CAUSAL INFERENCE FOR COMPLEX DATA: RANDOMIZATION INFERENCE FOR TREATMENT EFFECT HETEROGENEITY, NETWORK OUTCOMES, AND SUBGROUP SPECIFIC EFFECTS

BY

MARK M. FREDRICKSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Yuguo Chen, Co-Chair
Professor Jake Bowers, Co-Chair
Professor Ben Hansen
Professor Feng Liang

Abstract

This dissertation presents three new methodologies for analyzing randomized controlled trials using the researcher controlled randomization mechanism as the basis for inference. The first method extends inference for the “attributable effect”, the total of the difference of outcomes if the treatment group had instead been assigned to the control condition, to count and continuous data using a fast approximation algorithm. Alternative approaches are limited to binary data, require asymptotic approximations, or are computationally expensive. A refinement of the method to allow for including additional information is also included. The second method extends randomization inference to the study of network formation. Previous approaches either required strong parametric assumptions or only allowed for pre-treatment networks to be used. This approach develops several test statistics that can be used to test against common network formation models, based purely on the randomization of treatment. The final method improves inference in cluster randomized trials, where collections of individuals are assigned to treatment conditions simultaneously. Under the appealing assumption that larger clusters will have larger outcomes, on average, the method provides efficient, unbiased estimation of average treatment effects requiring minimal additional assumptions. All three of these methods demonstrate the relevance of randomized controlled trials to key areas of science and statistical development as well as the advantages of carefully crafting study design to fit the problem of interest. Data examples include a large scale field experiment involving health insurance, a gene-wide association study involving high dimensional outcomes, and a policy relevant study of parental social capital and student achievement in schools.

To my wife, Devin.

Acknowledgments

I wish to acknowledge the contributions of many individuals. Professors Jake Bowers and Ben Hansen have been my mentors throughout, and even extending slightly before, my graduate career. The ideas in this dissertation can trace their heritage to our previous joint work, and their development was greatly aided by discussions and feedback from Jake and Ben. I would also like to thank Professor Yuguo Chen, who has been positively extravagant in the amount of time given to help me develop my research. His careful reading has greatly increased the clarity and rigor in this dissertation. The support of other faculty in both the political science and statistics departments has been invaluable and is too numerous to list. Faculty support has been possible due to funding from many organizations, including the Bill and Melinda Gates Foundation, the National Science Foundation (grants SES-0753164, DMS-1406455), and the University of Illinois Campus Research Board.

I would also like to thank my family, particularly my wife. Ten years ago, she married me and moved with me to Illinois to start my graduate career. She has been with me through all the highs and lows. I rely on her patience and good humor more than I can express. My children also deserve a thank you. While they may not realize it now, they had a part to play as well. Nothing motivates me like the promise of a hug for dad when I get home.

Contents

Preface

Planned experiments have a long history within the field of statistics. Researchers appreciate the clarity of experiments, particularly in addressing causal questions. Many funding and regulatory agencies expressly require, or at least strongly favor, randomized controlled trials when possible in order to avoid the pitfalls of unmeasured confounding and self selection. Modern businesses scrupulously test new marketing and delivery mechanisms through randomized trials in order to determine how best to respond and serve clients. As new forms of data are generated, new questions asked, and the size and scope of inference increases, so too must randomized trials evolve. Statisticians have a role to play in this evolution of experiments by facilitating new analyzes and research designs. R. A. Fisher famously observed, “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.” More positively, statisticians can encourage researchers to be bold in their experimentation: gather new types of data, analyze data on a larger scale, ask new questions.

Fisher also maintained that randomization is the “reasoned basis” for inference. While models are often convenient approximations, the best method of analyzing data appeals not to opaque assumptions but to the known distribution of the treatment mechanism, controlled by the researcher. So called “randomization inference” has seen a rebirth in recent years due to increased attention to randomized experiments from many sources. This dissertation is composed of three developments in randomization inference in ways that make use of new types of data or address opportunities to improve randomization inference. The common thread linking these methods is that they take randomization as a primary basis for statistical inference: the known distribution of the randomly allocated treatment regime leads to stochastic implications for the data. In fact, the data themselves may be taken to be fixed, conditional on treatment. That we see some data and not others is a product of the treatment allocation, not a data generating process outside the control of the researcher. Consequently, these methods require fewer assumptions than competing approaches and focus attention on the importance of study design.

The first chapter considers a new method that expands a particular mode of randomization inference, “attributable effects,” to new types of data. Previous methods for attributable effects are largely limited to binary data or ordinal data, computationally burdensome, which limits application to small studies, or based on large sample approximations that may fail to hold when data exhibit highly non-normal distributions. This method expands the scope of inference for attributable effects in two ways. First, it casts the problem of testing an attributable effect in the language of optimization problems and shows the problem can be solved quickly using an approximation that appears to perform well in practice. Second, as the optimization is by construction a "worst case" bound, it shows how additional information can be included in the optimization process to improve inference if it seems like the data do not support the worst case scenario. In simulations and with real data, it is shown that the method performs particularly well for data that exhibit high skew or many zeros. Such data are frequently a problem for standard methods. The computational efficiency of the approach is also a large advantage when it is used to analyze a field experiment on health insurance with tens of thousands of participants.

The second chapter also expands randomization inference to a new type of data: networks. Graph formation models have existed since the seminal Erdös-Renyi model, but these often make strong parametric assumptions that are not well matched to the randomized allocation of treatment. Often these models do not include covariates for the nodes in the graph, such as a treatment assignment label, and condition outcomes for a single dyad on the rest of the graph, creating a de facto mediation analysis. While there has been work to extend randomization inference to allow incorporating existing networks, there has not been any work to consider using randomization inference to understand network formation. This approach considers perhaps the simplest model of network formation, that the network is entirely fixed with respect to the random treatment assignment, and develops several test statistics that are sensitive to certain types of deviations from this hypothesis, such as treatment induces clustering or makes treated nodes more or less central to the network. Through a close relationship between randomization tests simple randomization mechanisms and permutation tests, these methods can also be used to test hypotheses that a network is invariant to relabeling nodes in a network. Several existing permutation tests exist that create networks from high dimensional data, and the proposed method can be useful for a randomized trial with an extremely high-dimensional outcome, which is demonstrated using a gene-wide association study. While networks that are the result of randomized trials are not yet commonly studied, this methodology opens up new possibilities for researchers interested in testing how networks change in a causal way.

The final chapter in this dissertation takes a different approach than the previous two chapters by considering a common research design that has been under served: analyzing clustered randomized trials when the outcome of interest is at the unit rather than cluster level. Cluster randomized trials are frequently employed when treatment cannot easily or ethically be applied to individuals on a case-by-case basis. Common examples of clustering include students within classrooms, patients within clinics, and precincts within cities. When treatments cannot be meted out to individuals, for example it may not be impossible to teach different curricula to students in a single classroom, the level of the randomization and the level of the analysis can differ. Dating back 40 years or more, scholars have raised caution about analyzing cluster randomized trials as if they were in fact randomized at the individual level. Less has been offered, however, on methods to analyze individual outcomes without resorting to assumptions on functional forms and parametric distributions. Using a novel estimator, this chapter develops unbiased estimation of average treatment effects under a simple assumption that makes intuitive sense for cluster randomized trials: larger clusters will have larger outcomes than smaller clusters. This estimator also exhibits smaller variance than other unbiased methods, which is an important consideration when cluster randomized trials may have relatively few clusters compared to the number of subjects within the clusters. As an added benefit to this approach, it seamlessly handles the case of analyzing within subgroups defined at the individual level, traditionally a more difficult task for methods that require subgroups to be defined at the cluster level.

While the three chapters constituting the dissertation take on different types of data, different research designs, and different targets of inference, the common thread uniting them is a consistent commitment to using randomization to the fullest extent possible. As hinted at in the final chapter, however, models are useful tools that can provide useful insight to improve randomization inference. Careful incorporation of model assisted approaches is the natural next step for many of the methods proposed in this dissertation. Additional modeling constraints may further focus the attributable effects optimization problem to plausible outcomes, extending the current method of using a confidence interval for a certain variance quantity to other parameters. In the networks chapter, it was found that certain statistics were most powerful against different alternative hypotheses. Designing test statistics for specific alternative hypothesis of interest is a natural next step. The cluster randomized trial approach was able to use a simple model to find an unbiased estimator of the average treatment effect; no doubt other models would lead to other estimators, and providing an algorithm for deriving such estimators would be a valuable contribution. Looking further than the scope of these methods, this dissertation shows that randomized trials can remain relevant to key issues of modern statistics such as networks, high dimensional data, optimization problems, and asymptotic theory. Likewise, cutting edge science can be well served by carefully crafting randomized trials to fit the scientific question.

Chapter 1

Attributable effects for count and continuous data

1.1 Introduction

In 2008, the state of Oregon engaged in a lottery in which low income residents were selected to be allowed to apply for state funded Medicaid health insurance. Supporters of expanded state sponsored healthcare argue that offering medical insurance shifts incentives to use expensive emergency room care to less expensive scheduled clinical care. To address this argument, Finkelstein et al. (2012) contacted a subset of those assigned to both the health insurance arm and those who were not selected in the lottery to ascertain the amount spent on out of pocket medical costs. The $11,450$ households in the control condition reported a total of \$4.71 million spent on medical care in the previous six months (Finkelstein et al., 2012). On average, this translates to \$411.68 per control household, but this averaging obscures the fact that 49% of the control subjects reported spending zero dollars on out of pocket costs. For nearly half of the control subjects, the average is not very informative about the amount spent on medical care.

Instead of asking about average effects, it may be more useful to ask what portion of the costs can be attributed to the control subjects not being permitted to apply for Oregon's Medicaid program. Rosenbaum (2001) calls this quantity "the effect attributable to treatment." Many of the approaches for estimating and testing hypotheses about attributable effects have been focused on binary outcomes (Rosenbaum, 2001). Rosenbaum (2002a) extends his previous work to matched pair designs. Rigdon and Hudgens (2015) allow for attributable effects to both the treatment group and control group in order to get confidence intervals for the average treatment effect. Li and Ding (2016) improve the efficiency of these results. Fogarty et al. (2017) focus on the particular difficulties in observational studies that attempt to emulate randomized trials and propose numerical solutions that provide tests of effects on binary outcomes along with sensitivity analyses. Choi (2017) also provides optimization based techniques for solving attributable effects for binary

This chapter contains joint work with Professor Yuguo Chen and is currently under review for publishing.

data and includes methods for improving inference when information is available about interactions between subjects. Some progress has also been made on ordinal outcomes using well defined sequences of alternative hypotheses (Lu et al., 2015), bounds (Lu et al., 2016), or introducing nuisance parameters or latent variables (Volfovsky et al., 2015). Ding and Miratrix (2017) show how physical randomization, binary outcomes, and monotonicity combine to generate a multiple hypergeometric likelihood, which can be used for inference for both the attributable effect and the average treatment effect.

There are two notable exceptions to the focus on binary data. Hansen and Bowers (2009) present a survey sampling based approach for estimating attributable effects. While this approach expands the scope of data to include count and continuous outcomes, the method requires large sample approximations to hold. Feng et al. (2014) provide an exact test for continuous outcomes based on a complex optimization problem. The approach is based on the Mann-Whitney-Wilcoxon sum of ranks test statistic, which degrades in the presence of ties in the values of Y_i. For both methods, large numbers of zeros in the outcomes are difficult to handle.

In this chapter, we present a method that can be thought of as a hybrid between the existing exact tests for attributable effects and the survey sampling based estimation approach. We use a normal approximation as part of an optimization routine, but test the resulting hypothesis using exact methods. The method is computationally efficient, and simulations show that it performs well when the data contain large portions of zero values.

The rest of the chapter is organized as follows. Section 1.2 introduces the proposed method, along with notation and assumptions. Section 1.3 evaluates the accuracy of the key approximation and the statistical properties of the method through a variety of simulations. Section 1.4 returns to the Oregon Health Insurance Program experiment previously introduced to analyze several outcomes. Section 1.5 concludes with a discussion.

1.2 Methodology

1.2.1 Setting and notation

Consider N units in a study where n units are randomly assigned to the treatment condition and the remaining $m = N - n$ units are assigned to the control condition, writing $Z_i = 1$ for treatment and $Z_i = 0$ for control. For all subjects, we hypothesize *potential outcomes* to the different treatment conditions $y_i(1)$ when $Z_i = 1$ and $y_i(0)$ when $Z_i = 0$ (Neyman, 1923; Holland, 1986). The observed outcome Y_i is random in that it depends on Z_i: $Y_i = y_i(Z_i)$. Throughout, we shall use boldfaced symbols as vectors, so $\boldsymbol{Y} = \boldsymbol{y}(\boldsymbol{Z})$

defines the outcomes after treatment $(Y_1, \ldots, Y_N)' = (y_1(Z_1), \ldots, y_N(Z_N))'$. Implicit in this definition is an assumption that assignment to the treatment or control condition for unit i does not change the outcome of any unit j, often labeled as the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980).

Define the vector of individual effects $\boldsymbol{\tau} = \boldsymbol{y}(\mathbf{1}) - \boldsymbol{y}(\mathbf{0})$. Since we only observe $\boldsymbol{Y} = \boldsymbol{y}(\boldsymbol{Z})$, $\boldsymbol{\tau}$ is not identified. A sharp null hypothesis $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ states the values of $\boldsymbol{y}(\mathbf{1} - \boldsymbol{Z})$ and implies $\boldsymbol{\tau}_0$. After removing the hypothesized treatment effect from the treated units, $\tilde{\boldsymbol{y}} = \boldsymbol{y}(\boldsymbol{Z}) - \boldsymbol{\tau}_0 \cdot \boldsymbol{Z}$, the resulting data are independent of treatment assignment and a randomization test can be applied using a suitable test statistic (Fisher, 1935; Rosenbaum, 2002a, 2010). Examples of such tests include Fisher's exact test, the Wilcoxon-Mann-Whitney rank test, and many others. After selecting a test statistic $T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})$, its distribution under the hypothesis $H_0 : \boldsymbol{y}(\mathbf{0}) = \tilde{\boldsymbol{y}}$ is given by enumerating all possible ways of selecting n out of N units (Fisher, 1935). The p-value of the hypothesis is the proportion of randomizations that lead to a larger test statistic value than the observed value. Indexing all $J = \binom{N}{n}$ possible treatment assignments as $\boldsymbol{z}^{(j)}$, write the p-value as

$$p = \mathrm{P}\left(T(\boldsymbol{Z}, \tilde{\boldsymbol{y}}) \geq T(\boldsymbol{z}, \tilde{\boldsymbol{y}})\right) = J^{-1} \sum_{j=1}^{J} I(T(\boldsymbol{z}^{(j)}, \tilde{\boldsymbol{y}}) \geq T(\boldsymbol{z}, \tilde{\boldsymbol{y}})),$$

where $\boldsymbol{z}$ is the realized treatment assignment in the experiment and $I(\cdot)$ is the indicator function. One of the primary advantages of the Fisherian approach is that it does not rely on large sample approximations or distribution assumptions. The trade-off is that it requires hypothesizing the subject level treatment effects τ_i.

As an alternative to specifying the entire $\boldsymbol{\tau}_0$ vector, consider the attributable effect

$$A = \boldsymbol{Z}'\boldsymbol{\tau}. \tag{1.1}$$

Observe that a hypothesis of the form $H_0 : A = A_0$ is a *composite hypothesis* as it contains any $\boldsymbol{\tau}_0$ for which $\boldsymbol{Z}'\boldsymbol{\tau}_0 = A_0$. Theoretically, A_0 could be tested using a randomization test if one could find the $\boldsymbol{\tau}_0$ with the maximum p-value among the set $\{\boldsymbol{\tau}_0 : \boldsymbol{Z}'\boldsymbol{\tau}_0 = A_0\}$, as the p-value of the true $\boldsymbol{\tau}$ must be less than the maximum. For count data, enumerating all possible $\boldsymbol{\tau}_0$ is computationally intractable in most circumstances. For continuous data, such an enumeration is not even possible.

1.2.2 Approximating the largest p-value

Most of the previous approaches to attributable effects relied on "distribution free methods," in which the distribution of the test statistic did not depend the values of the outcomes themselves (Maritz, 1981). In these situations, the problem of finding the largest p-value is equivalent to finding the smallest test statistic value that results from an adjustment $\boldsymbol{\tau}_0$, as any adjustment $\boldsymbol{\tau}_0$ will result in the same null distribution. In this chapter, we take the opposite approach: the test statistic remains fixed while we search for a distribution that places the most mass above the test statistic value. While we have wide latitude selecting the test statistic T, a natural choice is the deviation of the treatment group's mean from the overall mean:

$$T(\boldsymbol{Z}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{N} Z_i y_i - \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{1.2}$$

This statistic has been widely studied in both the randomization and permutation literature as an analog of the parametric t-test (Lehmann and Romano, 2005, Chapter 5). For the present purposes, the primary advantages of this test statistic are the close alignment with the definition of the attributable effect and a convenient normal approximation.

Usefully, the value of the test statistic (1.2) evaluated at $\boldsymbol{z}$, the observed assignment vector, remains fixed under any possible $\boldsymbol{\tau}_0$ that is compatible with A_0. For observed treatment $\boldsymbol{z}$, observed data $\boldsymbol{y}$, and null hypothesis $\boldsymbol{\tau}_0 = (\tau_{0,1}, \tau_{0,2}, \ldots, \tau_{0,N})^T$ such that $\boldsymbol{z}'\boldsymbol{\tau}_0 = A_0$, define the adjusted data $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\tau}_0 \cdot \boldsymbol{z}$. The value of test statistic when applied to the adjusted data only depends on $\boldsymbol{\tau}_0$ through A_0:

$$T(\boldsymbol{z}, \tilde{\boldsymbol{y}}) = \frac{1}{n} \sum_{i=1}^{N} z_i (y_i - z_i \tau_{0,i}) - \frac{1}{N} \sum_{i=1}^{N} (y_i - z_i \tau_{0,i}) = T(\boldsymbol{z}, \boldsymbol{y}) - \frac{(m/n)}{N} A_0.$$

Therefore any hypothesis compatible with A_0 generates the same value of T.

While the observed test statistic remains unchanged, the distribution of $T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})$ depends on the particular $\tau_{0,i}$ values. Since $\tilde{\boldsymbol{y}}$ is a fixed quantity under the null that $A = A_0$, $T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})$ can be thought of as a sample average of n items drawn from a finite population of size N, centered on the true population mean $\mu_0 = N^{-1} \left(\sum_{i=1}^{N} y_i - A_0 \right)$. The mean and variance of $T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})$ follow from standard finite population sampling results (Cochran, 1999, Theorems 2.1, 2.2):

$$\begin{aligned}
\mathrm{E}\left(T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})\right) &= \frac{1}{n} \sum_{i=1}^{N} \mathrm{E}\left(Z_i\right) \tilde{y}_i - \mu_0 = 0, \\
\mathrm{Var}\left(T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})\right) &= \frac{m/n}{N(N-1)} \left[\sum_{i=1}^{N} z_i \left(\tau_{0,i} - y_i + \mu_0\right)^2 + \sum_{i=1}^{N} (1 - z_i) \left(y_i - \mu_0\right)^2 \right].
\end{aligned}$$

While the portion of the sum that depends on the control units is a constant, the portion depending on the treated units is a function of the exact $\tau_{0,i}$ values, even though $\sum_{i=1}^{N} Z_i \tau_{0,i} = A_0$ is fixed.

Under fairly mild conditions, T is approximately normally distributed in large samples (Hájek (1961);

(Lehmann, 1975, p. 353)). Consider a set of finite populations indexed by ν. For each population of N_ν subjects, n_ν are assigned to treatment and m_ν are assigned to control. For each population, the null hypothesis $\tau_{0,\nu}$ holds so adjusted values $\tilde{y}_{\nu,i}$ are fixed. The statistic $S_\nu = T_\nu / \mathrm{Var}\left(T_\nu\right)^{1/2}$ converges in distribution to $N(0,1)$ when $N_\nu, n_\nu, m_\nu \to \infty$ and

$$\frac{\max(\tilde{y}_{\nu,i} - \mu_\nu)^2}{\sum_{i=1}^{N_\nu}(\tilde{y}_{\nu,i} - \mu_\nu)^2} \max\left(\frac{n}{m}, \frac{m}{n}\right) \to 0 \quad \text{as } \nu \to \infty,$$

where $\mu_\nu = \frac{1}{N}\tilde{y}_{\nu,i}$. The first term requires that no individual $\tilde{y}$ be so large as to dominate the variance, while the second implies that neither the treated nor control group size become negligible, which seems particularly natural in the context of a series of increasingly larger experiments.

Let $c = \sum_{i=1}^{N}(1 - z_i)\left(y_i - \mu_0\right)^2$ be the control subjects' contribution to the variance of T. Under the regularity conditions above, squaring T leads to a scaled χ^2 distribution:

$$[T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})]^2 \sim \frac{m/n}{N(N-1)} \left[\sum_{i=1}^{N} z_i \left(\tau_{0,i} - y_i + \mu_0\right)^2 + c\right] \chi_1^2.$$

Recall that for any fixed value of A_0, the value of T^2 will be the same regardless of the particular values $\tau_{0,i}$. Therefore the vector of adjustments $\boldsymbol{\tau}_0$ that corresponds to the largest possible p-value consistent with A_0 can be found by maximizing the quantity $\sum_{i=1}^{N} z_i(\tau_{0,i} - y_i + \mu_0)^2$. In order to find the $\boldsymbol{\tau}_0$ that maximizes T, we make one of two possible assumptions for all i:

Assumption 1.1. $0 \leq y_i(0) \leq y_i(1)$,

or

Assumption 1.2. $0 \leq y_i(1) \leq y_i(0)$.

For the purpose of exposition, we focus on the case when Assumption 1.1 holds, but applying the methods when Assumption 1.2 holds simply requires substituting $\boldsymbol{W} = 1 - \boldsymbol{Z}$ for $\boldsymbol{Z}$ throughout.

Without loss of generality, we suppose that the first n units are the treated units (i.e., $z_i = 1$ for $i = 1, \ldots, n$ and $z_i = 0$ for $i = n + 1, \ldots, N$). Under Assumption 1.1, the following optimization problem finds the $\boldsymbol{\tau}_0$ with the largest p-value:

$$\begin{aligned}
(P) \text{ maximize: } g(\boldsymbol{\tau}_0) &= \sum_{i=1}^{n} \left(\tau_{0,i} - y_i + \mu_0\right)^2, \\
\text{subject to: } \sum_{i=1}^{n} \tau_{0,i} &= A_0, \\
0 \le \tau_{0,i} &\le y_i.
\end{aligned}$$

This optimization problem comes from the class of "quadratic convex maximization" problems (Floudas and Visweswaran, 1995). While maximizing a convex function over a convex set is generally an NP-hard problem, effectively equivalent to enumerating all possible vertices of the constraint space, the particular form of this problem allows for an efficient solution.

Theorem 1.1. *Let all $y_i \geq 0$. Sort the y_i such that,*

$$y_1 \geq y_2 \geq \cdots \geq y_n.$$

An optimal solution to P is given by:

$$\tau_{0,i} = \begin{cases} 0, & i < s, \\ A_0 - \sum_{i=s+1}^{n} y_i, & i = s, \\ y_i, & i > s, \end{cases}$$

where s is the largest integer such that $\sum_{i=s}^{n} y_i > A_0$.

A proof of Theorem 1.1 is given in the supplementary materials. As this solution can be implemented using a simple sort of the n treated units, followed by a linear pass through the data, so the complexity of the algorithm is $O(n \log n)$ using typical sorting routines. While Theorem 1.1 does not assume the data re either real values or integer values, the solution also applies to integer constrained $\boldsymbol{Y}$.

Corollary 1.1. *When A_0 is an integer and all y_i are integers, the solution to the integer constrained version of P is also given by Theorem 1.1.*

A proof is given in the supplemental materials.

It is important to note exactly what optimality guarantees Theorem 1.1 provides. Ultimately, we are seeking the $\boldsymbol{\tau}_0$ vector of adjustments that leads to the maximum p-value over all compatible $\boldsymbol{\tau}_0$ that sum to A_0. Theorem 1.1, however, finds the $\boldsymbol{\tau}_0$ vector that generates a null distribution for T with the *maximum* *variance*. When N is large, this distribution will be roughly normal, so the correspondence between maximum variance and maximum p-value will be close. For small samples, or when the normality approximation fails for other reasons such as high skew, this approximation may fail to find the the adjustment with the maximum p-value, despite having the largest variance. In Section 1.3.1, we investigate the fidelity of the approximation through a series of numerical studies and find it performs quite well, even in small or long tailed data.

1.2.3 Improving inference with estimated variance

The solution found by Theorem 1.1 sets $\tau_{0,i} = 0$ for $s-1$ of the treated units and $\tau_{0,i} = y_i$ for $n-s$ of the units, with a final "pivot" observation having a value that ensures the sum of treatment effects is A_0. Likewise, the adjusted data $\tilde{y}_1, \ldots \tilde{y}_n$ created after subtracting the treatment effects is composed of at least $n-s$ zeros. While this allocation of treatment effects is certainly the solution to the optimization problem as phrased, it may not be plausible for the reason that these data look very different than the observed responses of the control group. Unless we observe that the control group is also composed of many zeros and other values similar to the observed treated group, the solution $\boldsymbol{\tau}_0$ seems implausible as the true set of individual treatment effects, even if it is the case that $A = A_0$.

To make this intuition more formal, we focus on limiting the acceptable $\boldsymbol{\tau}_0$ to those that are compatible with reasonable estimates of the variance of all subjects' responses to the control condition: $\sigma_0^2 = N^{-1}\sum_{i=1}^{N}(y_i(0) - \bar{y}(0))^2$. Observe that for the control units $y_i(0)$ is observed, while for the treated units, the hypothesis $\boldsymbol{\tau} = \boldsymbol{\tau}_0$ allows recovering the control potential outcome for subjects that were treated. If we knew the true value of σ_0^2, we could limit the search of possible $\boldsymbol{\tau}_0$ vectors to those that were compatible with the equation $\sigma_0^2 = N^{-1}\sum_{i=1}^{N}(y_i - z_i\tau_{0,i} - \mu_0)^2$. As we do not know σ_0^2, we find a reasonable upper bound and account for the uncertainty from using the bound rather than the true value.

Specifically, we adopt the approach of Berger and Boos (1994) to test hypotheses for $\boldsymbol{\tau}_0$ over a confidence interval for σ_0^2. If an α-level hypothesis test of A_0 or a $1-\alpha$ confidence interval for A is desired, we first construct a $1-\gamma$ upper confidence bound $\bar{\sigma}_0^2$ and then test the individual hypotheses for A_0 at the $\alpha - \gamma$ level, conditional on $\bar{\sigma}_0^2$. As the probability of $\sigma_0^2 > \bar{\sigma}_0^2$ is less than γ, an $\alpha - \gamma$ level test of A_0 or a $1-(\alpha-\gamma)$ confidence interval will have size no greater than α and coverage no less than $1-\alpha$, respectively.

For the finite sampling context, O'Neill (2014) provides an upper confidence bound for the finite population variance of $y_i(0)$ based on simple random samples as

$$\left(\frac{m-1}{N-1} + \frac{n}{N-1}\frac{1}{F_{\gamma,d_1,d_2}}\right)s_0^2,$$

where s_0^2 is the sample variance for the control units and F_{γ,d_1,d_2} is the γ quantile of an F distribution with the degrees of freedom calculated as

$$d_1 = \frac{2m}{\kappa - (m-3)/(m-1)}, \quad d_2 = \frac{2n}{2 + (\kappa - 3)(1 - 2/N + 1/(Nm))}.$$

O'Neill (2014) motivates these calculations from a model in which the finite population is drawn from a super-population where κ is the fourth central moment. In the interest of simplicity, we take $\kappa = 3$, as would be the case if the superpopulation were normal, and the degrees of freedom calculations simplify to $d_1 = m - 1$ and $d_2 = n$. This parameter could also be estimated from the control responses.

As we noted in Section 1.2.2, the variance of the test statistic T depends on the finite population variance: $\text{Var}\,(T) = \frac{m/n}{N-1}\sigma_0^2$ (Cochran, 1999, Theorem 2.2). When $\bar{\sigma}_0^2$ is an upper bound for σ_0^2 and the variance implied by $\boldsymbol{\tau}_0$ is less than this bound, then it is both an optimal solution and compatible with the bound. We can then proceed to testing it at the α level. Alternatively, if the variance implied by $\boldsymbol{\tau}_0$ exceeds the bound, we need to find a solution that is compatible. In this case, any solution with a compatible variance such that $\sum_{i=1}^n \tau_i = A_0$ is equally valid, so we employ the survey sampling based approach to approximate the confidence interval that would be found for any such solution. In large samples, when the distribution of the test statistic is approximately normal, the set of A_0 that are not rejected at the $\alpha - \gamma$ level is

$$\sum_{i=1}^N Z_i Y_i - \frac{n}{m}\sum_{i=1}^N (1 - Z_i) Y_i \pm z_{1-(\alpha-\gamma)/2}\sqrt{N\frac{n}{m}\bar{\sigma}_0^2}, \tag{1.3}$$

where $z_{1-(\alpha-\gamma)/2}$ is the $1-(\alpha+\gamma)/2$ quantile of the standard normal distribution (additional details on this interval are given in supplementary materials). If A_0 is within this set, we can accept it at the overall α level, having accounted for the possible Type I error from using the upper bound $\bar{\sigma}_0^2$. This procedure can be seen as somewhere between the maximum variance method proposed in this chapter and the survey sampling approaches to estimating the attributable effect. As γ goes to zero, the upper bound $\bar{\sigma}_0^2$ will go to infinity, so that the maximum variance solution will always have a compatible variance and the procedures are equivalent. When γ gets close to α, the procedure will more frequently use a sampling based interval, which will be shown to be similar to a method introduced in Equation (1.4). In the next section, we investigate the performance of both the maximum variance method and limited variance methods in a series of simulations.

1.3 Simulations

1.3.1 Testing the normal approximation

For each hypothesized attributable effect A_0, there may be many compatible unit level sharp hypotheses $\boldsymbol{\tau}_0$ such that $\sum_{i=1}^{N} Z_i \boldsymbol{\tau}_i = A_0$. Rejecting A_0 at the α level implies that all compatible hypotheses must also be rejected at the α level. In the suggested methodology of Section 1.2.2, we propose using a normal approximation to the null distribution to find $\boldsymbol{\tau}_0$ with the largest p-value. We now present several simulations to assess how well the approximation works.

For n treated units and a hypothesis A_0, there are at most $\binom{n+A_0-1}{n}$ ways to allocate the A_0 to the n treated units when the potential outcomes $y_i(1)$ and $y_i(0)$ are integer values. For small experiments, these allocations can be explicitly enumerated to find the $\boldsymbol{\tau}_0$ vector with the largest p-value. This presents a way to compare how well the approximation holds in finding the largest p-value, at least for a sufficiently small experiments and effect sizes for which all $\binom{n+A_0-1}{n}$ possible allocations can be enumerated and checked.

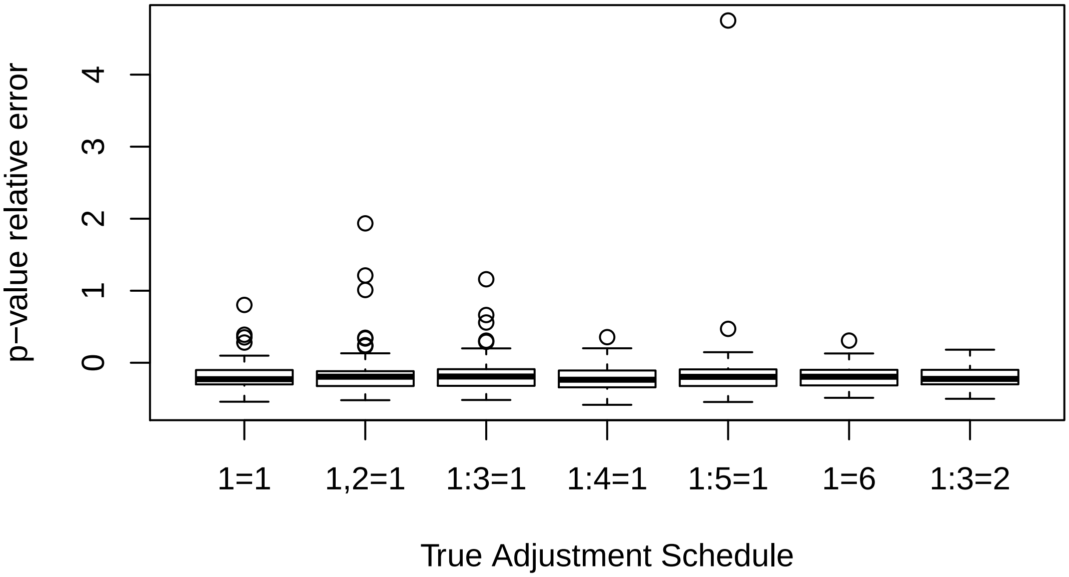


Figure 1.1: Boxplot comparing of normal approximation maximum p-value ($\hat{p}$) to true maximum p-value (p) using relative error $(p - \hat{p})/p$. The x-axis labels indicate the units that had positive τ_i values. For example, "1:3=2" indicates that $\tau_1 = \tau_2 = \tau_3 = 3$ and $\tau_i = 0$ for $i > 3$.

For a small experiment ($N = 10$, $n = 5$), we generated $\boldsymbol{y}(\mathbf{0})$ and then allocated A to the different units, with $A \in \{1, \ldots, 6\}$. The true A was either spread out or clustered it on only a few units. Additional details on this process can be found in the supplemental materials. Figure 1.1 shows the relative error of the p-value from the normal approximation comapred to p-value from complete enumeration. On the whole, the approximation works quite well, even for this small experiment. The approximation performed least well in these examples where the true treatment effect was larger and evenly distributed. Recalling that the solution to the approximation concentrates the adjustments to the smallest values, it makes sense that the approximation does not perform well in this situation.

As an additional check on the performance of the algorithm, the variance of T generated by the adjustment schedule found by the proposed algorithm was compared to the variances of T for all possible adjustments via enumeration. In all simulations, the adjustment selected had the largest variance of any possible solution. While this does not always imply the largest p-value, as seen in Figure 1.1, the algorithm is performing its job properly. As the sample size increases and the normal approximation improves, the accuracy with respect to finding the true maximum p-value should increase, which is shown in additional simulations reported in the supplemental materials.

1.3.2 Comparing to survey sampling approach

We now consider the performance of confidence intervals for A using the optimization routine detailed in Section 1.2.2, which we call the "variance maximization method," and the method that combines the variance maximization method with a bound for the variance in the control groups detailed in Section 1.2.3, which we label the "limited variance method." As a benchmark method for comparison, we use the survey sampling based approach of Hansen and Bowers (2009) (HB). Observe that the attributable effect A can be decomposed as

$$A = \sum_{i=1}^{N} Z_i(y_i(1) - y_i(0)) = \sum_{i=1}^{N} Z_i y_i(1) - \left(\sum_{i=1}^{N} y_i(0) - \sum_{i=1}^{N} (1 - Z_i) y_i(0) \right).$$

The quantities $\sum_{i=1}^{N} Z_i y_i(1)$ and $\sum_{i=1}^{N}(1 - Z_i) y_i(0)$ are completely observed as the totals in the treatment and control groups, respectively. While the total $\sum_{i=1}^{N} y_i(0)$ is not observed, it can be estimated using standard sample survey techniques (Hansen and Bowers, 2009; Sekhon and Shem-Tov, 2017). This leads to a large sample confidence interval for A:

$$\hat{A} \pm t_{1-\alpha/2} \sqrt{N \frac{n}{m} s_0^2} = \sum_{i=1}^{N} Z_i Y_i - \frac{n}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i \pm t_{1-\alpha/2} \sqrt{N \frac{n}{m} s_0^2}, \tag{1.4}$$

where s_0^2 is the sample variance for the control units and $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile from a Student's t-distribution with $m - 1$ degrees of freedom (additional details on the derivation of this interval are given in the supplemental materials).

In these simulations, we vary the total experiment size (N), the proportion of the $y_i(0)$ that are zero (p), and the true effect size (e). In each simulation, we compare the coverage rate of a 95% confidence interval as well as the ratio of interval lengths for the proposed methods and that of HB. For the limited variance method that requires splitting the α-level across two tests, we apply the method with both 99.9% and 99.0% upper confidence bounds for σ_0^2. Detailed information on the simulation process is given in the supplemental materials.

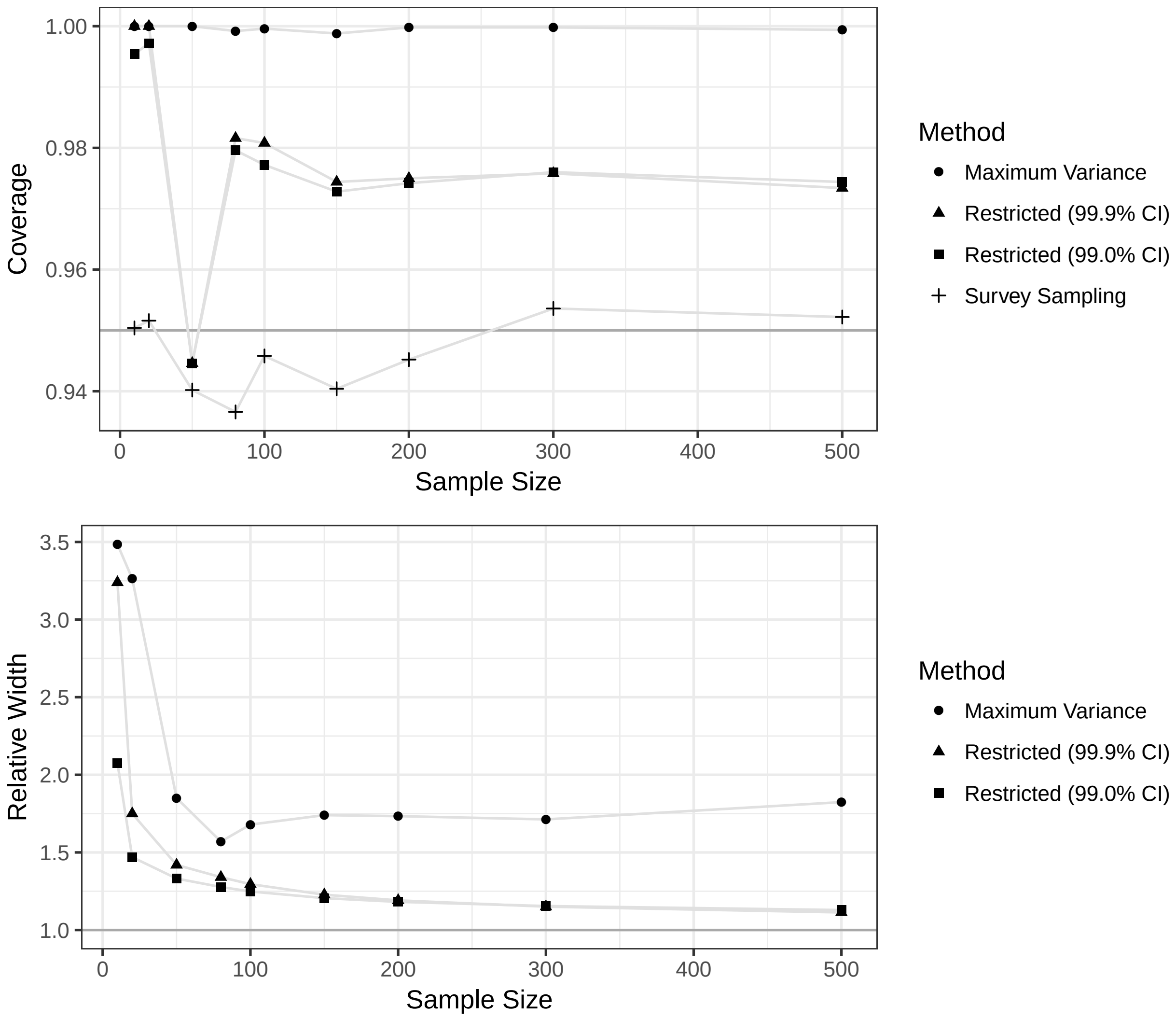


Figure 1.2: Sample size simulation for A. For 5000 replications, the experimental population (N) is varied from 10 to 500, while the other simulation parameters remain fixed ($p = 0.1$, $e = 1$, Proportion treated = 0.5).

In the first simulation, we varied the experimental population size between 10 and 500, treating half of the population in each experiment. Figure 1.2 shows the results of the simulations. The top panel reports the proportion of the simulations in which the generated confidence intervals covered the true A. As the figure shows, the basic variance maximization method has conservative coverage over the entire range of sample sizes. The two limited variance methods are also conservative, but less so than the maximum variance method. The survey sampling method under-covers at lower samples sizes but achieves nominal level for the larger sample sizes. In the lower panel, the width of the confidence intervals for the variance maximization and limited variance methods is compared to the width of the survey sampling method. For the size of the intervals, the survey sampling method is always the smallest, on average, though it is often under-covering for sample sizes less than 200. As the sample sizes get large, the limited methods approach the size of the survey sampling based method. The variance maximization method has the largest intervals, which is unsurprising given its large coverage rate.

In the second simulation, the parameter p, the probability of $y_i(0)$ being zero, varied from 0 to 0.95. Figure 1.3 again shows the 95% confidence interval coverage and relative widths. With a sample size of 100, the survey sampling based (HB) method has modest under-coverage for several values of p. As the proportion of zeros increases, the variance maximizing method becomes less conservative, but never falls below its nominal level. Recall that the solution to the variance maximization optimization problem sets the hypothesized $y_i(0)$ to zero (i.e., $\tilde{y}_i(0) = 0$) for the smallest observed treated units. As the proportion of zeros increases, this solution approaches the true $\boldsymbol{\tau}_0$, whereas in general cases the solution is only guaranteed to generate a p-value larger than that of the true $\boldsymbol{\tau}_0$. Consequently, the method performs particularly well in the case of zero-inflated outcomes. The relative width of the variance maximization interval tends to approach the width of the HB interval; however, the trend reverses for $p = 0.95$, suggesting there may be a limit to the proportion of zeros that this method can efficiently handle. The limited variance method performs well, maintaining good coverage for most values of p while also having interval lengths that are competitive with the sampling approach.

The third simulation varies the total effect size $\boldsymbol{\mathcal{T}} = \sum_{i=1}^{N} \tau_i$ as a function of the standard deviation of the randomly generated $y_i(0)$. Figure 1.4 shows the results as the effect size was varied between zero and two standard deviations. As the top panel of Figure 1.4 shows, the proposed methods maintain consistently conservative coverage rates across the different effect sizes, while the survey sampling method under covers somewhat. Interestingly, the relative size of the intervals for the proposed methods tended to increase as the effect size increased. In particular for the variance maximization method, this scenario represents the opposite of the zero-inflated situation: as the total effect increases, the true $\boldsymbol{y}(0)$ is less and less like $\tilde{\boldsymbol{y}}(0)$ as

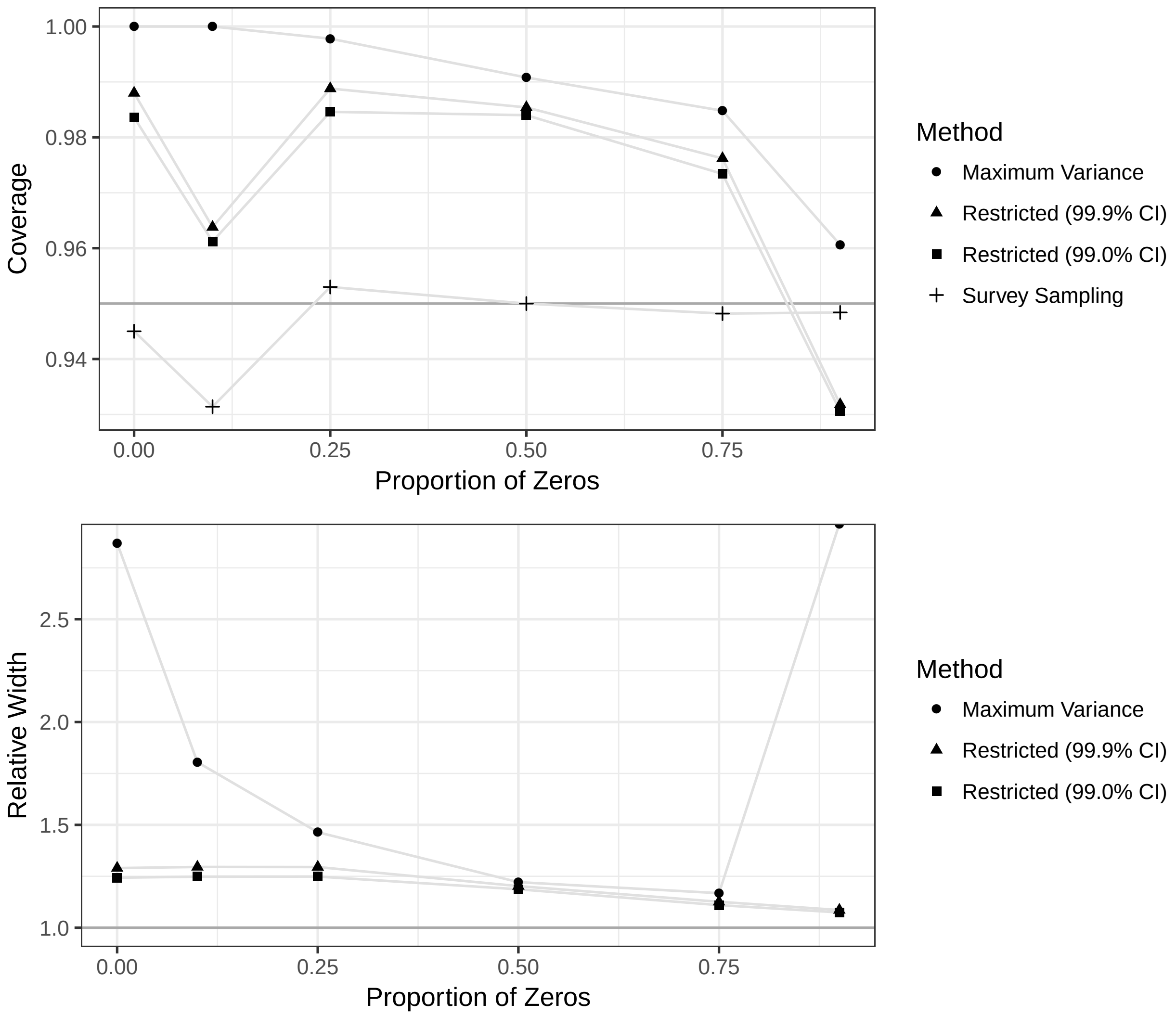


Figure 1.3: Proportion of zeros in $y(0)$ simulation for A. For 5000 replications, the parameter controlling the proportion of $y_i(0) = 0$ was varied between 0 and 0.95, while the other simulation parameters remain fixed ($e = 1$, $N = 100$, $n = 50$).

the large effect size makes more and more of the τ_i large.

Looking across these simulations, the overall pattern emerges that, at least on these data, the proposed methods appear to work well in small samples and when there is a great degree of treatment heterogeneity. The variance maximization method is almost always conservative in its coverage rates and has reasonably small interval widths in small samples or when there are many zeros in the data. The refinement that uses a confidence interval for σ_0^2 inherits many of these nice properties, while also having generally smaller confidence intervals in a wider variety of situations. One interesting result is that the choice of the level of the confidence interval used for σ_0^2 did not have much effect, at least for the values selected.

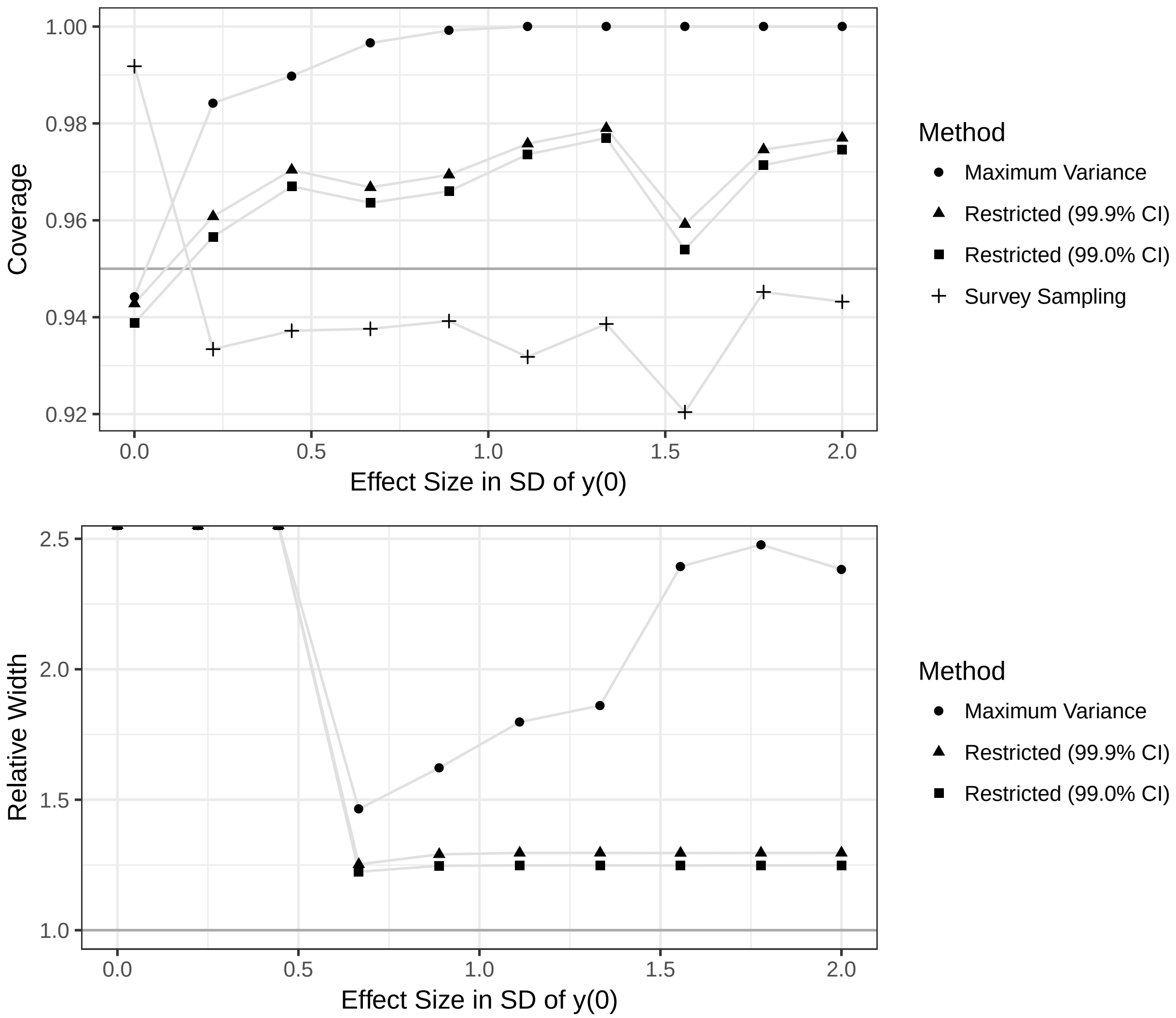


Figure 1.4: Effect size simulation for A. For 5000 replications, the size of the total effect $\boldsymbol{\mathcal{T}} = \sum_{i=1}^{N} \tau_i$ is varied from zero to two standard deviations of $y_i(0)$, while the other simulation parameters remain fixed (p = 0.1, $N = 100$, $n = 50$).

The simulations used thus far randomly assign treatment effects to individuals. Whether treatment effects are correlated with $y_i(0)$ can also influence the power of the test, particularly for the variance maximization method. Instead of randomly assigning treatment effects, Figure 1.5 shows the cumulative distribution functions for p-value of the test when the largest treatment effects are allocated to either the subjects with the largest $y_i(0)$ or smallest $y_i(0)$. To perform this simulation, $y_i(0)$ and treatment effects are generated using the simulation default settings. The $y_i(0)$ are sorted from largest to smallest and treatment effects are sorted in either increasing or decreasing order. When the effects are decreasing, the treatment helps the subjects that already have large $y_i(0)$ values; when the effects are increasing, the effects help those with the

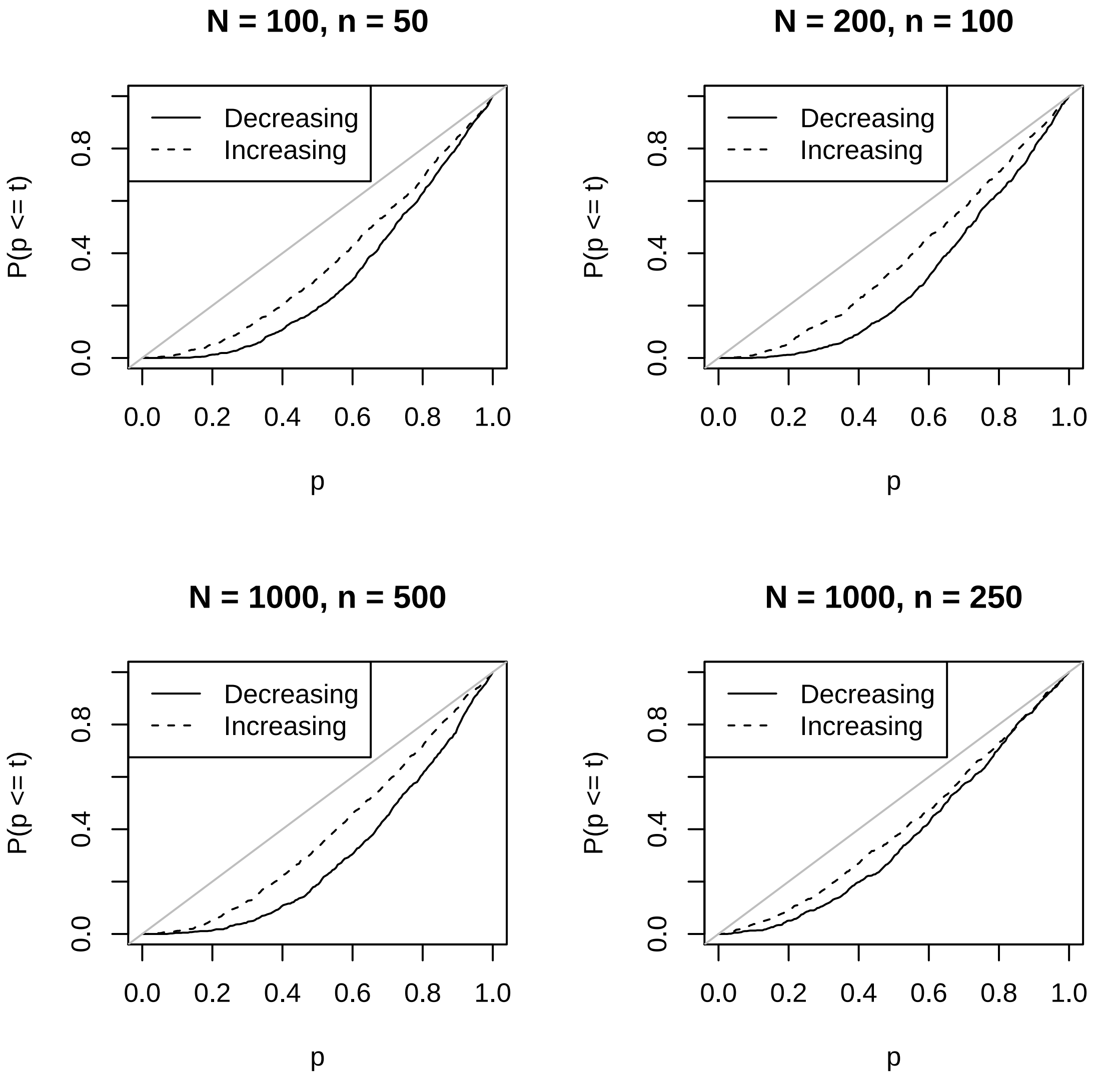


Figure 1.5: Cumulative distribution function of p-values when testing a true null hypothesis about A. All simulation parameters held at defaults. Potential outcomes to control are sorted such that $y_1(0) \geq y_2(0) \geq \cdots \geq y_N(0)$ and treatment effects are sorted in either increasing ($\tau_1 \leq \tau_2 \leq \cdots \leq \tau_N$) or decreasing ($\tau_1 \geq \tau_2 \geq \cdots \geq \tau_N$) order.

lowest $y_i(0)$. While the test is conservative for both sorting methods, it is less conservative when the largest effects are given to those with the smallest $y_i(0)$. As the optimization routine tests a hypothesis in which the treatment effects are concentrated on subjects with $y_i(0)$, it is unsurprising that the test is most powerful when the true treatment effect allocation is similar to the result of the optimization routine.

1.4 Oregon health insurance experiment

In 2008 the state of Oregon re-opened enrollment for Oregon Healthcare Plan (OHP) Standard, a medical insurance program for low-income households that were ineligible for the federal Medicare program (OHP Plus). As enrollment in this program had been closed for several years, officials anticipated a higher demand than could be accommodated under the available budget. To address the issue of over-subscription, state officials applied for, and received, a waiver from the Centers for Medicare and Medicaid Services to implement a lottery system to allocate opportunities to apply to the program. After an advertising program to solicit potential recipients, 74,922 individuals applied for the program. The initial solicitation did not require individuals show eligibility for the program, so being randomly selected into the program provided individuals the opportunity to complete an application, demonstrating eligibility in the program. Of the 74,922 applicants, 29,834 individuals were randomly selected to receive an invitation to apply for the program. Of these, 8,698 applied and were approved to enroll in OHP Plus. More details on the program and randomization process can be found in Finkelstein et al. (2012).

After 12 months, a portion of both the treated individuals (selected to complete an application) and the control individuals (not permitted to apply) were sent a survey requesting information on various health and economic questions. Of particular interest were the self-reported amounts of money spent out-of-pocket for medical care during the previous 6 months. Responses to the question included many zeros and were heavily right skewed. Of the subjects that responded to the questionnaire ($N = 22{,}766$), 53% claimed no out of pocket costs in the last 6 months, while 6 individuals reported out of pocket costs in excess of $100,000. Excluding subjects who reported zero out of pocket costs, the median cost reported was $250.

To analyze these questions, we first suppose that for all subjects Assumption 1.2 holds: $0 \leq y_i(1) \leq y_i(0)$. This assumption supposes that having medical insurance will not raise a subject's out of pocket costs. As the Medicaid program covers nearly all medical costs, this assumption seems plausible. Before applying the methods proposed in this chapter, we first create a dichotomous variable indicating whether a subject reported spending more than zero dollars on health care. Applying the method of Rosenbaum (2001) to predict the attributable effect yields a 95% prediction interval of $[597, 889]$. This result suggests that had the control subjects had the opportunity to apply for state sponsored health care, between 10.3 and 15.4 percent who had out of pocket costs would have been able to avoid them.

While such savings may be beneficial no matter the overall amount spent, by dichotomizing the cost, this analysis may confuse substantive changes in the amount the control group spent for small, but consistent changes. To answer this question, we apply the proposed method in this chapter to predict the attributable effect on the dollar scale and compare it to the survey sampling method of Hansen and Bowers (2009).

Method	Lower 95% Pred. Int.	Upper 95% Pred. Int.
Hansen and Bowers	0 (0)	4,704,329 (100)
Maximum Variance	0 (0)	1,283,000 (27)

Table 1.1: 95% prediction intervals for the attributable effect of not having the opportunity to apply for OHP Standard on total out of pocket costs. Numbers in parentheses represent percentage of maximum attributable effect, which is the sum of all control units' outcomes.

Method	Lower 95% Pred. Int.	Upper 95% Pred. Int.
Assuming $0 \leq y_i(0) \leq y_i(1)$		
Hansen and Bowers	0 (0%)	1068 (10.7%)
Maximum Variance	0 (0%)	1092 (10.9%)
Limited Variance (99.0)	0 (0%)	1100 (11%)
Assuming $0 \leq y_i(1) \leq y_i(0)$		
Hansen and Bowers	0 (0%)	213 (1.5%)
Maximum Variance	0 (0%)	160 (1.1%)
Limited Variance (99.0)	0 (0%)	160 (1.1%)

Table 1.2: 95% prediction intervals for the attributable effect of number of emergency department visits. Under the assumption that $0 \leq y_i(0) \leq y_i(1)$, the effect of having the opportunity to apply for Medicaid is identified. Under the assumption that $0 \leq y_i(1) \leq y_i(0)$, the effect for the control group is identified. Numbers in parentheses indicate the percentage of the observed total attributed to the treatment.

Table 1.1 shows the results of the different methods. While both methods include an attributable effect of zero in their estimates, the survey sampling method produces an interval that gives no information as it includes every possible value for the attributable effect. The maximum variance method, however, excludes attributable effects greater than 27 percent of the observed total in the control group (at 95% confidence). The limited variance method using 99% and 99.9% confidence intervals for σ_1^2 finds the same interval as the maximum variance method and is not reported in the table.

Replacing usage of emergency departments with scheduled medical visits is often touted as a justification for expanding government sponsored medical insurance. As emergency departments by law must provide care, regardless of the individual's lack of medical insurance, advocates argue that providing medical insurance can actually *decrease* overall spending as insured individuals can better take advantage of less expensive scheduled care. On the other hand, while emergency departments must treat subjects, they will still bill patients without medical insurance. Having access to Medicaid might incentivive individuals to consume more medical services, in particular emergency department visits, as their own costs will significantly decrease. To answer this controversy, subjects in the Portland, OR area were matched to hospital records to tabulate the number of emergency department visits per subject (Taubman et al., 2014). Overall, the subset of the experimental population included 24,646 subjects. Again, these data show a large portion of zero values (16,180) and strong right skew.

Unlike the out of pocket costs, for emergency room visits there is not a clear reason to assume that a subject's potential outcome to treatment is always at least as large as the potential outcome to control. We therefore consider both possible assumptions for monotonicity and predict the attributable effect for the treated subjects as well as the attributable effect for the control subjects. Table 1.2 provides the results of these tests, again comparing the survey sampling method to the maximum variance method and limited variance method with a 99% confidence interval for σ_1^2. All methods tend to favor Assumption 1.1 as the prediction interval contains a larger portion of the observed data, though all intervals include zero leaving the possibility of no effect or non-monotonic potential outcomes. These results bear some similarities to those reported in Taubman et al. (2014), where the authors dichotomized these data at several usage levels and found that treated subjects made more use of emergency facilities.

1.5 Discussion

In this chapter, we presented a novel method for testing hypotheses for the effect attributable to treatment, the sum of the individual effects of the subjects within the treatment group. This method expands the scope of attributable effects to count and continuous data, provided the researchers are able to assume that effects are non-negative and that responses under the treatment condition are no less than responses under the control condition (Assumption 1.1). Alternatively, this method can be applied to recover the sum of treatment effects for the control group when control responses are assumed to be greater than treatment responses (Assumption 1.2). We also presented a refinement of the method that uses the variance of the observed control responses to limit the search space for acceptable hypotheses. This method maintained many of the positive features of the simpler variance maximization method while being more efficient in the simulations. These methods are computationally efficient, and simulations show that using a normal approximation to the true null distribution adds little error compared to the true solution. From a statistical perspective, the methods appear to perform well in small samples or when there is a high degree of treatment effect heterogeneity. These methods might be most useful when combined with a prescreening method used to detect heterogeneity (e.g., Ding et al., 2016) and employed only if the constant treatment effect assumption seems to be a poor approximation to the true treatment effect distribution.

To evaluate the new methods, we compared them to the survey sampling based method of Hansen and Bowers (2009). While the new methods performed comparably, it should be noted that the survey sampling based estimator made use of neither the assumption of monotonicity nor the constraint that $0 \leq A \leq \sum_{i=1}^{N} Z_i Y_i$. While it lies outside the scope of the current chapter to amend the estimator to take advantage of these assumptions, there is a lengthy literature on both using monotonicity assumptions to derive bounds for average treatment effects (Manski, 1997; Kim, 2014; Demuynck, 2015; Frandsen and Lefgren, 2016; Huang et al., 2017) as well as estimating means under boundedness assumptions (Casella and Strawderman, 1981; Mandelkern, 2002; Evans et al., 2005), which could be combined to provide a more efficient estimator under the assumptions invoked for the proposed methods.

1.6 Proofs and Additional Simulations

1.6.1 Large sample prediction intervals for A

By definition, for any given $\boldsymbol{Z}$, the attributable effect of treatment can be decomposed as

$$A = \sum_{i=1}^{N} Z_i y_i(1) - \left(\sum_{i=1}^{N} y_i(0) - \sum_{i=1}^{N} (1 - Z_i) y_i(0) \right).$$

As $\sum_{i=1}^{N} Z_i y_i(1) = \sum_{i=1}^{N} Z_i Y_i$ and $\sum_{i=1}^{N}(1 - Z_i) y_i(0) = \sum_{i=1}^{N}(1 - Z_i) Y_i$ are observed quantities, we need only estimate $\sum_{i=1}^{N} y_i(0)$ using

$$\hat{Y}_0 = \frac{N}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i.$$

Plugging this estimator into the decomposition of A yields

$$\hat{A} = \sum_{i=1}^{N} Z_i Y_i - \left(\frac{N}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i - \sum_{i=1}^{N} (1 - Z_i) Y_i \right) = \sum_{i=1}^{N} Z_i Y_i - \frac{n}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i.$$

Under conditions stated in Section 2.2, when N is large, $\hat{Y}_0$ is approximately normal with mean $\sum_{i=1}^{N} y_i(0)$ and variance $N \frac{n}{m} \sigma_0^2$ (Cochran, 1999, Theorem 2.2), where σ_0^2 is the finite population variance of the $y_i(0)$. By estimating σ_0^2 with s_0^2, the sample variance of the control units, a $100 \times (1 - \alpha)\%$ prediction interval for A has the form:

$$\hat{A} \pm t_{1-(\alpha/2)} \sqrt{N \frac{n}{m} s_0^2} = \sum_{i=1}^{N} Z_i Y_i - \frac{n}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i \pm t_{1-\alpha/2} \sqrt{N \frac{n}{m} s_0^2},$$

where $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a t-distribution with $m - 1$ degrees of freedom.

Using the method of Berger and Boos (1994), we replace the t-distribution and s_0^2 with a normal distribution and a confidence interval for σ_0^2. As the largest prediction interval occurs at the upper bound for σ_0^2, which we notate $\bar{\sigma}_0^2$, the resulting prediction interval is

$$\sum_{i=1}^{N} Z_i Y_i - \frac{n}{m}\sum_{i=1}^{N}(1-Z_i)Y_i \pm z_{1-(\alpha-\gamma)/2)}\sqrt{N\frac{n}{m}\bar{\sigma}_0^2},$$

where γ is the amount of Type I error accorded to the upper confidence bound $\bar{\sigma}_0^2$ (i.e., it is a $100\times(1-\gamma)\%$ upper confidence bound) and $z_{1-(\alpha-\gamma)/2}$ is the $1-(\alpha-\gamma)/2$ quantile of a standard normal distribution.

1.6.2 Proof of Theorem 1.1

Proof. Recall that we wish to maximize:

$$\begin{aligned} g(\boldsymbol{\tau}_0) &= \sum_{i=1}^{n}\left(\tau_{0,i} - y_i + \mu_0)\right)^2 \\ &= \sum_{i=1}^{n}\tau_{0,i}^2 + \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n}\mu_0^2 - 2\sum_{i=1}^{n}\tau_{0,i}y_i + 2\mu_0\sum_{i=1}^{n}\tau_{0,i} - 2\mu_0\sum_{i=1}^{n} y_i \\ &= \sum_{i=1}^{n}(y_i - \tau_{0,i})^2 + \sum_{i=1}^{n}\mu_0^2 + 2\mu_0 A_0 - 2\mu_0\sum_{i=1}^{n} y_i. \end{aligned}$$

As the term $\sum_{i=1}^{n}\mu_0^2 + 2\mu_0 A_0 - 2\mu_0\sum_{i=1}^{n} y_i$ does not depend on τ_0, maximizing $g(\boldsymbol{\tau}_0)$ is equivalent to maximizing

$$h(\boldsymbol{\tau}_0) = \sum_{i=1}^{n}(y_i - \tau_{0,i})^2.$$

In other words, we can equivalently maximize the sum of squared remainders left after removing $\tau_{0,i}$. Writing $r_i = y_i - \tau_{0,i}$, rewrite the maximization problem as

$$\begin{aligned} (P') \text{ maximize: } & h(\boldsymbol{r}) = \sum_{i=1}^{n} r_i^2 \\ \text{subject to: } & \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} y_i - A_0 = R_0 \\ & 0 \le r_i \le y_i \end{aligned}$$

A simple greedy algorithm provides an optimal solution to P'. Sort the observations so that $y_1 \ge y_2 \ge \cdots \ge y_n$. Initialize $R_0^{(1)} = R_0$. For $i = 1, \ldots, n$, do:

1. If $y_i \ge R_0^{(i)}$, set $x_i = R_0^{(i)}$. For all $j > i$, set $r_j = 0$ and stop.

2. Otherwise, set $r_i = y_i$ and $R_0^{(i+1)} = R_0^{(i)} - y_i$.

3. If $i = n$, stop. Otherwise, update $i = i + 1$ and repeat the loop.

Let s be the largest integer such that $\sum_{i=1}^{s-1} y_i < R_0$. The result of the algorithm $\boldsymbol{r}$ has the form:

$$r_i = \begin{cases} y_i, & i < s, \\ R_0 - \sum_{i=1}^{s-1} y_i, & i = s, \\ 0, & i > s. \end{cases}$$

To show this is optimal, we show that we can transform any optimal solution into the greedy solution. Let $\boldsymbol{r}$ be the solution found by the greedy algorithm and $\tilde{\boldsymbol{r}}$ be any optimal solution. At each stage of the following algorithm, transform $\tilde{r}_i$ into r_i while maintaining the objective function value $h(\tilde{\boldsymbol{r}})$. At each state the proposed optimal solution has $\tilde{r}_j = r_j$ for $j < i$. Starting from $i = 1$,

1. If $\tilde{r}_i = r_i$, continue to $i + 1$.

2. Otherwise, consider the two possible values of r_i:

   (a) $r_i = R_0^{(i)}$: Observe that in this case $r_j = 0$ for $j > i$. As the solution $\tilde{\boldsymbol{r}}$ is feasible, it must be the case that $\sum_{j=i}^{n} \tilde{r}_j = R_0^{(i)} = r_i$. Since the $\tilde{r}_j$ are non-negative, this implies a contradiction that $h(\tilde{\boldsymbol{r}})$ is maximal:

   $$h(\boldsymbol{r}) - h(\tilde{\boldsymbol{r}}) = (R_0^{(i)})^2 - \sum_{j=i}^{n} \tilde{r}_j^2 = \left(\sum_{j=1}^{n} \tilde{r}_j\right)^2 - \sum_{j=i}^{n} \tilde{r}_j^2 = \sum_{j=i}^{n}\sum_{j'=i}^{n} \tilde{r}_{j'} > 0.$$

   Therefore, when $y_i \geq R_0^{(i)}$ the only optimal solution is the greedy one. At this point, we can stop, having found that the greedy solution is optimal.

   (b) $y_i < R_0^{(i)}$ and $r_i = y_i$. Since $\tilde{r}_i$ is also bounded by y_i, it must be the case that $\tilde{r}_i < r_i$. Again, since $\sum_{j=i}^{n} \tilde{r}_j = R_0^{(i)}$ and $\tilde{r}_j < r_i < R_0^{(i)}$, there must exist at least one $j > i$ such that $\tilde{r}_j > 0$. Let $\delta = \min(y_i - \tilde{r}_i, \tilde{r}_j)$. Then the solution $\hat{\boldsymbol{r}} = \tilde{r}_1, \ldots, \tilde{r}_i + \delta, \ldots, \tilde{r}_j - \delta, \ldots, \tilde{r}_n$ is also feasible. Comparing the difference of objective functions, we see:

   $$h(\hat{\boldsymbol{r}}) - h(\tilde{\boldsymbol{r}}) = (\tilde{r}_i + \delta)^2 - \tilde{r}_i^2 + (\tilde{r}_j - \delta)^2 - \tilde{r}_j^2 = 2\delta^2 + 2\tilde{r}_i\delta - 2\tilde{r}_j\delta.$$

   As δ is the lesser of $y_i - \tilde{r}_i$ or $\tilde{r}_j$, consider both cases:

   i. $\delta = y_i - \tilde{r}_i$: Then

   $$\delta^2 + \tilde{r}_i\delta - \tilde{r}_j\delta = y_i^2 - y_i\tilde{r}_i - \tilde{r}_j y_i + \tilde{r}_i\tilde{r}_j = (y_i - \tilde{r}_i)(y_i - \tilde{r}_j)$$

We already know that $y_i > \tilde{r}_i$. By the ordering of units, since $i > j$, we know that $y_i \geq y_j \geq \tilde{r}_j)$. Therefore $(y_i - \tilde{r}_i)(y_i - \tilde{r}_j) \geq 0$ so the solution $\hat{\boldsymbol{r}}$ is also optimal. Since $\delta = y_i - \tilde{r}_i$, then $\hat{r}_i = y_i = r_i$.

ii. $\delta = \tilde{r}_j$: Then

$$\delta^2 + \tilde{r}_i\delta - \tilde{r}_j\delta = \tilde{r}_j^2 + \tilde{r}_i\tilde{r}_j - \tilde{r}_j^2 = \tilde{r}_i\tilde{r}_j$$

As both $\tilde{r}_i \geq 0$ and $\tilde{r}_j \geq 0$, the solution $\hat{\boldsymbol{r}}$ is also optimal. As $\hat{r}_i = \tilde{r}_i + \tilde{r}_j < r_i$, it must be the case that some other unit j' is also non-zero and can be used to create $\delta' = \min(y_i - \hat{r}_i, \tilde{r}_{j'})$ and another optimal solution. This logic can be repeated until an optimal solution can be found that includes $\hat{r}_i = y_i$.

3. Update $\tilde{r}_i = \hat{r}_i = r_i$. At this point, $\tilde{r}_j = r_j$ for all $j \leq i$.

4. Continue for $i + 1$ and $R_0^{(i+1)} = R_0^{(i)} - r_i$.

At the end of this algorithm, $\tilde{\boldsymbol{r}} = \boldsymbol{r}$, the greedy solution, showing that any optimal solution can be transformed into the greedy solution while maintaining $h(\boldsymbol{r}) \geq h(\tilde{\boldsymbol{r}})$ at each step.

With a solution $\boldsymbol{r}$ to P', we can then translate back to P using the relationship $\boldsymbol{\tau}_0 = \boldsymbol{r} - \boldsymbol{y}$. Consequently, $\boldsymbol{\tau}_0$ has the form:

$$\tau_{0,i} = \begin{cases} 0, & i < s, \\ A_0 - \sum_{i=s+1}^{n} y_i, & i = s, \\ y_i, & i > s, \end{cases}$$

where s is the largest integer such that $\sum_{i=s}^{n} y_i > A_0$. □

1.6.3 Proof of Corollary 1.1

Proof. Observe that P is the continuous relation of the version of the problem for integer y_i. Let τ^* be the solution to P. For $i > s$, $\tau_i^* = y_i$, which are integer values. For $i = s$, $\tau_s^* = A_0 - \sum_{i=s+1}^{n} y_i$ is also an integer, as A_0 is an integer and the sum of any y_i values must also be an integer. For $i < s$, $\tau_i^* = 0$. Thus τ^* is an integer solution and must be the optimal solution to the integer constrained version of P. □

1.6.4 Additional simulation details

Testing the normal approximation

To test the suitability using the variance of the null distribution to approximate the p-values of the sharp null hypotheses, we simulated a small experiment with 10 units from which 5 were assigned to treatment.

First, the potential responses under control were simulated as:

$$y_i(0) = P + 20B,\ P \sim \text{Poisson}(7),\ B \sim \text{Binomial}(0.01, 2).$$

Next, the set of true treatment effects were added to the treated units' scores based on Table 1.3. The columns represent the treatment unit, and each row shows the individual effect of the treatment $\tau_{0,i}$. The true attributable effect for each row is the sum of row values. The first experiment adds one to the first treated unit, the second adds one to both the first and second, and so on. We also consider placing a much larger effect of six on the first unit and adding two to the first three units. For each allocation, the true attributable effect $A = \sum_{i=1}^{5} \tau_i$ was used to generate $\boldsymbol{y}(1)$ from $\boldsymbol{y}(0)$ and a hypothesis test of $A_0 = A$ was performed using the normal approximation strategy. Recall that the normal approximation is guaranteed to find the adjustment that leads to the largest variance of the null distribution of the test statistic T, but this may not correspond to the adjustment with the largest p-value, which is the true target. By enumerating all compatible allocations $\boldsymbol{\tau}_0$ and performing an exact randomization test, we can find the adjustment with maximum p-value and compare this p-value found by the normal approximation by computing the relative error $|p - \hat{p}|/p$, where p is the largest p-value and $\hat{p}$ is found from the method given in Section 1.2.2. In both cases, p-values were generated by completely enumerating all $\binom{10}{5}$ possible treatment allocations, generating the null distribution of the test statistic T^2, and comparing the observed test statistic to the null distribution. The simulations were repeated 100 times, each with a new $\boldsymbol{y}(0)$, for each true allocation.

	1	2	3	4	5
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1
6	6	0	0	0	0
7	2	2	2	0	0

Table 1.3: Strategies for allocating treatment effects used in small sample size simulations. Columns represent the true effect of treatment $\tau_{0,i} = y_i(1) - y_i(0)$ for each of 5 treated units. The attributable effect A is the sum of the row values.

In order to completely enumerate all possible treatment allocations compatible with a given A_0 as well as perform exact hypothesis tests, the simulations so far have been kept fairly small. To consider the effect of sample size on the performance of the variance maximization method, we repeated the simluations for larger experiments using 10 out of 20 treated and 15 out of 30 treated. For each experiment, the true treatment effect was 1 for 2 of the treated units and zero for the remainder. These experiments start to push the

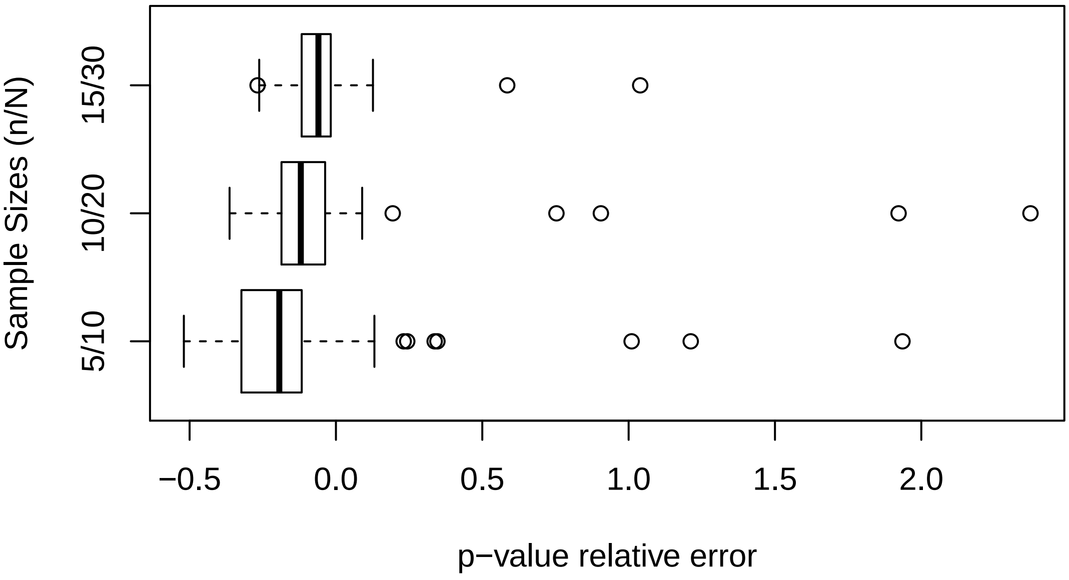


Figure 1.6: Boxplot of relative error when finding the largest p-value using the normal approximation method compared to complete enumeration. For $N = 10, n = 5$, p-values are computed exactly. For the simulations with 10 out of 20 and 15 out of 30 assigned to treatment, p-values are computed using 10,000 Monte Carlo samples.

boundaries of convenient computation when completely enumerating the entire randomization distribution, so a sample of 10,000 treatment assignments was used instead. If the method is working well, the distribution of p-values under the null should be approximately uniform when the null hypothesis is true. Figure 1.6 shows that the method performs reasonable well by this metric.

Coverage and confidence interval widths

The main chapter reports three simulations comparing the proposed methods to the survey sampling based method. Here we provide additional details on the simulation process.

For each simulation, the $\boldsymbol{y}(0)$ data were generated for N experimental subjects using a zero inflated binomial:

$$y_i(0) = (1 - P)B, P \sim \text{Bernoulli}(p), B \sim \text{Binomial}(100, 0.5).$$

This model was chosen to cover a range of conditions. When p is close to zero, the data are approximately bell shaped, which may show to the HB method's strengths. When p is increased, the data become bimodal and the HB method's approximations may cease to work well.

In order to create a full experiment, we must also generate $\boldsymbol{y}(\mathbf{1})$. To get the individual treatment effects τ_i, the population-level standard deviation σ_0 for the $\boldsymbol{y}(\mathbf{0})$ values are measured and a total effect computed as $\boldsymbol{\mathcal{T}} = \lfloor eN\sigma_0 \rfloor$, where e is the effect size multiplier. As the $\boldsymbol{y}(\mathbf{0})$ were discrete, the total treatment effect must be applied in integer amounts. There are $\binom{N+\boldsymbol{\mathcal{T}}-1}{N-1}$ possible ways to distribute the total effect $\boldsymbol{\mathcal{T}}$ to the N units. One was chosen uniformly at random and used to generate $\boldsymbol{y}(\mathbf{1})$.

For 5000 replications, a treatment assignment was generated and the observed data were created using the $y_i(1)$ for the treated units and the $y_i(0)$ for the control units. The true value of A was computed by subtracting the true $\boldsymbol{y}(\mathbf{0})$ from the observed data. For each replication, 95% confidence intervals were generated using the proposed method and the HB method. The interval widths were recorded as well as whether the intervals covered the true A value. To compute the p-value for the proposed method, 1000 Monte Carlo samples from the assignment mechanism were used.

Chapter 2

Randomization and permutation tests of network formation

2.1 Introduction

Recent years have seen renewed interest in "design-based" approaches to analyzing randomized controlled trials (RCTs). Design based inference, or randomization inference, relies on the known treatment assignment mechanism of a RCT to derive tests and estimators, rather than specifying parametric forms for outcomes. This paper is primarily concerned with the use of sharp null hypotheses. A sharp hypothesis specifies, exactly, the outcome that would be observed for all units under any possible treatment regime. The sharp null of no effects, as famously employed by Fisher (1935) in his "Lady Tasting Tea" experiment, states that the observed outcome would be identical under all possible treatment assignments, which allows the creation of a reference distribution for a suitable test statistic. By comparing the actually observed data to the distribution under the null, the researcher can answer the question, "How likely is that I would see data like this if treatment had no effect?" Sharp nulls can also be more expressive, allowing for constant additive effects, multiplicative effects, and even high degrees of heterogeneity (Rosenbaum, 2002b, 2010; Ding et al., 2016; Caughey et al., 2017).

Recently, methods have been developed to apply randomization inference to experiments with *spillover*, where treatment to one unit changes the outcome of another unit in the study (Rosenbaum, 2007; Bowers et al., 2013; Choi, 2017; Aronow and Samii, 2017; Athey et al., 2018). These studies use fixed networks, observed before treatment has been assigned, to test models stating how treatment to one node in the network changes the outcomes of other nodes. This paper considers a different question and asks if randomization inference can be used to analyze *network formation* itself. This type of analysis is typically the domain of parametric graph models that posit a distribution governing the generation of edges between nodes and then seek to estimate parameters for the models. Extensive coverage of these approaches can be found in

This chapter contains joint work with Professor Yuguo Chen.

Kolaczyk (2009); Goldenberg et al. (2010); Fienberg (2012); Hunter et al. (2012), and O'Malley (2013). One frequent application of graph modeling is to use estimated parameters to perform community detection, the identification of clusters within the graph of related nodes (Schaeffer, 2007; Fortunato, 2010; Coscia et al., 2011; Nascimento and de Carvalho, 2011; Fortunato and Castellano, 2012; Harenberg et al., 2014; Amelio and Pizzuti, 2014; Bedi and Sharma, 2016). While many of the community detection algorithms have been developed without a clear statistical model, they can often be thought of as maximum likelihood estimation for a suitably chosen model of link formation (Newman, 2013).

While the area of network analysis has seen much growth and continues to be a rapidly developing field, little has been done to engage with experimental design. Specifically, parametric models are problematic from a purely design-based perspective. First, the experimental design of an RCT does not typically justify the assumptions implicit in parametric models (Berk, 2004; Freedman, 2008a,c,b). Regression techniques fail to express the true stochastic nature of the experimental design, leading to possible bias in estimates and, even when estimates are correct, problematic standard errors. Second, the parametric models frequently condition on post-treatment outcomes when estimating parameters. Exponential random graph models, for example, require estimating probabilities of link formation conditional on the rest of the network rather than conditioning on the treatment assignment alone (Frank and Strauss (1986); Wasserman and Pattison (1996); though see Suesse (2012) for an alternative approach). As a result, these approaches are best classified as *mediation analysis*, as they condition a dyad's link formation on other post-treatment variables, namely the link formation of other dyads. While such mediation analyses may be interesting in their own right, they may not be the main question of interest and frequently require stronger assumptions in order to maintain causal interpretations (Robins and Greenland, 1992).

In this paper, we develop tests inspired by traditional network models, but built on a strong design-based foundation. All of the tests proposed in this paper derive their statistical justification from the random assignment of treatment to the nodes while still taking advantage of several ideas and algorithms pioneered in the existing network science literature. We classify our proposed methods as being based on local or global network structure. Local network structure tests closely follow the development of random graph models in that they are based on counting the presence of topological features (edges, triangles, etc.) within the treated and control groups. Global approaches, on the other hand, more closely follow graph partitioning and clustering approaches in that they first analyze the entire graph, irrespective of treatment assignment, and then use treatment assignment for inference. In Section 2.2.1 we review causal inference approaches to randomized trials that use the randomization procedure as the "reasoned basis" for inference (Fisher, 1935), in particular how this approach can be extended to networks. In Sections 2.2.2 and 2.2.3 we develop local and global tests that capture various ways in which treatments can influence network features. In Section 2.3 we evaluate the proposed methods in a variety of simulated network generation processes. In Section 2.4 we apply the methods to a gene wide association study after a randomized controlled trial and to a network of board members for a set of Norwegian companies. Section 2.5 concludes with a brief discussion.

2.2 Method

2.2.1 Causal inference for RCTs

Consider n units in a study with a random treatment $\boldsymbol{Z} \equiv (Z_1, Z_2, \dots, Z_n)'$, where $Z_i \in \{0, 1\}$. As $\boldsymbol{Z}$ is controlled by the researcher, the distribution of $\boldsymbol{Z}$ is known. Typically, and throughout this document, $\boldsymbol{Z}$ is generated by selecting n_1 units for treatment, with the remaining n_0 units receiving control, and with $\boldsymbol{Z}$ equally probable. To simplify later notation, for any binary variable write $\boldsymbol{Z}^{(0)} \equiv \boldsymbol{1} - \boldsymbol{Z}$.

In the potential outcomes framework (Neyman, 1923), each unit's response is a fixed value indexed by the treatment assignment: $Y_i = y_i(\boldsymbol{Z})$. Write $\boldsymbol{Y} = (y_1(Z_1), y_2(Z_2), \dots, y_n(Z_n))'$. Unit i would have response $y_i(\boldsymbol{z})$ if $\boldsymbol{Z} = \boldsymbol{z}$, but for some other treatment assignment $\boldsymbol{u}$, the response would be $y_i(\boldsymbol{u})$. We say that a treatment has an effect if $y_i(\boldsymbol{z}) \neq y_i(\boldsymbol{u})$ for at least one unit i. Write $\boldsymbol{y}(\boldsymbol{u}) = (y_1(\boldsymbol{u}), y_2(\boldsymbol{u}), \dots, y_n(\boldsymbol{u}))'$. By the fundamental problem of causal inference (Holland, 1986), we cannot observe both $y_i(\boldsymbol{z})$ and $y_i(\boldsymbol{u})$, so we must perform inference to determine if treatment has an effect. In this paper we use a sharp null hypothesis of no effect that states $H_0 : \boldsymbol{y}(\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{u})$ for all $\boldsymbol{z}$ and $\boldsymbol{u}$. Under this hypothesis, we write $\boldsymbol{y}$ for the common outcome that does not depend on treatment. The distribution of any statistic $T(\boldsymbol{Z}, \boldsymbol{y})$ can be determined by the distribution of $\boldsymbol{Z}$, which is a known randomization process (Fisher, 1935; Maritz, 1981; Rosenbaum, 2002b).

	$Z_i = 1$	$Z_i = 0$	Total
$Y_i = 1$	$\boldsymbol{Z}'\boldsymbol{Y}$	$\boldsymbol{Z}^{(0)\prime}\boldsymbol{Y}$	$\boldsymbol{Y}'\boldsymbol{Y}$
$Y_i = 0$	$\boldsymbol{Z}'\boldsymbol{Y}^{(0)}$	$\boldsymbol{Z}^{(0)\prime}\boldsymbol{Y}^{(0)}$	$\boldsymbol{Y}^{(0)\prime}\boldsymbol{Y}^{(0)}$
Total	$\boldsymbol{Z}'\boldsymbol{Z} = n_1$	$\boldsymbol{Z}^{(0)\prime}\boldsymbol{Z}^{(0)} = n_0$	n

Table 2.1: Cross tabulated outcomes under binary treatment experiment. By design, column totals are fixed. Under the sharp null hypothesis of no effect, the row totals are fixed.

When the outcome is binary, the result of the experiment can be summarized in a 2 × 2 table, as shown in Table 2.1. By design the column totals in this table are fixed. Fisher (1935) showed that if treatment has no effect, the row totals are also fixed, and the statistic $\boldsymbol{Z}'\boldsymbol{Y}$ follows a hypergeometric distribution. Since $\boldsymbol{Z}'\boldsymbol{Y}$ completely determines the table, when the margins are fixed we often say that the set of tables is distributed according to the hypergeometric distribution. For an alternative motivation of this distribution, one can imagine enumerating all possible $\boldsymbol{Z}$ and computing a test statistic $T(\boldsymbol{Z}, \boldsymbol{Y})$ for each one. If the test statistic is the total in the treated group, $T(\boldsymbol{Z}, \boldsymbol{Y}) = \boldsymbol{Z}'\boldsymbol{Y}$, this procedure is precisely Fisher's exact test.

This later formulation based on $T(\boldsymbol{Z}, \boldsymbol{Y})$ also introduces the role of the alternative to the test. If large values of T indicate evidence against the null hypothesis, we define the p-value of the test as

$$p(\boldsymbol{z}, \boldsymbol{y}) = \mathrm{P}\left(T(\boldsymbol{Z}, \boldsymbol{y}) \geq T(\boldsymbol{z}, \boldsymbol{y})\right) = \sum_{\boldsymbol{Z} \in \Omega} \mathrm{P}\left(\boldsymbol{Z}\right) I(T(\boldsymbol{Z}, \boldsymbol{y}) \geq T(\boldsymbol{z}, \boldsymbol{y})),$$

where $\boldsymbol{z}$ and $\boldsymbol{y}$ are the observed treatment and outcome and Ω is the sample space of possible assignments. In this paper we focus on complete random assignment for which $\mathrm{P}\left(\boldsymbol{Z} = \boldsymbol{z}\right) = (n_1!(n - n_1)!)/n!$, but the definition encompasses any treatment assignment mechanism. For a one-sided test of Table 2.1, using either $\boldsymbol{Z}'\boldsymbol{Y}$ or $-\boldsymbol{Z}'\boldsymbol{Y}$ as a test statistic is a straightforward approach. For two-sided tests, one might use $1 - h(\boldsymbol{Z}'\boldsymbol{Y})$, where h is the probability mass function of the hypergeometric distribution (Freeman and Halton, 1951), or $(\boldsymbol{Z}'\boldsymbol{Y} - \mu_0)^2$ where $\mu_0 = \boldsymbol{1}'\boldsymbol{Y} n_1/n$ is the mean of $\boldsymbol{Z}'\boldsymbol{Y}$ under the sharp null of no effects (Radlow and Alf, 1975). Gibbons and Pratt (1975) and Agresti (2013, Section 3.5.3) discuss the relative merits of these approaches. As we shall see in the subsequent questions, these two approaches suggest similar test statistics when testing a sharp null hypothesis of no effect for a network.

The randomization inference approach generalizes to networks in a natural way. Consider applying treatment to the n units and then measuring a simple network for those units (i.e., an undirected network with no self-loops). Rather than focus on the n subjects in the experiment, shift focus to the $m = n(n-1)/2$ possible connections between them. Each dyad (i, j), $i < j$, may have one of four possible treatment assignments: when i is treated and j is treated ($Z_i = 1, Z_j = 1$); when i is treated, but j is in the control condition ($Z_i = 1, Z_j = 0$); when j is treated, but i is in the control condition ($Z_i = 0, Z_j = 1$); and when both i and j are in the control condition ($Z_i = Z_j = 0$). If we assume that edge (i, j) would behave in the same fashion if either one of its endpoints were treated, we can state the treatment levels as $W_{ij} = Z_i + Z_j \in \{0, 1, 2\}$.

As with other types of outcomes, we can posit the existence of *potential networks* composed of *potential edges*, with respect to a treatment regime $\boldsymbol{W}$. For each dyad (i, j), let $y_{ij}(\boldsymbol{W}) = 1$ if units i and j have a link following treatment $\boldsymbol{W}$ and $y_{ij}(\boldsymbol{W}) = 0$ otherwise. If treatment had an effect, then $y_{ij}(\boldsymbol{W}) \neq y_{ij}(\boldsymbol{V})$ for at least one dyad ij for some $\boldsymbol{W} \neq \boldsymbol{V}$. If treatment had no effect, then we would observe the same network under all treatment assignments. Under this hypothesis, the sharp null of no effect, we can apply randomization inference to the network by selecting a test statistic $T(\boldsymbol{W}, \boldsymbol{Y})$. In the next section, we consider

T that operates on $\boldsymbol{Y}$ directly. In Section 2.2.3 we consider test statistics $T(\boldsymbol{Z}, g(\boldsymbol{Y}))$ that operate on $\boldsymbol{Y}$ through a function g that summarizes the network in some fashion.

2.2.2 Local approaches

In this section, we extend the logic of randomization tests for contingency tables to propose non-parametric tests of network formation as an alternative to parametric graph models. Indeed, this approach is based on the observation that networks can be collapsed into contingency tables, an observation that lead to the original fitting of graph models by logistic regression and approaches designed for log-linear models (Holland and Leinhardt, 1981; Fienberg and Wasserman, 1981b,a). Like existing graph models, the proposed non-parametric tests can be used to understand many aspects of network topology. Like Fisher's exact test, they require no assumptions beyond the design of the RCT itself.

	$W_{ij} = 2$	$W_{ij} = 1$	$W_{ij} = 0$	Total
$Y_{ij} = 1$	$\boldsymbol{W}^{(2)'}\boldsymbol{Y}$	$\boldsymbol{W}^{(1)'}\boldsymbol{Y}$	$\boldsymbol{W}^{(0)'}\boldsymbol{Y}$	$\boldsymbol{Y}'\boldsymbol{Y}$
$Y_{ij} = 0$	$\boldsymbol{W}^{(2)'}\boldsymbol{Y}^{(0)}$	$\boldsymbol{W}^{(1)'}\boldsymbol{Y}^{(0)}$	$\boldsymbol{W}^{(0)'}\boldsymbol{Y}^{(0)}$	$\boldsymbol{Y}^{(0)'}\boldsymbol{Y}^{(0)}$
Total	$m_2 = n_1(n_1 - 1)/2$	$m_1 = n_1 n_0$	$m_0 = n_0(n_0 - 1)/2$	$m = n(n-1)/2$

Table 2.2: Within and across group edge counts for an experiment on n nodes. Again, column totals are fixed by design, while row totals are fixed under the sharp null hypothesis.

As in the previous section, let $Y_{ij}(\boldsymbol{W}) = 1$ when there is a link in the network between i and j. Analogously to the binary case, define $\boldsymbol{W}^{(k)} \equiv (I(W_{12} = k), I(W_{13} = k), \ldots, I(W_{(n-1)n} = k))'$ and summarize the results of the treatment in Table 2.2. As with the 2 × 2 table, the column totals are fixed by the design. Under the sharp null of no effect (i.e., $Y_{ij}(\boldsymbol{W}) = Y_{ij}(\boldsymbol{V})$ for all $\boldsymbol{W}$, $\boldsymbol{V}$, i, and j), the row totals are fixed. As the row and column totals are fixed under the sharp null of no effects, the table is determined through two cells: $R_1 = \boldsymbol{W}^{(2)'}\boldsymbol{Y}$ (the number of edges within the treated group) and $R_2 = \boldsymbol{W}^{(0)'}\boldsymbol{Y}$ (the number of edges in the control group). Following the two-tailed tests of binary outcomes, we define

$$T_{\text{PMF}}(\boldsymbol{W}, \boldsymbol{Y}) = 1 - f(R_1, R_2), \tag{2.1}$$

where $R_1 = \boldsymbol{W}^{(2)'}\boldsymbol{Y}$, $R_2 = \boldsymbol{W}^{(0)'}\boldsymbol{Y}$, and f is the probability mass function of (R_1, R_2).

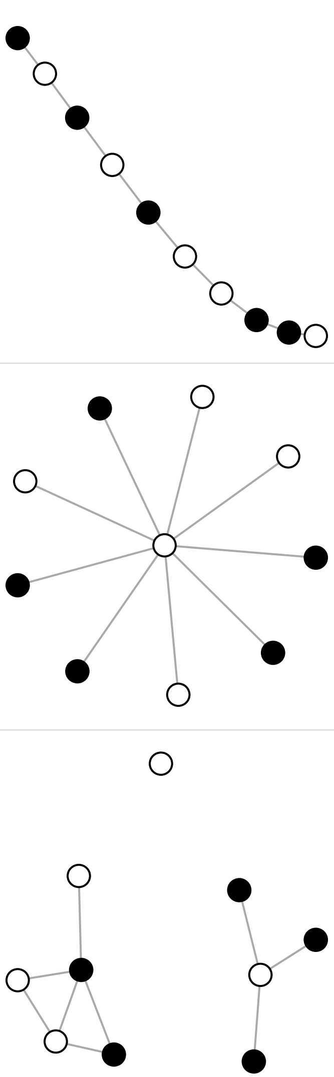
Fixed row and column totals might suggest that (R_1, R_2) follows a multiple hypergeometric distribution, but this is not the case. To illustrate this point, Figure 2.1 shows three simple networks: a line, a star, and an irregular network. All three have 10 nodes and 9 edges, so the row and column margins will all be the same for all three networks, but the distribution of tables is quite different. In this example, we will take 5 of the 10 nodes to be treated, leaving 5 in the control condition. There are $\binom{10}{5} = 252$ possible treatment assignments, each of which implies a particular table, though the table may be non-unique. The star network, for example, only has two unique tables depending on whether the central node is in the treatment or control conditions. For a particular treatment assignment, Figure 2.1 also shows the implied table for the network. Despite having the same number of nodes and edges, the distributions are very different, with the support increasing for more complex networks. Also observe that while the line and irregular network have the same table for the given randomization, the probability of observing this table when the network is a line is about 3 times larger than when the network has the pattern of the irregular network.

While the distribution function f is generally difficult to compute exactly, the first few moments of the distribution can be computed by combinatorial analysis (Frank, 1977, 1978; Chen and Friedman, 2017). Using these moments, we can create a mean centered statistic using the first two moments of (R_1, R_2):

$$T_{\mathrm{CF}}(\boldsymbol{W}, \boldsymbol{Y}) = \begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix}' \Sigma_{12}^{-1} \begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix} \tag{2.2}$$

The moments of R_1 and R_2 are given in Appendix 2.6. We label this test statistic as T_{CF} as it was first proposed by Chen and Friedman (2017), though variations appear in other places in the literature. Chen and Friedman propose T_{CF} for two sample tests for high dimensional data in which a graph has been formed over the combined samples. While there has been a long tradition of graph based permutation tests for high-dimensional or object data (Friedman and Rafsky, 1979, 1983; Schilling, 1986; Henze, 1988; Rosenbaum, 2005; Biswas et al., 2014), previous approaches required highly structured graphs as inputs such as minimum spanning trees, disjoint edges, or graphs in which all nodes had equal degree. To our knowledge, Chen and Friedman were the first to propose a test for arbitrary graphs, such as those that would be the result of an experiment. Through the close relationship between permutation tests and randomization tests, their test statistic also applies when analyzing network formation in randomized controlled trials.

Statistics based on R_1 have a longer history in the literature. Whaley (1983) links seemingly unrelated statistics, all of which are variations on a statistic proposed by Mantel (1967). Nyblom et al. (2003) apply a statistic based only on R_1 to the network context. For the more general matrix context, Baker and Hubert (1981) propose a method than can generate within and across group edge count statistics. Dow and de Waal (1989) uses this method to study "compactness" (within group edges) and "isolation" across group edges for a variety of networks. Additional discussion on the wide application of these types of statistics in the context of permutation tests can be found in Good (2005, Chapter 10).



	$W_{ij} = 2$	$W_{ij} = 1$	$W_{ij} = 0$
$Y_{ij} = 1$	1	1	7
$Y_{ij} = 0$	9	9	18

$P(T) = 0.1269841$

	0	1	2	3	4
0	2	4			
1	4	32	24		
2		24	72	24	
3			24	32	4
4				4	2

	$W_{ij} = 2$	$W_{ij} = 1$	$W_{ij} = 0$
$Y_{ij} = 1$	0	4	5
$Y_{ij} = 0$	10	6	20

$P(T) = 0.5$

	0	4
0		126
4	126	

	$W_{ij} = 2$	$W_{ij} = 1$	$W_{ij} = 0$
$Y_{ij} = 1$	1	1	7
$Y_{ij} = 0$	9	9	18

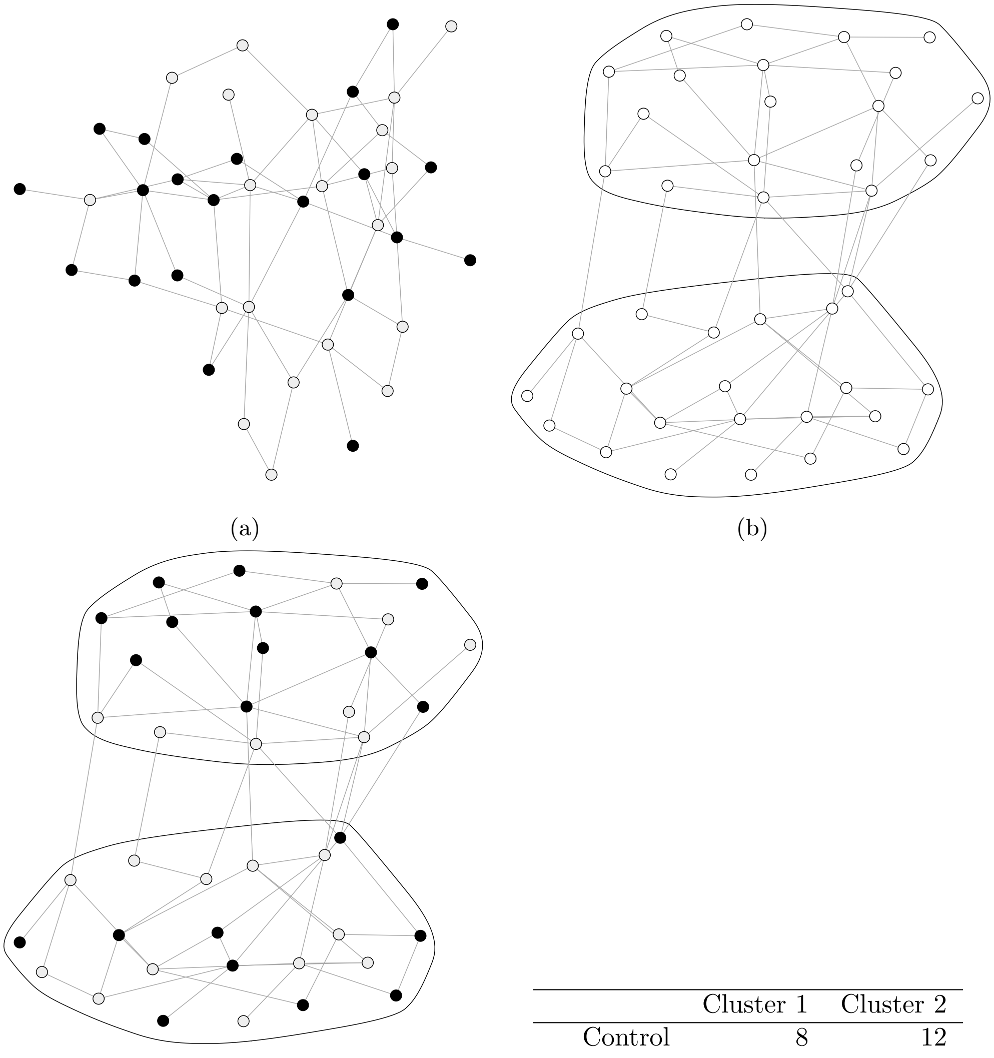
$P(T) = 0.03968254$

	0	1	2	3	4	5	6
0		1	6	12	8	1	
1	1	10	21	21			
2	6	21	36	19	6	3	
3	12	21	19	2	2	1	1
4	8		6	2			
5	1		3	1			
6				1			

Figure 2.1: Three networks with 10 nodes and 9 edges: line, star, and irregular (right) For an example randomization of 5 nodes to treatment (white) and control (black), the observed contingency table is shown (middle). The joint distribution (right) of statistics $R_1 = \boldsymbol{W}^{(2)'}\boldsymbol{Y}$ (rows) and $R_2 = \boldsymbol{W}^{(0)'}\boldsymbol{Y}$ (columns). The probability of observing a table with the given statistic is the cell value of the joint distribution table divided by 252.

2.2.3 Global approaches

In the previous section, randomization inference was applied to the observed network by first splitting the network into treatment and control subgraphs, and then comparing features of the two subgraphs, such as edge counts. In this section, we take the approach of analyzing the entire graph first and then applying randomization inference to the results of that analysis. Critically, the first step in this process, analyzing the entire graph, is done without respect to the observed treatment assignment. Only after performing this analysis is the treatment assignment information used to construct a randomization test.



(a)

(b)

(c)

	Cluster 1	Cluster 2
Control	8	12
Treatment	11	9

(d)

Figure 2.2: A graphical representation of using community detection to form a hypothesis test of the sharp null of no effects. Panel (a) shows an example network with treated (black) and control (gray) nodes. In panel (b), the treatment assignments are ignored and clustering is performed. In panel (c), treatment labels are returned and assignment-cluster totals are used to form panel (d).

Our first set of test statistics are constructed by applying community detection algorithms to the graph. Reviews of community detection algorithms can be found in Schaeffer (2007); Fortunato (2010); Coscia et al. (2011); Nascimento and de Carvalho (2011); Fortunato and Castellano (2012); Harenberg et al. (2014); Amelio and Pizzuti (2014), and Bedi and Sharma (2016). In this paper, we focus on algorithms for *non-overlapping clusters*, also known as *graph partitioning* algorithms. After applying the community detection algorithm, we have a set of k labels $\mathcal{C} \in \{1, \ldots, k\}$ for the n nodes. We use these labels to construct a $2 \times k$ table of treatment and control nodes in each cluster. In the simplest case when $k = 2$, we have another example of a binary classification problem, as in Table 2.1. As the counts in this table are *nodes* rather than *dyads* analysis can proceed via Fisher's exact test, using either a hypergeometric probability mass function based test statistic or a mean centered test statistic. For $k > 2$ clusters, numerous extensions exist that generalize the 2×2 methods to $2 \times k$ tables (Agresti, 1992; Hirji and Johnson, 1996).

Figure 2.2 provides a graphical representation of using clustering to create a hypothesis test. For a small simulated network and treatment assignment, the figure shows the initial network, clustering without treatment assignment labels, adding the labels back, and cross classifying the node-cluster counts. In this example, spectral clustering was performed to partition the graph into two blocks, though any other clustering procedure may be used. Using a two-sided Fisher's exact test, the p-value of this test is 0.527.

Researchers have identified many other global properties of graphs that can be used to describe their topology. Those methods that assign numerical or ordinal scores to nodes can be used to construct tests as well. One key area of inquiry in social network analysis is ranking nodes on their "centrality" to the network. There are several different measures of centrality (Freeman, 1978), typically based on either graph theoretic quantities such has the number of paths in which a node is present (Borgatti, 2005; Borgatti and Everett, 2006) or spectral decomposition of the graph (Bonacich, 1972, 2007). For this paper we focus a spectral method, the eigenvector of the largest eigenvalue λ of the adjacency matrix A:

$$A\boldsymbol{x} = \lambda \boldsymbol{x}.$$

For each node, x_i can be thought of as proportional to the sum of the centrality scores of i's neighbors, where λ is the constant of proportionality. Therefore, central nodes are those that are connected to other central nodes, on average. When there are multiple disconnected components to the graph, there will be multiple eigenvectors for λ, with the ith entry being non-zero for only one vector for each node i. In that case, we take x_i to be the non-zero entry for any of the matching eigenvectors.

To perform inference, the x_i values can be ranked to perform a Wilcoxon-Mann-Whitney (WMW) test of

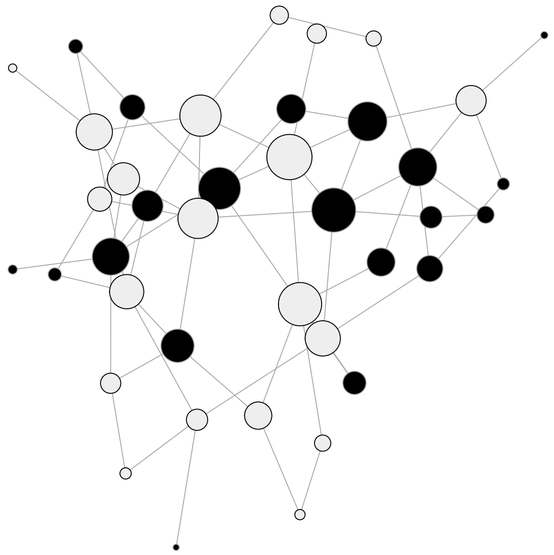


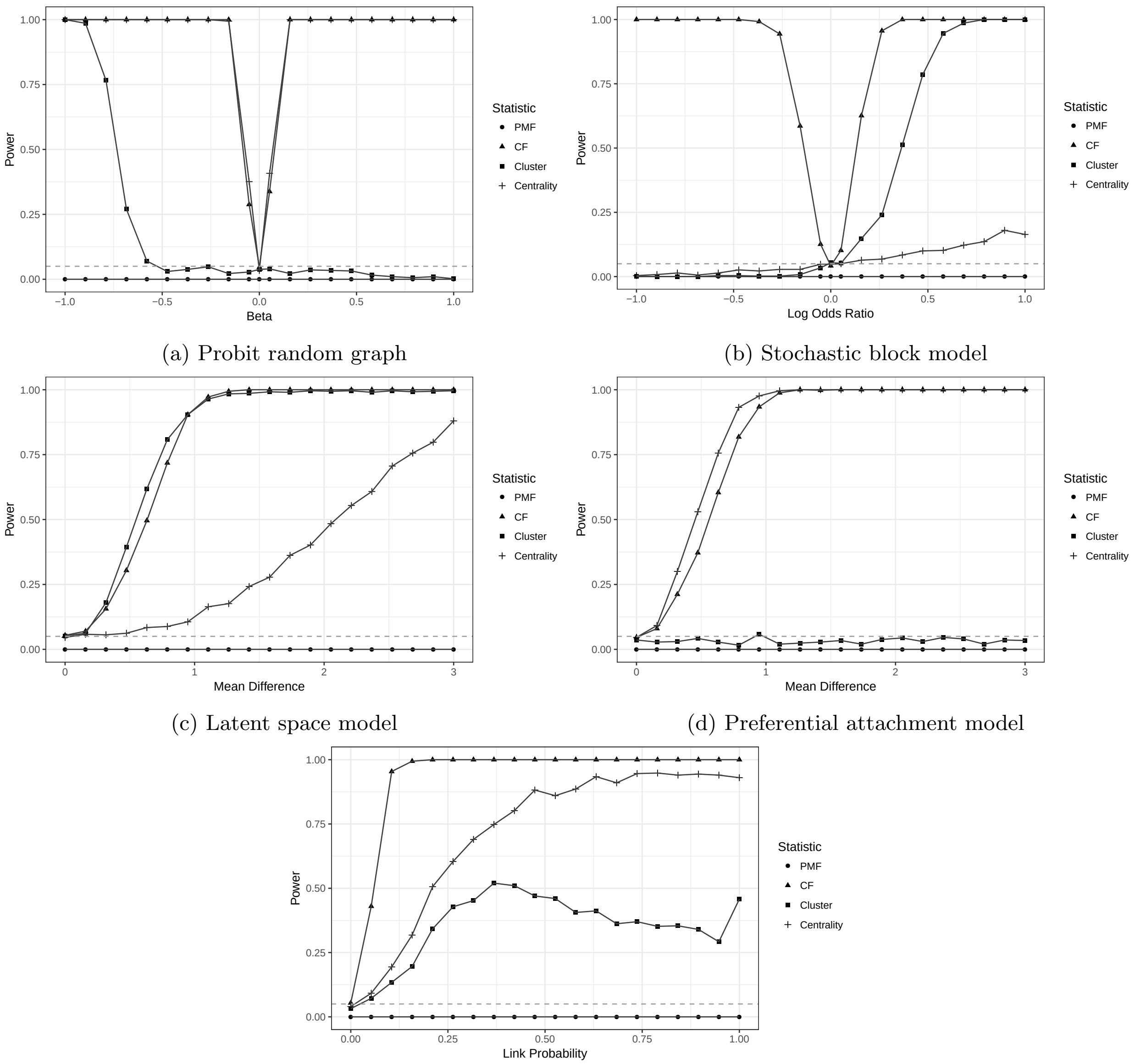
Figure 2.3: Example network from Figure 2.2 with node sizes proportional to the rank of eigenvector centrality.

the hypothesis that treatment had no effect on the network (Lehmann, 1975; Maritz, 1981). While any other randomization test for numerical data could be performed, such as a permutational t-test, distribution free methods like the WMW have a close connection to the randomization inference literature and make good choices when the scale of the data is of secondary importance to the relative contributions of the individual observations (Rosenbaum, 2002b, 2010). Figure 2.3 plots the network used in the previous example with node sizes proportional to the rank of the centrality of the node, as measured by eigenvector centrality. The p-value from the WMW test is 0.698.

2.3 Simulations

In the following simulations, we test how well the four discussed test statistics perform when the null hypothesis is false because the network depends on the treatment assignment in some way. All simulations contain 100 units with n_1 of those assigned to the treatment condition. For each of $k = 500$ replications, the network is generated and the strict null hypothesis of no effect is tested with each of the four statistics at the $\alpha = 0.05$ level.

The first simulation is performed on a random graph where the probability of each edge depends the number of nodes (0, 1, or 2) that are treated. In particular, $\mathrm{P}\left(Y_{ij} = 1\right) = \Phi(\beta(Z_i + Z_j) - 1)$, where Φ is the cumulative distribution function (CDF) of a standard normal variable and β is a coefficient that is varied



(a) Probit random graph

(b) Stochastic block model

(c) Latent space model

(d) Preferential attachment model

(e) Neighborhood closing model

Figure 2.4: Power plots for the four test statistics for a variety of data generation methods. Each model is parameterized by the x-axis. The y-axis is the probability of rejecting the null hypothesis at the $\alpha = 0.05$ level. Panel (a) is a random graph with link probability equal to $\Phi(\beta(Z_i + Z_j) - 1)$, where Φ is the standard normal CDF. Panel (b) is a stochastic block model parameterized on the log odds ratio of within block edges compared to across block edges. Panel (c) is a one-dimensional latent space model parameterized on the difference of cluster centers. Panel (d) is a preferential attachment model parameterized by the preference for treatment members. Panel (e) adds edges between the neighbors of treated units with probability p starting from a fixed network.

from -1 to 1. We label this model the "probit random graph" as it shares a similar functional form to probit generalized linear models. Figure 2.4(a) shows the results of the simulation in which both the centrality and CF statistics perform quite well. The clustering based statistic generally performs poorly. Only when the β

is small is it able to distinguish that the treatment is having an effect on the network.

The second simulation generates the network from a stochastic block model, also known as a planted partition model. In this model, the treatment and control groups define two latent communities. We take the simple approach in which all edges within communities occur with probability p_{within}, independently. All across group edges occur with probability p_{across}. We parameterize the simulation using an odds ratio parameter

$$\theta = \frac{p_{\text{within}}/(1-p_{\text{within}})}{p_{\text{across}}/(1-p_{\text{across}})}.$$

For these simulations, we draw a random across group probability from $[0.1, 0.5]$ and then fix the probabiilty of within group edge based on the equation:

$$p_{\text{within}} = \frac{\theta p_{\text{across}}}{1 + p_{\text{across}}(\theta - 1)}.$$

Figure 2.4(b) shows the results of these simulations with $\log(\theta)$ on the x-axis. Again, the CF statistic performs very well. As the probability of a within group edge gets larger than an across group edge, the clustering based statistic also performs well. Both the centrality and PMF statistic do not perform well in this simulation.

In the third simulation, we use a one-dimensional latent space model. Each node i is given a location on the real line x_i, and the probability of an edge between any two nodes i and j is given by $\exp(-(x_i - x_j)^2)$. The locations of the control nodes are drawn from a standard normal, while the treated units are drawn from a normal distribution with mean μ and unit variance. Figure 2.4(c) shows the best performance from the clustering and CF statistics, with some the centrality statistic achieving some power as the average distance between the treated and control groups increases. Again the PMF statistic performs below its nominal level.

In the fourth simulation, we generate a “scale free” network where few nodes have very high degree and most nodes have very low degree. Pairs of nodes (i, j) are drawn with probabilities p_i and p_j, respectively, and an edge is formed between i and j. The process is repeated until $n(n-1)/8$ edges are allocated. Nodes with higher probabilities of being sampled will have many more neighbors than those with low probabilities. To assign probabilities, we use the latent positions x_i from the previous simulation. All nodes are ranked such that $r_i = 1$ implies that node i has the highest x_i and $r_i = n$ implies i has the smallest x_i. Then $p_i \propto 1/r_i$, such that $\sum_{i=1}^{n} p_i = 1$. As the parameter μ, the difference between the center of the treatment and control latent positions, increases, the probability that most preferred nodes are treated nodes increases as well. Figure 2.4(d) shows that the centrality statistic performs the best for this data generating process, with CF also performing well. The cluster and PMF statistics have no power at any value of θ.

In the fifth simulation, we take an algorithmic approach to network generation. First, we generate an Erdös-Renyi random graph with edge probability 0.10. After assigning treatment and control labels to the nodes, for all nodes i and j such that there exists a node k that is a treated node and (i, k) and (j, k) are in the original graph, can form a link between i and j with probability p, which is varied from 0 to 1 over the simulation. Figure 2.4(e) shows that the CF statistic outperforms the others, though both the centrality and the clustering statistic have reasonable power.

While not an exhaustive list of ways in which networks could be generated, the five selected models cover many of the most common approaches used in network analysis. Looking across these simulations, we see that the CF statistic is frequently a very powerful statistic, often having the greatest power or nearly greatest power of the four statistics. If researchers suspect that treatment induces a stochastic block model or a preferential attachment model, the clustering or centrality statistics might prove a better choice.

2.4 Applications

2.4.1 Gene-wide association study

Tsavachidou et al. (2009) conducted a 2×2 factorial randomized controlled trial to test the effect of selenium and vitamin E to combat the progression of prostate cancer. Both selenium and vitamin E had been identified in a previous observational study of prostate cancer as having potentially positive benefits. Subjects were recruited from patients scheduled to undergo a prostatectomy due to existing prostate cancer. Overall, 39 patients were recruited. After 3 to 6 weeks of treatment (placebo, selenium, vitamin E, or both), 39 subjects underwent surgery to remove their prostates. Cells were collected and subjected to expression assay. The original study selected cells in three different regions of the excised prostate: epithelial cells, stroma cells, and tumor cells. As only epithelial cells assays are available for all 39 patients, we focus on only those data in this analysis.

After collecting the microarray expression data, Tsavachidou et al. (2009) fit two-way ANOVA models for the two main effects as well as the interaction effect, assuming Normally distributed error terms. With nearly 14,000 genes under study, the researchers applied a beta-uniform mixture model to control the false discovery rate at the 2% level.Comparing the placebo to selenium, vitamin E, and combination treatments, the researchers found 2109 differentially expressed genes, with 1329 of those significant comparisons coming from the selenium-placedbo contrasts. Along with a unsupervised cluster analysis of gene expressions, Tsavachidou et al. (2009) concluded that there were significant differences between the treatment conditions with respect to gene expression.

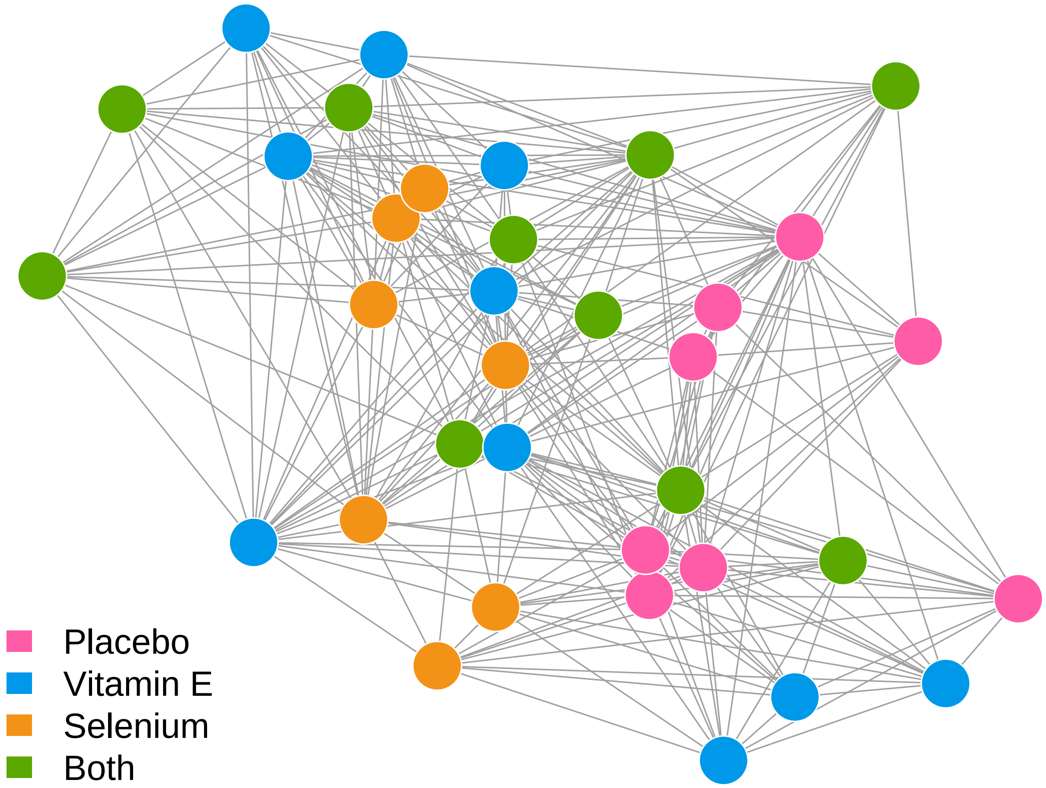


Figure 2.5: The network derived from the gene expression data described in Tsavachidou et al. (2009) based on similarity of expression.

As an alternative to the parametric methods employed in the original publication, we will now apply a network based approach that will have the dual benefits of requiring no asymptotic or parametric assumptions as well as providing an omnibus test of the hypothesis that the treatment groups have the same pattern of gene expressions. To this end, we create a gene co-expression network in which nodes are subjects and edges are present between subjects that have a similar pattern of gene expression, looking across all genes in the microarray assay.

For each subject, we rank all genes by expression level. Overall rates of expression may vary for subjects for idiosyncratic reasons; transforming expression levels into ranks within subjects allows for a common scale. From these ranks, an edge is added between i and j if either i or j is in the other's top ten ranks. Figure 2.5 shows the resulting network for the 39 subjects and 74 edges. Nodes are color coded by their treatment assignment.

After collapsing the treatment categories to subjects that received any selenium (the selenium and combination therapy groups) and those that did not (the vitamin E and pure control groups), we test the null hypothesis of no effect on the network using the four proposed test statistics. The strongest result was found for the clustering statistic. Figure 2.6 shows the clusters found when using the clustering statistic. Visually, the treated units (black circles) largely separate from the control units (white squares). The p-value of 0.0409 quantifies that this type of pattern would occur in very few random assignments, providing evidience against the sharp null of no effects. This evidence is further reflected in the local statistics, with the CF statistic having a p-value 0.059 and the PMF statistic having a p-value of 0.0648. The centrality statistic provided little evidence against the null with a a p-value of 0.8834. That the centrality statistic was not particularly powerful is consistent with the manner in which this network was constructed, with every node having at least a degree of 10.

2.4.2 Female representation on corporate boards

Seierstad and Opsahl (2011) studied female representation on 384 corporate boards in Norway over the period of May 2002 to August 2011. On alternating months, they compiled lists of corporate boards and matched first names to lists of names that have clear gender reference. Names that could not be easily matched were assigned a gender by investigating corporate web sites. Seierstad and Opsahl (2011) created two mode networks that linked individuals to boards and one mode projections linking individuals who served on the same board in the same month.

We investigate the network created by the union of networks of board members for the period of October 2010 to August 2011, comprising six individual networks. Figure 2.7 shows the network with female members in white and male members in black. We test the null hypothesis that gender labels can be shuffled at random. The local statistics showed fairly strong evidence against the null with the p-value for the Chen and Friedman statistic being equal to 1 in 100,001, which is the number of Monte Carlo samples used, and the p-value for the PMF statistic being 0.076. The global statistics showed less evidence against the null with the cluster and centrality statistics returning p-values of 0.784 and 0.441, respectively.

2.5 Discussion

In this paper we considered a rigorous application of randomization inference to the setting of analyzing networks that are the result of a randomized controlled trial. This approach makes no assumptions how networks are formed, but allows researchers to select test statistics that are sensitive to different network formation processes. From the simulations, we see edge count statistics are typically good choices for any data generating problem, but for stochastic block models or preferential attachment networks, the cluster or centrality statistic can be more powerful.

The flexibility to select a test statistic for a particular alternative is a primary advantage of this method. Creating statistics beyond the four proposed is also possible. Instead of counting edges, within and across treatment group counts of triangles could replace edge counts. There are several other measures of node centrality that may be useful for particular problems. It is also possible to take a hybrid approach in which the treatment and control groups are partitioned in the manner of a local test, but these subgroups are then analyzed using techniques more in line with global approaches.

Throughout this paper, we considered the case when a fixed number of units n_1 was selected for treatment with equal probability, but the method is also applicable to other assignment mechanisms. Provided one can sample from the randomization distribution, Monte Carlo methods can be used to get estimates of the null distributions. The moments from the T_{CF} statistic can be re-computed directly from the treatment assignment, or estimated from the same Monte Carlo process.

2.6 Moments of R_1, R_2

The test statistic T_{CF} requires computing the expected number of edges within the treated group (μ_1) and the control group (μ_2), as well as the variance-covariance matrix Σ, with entries σ_{ij}. Let $|E|$ be the number of edges in the network, n the number of nodes, n_1 the number of units assigned to the treatment condition, and $n_0 = n - n_1$ the number of units assigned to control.

$$\begin{aligned}
\mu_1 &= |E|\frac{n_1(n_1-1)}{n(n-1)}, \\
\mu_2 &= |E|\frac{n_0(n_0-1)}{n(n-1)}, \\
\sigma_1^2 &= \mu_1(1-\mu_1) + C\frac{n_1(n_1-1)(n_1-2)}{n(n-1)(n-2)} + (|E|(|E|-1)-C)\frac{n_1(n_1-1)(n_1-2)(n_1-3)}{n(n-1)(n-2)(n-3)}, \\
\sigma_2^2 &= \mu_2(1-\mu_2) + C\frac{n_0(n_0-1)(n_0-2)}{n(n-1)(n-2)} + (|E|(|E|-1)-C)\frac{n_0(n_0-1)(n_0-2)(n_0-3)}{n(n-1)(n-2)(n-3)} \\
\sigma_{12} &= (|E|(|E|-1)-C)\frac{n_1 n_0(n_1-1)(n_0-1)}{n(n-1)(n-2)(n-3)} - \mu_1\mu_2,
\end{aligned}$$

Where $C = \sum_{i=1}^n d_i^2 - \sum_{i=1}^n d_i$ and d_i is the degree of node i. A proof of these quantities is given in Chen and Friedman (2017) Appendix A.1.

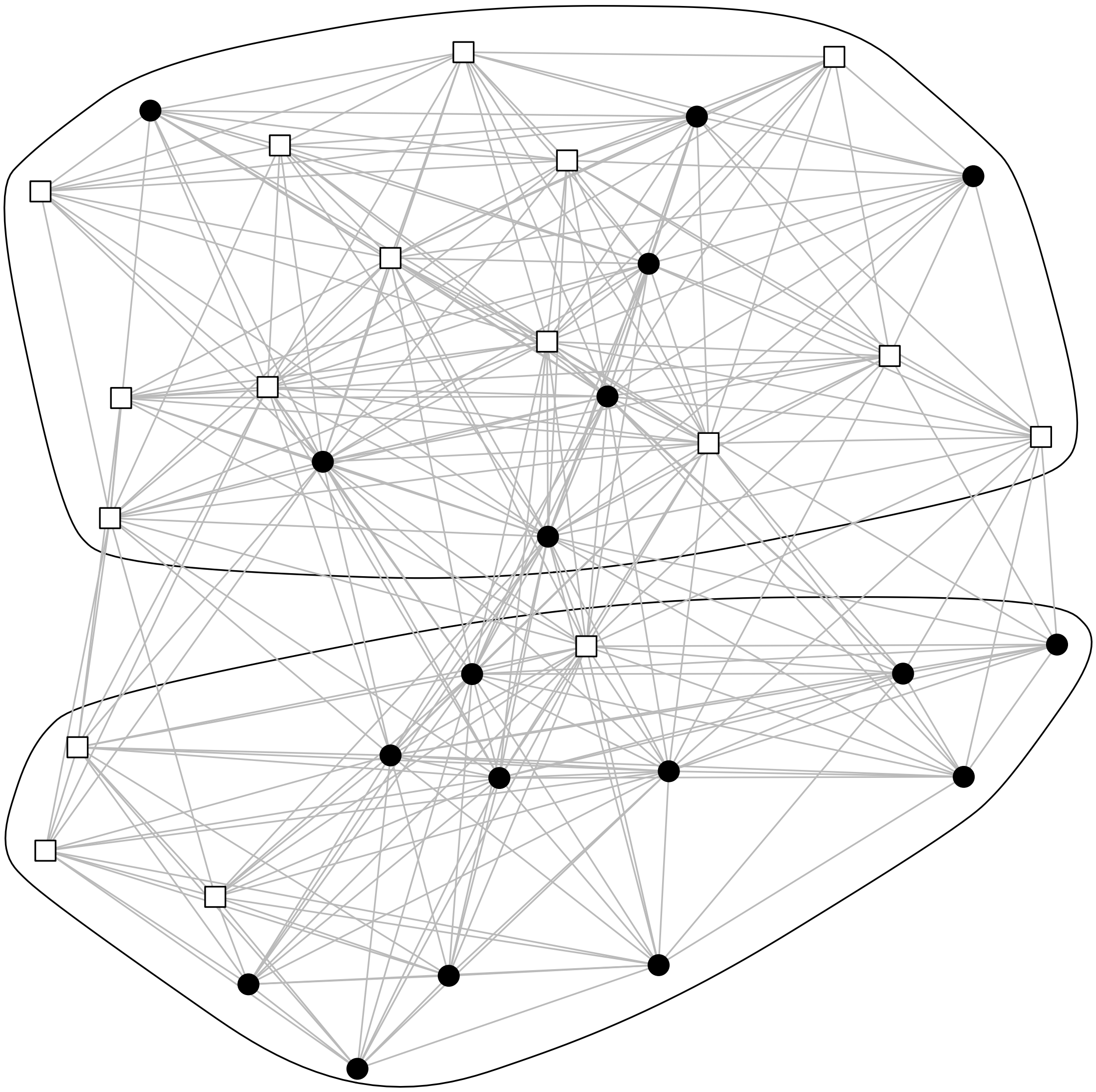


Figure 2.6: Network of selenium and placebo subjects with clusters identified. Black circles are control units, white squares are treated units.

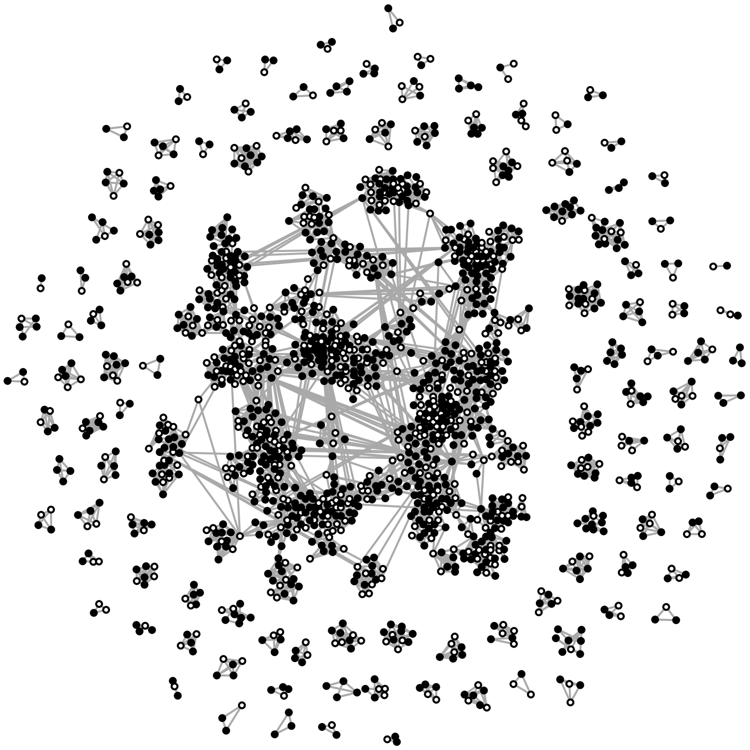


Figure 2.7: Gender co-membership on publicly listed boards in Norway in 2010 and 2011. White nodes are female members and black nodes are male members. Board members share an edge if both members served on the same board at some point during the study period.

Chapter 3

Average treatment effects for cluster randomized trials

3.1 Introduction

Cluster randomization occurs when, for reasons of policy or study implementation, groups of subjects must be randomized to the same treatment condition simultaneously. Cluster randomization is common in many areas of investigation: students within schools, patients within clinics, products within factories, voters within precincts. In each case it may only be feasible or desirable to apply a certain treatment across all subjects within the cluster (a school start time, a patient in-take procedure, a manufacturing technique, a voter mobilization campaign) even though researchers are primarily interested in outcomes at the subject level.

Existing literature on clustered assignment has rightly warned researchers that analyzing unit level outcomes as if each unit had been separately randomized leads to invalid inference (Cornfield, 1978; Donner and Klar, 1994; Schochet, 2013; Middleton and Aronow, 2015). This literature clearly demonstrates that treating cluster randomized trials as if the subjects were the level of randomization nullifies the desirable statistical properties usually conferred by randomization. The downside of this literature, however, is that it has cast the problem of analyzing cluster randomized designs as one of emulating subject level randomization. From this perspective, clustering is a nuisance that would be avoided if possible, not a useful and interesting facet of the design.

This is particularly apparent in discussions of covariance adjustment for cluster randomized trials. In the randomization inference paradigm, covariance adjustment endeavors to remove noise from the outcome while still relying on randomization as the only basis for inference. This approach contrasts with regression modeling approaches in which covariance adjustment is part of model building. The danger in this latter approach is that assumptions necessary for model building are rarely provided by randomization (Freedman,

This chapter contains joint work with Professor Ben Hansen.

2008a,c,b; Samii and Aronow, 2012; Berk et al., 2013; Lin, 2013; Middleton, 2008). In the context of cluster randomized trials, randomization inference approaches to covariance adjustment tend to emphasize clustering in discussing how to construct proper test statistic or sampling distributions, but make little use of the clustering in the actual covariance adjustment itself (Small et al., 2008; Aronow and Middleton, 2013).

What is lost in this discussion is that clusters have a unique covariate not present in subject level designs: cluster size. It is generally difficult to state which covariates are important to adjust in a randomized trial, because the relationship between background variables and potential outcomes is not known. With cluster size, however, the intuition that large clusters will have larger total outcomes in many cases leads immediately to the idea of basing at least some of the covariance adjustment on cluster size.

We are certainly not the first to recognize this feature of cluster randomized designs. From a design perspective, covariance adjustment can be see as an attempt to balance covariates across treatment conditions (Morgan and Rubin, 2012). If treatment assignment is proportional to cluster size, all variance in the number of subjects assigned to each treatment condition will disappear, which forms a type of covariance adjustment. On the analysis side, Middleton and Aronow (2015) included cluster size as the first of possibly several covariates in adjusting cluster outcomes in a randomization inference framework. In this approach, the cluster sizes included in each treatment condition is included in the estimation strategy as a correction to the imbalance.

The method proposed in this chapter falls somewhat between the probability proportional to size and post-hoc covariance adjustment approaches for incorporating cluster size. On one hand, it describes the price paid by allowing variation the number of subjects assigned to treatment, suggesting designs that minimize variation of cluster size. On the other hand, the approach allows for adjusting several different analyses derived from a single design, such as analyzing subgroups that vary in their numbers across clusters such that no single design could possibly assign treatment proportional to each subgroup. Subgroups in clustered assignment are a particularly vexing issue as common design fixes such as stratifying individuals by subgroup is not typically possible when clusters varying in distribution of subgroup members. Methods that post-stratify based on subgroup membership are typically incompatible with clustered assignment (Miratrix et al., 2013). Moreover, designs that randomized at the subject level, but then analyze on subgroups, effectively generate cluster randomized trials where the cluster sizes are either zero or one. Even if researchers only wish to analyze an outcome for the entire study, missing data and non-compliance with treatment assignment often force analysis within subgroups defined by having data or complying with the treatment.

Our method also demonstrates the advantages of fully embracing the key feature of causal inference: that data form two or more samples from different potential outcomes. It has been known since Neyman (1923) that estimation of treatment effects can be described as taking samples from the hypothetical population of all subjects exposed to the treatment condition and, separately, the hypothetical population of all subjects exposed to the control condition. This paradigm allows for applying many results from the survey sampling literature when estimating average treatment effects, defined as the difference in population averages under different treatment conditions. What this approach sometimes fails to highlight is that causal inference is frequently more than just analyzing k-samples. The dependence between samples (i.e. treatment conditions) is more than just a nuisance in computing covariances, it can also be a useful feature that can be profitably exploited. As we show in this chapter, while a natural application of survey sampling methodology leads an efficient, yet biased, estimator, a small adjustment to the estimator, which is only possible in the causal inference context, eliminates the bias when cluster size is roughly proportional to cluster totals.

In Section 3.2, we describe the setting and notation and introduce several estimators for average treatment effects in cluster randomized trials. Section 3.3, we apply our method to a field study that randomized schools to receive a community building intervention. While randomization was applied at the school level, we analyze outcomes for families and students clustered within schools. Finally, in Section 3.4, we conclude with a discussion.

3.2 Methods

3.2.1 Potential outcomes

Consider a cluster randomized trial with c clusters, each with size w_i. Thus the total experimental population of units is $n = \sum_{i=1}^{c} w_i$. If unit level subgroups can be identified, for example particular demographic categories for students clustered in schools, identify the total number of units in subgroup g in a cluster as $w_i^{(g)}$. As special cases, we can think of a completely randomized study of units as one in which each cluster has size 1, in which case all cluster subgroup counts will be either zero or one.

Each cluster is randomly assigned to one of several possible treatments. The vector $\boldsymbol{Z}$ indicates treatment assignment such that, $Z_i = k$ if the ith cluster is assigned to the kth treatment level. Following the potential outcomes framework (Neyman, 1923; Holland, 1986), we posit the existence of a set of outcomes for each subject $y_{ij}(k)$, indexed by treatment assignment. We aggregate potential outcomes by cluster such that $y_i(k) = \sum_{i=1}^{w_i} y_{ij}(k)$. The observed outcome is thus the potential outcome indexed by the realized treatment of cluster i: $Y_i = y_i(Z_i)$. As this notation suggests, we assume that the stable unit treatment value assumption (Rubin, 1980) holds at the cluster level. This assumption states: (a) all clusters assigned to the same treatment receive the same treatment (i.e., $Y_i = y_i(k) \iff Z_i = k, \forall i, k$) and (b) that the assignment of cluster i is not a function of any other unit's assignment (i.e., for the complete assignment vector $\boldsymbol{Z}$, $Y_i = y_i(\boldsymbol{Z}) = y_i(Z_i)$).

In this chapter, we are primarily concerned with unit level average treatment effects (ATEs), defined as the average of the difference of potential outcomes for each unit for under treatments k and l. For the entire population study population, this is defined as

$$\rho_k - \rho_l = \frac{\sum_{i=1}^{c} \sum_{j=1}^{w_i} y_{ij}(k)}{n} - \frac{\sum_{i=1}^{c} \sum_{j=1}^{w_i} y_{ij}(l)}{n} = \frac{\sum_{i=1}^{c} y_i(k)}{\sum_{i=1}^{c} w_i} - \frac{\sum_{i=1}^{c} y_i(l)}{\sum_{i=1}^{c} w_i}$$

In the subsequent sections, we introduce several estimators of $\rho_k - \rho_l$. We begin with the well-known Horvitz-Thompson estimator, which has the desirable quality of being unbiased for the ATE. Then we consider a ratio estimator that uses the observed totals of w_i for treatment levels k and l. This estimator can be unbiased in fairly narrow circumstances, and we bound the amount of bias. Finally, we introduce an estimator that builds on the ratio estimator but remains unbiased in a wider range of situations.

3.2.2 Horvitz-Thompson estimators

A straightforward estimator of the treatment effect of k versus l can be constructed using Horvitz-Thompson (HT) estimators for the total of potential responses divided by the known total of weights. Construct the inverse propensity weighted indicators $K_i = I(Z_i = k) / \mathrm{P}(Z_i = k)$ and $L_i = I(Z_i = l) / \mathrm{P}(Z_i = l)$ in order to define the Horvitz-Thompson estimator of $\rho_k - \rho_l$ as

$$\frac{H_k(\boldsymbol{Y})}{n} - \frac{H_l(\boldsymbol{Y})}{n} = \frac{\sum_{i=1}^{c} K_i Y_i}{n} - \frac{\sum_{i=1}^{c} L_i Y_i}{n} = \frac{\boldsymbol{K}'\boldsymbol{Y}}{n} - \frac{\boldsymbol{L}'\boldsymbol{Y}}{n}$$

Observing that $K_i Y_i = 0$ when i is not assigned to k, we can replace Y_i by $y_i(k)$ to write $K_i Y_i = K_i y_i(k)$. Since $\mathrm{E}(K_i) = \mathrm{E}(Z_i = k) / \mathrm{P}(Z_i = k) = 1$, the Horvitz-Thompson estimator is unbiased for $\rho_k - \rho_l$:

$$\mathrm{E}\left(\frac{H_k(\boldsymbol{Y})}{n} - \frac{H_l(\boldsymbol{Y})}{n}\right) = \frac{\sum_{i=1}^{c} \mathrm{E}(K_i)\, y_i(k)}{n} - \frac{\sum_{i=1}^{c} \mathrm{E}(L_i)\, y_i(l)}{n} = \rho_k - \rho_l$$

The variance of the estimator can be decomposed as

$$\frac{1}{n^2}\left[\mathrm{Var}\left(H_k(\boldsymbol{Y})\right) + \mathrm{Var}\left(H_l(\boldsymbol{Y})\right) - 2\mathrm{Cov}\left(H_k(\boldsymbol{Y}), H_l(\boldsymbol{Y})\right)\right]$$

Let $\pi_i(k) = \mathrm{P}\left(Z_i = k\right)$ and $\pi_{ij}(k, l) = \mathrm{P}\left(Z_i = k, Z_j = l\right)$. When $\pi_{ij}(k, l) > 0$ for all units, standard survey sampling results give unbiased estimators of $H_k(\boldsymbol{Y})$ and $H_l(\boldsymbol{Y})$ as (Cochran, 1999):[1]

$$\hat{V}(H_k(\boldsymbol{Y})) = \sum_{i=1}^{c} K_i \frac{1 - \pi_i(k)}{\pi_i(k)} Y_i^2 + \sum_{i \neq j} K_i K_j \frac{\pi_{ij}(k,k) - \pi_i(k)\pi_j(k)}{\pi_{ij}(k,k)} Y_i Y_j$$
$$\hat{V}(H_l(\boldsymbol{Y})) = \sum_{i=1}^{c} L_i \frac{1 - \pi_i(l)}{\pi_i(l)} Y_i^2 + \sum_{i \neq j} L_i L_j \frac{\pi_{ij}(l,l) - \pi_i(l)\pi_j(l)}{\pi_{ij}(l,l)} Y_i Y_j$$

As the term $\mathrm{Cov}\left(H_k(\boldsymbol{Y}), H_l(\boldsymbol{Y})\right)$ depends on the joint distribution of the potential outcomes, which is never observed for any cluster, unbiased estimation of this term is not possible. It is, however, possible to estimate a quantity that is smaller so that the overall variance estimator is conservative in expectation. Aronow and Samii (2017) provide such a conservative estimator for $\mathrm{Cov}\left(H_k(\boldsymbol{Y}), H_l(\boldsymbol{Y})\right)$ as

$$\hat{C}(H_k(\boldsymbol{Y}), H_l(\boldsymbol{Y})) = \sum_{i \neq j} K_i L_j \frac{\pi_{ij}(k,l) - \pi_i(k)\pi_j(l)}{\pi_{ij}(k,l)} Y_i Y_j - \frac{1}{2} \sum_{i=1}^{n} (K_i + L_i) Y_i^2$$

Putting these estimators together, we get:

$$\begin{aligned} \hat{V}\left(\frac{H_k(\boldsymbol{Y})}{n} - \frac{H_l(\boldsymbol{Y})}{n}\right) &= \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{K_i}{\pi(k)} + \frac{L_i}{\pi_i(l)}\right) Y_i^2 \\ &\quad - \frac{1}{n^2} \sum_{i \neq j} \left(\frac{K_i K_j \pi_i(k)\pi_j(k)}{\pi_{ij}(k,k)} + \frac{L_i L_j \pi_i(l)\pi_j(l)}{\pi_{ij}(l,l)} - 2\frac{K_i L_j \pi_i(k)\pi_j(l)}{\pi_{ij}(k,l)}\right) Y_i Y_j \end{aligned}$$

3.2.3 Hájek estimators

While the HT estimator works well for any distribution of potential outcomes, we could exploit known structure in the potential outcomes to generate a more efficient estimator. For cluster randomized trials, it is often reasonable to make the assumption that that *on average large clusters have large outcomes.* Consider the following reparameterization of the potential outcomes with some treatment level k:

$$y_i(k) = \alpha_k w_i + r_i(k)$$

As we are free to select α_k as any value we please, picking $\alpha_k = \rho_k$ has a useful consequence for the residual terms.

Lemma 3.1. *The sum of residuals $\sum_{i=1}^{c} r_i(k) = 0$ if and only if $\alpha_k = \rho_k$.*

[1] Aronow and Samii (ress) provide conservative estimators for designs for which $\pi_{ij} = 0$ for some i and j (e.g., matched pairs). Aronow and Samii (2017) extend these results for the covariance estimation as well.

Proof. By definition $\rho_k = (\sum_{i=1}^{c} y_i(k))/(\sum_{i=1}^{c} w_i)$. Substituting the decomposition of $y_i(k)$ yields:

$$\rho_k = \frac{\alpha_k \sum_{i=1}^{c} w_i}{\sum_{i=1}^{c} w_i} + \frac{\sum_{i=1}^{c} r_i(k)}{\sum_{i=1}^{c} w_i} = \alpha_k + \frac{\sum_{i=1}^{c} r_i(k)}{\sum_{i=1}^{c} w_i}$$

□

Lemma 3.1 allows writing potential outcomes of any treatment level k as:

$$y_i(k) = \rho_k w_i + r_i(k), \quad \sum_{i=1}^{c} r_i(k) = 0 \tag{3.1}$$

A typical approach to outcomes proportional to size from the sampling literature is to use a ratio estimator. In particular, we could consider a "Hájek estimator" that replaces the known total of cluster sizes n with a Horvitz-Thompson estimator of the total (Hájek, 2011). The result is a difference of two ratios of Horvitz-Thompson estimators:

$$R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) = \frac{H_k(\boldsymbol{Y})}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{Y})}{H_k(\boldsymbol{w})} = \frac{\boldsymbol{K}'\boldsymbol{Y}}{\boldsymbol{K}'\boldsymbol{w}} - \frac{\boldsymbol{L}'\boldsymbol{Y}}{\boldsymbol{L}'\boldsymbol{w}} \tag{3.2}$$

If the potential outcomes were exactly proportional to cluster size (i.e., $y_i(k) = \rho_k w_i$), the ratio estimator is unbiased. For any treatment level k,

$$\mathrm{E}\left(R_k(\boldsymbol{Y})\right) = \mathrm{E}\left(\frac{\sum_{i=1}^{c} K_i(\rho_k w_i)}{\sum_{i=1}^{c} K_i w_i}\right) = \rho_k \mathrm{E}\left(\frac{H_k(\boldsymbol{w})}{H_k(\boldsymbol{w})}\right) = \rho_k$$

so the difference $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$ is also unbiased. This assumption may be difficult to justify in most cases, and the Hájek estimator will exhibit bias of the form:

$$\mathrm{E}\left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) - (\rho_k - \rho_l)\right) = \mathrm{E}\left(\frac{H_k(\boldsymbol{r}(k))}{H_k(\boldsymbol{w})}\right) - \mathrm{E}\left(\frac{H_l(\boldsymbol{r}(l))}{H_l(\boldsymbol{w})}\right)$$

Observe that the bias term depends on both the structure of the potential outcomes and randomization mechanism for $\boldsymbol{K}$ and $\boldsymbol{L}$. Notably, when covariance of $H_k(\boldsymbol{r}(k)$ and $H_l(\boldsymbol{w})$ is the same as the covariance of $H_l(\boldsymbol{r}(l))$ and $H_k(\boldsymbol{w})$, the bias can be described by bound that is the product of the variance of the estimator with a purely design based quantity.

Proposition 3.1. *When $Cov(H_k(\boldsymbol{r}(k)), H_l(\boldsymbol{w})) = Cov(H_l(\boldsymbol{r}(l)), H_k(\boldsymbol{w}))$, the squared bias of the Hájek* *estimator is bounded by*

$$[E(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) - (\rho_k - \rho_l))]^2 \leq Var(C_{k,l})\ Var(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})) \tag{3.3}$$

where

$$C_{k,l} = \frac{H_k(\boldsymbol{w})H_l(\boldsymbol{w})}{E(H_k(\boldsymbol{w})H_l(\boldsymbol{w}))}$$

Proof. By Lemma 3.1 and the unbiasedness of Horvitz-Thompson estimators, $\mathrm{E}\left(H_k(\boldsymbol{r}(k))\right) = \mathrm{E}\left(H_l(\boldsymbol{r}(l))\right) = 0$. Then the assumption $\mathrm{Cov}\left(H_k(\boldsymbol{r}(k)), H_l(\boldsymbol{w})\right) - \mathrm{Cov}\left(H_l(\boldsymbol{r}(l)), H_k(\boldsymbol{w})\right) = 0$ is equivalent to

$$\mathrm{E}\left(H_k(\boldsymbol{r}(k))H_l(\boldsymbol{w})\right) - \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{r}(l))\right) = 0$$

Multiply and divide by $H_k(\boldsymbol{w})H_l(\boldsymbol{w})$ to get

$$\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\left[\frac{H_k(\boldsymbol{r}(k))}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{r}(l))}{H_l(\boldsymbol{w})}\right]\right) = 0$$

Therefore, the covariance of these terms is equal to the negative product of their expectations,

$$\begin{aligned}\mathrm{Cov}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w}), \frac{H_k(\boldsymbol{r}(k))}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{r}(l))}{H_l(\boldsymbol{w})}\right) &= -\mathrm{E}\left(\frac{H_k(\boldsymbol{r}(k))}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{r}(l))}{H_l(\boldsymbol{w})}\right)\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right) \\ &= -\mathrm{E}\left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) - (\rho_k - \rho_l)\right)\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)\end{aligned}$$

Applying the Cauchy-Schwartz inequality yields,

$$[\mathrm{E}\left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) - (\rho_k - \rho_l)\right)]^2 \leq \frac{\mathrm{Var}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)}{[\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)]^2}\mathrm{Var}\left(\frac{H_k(\boldsymbol{r}(k))}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{r}(l))}{H_l(\boldsymbol{w})}\right)$$

As $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) = \rho_k - \rho_l + \frac{H_k(\boldsymbol{r}(k))}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{r}(l))}{H_l(\boldsymbol{w})}$, the last term can be written as $\mathrm{Var}\left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})\right)$. □

This bound shows that the maximum ratio of bias to variance is given by the standard deviation of the variable $C_{k,l}$ or equivalently the coefficient of variation for the quantity $H_k(\boldsymbol{w})H_l(\boldsymbol{w})$. As the design and the vector $\boldsymbol{w}$ is known, the researcher can make decisions regarding the magnitude of bias without having to look at the actual data. Cox and Hinkley (1974, chapter 8) suggest for a ratio of less than 1, bias has little impact an inferences, noting that "an estimate of small bias and small variance will for most purposes be preferable to one with no bias and appreciable variance" (p. 266). Särndal et al. (1992, section 4.2) suggest the more stringent bound 0.1, but allow that "the distorting effect is not extremely pronounced" even if the bias ratio is large as 0.5.

Proposition 3.2. *The variance of the Hájek estimator is given by:*

$$Var(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})) = \frac{1}{n^2} Var(H_k(\boldsymbol{r}(k)) - H_l(\boldsymbol{r}(l)))$$

Proof. We take an estimating equations approach to finding the variance of $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$. Consider estimating the ratio ρ_k from pairs (Y_i, w_i). Taking expectation is taken with respect to the finite population of clusters, since $\mathrm{E}\left(r_i(k)\right) = 0$,

$$\mathrm{E}\left(y_i(k) - \rho_k w_i\right) = \mathrm{E}\left(\psi(y_i(k), w_i, k)\right) = 0$$

Then the finite sample estimating equation (Godambe and Thompson, 1986; Binder and Patak, 1994) for ρ_k is a Horvitz-Thompson estimator of ψ:

$$H_k(\psi(\boldsymbol{Y}, \boldsymbol{w}, \rho_k)) = \sum_{i=1}^{c} K_i\, \psi(Y_i, w_i, \rho_k) = \sum_{i=1}^{c} K_i(Y_i - \rho_k w_i) = \sum_{i=1}^{c} K_i Y_i - \rho_k \sum_{i=1}^{c} K_i w_i$$

Setting the estimating equation equal to zero and solving for ρ_k yields $R_k(\boldsymbol{Y})$. Stacking estimating equations for $\theta = (\rho_k, \rho_l, \rho_k - \rho_l)'$, gives:

$$\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{w}, \theta) = \begin{pmatrix} H_k(\psi(\boldsymbol{Y}, \boldsymbol{w}, \rho_k)) \\ H_l(\psi(\boldsymbol{Y}, \boldsymbol{w}, \rho_l)) \\ \rho_k - \rho_l \end{pmatrix}$$

Basic M-estimation theory states that the variance of the estimating equations can be written in "sandwich" form $A(\theta)^{-1}B(\theta)\left[A(\theta)^{-1}\right]'$ (Stefanski and Boos, 2002), where

$$A(\theta) = \mathrm{E}\left(-\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{w}, \theta)\right)$$
$$B(\theta) = \mathrm{E}\left(\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{w}, \theta)\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{w}, \theta)'\right)$$

For the Hájek estimator $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$ under the working model, these matrices are of the form:

$$A^{-1} = \begin{pmatrix} n^{-1} & 0 & 0 \\ 0 & n^{-1} & 0 \\ n^{-1} & -n^{-1} & 1 \end{pmatrix}, B = \begin{pmatrix} \mathrm{Var}\left(H_k(\boldsymbol{r}(k))\right) & \mathrm{Cov}\left(H_k(\boldsymbol{r}(k)), H_l(\boldsymbol{r}(l))\right) & 0 \\ \mathrm{Cov}\left(H_k(\boldsymbol{r}(k)), H_l(\boldsymbol{r}(l))\right) & \mathrm{Var}\left(H_l(\boldsymbol{r}(l))\right) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

So the variance of the estimator is

$$\begin{aligned} A^{-1}B\left[A^{1}\right]' &= \frac{1}{n^2}\left(\mathrm{Var}\left(H_k(\boldsymbol{r}(k))\right) + \mathrm{Var}\left(H_l(\boldsymbol{r}(l))\right) - 2\mathrm{Cov}\left(H_k(\boldsymbol{r}(k)), H_l(\boldsymbol{r}(l))\right)\right) \\ &= \frac{1}{n^2}\mathrm{Var}\left(H_k(\boldsymbol{r}(k)) - H_l(\boldsymbol{r}(l))\right) \end{aligned}$$

□

To estimate $\frac{1}{n^2}\mathrm{Var}\left(H_k(\boldsymbol{r}(k)) - H_l(\boldsymbol{r}(l))\right)$, observe that this is the variance of a Horvitz-Thompson estimator for the average difference of residuals. While we do not observe the residuals directly, we plug in estimated residuals

$$\hat{r}_i = Y_i - w_i R_{Z_i}(\boldsymbol{Y})$$

into the Horvitz-Thompson variance estimator:

$$\begin{aligned} \hat{V}\left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{K_i}{\pi(k)} + \frac{L_i}{\pi_i(l)}\right)\hat{r}_i^2 \\ &- \frac{1}{n^2}\sum_{i\neq j}\left(\frac{K_iK_j\pi_i(k)\pi_j(k)}{\pi_{ij}(k,k)} + \frac{L_iL_j\pi_i(l)\pi_j(l)}{\pi_{ij}(l,l)} - 2\frac{K_iL_j\pi_i(k)\pi_j(l)}{\pi_{ij}(k,l)}\right)\hat{r}_i\hat{r}_j \end{aligned}$$

3.2.4 Unbiased estimation of proportional outcomes

In many cases the Hájek estimator will have better efficiency than the Horvitz-Thompson estimator, such that if the bias is well controlled the Hájek estimator will still have lower mean squared error than the Horvitz-Thompson estimator. Nevertheless, researchers may still prefer an estimator that is unbiased under a wider set of circumstances. Researchers may be analyzing data for which the possible bias is large. Even when designing a study it may not always be possible to exert sufficient control over the design to make bias negligible, particularly if clusters vary widely in the number of subgroup members. Moreover, if some clusters contain zero subgroup members, the Hájek estimator may not be defined for all possible randomizations.

In the sampling literature, the ratio estimator is often motivated by capturing and smoothing the variance in the numerator by dividing by the denominator. If the numerator and denominator are positively dependent, the resulting ratio should have significantly smaller variance than the numerator alone. If *dividing* by a positively dependent quantity is useful, it would also be beneficial to *multiply* by a negatively dependent quantity. In classical sampling motivations, based on a single sample, such a quantity may be difficult to construct. In the context of causal inference, on the other hand, we can use information about clusters assigned to l to assist inferences about clusters assigned to k. Observe that for any randomization scheme such that $\mathrm{P}\left(Z_i = k, Z_j = l\right) \leq \mathrm{P}\left(Z_i = k\right) \mathrm{P}\left(Z_j = l\right)$ for all i and j , we have that

$$\begin{aligned}
\operatorname{Cov}\left(H_k(\boldsymbol{w}), H_l(\boldsymbol{w})\right) &= \mathrm{E}\left(H_k(\boldsymbol{w}) H_l(\boldsymbol{w})\right) - \left(\sum_{i=1}^{c} w_i\right)^2 \\
&= \sum_{i=1}^{c} \sum_{j=1}^{c} \mathrm{E}\left(K_i L_j\right) w_i w_j - \sum_{i=1}^{c} \sum_{j=1}^{c} w_i w_j \\
&= \sum_{i \neq j} \mathrm{E}\left(K_i L_j\right) w_i w_j - \sum_{i=1}^{c} \sum_{j=1}^{c} w_i w_j \leq 0
\end{aligned}$$

This suggests that the product of $H_k(\boldsymbol{Y})$ and $H_l(\boldsymbol{w})$ may behave similarly to a ratio estimator.

As we are multiplying by an estimator of $\sum_{i=1}^{c} w_i$ instead of dividing by this quantity, we need to divide by something that is roughly $(\sum_{i=1}^{c} w_i)^2$. A natural choice would be $\mathrm{E}\left(H_k(\boldsymbol{w}) H_l(\boldsymbol{w})\right)$:

$$S_{k,l}(\boldsymbol{Y}) = \frac{H_l(\boldsymbol{w}) H_k(\boldsymbol{Y}) - H_k(\boldsymbol{w}) H_l(\boldsymbol{Y})}{\mathrm{E}\left(H_k(\boldsymbol{w}) H_l(\boldsymbol{w})\right)}$$

A simple rewriting of terms shows that $S_{k,l}$ can be thought of a "corrected" version the Hájek estimator:

$$S_{k,l}(\boldsymbol{Y}) = \frac{H_k(\boldsymbol{w}) H_l(\boldsymbol{w})}{\mathrm{E}\left(H_k(\boldsymbol{w}) H_l(\boldsymbol{w})\right)} \left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})\right) = C_{k,l} \left(R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})\right)$$

As an alternative motivation of $S_{k,l}$, observe that we can write $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$ with a common denominator as:

$$R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y}) = \frac{H_k(\boldsymbol{Y})}{H_k(\boldsymbol{w})} - \frac{H_l(\boldsymbol{Y})}{H_l(\boldsymbol{w})} = \frac{H_l(\boldsymbol{w}) H_k(\boldsymbol{Y}) - H_k(\boldsymbol{w}) H_l(\boldsymbol{Y})}{H_k(\boldsymbol{w}) H_l(\boldsymbol{w})}$$

As ratios can be difficult to analyze, a reasonable approximation to $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$ would be to replace the denominator with it's expectation (see Hansen and Bowers, 2008, for a another use of this technique). The resulting estimator is precisely $S_{k,l}(\boldsymbol{Y})$ as defined.

Perhaps unsurprisingly, the term $C_{k,l}$ was also used in Proposition 3.1 in forming a bound on the bias of $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$. As we show below, multiplying $R_k(\boldsymbol{Y}) - R_l(\boldsymbol{Y})$ by $C_{k,l}$ removes the bias from the ratio estimator under similar conditions used in Proposition 3.1.

Proposition 3.3. *Define $\boldsymbol{K}$ and $\boldsymbol{L}$ be propensity scaled treatment indicators such that $K_i = I(Z_i = k)/P(Z_i = k)$ and $L_i = I(Z_i = l)/P(Z_i = l)$. For any design that guarantees*

1. *$E(\boldsymbol{K}\boldsymbol{L}') = E(\boldsymbol{L}\boldsymbol{K}')$*

2. *$Cov(H_k(\boldsymbol{r}(k)), H_l(\boldsymbol{w})) = Cov(H_k(\boldsymbol{w}), H_l(\boldsymbol{r}(l)))$*

the estimator $S_{k,l}(\boldsymbol{Y})$ is unbiased for $\rho_k - \rho_l$.

Proof. Under the assumption $\mathrm{E}\left(\boldsymbol{K}\boldsymbol{L}'\right) = \mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)$, the expectation of the numerator of $S_{k,l}(\boldsymbol{Y})$ is

$$\begin{aligned}
\mathrm{E}\left(H_l(\boldsymbol{w})H_k(\boldsymbol{Y}) - H_k(\boldsymbol{w})H_l(\boldsymbol{Y})\right) &= \mathrm{E}\left(\boldsymbol{w}'\boldsymbol{L}\boldsymbol{K}'\boldsymbol{y}(k) - \boldsymbol{w}'\boldsymbol{K}\boldsymbol{L}'\boldsymbol{y}(l)\right) \\
&= \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)\boldsymbol{y}(k) - \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{K}\boldsymbol{L}'\right)\boldsymbol{y}(l) \\
&= \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)\left(\boldsymbol{y}(k) - \boldsymbol{y}(l)\right) \\
&= \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)\left((\rho_k - \rho_l)\boldsymbol{w} + \boldsymbol{r}(k) - \boldsymbol{r}(l)\right) \\
&= (\rho_k - \rho_l)\boldsymbol{w}'\mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)\boldsymbol{w} + \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)\boldsymbol{r}(k) - \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{K}\boldsymbol{L}'\right)\boldsymbol{r}(l) \\
&= (\rho_k - \rho_l)\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right) + \mathrm{E}\left(H_l(\boldsymbol{w})H_k(\boldsymbol{r}(k))\right) - \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{r}(l))\right)
\end{aligned}$$

Dividing by $\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)$ makes the first term $\rho_k - \rho_l$. By the assumption of equal covariances, $\boldsymbol{w}'\mathrm{E}\left(\boldsymbol{L}\boldsymbol{K}'\right)\boldsymbol{r}(k) - \boldsymbol{w}'\mathrm{E}\left(\boldsymbol{K}\boldsymbol{L}'\right)\boldsymbol{r}(l) = 0$. □

To estimate the variance of $S_{k,l}$, observe that it can be viewed as a Horvitz-Thompson estimator of the following population quantity (scaled by $\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)$):

$$\sum_{(i,j)} \frac{\pi_{ij}(k,l)}{\pi_i(k)\pi_j(l)}\left(w_j y_i(k) - w_i y_j(l)\right)$$

The "population" is the set of pairs of clusters $\{(i,j) : i \neq j, i, j \in \{1, \ldots, c\}\}$ (the sum in the previous equation is defined over this set). The natural Horvitz-Thompson estimator of this quantity is

$$\sum_{(i,j)} \frac{I(Z_i = k)I(Z_j = l)}{\pi_{ij}(k,l)} \frac{\pi_{ij}(k,l)}{\pi_i(k)\pi_j(l)}\left(w_j Y_i - w_i Y_j\right) = \sum_{(i,j)} K_i L_j \left(w_j Y_i - w_i Y_j\right) = H_l(\boldsymbol{w})H_k(\boldsymbol{Y}) - H_k(\boldsymbol{w})H_l(\boldsymbol{Y}).$$

Scaling by $\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)$ yields $S_{k,l}$.

As Horvitz-Thompson type estimator, the variance of the estimator takes can be found using standard equations.

$$
\begin{aligned}
\operatorname{Var}\left(S_{k,l}\right) = \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)^{-2} \times \\
\Bigg\{ & \sum_{(i,j)} \frac{\pi_{ij}(k,l)(1-\pi_{ij}(k,l))}{\pi_i(k)^2\pi_i(l)^2}(w_j y_i(k) - w_i y_j(l))^2 \\
& + \sum_{(i,j,f,g)} \frac{\pi_{ijfg}(k,l,k,l) - \pi_{ij}(k,l)\pi_{fg}(k,l)}{\pi_i(k)\pi_j(l)\pi_f(k)\pi_g(l)}(w_j y_i(k) - w_i y_j(l))(w_g y_f(k) - w_f y_g(l)) \\
& + \sum_{(i,j,f)} \frac{\pi_{ijf}(k,l,l) - \pi_{ij}(k,l)\pi_{if}(k,l)}{\pi_i(k)^2\pi_j(l)\pi_f(l)}(w_j y_i(k) - w_i y_j(l))(w_f y_i(k) - w_i y_f(l)) \\
& + \sum_{(i,j,f)} \frac{\pi_{ijf}(k,l,k) - \pi_{ij}(k,l)\pi_{fj}(k,l)}{\pi_i(k)\pi_j(l)^2\pi_f(k)}(w_j y_i(k) - w_i y_j(l))(w_j y_f(k) - w_f y_j(l)) \\
& - \sum_{(i,j,f)} \frac{\pi_{ij}(k,l)\pi_{fi}(k,l)}{\pi_i(k)\pi_i(l)\pi_j(l)\pi_f(k)}(w_j y_i(k) - w_i y_j(l))(w_i y_f(k) - w_f y_i(l)) \\
& - \sum_{(i,j,f)} \frac{\pi_{ij}(k,l)\pi_{jf}(k,l)}{\pi_i(k)\pi_j(l)\pi_j(k)\pi_f(l)}(w_j y_i(k) - w_i y_j(l))(w_f y_j(k) - w_j y_f(l)) \\
& - \sum_{(i,j)} \frac{\pi_{ij}(k,l)\pi_{ij}(l,k)}{\pi_i(k)\pi_i(l)\pi_j(k)\pi_j(l)}(w_j y_i(k) - w_i y_j(l))(w_i y_j(k) - w_j y_i(l)) \quad \Bigg\}
\end{aligned}
$$

Here again, the notation (i,j), (i,j,f), and (i,j,f,g) denotes all 2-, 3-, and 4-tuples with distinct components. We provide simplified versions for complete randomization and blocked randomization in a subsequent section.

To estimate this variance we make a simplifying assumption to impute the unobserved potential outcomes, for example $y_i(l)$ for subjects in the k treatment level. We assume that at the cluster level, the residual terms $r_i(k)$ and $r_i(l)$ are equal. Under this assumption, we can impute the missing potential outcome as $y_i(l) = y_i(k) - w_i S_{k,l}$ or $y_i(k) = y_i(l) + w_i S_{k,l}$. Under this assumption, we can calculate the variance directly and use it as an estimate of the true variance of $S_{k,l}$.

3.2.5 Subgroups and non-compliance

Using the cluster total notation, $y_i(k) = \sum_i^c y_{ij}(k)$, the unit level ATE can be defined in terms of cluster level totals. In addition to the ATE for the entire experimental population, we are often interested in average treatment effects for subsets of the populations which we call "subgroups." Let $g_{ij} = 1$ if a subject is in subgroup g and $g_{ij} = 0$ otherwise. Let $n^{(g)}$ be the total number of subgroup members. Then subgroup specific effect for g is

$$\rho_k^{(g)} - \rho_l^{(g)} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{w_i} g_{ij} y_{ij}(k)}{n^{(g)}} - \frac{\sum_{i=1}^{c}\sum_{j=1}^{w_i} g_{ij} y_{ij}(l)}{n^{(g)}} = \frac{\sum_{i=1}^{c} y_i^{(g)}(k)}{\sum_{i=1}^{c} w_i^{(g)}} - \frac{\sum_{i=1}^{c} y_i^{(g)}(l)}{\sum_{i=1}^{c} w_i^{(g)}}$$

The general case presented so far can be thought of the subgroup in which $g_{ij} = 1$ for all subjects.

The HT and Hájek estimators can suffer from complications in the presence of subgroups. For the HT estimator when $w_i^{(g)} = 1$ and for some designs, such as the complete random assignment discussed in the section, it is the case that the HT estimator is equivalent to the difference of means of the treatment and control groups. For example, for complete random assignment (discussed in more detail in the next section), $n = \sum_{i=1}^{c} 1 = c$, $\pi_i(k) = c_k/c$ and $pi_i(l) = c_l/c$ and

$$\frac{H_k(\boldsymbol{Y})}{c} - \frac{H_l(\boldsymbol{Y})}{c} = \frac{c}{c_k}\frac{\sum_{i=1}^{c} I(Z_i = k)Y_i}{c} - \frac{c}{c_l}\frac{\sum_{i=1}^{c} I(Z_i = l)Y_i}{c} = \frac{\sum_{i=1}^{c} I(Z_i = k)Y_i}{c_k} - \frac{\sum_{i=1}^{c} I(Z_i = l)Y_i}{c_l}$$

Notably, in this expression the number of subjects in the treatment group $\sum_{i=1}^{c} I(Z_i = k)w_i$ is a constant for all $\boldsymbol{Z}$. When the $w_i^{(g)}$ are not constant, however, such as when analyzing a subgroup, while $n^{(g)}$ is still a constant, it is no longer the case that $\sum_{i=1} I(Z = k)w_i$ is a constant. Dividing by the observed number of treated or control results in a ratio of random variables instead of a random variable divided by a constant. This is easily addressed by dividing by $n(^g)$ rather than the observed number, but it is an common mistake to make.

The Hájek estimator suffers as well when $w_i^{(g)} = 0$ for a sufficient number of clusters. In that case the denominator $\sum_{i=1}^{c} K_i w_i^{(g)}$ can be equal to zero for some randomizations. In those cases, the Hájek estimator is not defined. Work-arounds for this issue require conditioning on the observed $w_i^{(g)}$ in the treatment levels (Morgan and Rubin, 2012), but can be avoided by choosing a more robust estimator such as $S_{k,l}$.

3.2.6 Completely randomized and stratified designs

Having relied on the assumptions that $\mathrm{E}(\boldsymbol{K}\boldsymbol{L}') = \mathrm{E}(\boldsymbol{L}\boldsymbol{K}')$ and $\mathrm{E}(H_k(\boldsymbol{r}(k))H_l(\boldsymbol{w})) = \mathrm{E}(H_k(\boldsymbol{w})H_l(\boldsymbol{r}(l)))$, it is worth asking if these properties hold under any useful designs and what additional assumptions are necessary on the potential outcomes. Conveniently, these assumptions are met under stratified designs when the residuals r_i and cluster totals w_i are correlated by the same amount for all potential outcomes.

Proposition 3.4. *For the potential outcomes to k and l, define $r_i(k) = y_i(k) - \rho_k w_i$ and $r_i(l) = y_i(l) - \rho_l w_i$. Suppose the c clusters are partitioned into b strata such that within each stratum s, there are c_s clusters, with $c_{s,k}$ assigned to k and $c_{s,l}$ to l. Furthermore, suppose that either:*

1. *All stratum sizes are the same and*

$$\sum_{i=1}^{c} w_i r_i(k) = \sum_{i=1}^{c} w_i r_i(l)$$

2. *Stratum sizes vary, but for all strata s:*

$$\sum_{i=1}^{c_s} w_{si} r_{si}(k) = \sum_{i=1}^{c_s} w_{si} r_{si}(l)$$

   and

$$\sum_{i=1}^{c_s} r_{si}(k) = \sum_{i=1}^{c_s} r_{si}(l) = 0$$

Then the properties $E(\boldsymbol{KL}') = E(\boldsymbol{LK}')$ and $E(H_k(\boldsymbol{r}(k))H_l(\boldsymbol{w})) = E(H_k(\boldsymbol{w})H_l(\boldsymbol{r}(l)))$ hold.

Proof. When $i \neq j$ are in the same block s,

$$\mathrm{E}\left(K_i L_j\right) = \frac{\mathrm{P}\left(Z_i = k, Z_j = l\right)}{\mathrm{P}\left(Z_i = k\right)\mathrm{P}\left(Z_j = k\right)} = \frac{c_{s,k}c_{s,l}}{c_s(c_s-1)}\frac{c_s^2}{c_{s,k}c_{s,l}} = \frac{c_s}{c_s-1} = \mathrm{E}\left(K_j L_i\right)$$

When i and j are in separate blocks, treatment assignment is entirely independent

$$\mathrm{E}\left(K_i L_j\right) = \frac{\mathrm{P}\left(Z_i = k, Z_j = l\right)}{\mathrm{P}\left(Z_i = k\right)\mathrm{P}\left(Z_j = k\right)} = \frac{\mathrm{P}\left(Z_i = k\right)\mathrm{P}\left(Z_j = l\right)}{\mathrm{P}\left(Z_i = k\right)\mathrm{P}\left(Z_j = k\right)} = 1 = \mathrm{E}\left(K_j L_i\right)$$

Therefore $\mathrm{E}\left(\boldsymbol{KL}'\right) = \mathrm{E}\left((\boldsymbol{KL}')'\right) = \mathrm{E}\left(\boldsymbol{LK}'\right)$.

Let $\boldsymbol{w}_s$ be the sum of w_i for clusters in stratum s. Since $\mathrm{E}\left(K_i L_i\right) = 0$ and $\mathrm{E}\left(K_i L_j\right) = \mathrm{E}\left(L_i K_j\right) = 1$ for

i and j in different blocks, we have

$$\begin{aligned}
&\mathrm{E}\left(H_l(\boldsymbol{w})H_k(\boldsymbol{r}(k))\right) - \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right) \\
&= \sum_{s=1}^{b} \boldsymbol{w}_s' \mathrm{E}\left(\boldsymbol{L}_s \boldsymbol{K}_s'\right) \boldsymbol{r}_s(k) - \boldsymbol{w}_s' \mathrm{E}\left(\boldsymbol{K}_s \boldsymbol{L}_s'\right) \boldsymbol{r}_s(l) + \sum_{s \neq t} \boldsymbol{w}_s' \mathrm{E}\left(\boldsymbol{L}_s \boldsymbol{K}_t'\right) \boldsymbol{r}_t(k) - \boldsymbol{w}_s' \mathrm{E}\left(\boldsymbol{K}_s \boldsymbol{L}_t'\right) \boldsymbol{r}_t(l) \\
&= \sum_{s}^{b} \frac{c_s}{c_s - 1} \sum_{i \neq j}^{c_s} \left(w_i r_j(k) - w_i r_j(l)\right) + \sum_{s \neq t} \boldsymbol{w}_s' \boldsymbol{r}_t(k) - \boldsymbol{w}_s' \boldsymbol{r}_t(l) \\
&= \sum_{s}^{b} \frac{c_s}{c_s - 1} \sum_{i \neq j}^{c_s} \left(w_i r_j(k) - w_i r_j(l)\right) + \sum_{s \neq t} \boldsymbol{w}_s' \boldsymbol{r}_t(k) - \boldsymbol{w}_s' \boldsymbol{r}_t(l) + \boldsymbol{w}' \boldsymbol{r}(k) - \boldsymbol{w}' \boldsymbol{r}(l) \\
&= \sum_{s}^{b} \frac{c_s}{c_s - 1} \sum_{i \neq j}^{c_s} \left(w_i r_j(k) - w_i r_j(l)\right) + \sum_{s=1}^{b} \sum_{t=1}^{b} \boldsymbol{w}_s' \boldsymbol{r}_t(k) - \boldsymbol{w}_s' \boldsymbol{r}_t(l) \\
&= \sum_{s}^{b} \frac{c_s}{c_s - 1} \sum_{i \neq j}^{c_s} \left(w_i r_j(k) - w_i r_j(l)\right) + \left(\sum_{i=1}^{c} w_i\right)^2 \left(\sum_{i=1}^{c} r_i(k)\right)^2 + \left(\sum_{i=1}^{c} w_i\right)^2 \left(\sum_{i=1}^{c} r_i(l)\right)^2 \\
&= \sum_{s}^{b} \frac{c_s}{c_s - 1} \sum_{i \neq j}^{c_s} \left(w_i r_j(k) - w_i r_j(l)\right)
\end{aligned}$$

Since $\sum_{i=1}^{c} r_i(k) = \sum_{i=1}^{c} r_i(l) = 0$ by Lemma 3.1.

When all strata are the same size,

$$\begin{aligned}
&\mathrm{E}\left(H_l(\boldsymbol{w})H_k(\boldsymbol{r}(k))\right) - \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right) \\
&= \frac{c_s}{c_s - 1} \sum_{s}^{b} \sum_{i \neq j}^{c_s} w_i r_j(k) - w_i r_j(l) \\
&= \frac{c_s}{c_s - 1} \sum_{s}^{b} \sum_{i \neq j}^{c_s} w_i r_j(k) - w_i r_j(l) + \frac{c_s}{c_s - 1} \sum_{s}^{b} \sum_{i}^{c_s} w_i r_i(k) - w_i r_i(l) \\
&= \frac{c_s}{c_s - 1} \left[\left(\sum_{i=1}^{c} w_i\right)^2 \left(\sum_{i=1}^{c} r_i(k)\right)^2 + \left(\sum_{i=1}^{c} w_i\right)^2 \left(\sum_{i=1}^{c} r_i(l)\right)^2 \right] \\
&= 0
\end{aligned}$$

Alternatively, if within blocks $\boldsymbol{w}_s'\boldsymbol{r}_s(k) - \boldsymbol{w}_s'\boldsymbol{r}_s(l) = 0$ and $\mathbf{1}_s'\boldsymbol{r}_s(k) = \mathbf{1}_s'\boldsymbol{r}_s(l) = 0$,

$$\begin{aligned}
&\mathrm{E}\left(H_l(\boldsymbol{w})H_k(\boldsymbol{r}(k))\right) - \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right) \\
&= \sum_s^b \frac{c_s}{c_s - 1} \sum_{i \neq j}^{c_s} w_i r_j(k) - w_i r_j(l) \\
&= \sum_s^b \frac{c_s}{c_s - 1} \left(\sum_{i \neq j}^{c_s} w_i r_j(k) - w_i r_j(l) + \sum_{i=1}^{c_s} w_i r_i(k) - w_i r_i(l) \right) \\
&= \sum_s^b \frac{c_s}{c_s - 1} \left[\left(\sum_{i=1}^{c_s} w_i \right)^2 \left(\sum_{i=1}^{c_s} r_i(l) \right)^2 - \left(\sum_{i=1}^{c_s} w_i \right)^2 \left(\sum_{i=1}^{c_s} r_i(l) \right)^2 \right] \\
&= 0
\end{aligned}$$

In either case, $\mathrm{E}\left(H_k(\boldsymbol{r}(k))H_l(\boldsymbol{w})\right) = \mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{r}(l))\right)$ holds. □

Simplified Estimators

For a completely randomized design with c_k clusters assigned to k and c_l clusters assigned to l, many of estimators simplify. Interestingly, $\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right)$ does not depend on k or l:

$$\mathrm{E}\left(H_k(\boldsymbol{w})H_l(\boldsymbol{w})\right) = \frac{c}{c-1} \left[\left(\sum_{i=1}^{c} w_i \right)^2 - \sum_{i=1}^{c} w_i^2 \right]$$

Many of these simplifications are easier to express as sums within the treated and control groups. Define the functions $\mathcal{K}(w^a Y^b) = \sum_{i=1}^c I(Z_i = k) w_i^a Y_i^b$ and $\mathcal{L}(w^a Y^b) = \sum_{i=1}^c I(Z_i = l) w_i^a Y_i^b$.

$$S_{k,l} = \frac{c(c-1)}{c_k c_l} \frac{\mathcal{L}(w)\mathcal{K}(Y) - \mathcal{K}(w)\mathcal{L}(Y)}{\left(\sum_{i=1}^c w_i\right)^2 - \sum_{i=1}^c w_i^2}$$

The variance similarly simplifies to:

$$\begin{aligned}
\mathrm{Var}\left(S_{k,l}\right) = \left[\left(\sum_i^c w_i\right)^2 - \sum_i^c w_i^2\right]^{-2} \times \Bigg\{ & \left(\frac{c(c-1) - c_k c_l}{c^2 c_k c_l}\right) \sum_{(i,j)} (w_j y_i(k) - w_i y_j(l))^2 \\
& + \left(\frac{c(c-1)(c_k-1)(c_l-1)}{c_k c_l (c-2)(c-3)} - 1\right) \sum_{(i,j,f,g)} (w_j y_i(k) - w_i y_j(l))(w_g y_f(k) - w_f y_g(l)) \\
& + \left(\frac{c(c-1)(c_l-1)}{c_k c_l (c-2)} - 1\right) \sum_{(i,j,f)} (w_j y_i(k) - w_i y_j(l))(w_f y_i(k) - w_i y_f(l)) \\
& + \left(\frac{c(c-1)(c_k-1)}{c_k c_l (c-2)} - 1\right) \sum_{(i,j,f)} (w_j y_i(k) - w_i y_j(l))(w_j y_f(k) - w_f y_j(l)) \\
& - \sum_{(i,j,f)} (w_j y_i(k) - w_i y_j(l))(w_i y_f(k) - w_f y_i(l)) \\
& - \sum_{(i,j,f)} (w_j y_i(k) - w_i y_j(l))(w_f y_j(k) - w_j y_f(l)) \\
& - \sum_{(i,j)} (w_j y_i(k) - w_i y_j(l))(w_i y_j(k) - w_j y_i(l)) \Bigg\}
\end{aligned}$$

3.3 FAST social capital experiment

Gamoran et al. (2012) describes an experiment conducted in schools in Texas and Arizona designed to increase the social capital of first grade students and their families. Participating schools in three districts in Phoenix, AZ and one district in San Antonio, TX were randomly assigned to provide a series of community building exercises using the Families and Schools Together (FAST) curriculum in which families shared meals and played games. Control schools recruited participant families, but provided no additional community engagement. Relevant outcomes included measures of social capital for the families, student classroom behavior, and student academic achievement. Outcomes of Hispanic families were particularly important to the researchers, as these families, many of whom were immigrants, were traditionally less engaged with the school community.

	City	Schools	Min Size	Max Size	Avg Size	SD Size
1	Phoenix	6	14	93	69.00	29.37
2	Phoenix	8	43	79	60.62	14.68
3	Phoenix	12	20	63	48.17	12.60
4	San Antonio	12	34	90	58.58	17.41
5	San Antonio	14	44	94	64.57	15.46

Table 3.1: Within-block distribution of cluster sizes for the FAST study.

There were 52 schools recruited to the study. For the randomization, schools were blocked within city. In

Phoenix, schools were additionally blocked within one of three districts, while in San Antonio two blocks were formed by the researchers to match the schools based on percentage of “free or reduced lunch” qualifying students. Table 3.1 shows the number of schools per block, along with the average and standard deviation of the number of students per school within the block. As we can see, the variation on cluster size, particularly in the first Phoenix district, is quite large.

	English Proficient	English Language Learner
Native American	0.0123	0.0006
Asian or Pacific Islander	0.0117	0.0036
Black	0.0743	0.0019
Hispanic	0.4912	0.2416
White	0.1294	0.0036

Table 3.2: Proportion of study particiapants categorized by school district reported race/ethnicity and English language learner status.

Overall, there were 3084 students enrolled in the study, with 1592 in the community building schools and 1492 in regular practice schools. Table 3.2 shows the proportion of students categorized by the district’s reported ethnicity and whether students were considered “English language learners” (i.e., students with difficulties learning in an English environment) when entering the first grade. Researchers targeted districts with large Hispanic populations, and it is no surprise that Hispanic students make up nearly three quarters of the study population. Hispanic students also make up nearly all of the students classified English language learners.

Gamoran et al. (2012) analyzed outcomes using cluster aggregated values, which address how the average response of a school is affected by the treatment, but does not address the student and family level treatment effects as the school sizes vary a great deal (Schochet, 2013). Turley et al. (2017) added longitudinal data for the study and provided analysis using a hierarchical linear model (Raudenbush, 1997). We recreate two of the analysis performed in Turley et al. (2017) comparing treated and control families and students on parents’ self-reported social capital and an index of teach reported pro-social behavior by students. Additionally, we create an index of academic performance combining teacher reported abilities in the first year of the treatment and performance on state-wide math and reading exams in the third year following treatment. Figure 3.1 shows the distributions of these three outcomes broken down by treatment and control groups.

All three distributions exhibit some amount of missing data. In the parents’ social capital measure, only approximately 60% of parents completed surveys at the end of the first year of the study and not all parents answered all questions. The pro-social behavior measures came from teacher surveys, which had much higher completion rates. The academic outcomes blend teacher reports with end of year tests in the third year of the study. Children who moved out of the districts or otherwise did not complete year end tests with

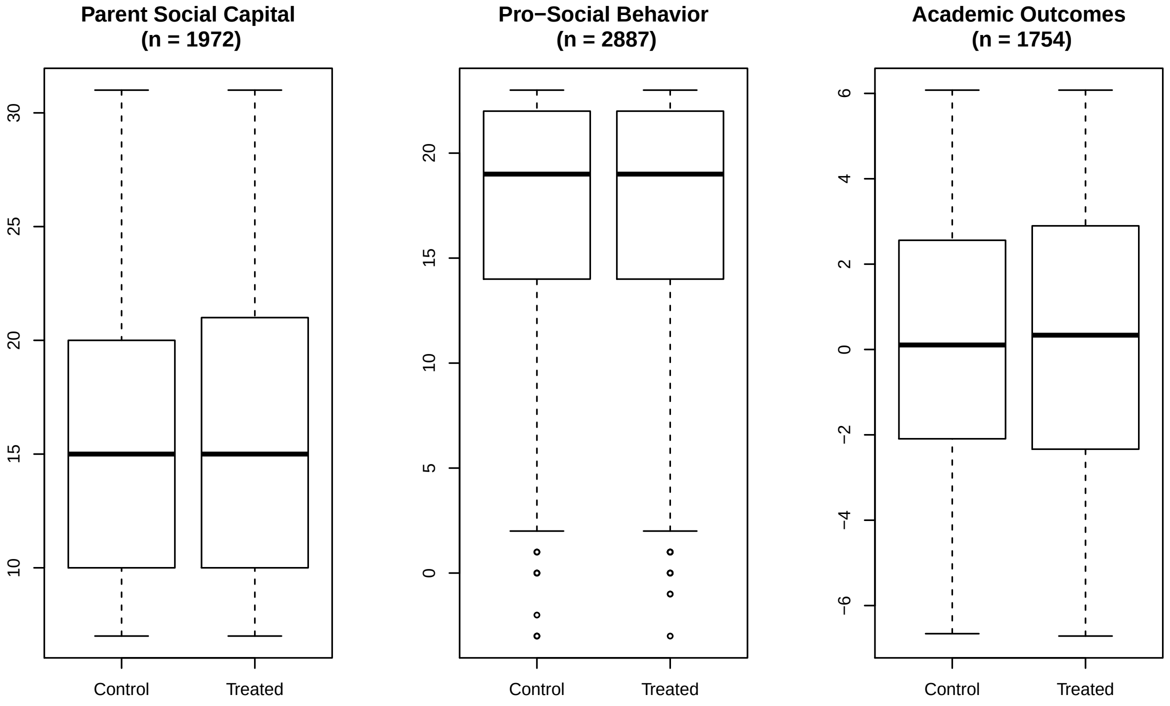


Figure 3.1: Box plots comparing treated and control families and students on composite measure of parental social capital, teacher reported student pro-social behavior, and academic achievement one and three years after treatment.

their cohort were thus excluded from these measures. To address missingness concerns, we impute a naïve estimate of ρ_k and ρ_l using the within treated and control group means.

Table 3.3 includes the overall average treatment effect estimates for the social capital, pro-social behavior, and academic outcomes. For each of the outcomes, we apply the Horvitz-Thompson estimator, the Hájek ratio estimator, and the corrected ratio estiamtor $S_{k,l}$ to compare the treated and control subjects. In addition the ATE estimates, the table also includes variance estimates and 95% confidence intervals. Overall, the Hájek and corrected estimators were consistently close, while the Horvitz-Thompson estimator varied from the other two more. Including w_i via the Hájek estimator or $S_{k,l}$ significantly improved precision, leading to much smaller confidence intervals. Nevertheless, none of the 95% confidence intervals exclude zero so that these results are compatible with the treatment having no effect on any of these outcomes. Turley et al. (2017) reported similar findings for the overall groups, though they noted that many families were assigned to the FAST treatment but never attended any classes. In an analysis of the "treatment-on-the-treated" average effect, modeling compliance for the control group, they found significant results for subjects that did imply with the treatment. Such an analysis is outside the scope of this chapter, but could

Outcome	Estimator	$\hat{\rho}_1 - \hat{\rho}_0$	$\hat{V}$	Lower 95%	Upper 95%
Social Capital	Horvitz-Thompson	1.4138	1.9290	-1.3095	4.1370
	Hájek	0.3968	0.0814	-0.1626	0.9561
	$S_{k,l}$	0.3971	0.0539	-0.0582	0.8523
Pro-Social Behavior	Horvitz-Thompson	1.0890	2.2406	-1.8459	4.0239
	Hájek	-0.0335	0.1190	-0.7098	0.6427
	$S_{k,l}$	-0.0336	0.0818	-0.5944	0.5273
Academic Outcomes	Horvitz-Thompson	0.1921	0.0296	-0.1453	0.5294
	Hájek	0.1803	0.0288	-0.1524	0.5129
	$S_{k,l}$	0.1804	0.0198	-0.0952	0.4560

Table 3.3: Results of analyzing three different outcomes from the Gamoran et al. (2012) social captial study. For each outcome, we apply the Horvitz-Thompson, Hájek, and bias corrected estimator $S_{k,l}$. The table shows the estimate, the estimated variance, and the bounds of a 95% confidence interval.

be incorporated by defining subgroups based on compliance status, with the class membership being defined by a model for the control subjects.

Estimator	Subgroup	$\rho_k - \rho_l$	$\hat{V}$	Lower 95%	Upper 95%
Horvitz-Thompson	Native American	4.49	28.54	-5.99	14.97
Horvitz-Thompson	Asian or Pacific Islander	11.29	74.77	-5.66	28.25
Horvitz-Thompson	Black	0.60	18.42	-7.82	9.01
Horvitz-Thompson	Hispanic	0.93	4.16	-3.07	4.93
Horvitz-Thompson	White	3.53	34.43	-7.98	15.03
Hájek	Native American	0.05	0.01	-0.09	0.19
Hájek	Asian or Pacific Islander	0.16	0.02	-0.10	0.42
Hájek	Black	-0.03	0.10	-0.65	0.60
Hájek	Hispanic	-0.07	0.97	-2.00	1.86
Hájek	White	0.33	0.62	-1.21	1.88
$S_{k,l}$	Native American	0.05	0.00	-0.07	0.17
$S_{k,l}$	Asian or Pacific Islander	0.16	0.01	-0.05	0.37
$S_{k,l}$	Black	-0.03	0.07	-0.54	0.49
$S_{k,l}$	Hispanic	-0.07	0.71	-1.72	1.58
$S_{k,l}$	White	0.34	0.42	-0.94	1.61

Table 3.4: Subgroup specific effects for parent social capital outcomes.

Similar results appear in the subgroup specific analyses. Subgroup specific effects for the three outcomes are given in Table 3.4, Table 3.5, and Table 3.6. For none of the groups did the treatment appear to be significant. While it is not directly tested, given the overlap in confidence intervals, it seems that treatment did not improve average outcomes any of the subgroups more than any of the others.

3.4 Discussion

In this chapter we have considered the issue of estimating average treatment effects in cluster randomized studies. Like some previous approaches, we argue that cluster size (w_i) is a critical variable in understanding

Estimator	Subgroup	$\rho_k - \rho_l$	$\hat{V}$	Lower 95%	Upper 95%
Horvitz-Thompson	Native American	8.41	41.48	-4.21	21.04
Horvitz-Thompson	Asian or Pacific Islander	10.73	95.02	-8.38	29.84
Horvitz-Thompson	Black	-0.12	27.54	-10.41	10.17
Horvitz-Thompson	Hispanic	0.66	4.64	-3.56	4.89
Horvitz-Thompson	White	2.97	41.98	-9.74	15.67
Hájek	Native American	0.10	0.01	-0.07	0.28
Hájek	Asian or Pacific Islander	0.15	0.02	-0.15	0.44
Hájek	Black	-0.09	0.15	-0.86	0.68
Hájek	Hispanic	-0.35	0.98	-2.29	1.60
Hájek	White	0.25	0.77	-1.46	1.97
$S_{k,l}$	Native American	0.10	0.01	-0.04	0.25
$S_{k,l}$	Asian or Pacific Islander	0.15	0.02	-0.10	0.39
$S_{k,l}$	Black	-0.09	0.11	-0.73	0.55
$S_{k,l}$	Hispanic	-0.35	0.73	-2.02	1.33
$S_{k,l}$	White	0.25	0.52	-1.17	1.67

Table 3.5: Subgroup specific effects for student prosocial behavior outcomes.

the operating characteristics of estimators of average treatment effects in cluster randomized trials. The standard Horvitz-Thompson estimator does not directly include the variation in w_i in the estimator, though this variation directly contributes to the variance of the estimator. Ratio estimators such as the Hájek estimator include the observed variation in w_i across the treatment conditions in the estimator but are not unbiased expected in narrow circumstances.

Under fairly broad assumptions, the bias in the Hájek estimator can be bounded in proportion to an estimable variance term. Moreover, a term in constant of proportionality in the bound on the bias can be used as a multiplicative term to develop an estimator that is unbiased under the same broad assumptions used in the derivation of the bound. We showed how this new, corrected estimator $S_{k,l}$ can be interpreted of as scaled version of the Hájek estimator, an approximation to the Hájek estimator with the denominator held at its expectation, or as a Horvitz-Thompson estimator of a particular population quantity. Using this last representation, we a derived variance estimator for $S_{k,l}$. We also showed that several common designs meet our estimator's requirements and provided simplified versions of $S_{k,l}$ and its variance estimator for those designs.

Estimator	Subgroup	$\rho_k - \rho_l$	$\hat{V}$	Lower 95%	Upper 95%
Horvitz-Thompson	Native American	0.30	0.41	-0.96	1.55
Horvitz-Thompson	Asian or Pacific Islander	0.35	0.99	-1.61	2.30
Horvitz-Thompson	Black	-0.26	0.12	-0.95	0.43
Horvitz-Thompson	Hispanic	0.17	0.03	-0.14	0.49
Horvitz-Thompson	White	0.51	0.30	-0.56	1.59
Hájek	Native American	0.00	0.00	-0.01	0.02
Hájek	Asian or Pacific Islander	0.00	0.00	-0.03	0.03
Hájek	Black	-0.02	0.00	-0.07	0.03
Hájek	Hispanic	0.12	0.01	-0.11	0.35
Hájek	White	0.06	0.01	-0.08	0.21
$S_{k,l}$	Native American	0.00	0.00	-0.01	0.02
$S_{k,l}$	Asian or Pacific Islander	0.00	0.00	-0.02	0.03
$S_{k,l}$	Black	-0.02	0.00	-0.06	0.02
$S_{k,l}$	Hispanic	0.12	0.01	-0.07	0.32
$S_{k,l}$	White	0.06	0.00	-0.06	0.18

Table 3.6: Subgroup specific effects for student academic outcomes.

Bibliography

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Agresti, A. (2013). *Categorical Data Analysis*. John, third edition edition.

Amelio, A. and Pizzuti, C. (2014). *Overlapping Community Discovery Methods: A Survey*, pages 105–125. Springer Vienna, Vienna.

Aronow, P. M. and Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135–154.

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. Working paper.

Aronow, P. M. and Samii, C. (In press). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology*.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.

Baker, F. B. and Hubert, L. J. (1981). The analysis of social interaction data: A nonparametric technique. *Sociological Methods & Research*, 9(3):339–361.

Bedi, P. and Sharma, C. (2016). Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):115–135.

Berger, R. L. and Boos, D. B. (1994). P-values maximized over a confidence for the nuisance parameter. 89(427):1012 – 1016.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37(3-4):170–196.

Berk, R. A. (2004). *Regression Analysis: A constructive criticism*. Sage Publications, Inc., Thousand Oaks, CA.

Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89(427):1035–1043.

Biswas, M., Mukhopadhyay, M., and Ghosh, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101(4):913–926.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113–120.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555 – 564.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1):55 – 71.

Borgatti, S. P. and Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466 – 484.

Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97 – 124.

Casella, G. and Strawderman, W. E. (1981). Estimating a bounded normal mean. *The Annals of Statistics*, 9(4):870–878.

Caughey, D., Dafoe, A., and Miratix, L. (2017). Beyond the sharp null: Permutation tests actually test heterogeneous effects.

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 0(ja):1–41.

Choi, D. S. (2017). Estimation of monotone treatment effects in network experiments. 112(519):1147–1155.

Cochran, W. (1999). *Sampling Techniques*. John Wiley & Sons, third edition.

Cornfield, J. (1978). Randomization by group: a formal analysis. *American Journal of Epidemiology*, 108(2):100–102.

Coscia, M., Giannotti, F., and Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(5):512–546.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall/CRC Press.

Demuynck, T. (2015). Bounding average treatment effects: A linear programming approach. *Economics Letters*, 137(Supplement C):75 – 77.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B*, 78(3):655–671.

Ding, P. and Miratrix, L. W. (2017). Model-free causal inference of binary experimental data. *ArXiv e-prints*.

Donner, A. and Klar, N. (1994). Cluster randomization trials in epidemiology: theory and application. *Journal of Statistical Planning and Inference*, 42(1):37 – 56.

Dow, M. M. and de Waal, F. B. (1989). Assignment methods for the analysis of network subgroup interactions. *Social Networks*, 11(3):237 – 255. Special Issue on Non-Human Primate Networks.

Evans, S. N., Hansen, B. B., and Stark, P. B. (2005). Minimax expected measure confidence sets for restricted location parameters. *Bernoulli*, 11(4):571–590.

Feng, X., Feng, Y., Chen, Y., and Small, D. S. (2014). Randomization inference for the trimmed mean of effects attributable to treatment. *Statistica Sinica*, 24(2):773–797.

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839.

Fienberg, S. E. and Wasserman, S. S. (1981a). Categorical data analysis of single sociometric relations. *Sociological Methodology*, 12:156–192.

Fienberg, S. E. and Wasserman, S. S. (1981b). An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association*, 76(373):54–57.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., and Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Floudas, C. A. and Visweswaran, V. (1995). Quadratic optimization. In Horst, R. and Pardalos, P. M., editors, *Handbook of Global Optimization*, pages 217–269. Springer US, Boston, MA.

Fogarty, C. B., Shi, P., Mikkelsen, M. E., and Small, D. S. (2017). Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. *Journal of the American Statistical Association*, 112(517):321–331.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75 – 174.

Fortunato, S. and Castellano, C. (2012). *Community Structure in Graphs*, pages 490–512. Springer New York, New York, NY.

Frandsen, B. R. and Lefgren, L. J. (2016). Partial identification of the distribution of treatment effects. Unpublished manuscript.

Frank, O. (1977). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4(2):81–89.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1(1):91 – 101.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, 2(1):176 – 196.

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180 – 193.

Freedman, D. A. (2008c). Randomization does not justify logistic regression. *Statistical Science*, 23(2):237 – 249.

Freeman, G. H. and Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38(1/2):141–149.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

Friedman, J. H. and Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, pages 377–391.

Gamoran, A., Turley, R. N. L., Turner, A., and Fish, R. (2012). Differences between hispanic and non-hispanic families in social capital and child development: First-year findings from an experimental study. *Research in Social Stratification and Mobility*, 30(1):97 – 112. Inequality across the Globe.

Gibbons, J. D. and Pratt, J. W. (1975). P-values: Interpretation and methodology. *The American Statistician*, 29(1):20–25.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review / Revue Internationale de Statistique*, 54(2):127–138.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233.

Good, P. I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer, New York, third edition.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics*, 32(2):506–523.

Hájek, J. (2011). Comment to "An essay on the logical foundations of survey sampling, part one" by D. Basu. In DasGupta, A., editor, *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics, page 200. Springer New York.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23:219.

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

Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., and Samatova, N. (2014). Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):426–439.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783.

Hirji, K. F. and Johnson, T. D. (1996). A comparison of algorithms for exact analysis of unordered 2 × k contingency tables. *Computational Statistics & Data Analysis*, 21(4):419 – 429.

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

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Huang, E. J., Fang, E. X., Hanley, D. F., and Rosenblum, M. (2017). Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics*, 18(2):308–324.

Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882.

Kim, J. H. (2014). Identifying the distribution of treatment effects under support restrictions. Unpublished manuscript.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data*. Springer New York.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Inc., San Francisco.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Li, X. and Ding, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*, 35(6):957–960.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Lu, J., Ding, P., and Dasgupta, T. (2015). Construction of alternative hypotheses for randomization tests with ordinal outcomes. *Statistics & Probability Letters*, 107:348 – 355.

Lu, J., Ding, P., and Dasgupta, T. (2016). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. ArXiv e-print.

Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statist. Sci.*, 17(2):149–172.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, 65(6):1311–1334.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Maritz, J. S. (1981). *Distribution-Free Statistical Methods.* Chapman and Hall, London.

Middleton, J. A. (2008). Bias of the regression estimator for experiments using clustered random assignment. *Statistics & Probability Letters*, 78(16):2654 – 2659.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6.

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B*, 75(3):369 – 396.

Morgan, K. L. and Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.

Nascimento, M. C. and de Carvalho, A. C. (2011). Spectral methods for graph clustering – a survey. *European Journal of Operational Research*, 211(2):221 – 231.

Newman, M. E. J. (2013). Spectral methods for community detection and graph partitioning. *Phys. Rev. E*, 88:042822.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465 – 480. (Originally in Roczniki Nauk Tom X (1923) 1 – 51 (Annals of Agricultural Sciences). Translated from original Polish by Dambrowska and Speed.).

Nyblom, J., Borgatti, S., Roslakka, J., and Salo, M. A. (2003). Statistical analysis of network data—an application to diffusion of innovation. *Social Networks*, 25(2):175 – 195.

O'Malley, A. J. (2013). The analysis of social network data: an exciting frontier for statisticians. *Statistics in Medicine*, 32(4):539–555.

O'Neill, B. (2014). Some useful moment results in sampling problems. *The American Statistician*, 68(4):282–296.

Radlow, R. and Alf, E. F. (1975). An alternate multinomial assessment of the accuracy of the chi-squared test of goodness of fit. *Journal of the American Statistical Association*, 70(352):811–813.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173–185.

Rigdon, J. and Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924 – 935.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(2):219 – 231.

Rosenbaum, P. R. (2002a). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192.

Rosenbaum, P. R. (2002b). *Observational Studies.* Springer, 2^{nd} edition.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191 – 200.

Rosenbaum, P. R. (2010). *Design of Observational Studies.* Springer, New York.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Samii, C. and Aronow, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics and Probability Letters*, 82(2):365 – 370.

Särndal, C.-E., Wretman, J. H., and Swensson, B. (1992). *Model assisted survey sampling.* Springer-Verlag.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27 – 64.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

Schochet, P. Z. (2013). Estimators for clustered education rcts using the neyman model for causal inference. *Journal of Educational and Behavioral Statistics*, 38(3):219–238.

Seierstad, C. and Opsahl, T. (2011). For the few not the many? the effects of affirmative action on presence, prominence, and social capital of women directors in norway. *Scandinavian Journal of Management*, 27(1):44 – 54.

Sekhon, J. S. and Shem-Tov, Y. (2017). Efficient estimation of sample average treatment effects. Unpublished manuscript.

Small, D. S., Ten Have, T. R., and Rosenbaum, P. R. (2008). Randomization inference in a group-randomized trial of treatments of depression: covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association*, 103(481):271 – 279.

Stefanski, L. A. and Boos, D. B. (2002). The calculus of m-estimation. *The American Statistician*, 56:29 – 38.

Suesse, T. (2012). Marginalized exponential random graph models. *Journal of Computational and Graphical Statistics*, 21(4):883–900.

Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K., and Finkelstein, A. N. (2014). Medicaid increases emergency-department use: Evidence from Oregon's Health Insurance Experiment. *Science*, 343(6168):263–268.

Tsavachidou, D., McDonnell, T. J., Wen, S., Wang, X., Vakar-Lopez, F., Pisters, L. L., Pettaway, C. A., Wood, C. G., Do, K.-A., Thall, P. F., Stephens, C., Efstathiou, E., Taylor, R., Menter, D. G., Troncoso, P., Lippman, S. M., Logothetis, C. J., and Kim, J. (2009). Selenium and vitamin e: Cell type – and intervention-specific tissue effects in prostate cancer. *Journal of the National Cancer Institute*, 101(5):306–320.

Turley, R. N. L., Gamoran, A., McCarty, A. T., and Fish, R. (2017). Reducing children's behavior problems through social capital: A causal assessment. *Social Science Research*, 61:206 – 217.

Volfovsky, A., Airoldi, E. M., and Rubin, D. B. (2015). Causal inference for ordinal outcomes. ArXiv e-prints.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. an introduction to Markov graphs and p. *Psychometrika*, 61(3):401–425.

Whaley, F. S. (1983). The equivalence of three independently derived permutation procedures for testing the homogeneity of multidimensional samples. *Biometrics*, 39(3):741–745.