\dot{Y}^{(k)}_{\mathrm{GR},p}$ is not random, conditioned on the sample.

If we only look at the parts of the linear approximation that are random conditioned on the sample, we find that

$$\begin{aligned} f(\hat{Y}^{(k)}_{\mathrm{GR},1}, \ldots, \hat{Y}^{(k)}_{\mathrm{GR},P}) &\approx a^{(k)} + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p}(\dot{Y}^{(k)}_{\mathrm{GR},1}, \ldots, \dot{Y}^{(k)}_{\mathrm{GR},P}) \right] \hat{Y}^{(k)}_{\mathrm{GR},p} \\ &\approx a^{(k)} + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{T}_{hi} \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p}(\dot{Y}^{(k)}_{\mathrm{GR},1}, \ldots, \dot{Y}^{(k)}_{\mathrm{GR},P}) \right] (\hat{Y}_{hi+} - \hat{\boldsymbol{x}}' \dot{\boldsymbol{b}}^{(k)}). \end{aligned}$$

Now if we let

$$\hat{z}_{hi}^{(k)} = \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p} (\dot{Y}_{\mathrm{GR},1}^{(k)}, \ldots, \dot{Y}_{\mathrm{GR},P}^{(k)}) \right] (\hat{Y}_{hi+} - \hat{\boldsymbol{x}}' \dot{\boldsymbol{b}}^{(k)}),$$

we can calculate the variance using standard arguments.

For the randomization distribution, a similar argument works to derive

$$f(\tilde{Y}_{\mathrm{GR},1}^{(k)}, \ldots, \tilde{Y}_{\mathrm{GR},P}^{(k)}) \approx a + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \check{R}_{hi} \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p} (\tilde{Y}_{\mathrm{GR},1}^{(\bullet)}, \ldots, \tilde{Y}_{\mathrm{GR},P}^{(\bullet)}) \right] (\hat{Y}_{hi+} - \hat{\boldsymbol{x}}' \hat{\boldsymbol{b}}),$$

with

$$\tilde{Y}_{\mathrm{GR},p}^{(\bullet)} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \boldsymbol{x}'_{hi+} \hat{\boldsymbol{b}} + \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} (\hat{Y}_{hi+} - \hat{\boldsymbol{x}}'_{hi+} \hat{\boldsymbol{b}}).$$

The variance can then be derived by standard arguments using

$$\tilde{z}_{hi} = \sum_{p=1}^{P} \left[\frac{\partial f}{\partial y_p} (\tilde{Y}_{\mathrm{GR},1}^{(\bullet)}, \ldots, \tilde{Y}_{\mathrm{GR},P}^{(\bullet)}) \right] (\hat{Y}_{hi+} - \hat{\boldsymbol{x}}'_{hi} \hat{\boldsymbol{b}}).$$

6.5 Difficulties of Real Surveys

In the previous chapters of this dissertation, we have ignored several important aspects of surveys, including nonsampling errors (nonresponse, coverage errors, etc.) and modern survey designs (including adaptive or responsive designs).

With undercoverage error, care will be needed when generalizing results from the sample to the population, because some members of the population cannot be in the sample. For determining inference from the experiment to the sample, however, undercoverage does not present any additional issues because the coverage error only affects inference from the sample to the population. Nonresponse errors can cause issues by removing some experimental units from the sample, leading to a missing data problem. Some weighting approaches to deal with this issue are presented in Section 6.5.1. Another potential issue is measurement error. Measurement error could occur with experiments in surveys when an experimental treatment changes a respondent's response. In

some cases (for example, chapters 7, 8, and 17–20 of Lavrakas et al. [2019]), this measurement error could be what the experiment is trying to measure. However, as the goal of many surveys is to get accurate measurements of the population, this is something to be aware of when considering embedding experiments in a survey.

6.5.1 Adjusting for Nonresponse

The theory presented thus far in the dissertation, like much theory in survey statistics, was derived under the situation of perfect response from the sample. If the experiment is based on estimating response rates or survey costs per response, then all the relevant information may be obtained whether or not individuals respond to the survey. However, in many embedded experiments, nonresponse leads to loss of information, so additional assumptions will need to be formulated to conduct any analysis on survey experiments that involve major nonresponse. Before starting to discuss nonresponse, one point needs to be made. If nonresponse occurs below the level of the experimental units, then that nonresponse can be taken into account in the EU total estimates, which will not have any effect on the methodology described thus far in this dissertation. However, in this case, an analysis of this differential nonresponse would be useful in understanding the effects of the treatments. In the remainder of this subsection, we focus on estimators for the randomization distribution in the case of EU level nonresponse, which can be a problem if the experimental units are households or individuals.

To motivate the approaches for nonresponse adjustment, we consider a simple example. Suppose there is a block with 20 units, 10 of which were assigned the treatment and 10 of which were assigned a control. Suppose that out of the treatment group, there were 8 responses, and out of the control group, there were 4 responses. If we are to perform a randomization test on data including nonresponse, we would need to condition on who responds in some sense. There are still two options in how to adjust for response: we could ignore the relationship between the treatment and response, or condition on the number of responses within each treatment.

One method to deal with nonresponse is to treat nonresponse as a factor that is fixed with treatment, but unknown at the time of original assignment. In this method, the treatment randomization will be treated at the level of the sample, but only the set of respondents will be used in the analysis. In the example started above, this would involve randomizing the treatments to all 20 units, and then limiting each randomized dataset to the 12 units that responded. Another method is to condition on the number of responses within each treatment, and only randomize treatments on the set of respondents. In this method, we would consider only the 12 units that responded, and consider randomizations that assign 8 of these units to treatment, and 4 to control. In the remainder of these sections, we will explore the mathematical properties of these methods. We will call the first method described in this paragraph the "sample randomization method" and the second method described in the paragraph the "response set randomization method."

To aid in the mathematical discussion, we introduce some new notation for the randomization test with these methods used for nonresponse adjustment. We let S^r denote the set of respondents from the population, m_h^r be the total number of respondents within each block h, and $m_h^{r(k)}$ be the number of respondents within each treatment/block combination. We also introduce a version of the treatment randomization indicator for randomization of treatment label on the response set: $\check{\boldsymbol{R}}_{hi}^r = \frac{m_h^r}{m_h^{r(k)}} \boldsymbol{R}_{hi}$. We additionally include two nonresponse adjustments. We introduce a nonresponse-adjusted weight to weight from the response set to the finite population, denoted $w_{hi}^r(S^r)$, and we allow a nonresponse adjustment to weight from the response set within a given set of EUs labeled as receiving the same treatment to all such EUs in the sample, denoted $q_{hi}(S^r, \mathbf{R})$. $\mathbf{R}$ is the matrix containing the randomized treatment labels of all EUs in the sample (regardless of whether they responded). The former weight is a nonresponse-adjusted weight for the sample, while the latter is an adjustment factor to ensure the effects of nonresponse between the treatments are similar in all randomizations. Before showing the two estimators using the above notation, we will show the estimator ignoring nonresponse for comparison:

$$\hat{\boldsymbol{Y}} = \sum_{h=1}^{H} \sum_{i \in S_h} w_{hi} \hat{\boldsymbol{Y}}_{hi} \circ \check{\boldsymbol{R}}_{hi}.$$

The sample randomization method yields the estimator

$$\hat{\boldsymbol{Y}}_{\text{samp}} = \sum_{h=1}^{H} \sum_{i \in S_h} I(hi \in S^r) w_{hi} q_{hi}(S^r, \mathbf{R}) \hat{\boldsymbol{Y}}_{hi} \circ \check{\boldsymbol{R}}_{hi},$$

and the response set randomization method yields the estimator

$$\hat{\boldsymbol{Y}}_{\text{resp}} = \sum_{h=1}^{H} \sum_{i \in S_h^r} w_{hi}^r(S^r) \hat{\boldsymbol{Y}}_{hi} \circ \check{\boldsymbol{R}}_{hi}^r.$$

The sample randomization estimator takes the original estimate, ignoring nonresponse, and then provides a nonresponse adjustment within each treatment label group for each randomization. The response set randomization estimator conditions on the response set, and uses a nonresponse adjustment that ignores experimental treatment randomization, but uses an implicit nonresponse adjustment by adjusting the sizes of the response sets within each treatment.

The sample randomization method allows for the examination of differences in nonresponse-adjusted estimators between the treatments. It allows for different nonresponse adjustment on each randomization, so it sees how the estimates will change based on rerunning the experiment on the sample many times. One disadvantage of the sample randomization method is that it can lead to situations where there are no respondents for a given treatment in a block for the randomized complete block design. Allowing for a nonresponse adjustment for each treatment randomization can also lead to complexity, but this complexity could potentially be overcome by easily linearizable calibration estimator like GREG.

The response set randomization method is a good simple method that can work on any response set that still has complete blocks. In this case, response set is analyzed as if it were the sample, and the treatment design is analyzed as if it were designed on the response set. To go from the overall responses to the responses within a treatment, the totals are multiplied by their inclusion probabilities based on their proportions in the response set. This acts like a built-in nonresponse adjustment across treatments. This can detect differences in the responses between both treatments. This is useful for determining differences in response behavior between each treatment, but is less

useful if the goal is to determine differences between nonresponse adjusted estimators between the treatments.

There is a case, however, where the response set randomization estimator gives the same answer as the sample randomization estimator with a nonresponse adjustment within each treatment, provided that we additionally restrict the sample randomization estimator to cases where the number of units in the response set assigned each treatment remains unchanged. If the nonresponse adjustment for the sample involves a weighting cell adjustment where the cells are the blocks used in the experimental design, this method will automatically adjust that cell adjustment to as if it were done within both blocks and treatments.

To better explain how the response set randomization estimator can mimic a weighting cell adjustment, we return to the example started above. Recall that we were assuming a block of 20 units, with responses from 8 out of 10 units assigned treatment and 4 out of 10 units assigned control. We will perform the nonresponse adjustment in two ways. The first will be performing a nonresponse adjustment within block and treatment, followed by assigning treatment weights based on how the treatments were assigned in the sample, as in the sample randomization estimator, in the case where the number of responses within treatment and control stays at 8 and 4 respectively. The second will involve performing a weighting cell adjustment within the block, ignoring treatment, and then performing the treatment adjustment on the EUs in the response set, using the response set randomization estimator. For mathematical simplicity, we assume the sampling design was self-weighting, so $w_{hi} = 1$ for all EUs.

The weighting cell adjustment within block and treatment gives a weight of $10/8 = 5/4$ in the treatment group (to weight from 8 respondents to 10 sampled units assigned the treatment) and $10/4 = 5/2$ in the control group. Weighting again for treatment multiplies both of these weights by 2 as half of the sampled EUs were assigned each treatment, giving final weights for the sample randomization estimator (if treatment assignment levels are maintained) of $w_{hi}q_{hi}(S^r, \mathbf{R})\check{R}_{hi}^{(1)} = 5/2$ for responding EUs in the treatment group ($k = 1$) and $w_{hi}q_{hi}(S^r, \mathbf{R})\check{R}_{hi}^{(0)} = 5$ for responding EUS in the control group ($k = 0$). The overall block weighting cell adjustment gives weights in

each block of $20/12 = 5/3$. The weighting cell adjustment within each treatment would then be $12/8 = 3/2$ in the treatment group and $12/4 = 3$ in the control group. This gives final weights of $w_{hi}^r(S^r)\check{R}_{hi}^{(1)} = 5/2$ in the treatment group ($k = 1$) and $w_{hi}^r(S^r)\check{R}_{hi}^{(0)} = 5$ in the control group ($k = 0$), which are the same as the results as the sample randomization estimator.

It is also worth mentioning that in the above discussion, we restricted the sample randomization method to cases where the number of responses in both treatments was the same as the original study. It is a reasonable guess that this method will always give the same result as the response set randomization estimator, because this new restriction results in the same sample sizes of treatment and control in both cases. However, in the above discussion, we were using a weighting cell adjustment, which only depends on the number of responses in each block or each treatment. For more complicated nonresponse adjustments that use auxiliary information on each individual, the estimators will not necessarily give the same answer, even if the number of responses within each treatment in the sample randomization estimated is restricted.

Our final recommendation is that the method for nonresponse adjustment can depend on the problem. If you are analyzing an experiment where you really want to see the effect on nonresponse adjusted estimators between two treatments, it may be worth it to use the sample randomization estimator. This is especially true if the nonresponse adjustment is based on an easily linearizable calibration estimator like the GREG. If computing time is a factor, or the weighting method involves weighting class adjustment at the level of blocks, then the response set randomization estimator may make more sense. More investigation will be needed in the cases where nonresponse is significant enough to remove a treatment from a block. Another important avenue for future work would be simulation studies on methods of dealing with nonresponse to determine which methods are more robust to extreme nonresponse, and to see how they perform at solving different problems.

6.6 Conclusions

In this chapter, we extended the methodology of the randomization test for use with the GREG estimator, and for nonlinear statistics. While there was no theory developed in this chapter, this chapter helps illuminate the broad applicability of this method.

As experiments and surveys become more complicated, and the focus on data quality becomes more important, it is useful to have methods that can detect differences in any situation where the randomization can be described. A Monte-Carlo randomization test is a flexible method that can be applied to many situations.

Additionally, there may be situations where researchers need a simple method to understand results. Using linearization methods can be useful to give valid results that do not require additional summarization of results.

More work should be done on understanding how randomization methods work in the presence nonresponse. Given that randomization methods are based on the initial randomization of the sample to different treatments, nonresponse is likely to be a major problem. More simulations investigating the robustness of randomization methods to nonresponse, and the potential improvements of different nonresponse adjustment methods, will be crucial.

Chapter 7

Conclusions and Future Work

In this dissertation, we introduced randomization tests for experiments embedded within complex surveys. We also derived a linearization variance approximation for the nondifferentiable dissimilarity index. We provided applications of our new methodologies in simulation studies, and in practice with a survey conducted by the Health District of Northern Larimer County.

While a design-based test for experiments in surveys does exist (van den Brakel [2001]), it is important to have other options. While there are situations where the randomization test has less power than the design-based test, in some cases, the power loss is not very large. Also, the randomization test admits a Monte Carlo version that can be used in situations where an asymptotic approximation may not be applicable or desirable.

Our findings confirm what has been said by van den Brakel [2001] (Section 2.5) and Smith [1983]. When the goal of an experiment is to determine information about the conduct of the survey or to determine effectiveness on measures of data quality, a model-based analysis can work well. However, when the goal is to compare population statistics conditioned on the sample, an approach that explicitly accounts for the sampling design is preferable.

The simulation results also provided valuable insight for randomization tests of experiments embedded in complex surveys in practice. The results directly examining a contrast of totals revealed that while examining results comparing uncalibrated estimates of totals can be useful when deriving theory, such tests do not always work well in practice. The test comparing totals directly for the CRD had almost no power for the sizes of differences tested (Figure 4.4), though it performed as well as the test of means in the RCBD (Figure 5.2). It is thus useful to use some sort of calibration when comparing statistics in practice, so that the variation in the estimated population size between the treatments does not overwhelm the treatment effect. This calibration could involve normalization to the estimated population size $\hat{N}_+$ (as in the ratio estimator), or using the GREG estimator.

This dissertation has created an option for randomization theory to be used in survey experiments, more development of the theory can be made. One avenue is developing theory for different experimental designs. Examples include split-plot, fractional factorial, Latin square, or incomplete block designs. It could also valuable to develop theory for analyzing the experiments designed with more complicated adaptive procedures. In these cases, even the Monte Carlo test may not be obvious (for example, Example 8 in Proschan and Dodd [2019]).

Another avenue for development is to drop or relax assumptions. One assumption to potentially relax is the assumption of additivity (4.6 and 5.1). One may consider cases where there is variation of the treatment effects between individuals, like what we used in our simulation studies (Sections 4.8 and 5.7). Another assumption to relax is the assumption used for power calculations that the EUs are drawn with probability proportional to size (4.11,5.2). Another way to state this assumption is that there is no variance in the weighted EU sizes. One could naturally consider cases where this variation is not very large. A third example is the assumption that survey weights are uniformly bounded (Assumption *A2* or *B3*). As people react differently to different types of treatment, and nonresponse issues often create fairly unbalanced weights in surveys, relaxing this strict bound would be very useful for many practical problems.

Another avenue for advancement is in simulation studies. One very realistic situation is differential nonresponse between the two treatments. This situation would be of interest to maintaining data quality for a survey switching modes. It would be useful for survey practice to know if the test can detect when there is no difference in nonresponse adjusted estimators.

It would also be valuable to develop theory about how the Monte Carlo test extends to the general population. The test proposed is a test conditioned on the chosen sample. Our current results show that, under some conditions, asymptotically normal test results can be extended to the finite population. However, some questions of potential interest may not have asymptotically normal estimators. In these cases, the Monte Carlo test is a valuable tool that can provide inference for the sample. However, there is no general theory extending this inference to the finite population. Such a result will be valuable for extending these results.

Lastly, it would be valuable to have a literature review on methods for evaluating experiments embedded in surveys. There can be many different types of response variables, and many ways to analyze the outcomes. It would be useful to understand when model-based inferences may be acceptable, and when randomization or design-based methods are preferred.

Bibliography

Rebecca Allen, Simon Burgess, Russell Davidson, and Frank Windmeijer. More reliable inference for the dissimilarity index of segregation. *Econometrics Journal*, 18:40–66, 2015. ISSN 1368423X. doi: 10.1111/ectj.12039.

Olga A. Avdeyeva and Richard E. Matland. An experimental test of mail surveys as a tool for social inquiry in russia. *International Journal of Public Opinion Research*, 25:173–194, 6 2013. ISSN 09542892. doi: 10.1093/ijpor/eds020.

D. Basu. Randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75:575–595, 1980. doi: 10.1007/978-1-4419-5825-9_28.

Joseph C. Bevis. American association for public opinion research economical incentive used for mail questionnaire. *Public Opinion Quarterly*, 12:492–493, 1948. URL https://www.jstor.com/stable/2745355.

Hind Beydoun, Audrey F. Saftlas, Kari Harland, and Elizabeth Triche. Combining conditional and unconditional recruitment incentives could facilitate telephone tracing in surveys of postpartum women. *Journal of Clinical Epidemiology*, 59:732–738, 2006. ISSN 08954356. doi: 10.1016/j.jclinepi.2005.11.011.

Paul P Biemer, Joe Murphy, Stephanie Zimmer, Chip Berry, Grace Deng, and Katie Lewis. Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, 6:240–261, 6 2018. ISSN 2325-0984. doi: 10.1093/jssam/smx015.

F. Jay Breidt, Richard A. Davis, Nan-Jung Hsu, and Murray Rosenblatt. Pile-up probabilities for the laplace likelihood estimator of a non-invertible first order moving average. *Time Series and Related Topics*, 52:1–19, 2006. doi: 10.1214/074921706000000923.

J. Michael Brick and Douglas Williams. Explaining rising nonresponse rates in cross-sectional surveys. *Annals of the American Academy of Political and Social Science*, 645:36–59, 2013. ISSN 00027162. doi: 10.1177/0002716212456834.

Ashis Kumar Chakraborty and Moutushi Chatterjee. On multivariate folded normal distribution. *Source: Sankhyā: The Indian Journal of Statistics, Series B*, 75:1–15, 2013.

Allan H Church. Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57:62–79, 1993. URL https://academic.oup.com/poq/article-abstract/57/1/62/1833464.

Jared Coopersmith, Lisa Klein Vogel, Timothy Bruursema, and Kathleen Feeney. Effects of incentive amount and type of web survey response rates. *Survey Practice*, 9:1–10, 2016. ISSN 2168-0094. doi: 10.29115/sp-2016-0002.

Charles F. Cortese, R. Frank Falk, and Jack K. Cohen. Further considerations on the methodological analysis of segregation indices. *American Soc*, 41:630–637, 1976. URL https://www.jstor.org/stable/2094840.

Jean-Claude Deville and Carl-Erik Sarndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

Peng Ding. A paradox from randomization-based causal inference. *Statistical Science*, 32:331–345, 2017. doi: 10.1214/16-STS571.

Otis Duncan and Beverly Duncan. A methodological analysis of segregation indexes. *American Sociological Review*, 20:210–217, 1955. URL https://www.jstor.com/stable/2088328.

Philip James Edwards, Ian Roberts, Mike J. Clarke, Carolyn DiGuiseppi, Reinhard Wentz, Irene Kwan, Rachel Cooper, Lambert M. Felix, and Sarah Pratap. Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, 3, 2009. ISSN 1469493X. doi: 10.1002/14651858.MR000008.pub4.

Michael R. Elliott and Richard Valliant. Inference for nonprobability samples. *Statistical Science*, 32:249–264, 2017. ISSN 08834237. doi: 10.1214/16-STS598.

Michael D. Ernst. Permutation methods: A basis for exact inference. *Statistical Science*, 19: 676–685, 2004. ISSN 08834237. doi: 10.1214/088342304000000396.

Stephen E. Fienberg and Judith M. Tanur. Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review / Revue Internationale de Statistique*, 55:75, 1987. ISSN 03067734. doi: 10.2307/1403272.

Stephen E. Fienberg and Judith M. Tanur. From the inside out and the outside in: Combining experimental and sampling structures. *Canadian Journal of Statistics*, 16:135–151, 1988. ISSN 03195724. doi: 10.2307/3314634.

Ronald Aylmer Fisher. *The Design of Experiments*. Hafner Press, 9 edition, 1971.

Wayne A. Fuller. *Sampling Statistics*. John Wiley & Sons, 2009. ISBN 9780470523551. doi: 10.1002/9780470523551.

M. H. Gail, W. Y. Tan, and S. Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75:57–64, 1988. ISSN 00063444. doi: 10.1093/biomet/75.1.57.

Paul H Garthwaite. Confidence intervals from randomization tests. *Biometrics*, 52:1387–1393, 1996.

Robert M. Groves. Nonresponse rates and nonresponse bias in household surveys: What do we know about the linkage between nonresponse rates and nonresponse bias? *Public Opinion Quarterly*, 70:646–675, 2006.

Robert M. Groves and Steven G. Heeringa. Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 169:439–457, 2006. ISSN 09641998. doi: 10.1111/j.1467-985X.2006.00423.x.

Robert M. Groves and Emilia Peytcheva. The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72:167–189, 2008. ISSN 0033362X. doi: 10.1093/poq/nfn011.

Robert M. Groves, Mick P. Couper, Stanley Presser, Eleanor Singer, and Roger Tourangeau. Experiments in producing nonresponse bias giorgina piani acosta. *Public Opinion Quarterly*, 70: 720–736, 2006.

Robert M. Groves, J. Michael Brick, Mick P. Couper, William Michael Kalsbeek, Brian Michael Harris-Kojetin, Frauke Michael Kreuter, Beth-Ellen Michael Pennell, Trivellore Michael Raghunathan, Barry Michael Schouten, Tom Michael Smith, Roger Michael Tourangeau, Ashley Bowers, Matthew Jans, Courtney Kennedy, Rachel Levenstein, Kristen Olson, Emilia Peytcheva, Sonja Ziniel, and James Wagner. Issues facing the field: Alternative practical measures of representativeness of survey respondent pools. *Survey Practice*, 1:1–6, 2008. ISSN 2168-0094. doi: 10.29115/sp-2008-0013.

Steven G. Heeringa, Brady T. West, and Patricia A. Berglund. *Applied Survey Data Analysis*. CRC Press, 2 edition, 2017.

Wassily Hoeffding. A combinatorial central limit theorem. *Source: The Annals of Mathematical Statistics*, 22:558–566, 1951.

Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192, 1952. doi: 10.1007/978-1-4612-0865-5_13.

Daniel G Horvitz and D. J. Thompson. Sampling from a discrete universe: I. sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952. ISSN 14230062. doi: 10.1159/000150726.

Julius Jahn, Calvin F Schmid, and Clarence Schrag. The measurement of ecological segregation. *American Sociological Review*, 12:293–303, 1947.

Leslie Kanuk and Conrad Berenson. Mail surveys and response rates: A literature review. *Source: Journal of Marketing Research*, 12:440–453, 1975.

Oscar Kempthorne. The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50:946–967, 1955. URL https://www.jstor.org/stable/2281178.

Oscar Kempthorne. Response to "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75:584–587, 1980.

Oscar Kempthorne and T. E. Doerfler. The behaviour of some significance tests under experimental randomization. *Biometrika*, 56:231, 1969. ISSN 00063444. doi: 10.2307/2334417.

Phillip S. Kott. The delete-a-group jackknife. *Journal of Official Statistics*, 17:521–526, 2001.

David A. Lane. Response to "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75:587–589, 1980.

Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady T. West, editors. *Experimental Methods in Survey Research*. John Wiley & Sons, Inc., 2019.

F C Leone, L S Nelson, and R B Nottingham. The folded normal distribution. *Technometrics*, 3: 543–550, 1961.

Peter Lundquist and Carl Erik Särndal. Aspects of responsive design with applications to the swedish living conditions survey. *Journal of Official Statistics*, 29:557–582, 2013. ISSN 0282423X. doi: 10.2478/jos-2013-0040.

P. C. Mahalanobis. Recent experiments in statistical sampling in the indian statistical institute. *Sankhya: The Indian Journal of Statistics*, 20:329–398, 1958.

Elvira Mauz, Elena von der Lippe, Jennifer Allen, Ralph Schilling, Stephan Müters, Jens Hoebel, Patrick Schmich, Matthias Wetzstein, Panagiotis Kamtsiuris, and Cornelia Lange. Mixing

modes in a population-based interview survey: Comparison of a sequential and a concurrent mixed-mode design for public health research. *Archives of Public Health*, 76:1–17, 2018. ISSN 20493258. doi: 10.1186/s13690-017-0237-1.

Benjamin L. Messer and Don A. Dillman. Surveying the general public over the internet using address-based sampling and mail contact procedures. *Public Opinion Quarterly*, 75:429–457, 2011. ISSN 0033362X. doi: 10.1093/poq/nfr021.

Morgan M. Millar and Don A. Dillman. Improving response to web and mixed-mode surveys. *Public Opinion Quarterly*, 75:249–269, 2011. ISSN 0033362X. doi: 10.1093/poq/nfr003.

Diana C. Mutz. *Population-Based Survey Experiments*. Princeton University Press, 2011.

R. D. Narain. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174, 1951.

National Research Council. *Nonresponse in social science surveys: A research agenda*. The National Academies Press, 2013. ISBN 0309272475. doi: 10.17226/18293. URL https://doi.org/10.17226/18293.

Jerzy Neyman, K. Iwaszkiewicz, and St. Kolodziejczyk. Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 2: 107–180, 1935. URL https://www.jstor.org/stable/2983637.

Patrick Onghena. Randomization tests or permutation tests? a historical and terminological clarification. *Randomization, Masking, and Allocation Concealment*, pages 209–228, 2017. doi: 10.1201/9781315305110.

Megan E. Patrick, Mick P. Couper, Virginia B. Laetz, John E. Schulenberg, Patrick M. O'Malley, Lloyd D. Johnston, and Richard A. Miech. A sequential mixed-mode experiment in the u.s. national monitoring the future study. *Journal of Survey Statistics and Methodology*, 6:72–97, 2018. ISSN 23250992. doi: 10.1093/jssam/smx011.

Andy Peytchev. Consequences of survey nonresponse. *Annals of the American Academy of Political and Social Science*, 645:88–111, 2013. ISSN 00027162. doi: 10.1177/0002716212461748.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4:119–130, 1937a. URL https://www.jstor.org/stable/2984124.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4: 225–232, 1937b. URL https://www.jstor.org/stable/2983647.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations iii .* the analysis of variance test. *Biometrika*, 29:322–335, 1938. URL https://doi.org/10.1093/biomet/29.3-4.322.

Michael A. Proschan and Lori E. Dodd. Re-randomization tests in clinical trials. *Statistics in Medicine*, 38:2292–2302, 5 2019. ISSN 10970258. doi: 10.1002/sim.8093.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Kevin Andrew Rader. *Methods for Analyzing Survival and Binary Data in Complex Surveys*. PhD thesis, Harvard University, 5 2014.

Michael R. Ransom. Sampling distributions of segregation indexes. *Sociological Methods & Research*, 28:454–475, 2000.

Neomi Rao. Cost effectiveness of pre-and post-paid incentives for mail survey response survey practice. *Survey Practice*, 13, 2020. doi: 10.29115/SP-2020-0004. URL https://doi.org/10.29115/SP-2020-0004.

J. Robinson. The large-sample power of permutation tests for randomization models. *The Annals of Statistics*, 1:291–296, 1973.

Donald B. Rubin. Response to "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, pages 591–593, 1980.

Barry Schouten. Statistical inference based on randomly generated auxiliary variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80:33–56, 2018. ISSN 14679868. doi: 10.1111/rssb.12242.

Barry Schouten, Fannie Cobben, and Jelke Bethlehem. Indicators for the representativeness of survey response. *Survey Methodology*, 35:101–113, 2009. ISSN 07140045.

Barry Schouten, Andy Peytchev, and James R. Wagner. *Adaptive Survey Design*. Chapman and Hall/CRC, 1 edition, 2018. ISBN 9781498767873.

Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

Natalie Shlomo, Chris Skinner, and Barry Schouten. Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, 142:201–211, 2012. ISSN 03783758. doi: 10.1016/j.jspi.2011.07.008. URL http://dx.doi.org/10.1016/j.jspi.2011.07.008.

Eleanor Singer and Cong Ye. The use and effects of incentives in surveys. *Annals of the American Academy of Political and Social Science*, 645:112–141, 1 2013. ISSN 00027162. doi: 10.1177/0002716212458082.

T M F Smith. The foundations of survey sampling: A review. *Journal of the Royal Statistical Society. Series A (General)*, 139:183–204, 1976.

T M F Smith. On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General)*, 146:394–403, 1983. URL https://about.jstor.org/terms.

Jerzy Splawa-Neyman, Dorota M. Dabrowska, and Terry P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5: 465–480, 1990. URL https://www.jstor.org/stable/2245382.

Charlotte G. Steeh. Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45:40–57, 1981.

Carl Erik Särndal. The 2010 morris hansen lecture dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27:1–21, 2011. ISSN 0282423X.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, 1992.

Daniell Toth. A permutation test on complex sample data. *Journal of Survey Statistics and Methodology*, 8:772–791, 2020. ISSN 23250992. doi: 10.1093/jssam/smz018.

U.S. Census Bureau. Acs public use microdata sample (pums) overview., 2021. URL https://data.census.gov/mdat/#/.

Jan A. van den Brakel. Design-based analysis of embedded experiments with applications in the dutch labour force survey. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 171:581–613, 2008. ISSN 09641998. doi: 10.1111/j.1467-985X.2008.00532.x.

Jan A. van den Brakel. *Design-Based Analysis of Experiments Embedded in Probability Samples*, pages 457–479. Wiley, 2019.

Jan A. van den Brakel and Robbert H. Renssen. Analysis of experiments embedded in complex sampling designs. *Survey Methodology*, 31:23–40, 2005.

Jan Arie van den Brakel. *Design and Analysis of Experiments Embedded in Complex Sample Surveys*. PhD thesis, Erasmus University of Rotterdam, 2001.

Jan Arie van den Brakel. Design-based analysis of factorial designs embedded in probability samples. *Survey Methodology*, 39:323–349, 2013.

Jan Arie van den Brakel and R. Renssen. Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14:277–295, 1998.

Jan Arie van den Brakel and C A M van Berkel. A design-based analysis procedure for two-treatment experiments embedded in sample surveys . an application in the dutch labor force survey. *Journal of Official Statistics*, 18:217–231, 2002.

James Wagner. The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, 74:223–243, 2010. ISSN 0033362X. doi: 10.1093/poq/nfq007.

James Wagner and Frost Hubbard. Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology*, 2:323–342, 2014. ISSN 23250992. doi: 10.1093/jssam/smu009.

James R. Wagner. *Adaptive survey design to reduce nonresponse bias*. PhD thesis, University Of Michigan, 2008. URL http://books.google.com/books?hl=en%7B\&%7Dlr=%7B\&%7Did=j1VbF2qP1TgC%7B\&%7Doi=fnd%7B\&%7Dpg=PR3%7B\&%7Ddq=Adaptive+Survey+Design+to+Reduce+Nonresponse+Bias+by%7B\&%7Dots=jrCI0ZItIQ%7B\&%7Dsig=xsBukpJKxOz8Cwt2LUcPyIx2GoI.

Yanying Wang and William F. Rosenberger. Randomization-based interval estimation in randomized clinical trials. *Statistics in Medicine*, 39:2843–2854, 2020. ISSN 10970258. doi: 10.1002/sim.8577.

B. L. Welch. On the z-test in randomized blocks and latin squares. *Biometrika*, 29:21–52, 1937. URL https://www.jstor.org/stable/2332405.

Josephine J. Williams. Another commentary on so-called segregation indices. *American Sociological Review*, 13:298–303, 1948.

Changbao Wu and Mary E. Thompson. *Sampling Theory and Practice*. Springer, 2020.

Chan Zhang, James M. Lepkowski, and Lirui He. Exploring the feasibility of mail surveys in urban china. *Field Methods*, 30:263–276, 2018. ISSN 15523969. doi: 10.1177/1525822X18783951.