Advancing Design and Inference in a Causal Framework

Advancing Design and Inference in a Causal Framework

A dissertation presented

by

Nicole E. Pashley

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

April 2020

Dissertation Advisors:
Professor Luke W. Miratrix
Professor Tirthankar Dasgupta

Author:
Nicole E. Pashley

Advancing Design and Inference in a Causal Framework

Abstract

Causal inference, or the assessment of the effect of interventions on outcomes of interest, is ubiquitous in many fields. Often causal inference is made on the basis of the randomization of units to treatments, focusing on the design of an experiment or, in observational studies, how to approximate an experiment. The focus on randomization has allowed for results that do not depend on the classic structural assumptions needed in, for instance, linear regression. However, the beautifully simple idea of using randomization as the basis for inference can induce many subtle problems. This dissertation examines three such problems in causal inference. First, we explore the gains of blocked designs, as compared to non-blocked designs. Conservative variance estimators for the case of many blocks of variable size are built, leading to more general inference tools than previously established. Next, we examine the use of conditioning in causal inference, drawing parallels to analyzing an experiment as if another experiment had been run. Finally, we identify challenges in the analysis of observational data with multiple treatments and provide a framework on which to build inference.

Contents

List of Tables

List of Figures

Citations to Previous Work

Chapter 1 is based on Pashley and Miratrix (2017), submitted to arxiv.org:
Pashley, N. E. and Miratrix, L. W. (2017). Insights on variance estimation for blocked and matched pairs designs. *arXiv preprint arXiv:1710.10342*

Chapter 2 is based on work with Guillaume W. Basse and Luke W. Miratrix.

Chapter 3 is based on Pashley and Bind (2019), submitted to arxiv.org:
Pashley, N. E. and Bind, M.-A. C. (2019). Causal inference for multiple non-randomized treatments using fractional factorial designs. *arXiv preprint arXiv:1905.07596*

Acknowledgments

My PhD work would not have been possible without the help and support of many individuals. First I would like to thank my primary advisor, Luke Miratrix. Thank you for all of your guidance, wisdom, and support over the years. Thank you also to my two other committee members and advisors, Tirthankar Dasgupta and Kosuke Imai, for helping me grow as a statistician through illuminating discussions and helpful advice. To my other collaborators, including Guillaume Basse, Marie-Abèle Bind, Matthew Blackwell, and Peter Schochet, thank you for enriching my PhD with interesting work and vibrant conversations. There are many faculty members in the Harvard Department of Statistics who deserve my great appreciation; thank you all for making the department a community where I could grow. I would be remiss if I did not acknowledge the many other educators along the way, especially those at Queen's University including Chunfang Devon Lin and Wenyu Jiang, who aided my academic journey.

I would not have made it through the PhD without the kind support of my peers and friends. I would like to give a special thank you to Kristen Hunter, Katy McKeough, and Sanqian Zhang. Thank you for making the PhD an enjoyable and memorable experience. Thank you to the other Harvard statistics students who have enhanced my PhD, including Espen Bernton, Niloy Biswas, Zach Branson, Luis Campos, Kin Wai Chan, Ambarish Chattopadhyay, Jonathan Che, Chenguang Dai, Juan Diaz, Ruobin Gong, Dongming Huang, Phyllis Ju, Maxime Rischard, Stephane Shao, Albert Wu, and Lo-Hua Yuan. Thank you to the entire Miratrix C.A.R.E.S. Lab for teaching me so much and for helping me to develop as a researcher. There are many other academic colleagues who I owe thanks for great camaraderie and interesting conversations.

I would like to additionally acknowledge my family for all of their support. Thank you to my parents for always encouraging my mathematical endeavors and thank you to my sisters, Charlotte and Suzanne, for always making me laugh. Finally, the biggest thank you goes to Oliver, for everything you do for me. I would not have made it through without you by my side.

In addition to these thank you's, I gratefully acknowledge support from the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

To Oliver and my family.

Chapter 0

Introduction

Casual inference studies the quantification of the effect of an intervention(s) on an outcome of interest. Stemming from the work of Fisher and Neyman (e.g., Fisher, 1935; Splawa-Neyman et al., 1923/1990), there is a tradition of using the randomization of units to treatment or control conditions as the basis of inference in this setting. This intrinsically links the design of an experiment to its analysis. This simple idea of randomizing units to treatments and then basing analyses off of that randomness is powerful, but also introduces many subtle problems. In this dissertation, we explore three such problems. The first problem has to do with the analysis of blocked designs, and in particular variance estimation under such designs, with randomization based inference. The second problem concerns when it is valid to analyze an experiment as if another experiment had been run, exploring how conditioning can provide a randomization based justification. Finally we move to an observational setting with multiple treatments, where the recreation of a hypothetical experiment is made difficult by data sparsity with respect to treatment combinations.

Throughout this work, the careful focus on the random assignment of units to treatments has allowed for results that are designed based and free of structural assumptions. The following chapters involve building some foundational work to capture how estimators behave given different assumptions regarding assignment and sampling, and then extending this foundation. Furthermore, this dissertation provides rigorous and usable tools that are

also transparent to applied researchers.

Insights on Variance Estimation for Blocked and Matched Pairs Designs

In the causal inference literature, evaluating blocking from a potential outcomes perspective has two main branches of work. The first focuses on larger blocks, with multiple treatment and control units in each block. The second focuses on matched pairs, with a single treatment and control unit in each block. These literatures not only provide different estimators for the standard errors of the estimated average impact, but they are also built on different sets of assumptions. Additionally, neither literature handles cases with blocks of varying size that contain singleton treatment or control units, a case which can occur with different forms of matching or post-stratification. Differences in the two literatures have also created some confusion regarding the benefits of blocking in general. In this chapter, we first reconcile the literatures by carefully examining the performance of different estimators of treatment effect and of associated variance estimators under several different frameworks. We then use these insights to derive novel variance estimators for experiments containing blocks of different sizes. We also assess in which situations blocking is not guaranteed to reduce precision.

Conditional As-If Analyses in Randomized Experiments

The injunction to 'analyze the way you randomize' is well-known to statisticians since Fisher advocated for randomization as the basis of inference. Yet even those convinced by the merits of randomization-based inference seldom follow this injunction to the letter. Bernoulli-randomized experiments are often analyzed as completely randomized experiments, and completely randomized experiments are analyzed as if they had been stratified; more generally, it is not uncommon to analyze an experiment 'as-if' it had been randomized differently. This chapter examines the theoretical foundation behind this practice within a randomization-based framework. Specifically, we ask when is it legitimate to analyze an experiment randomized according to a design η_0 as if it had been randomized according to some other design, η. We show that a sufficient condition for this type of analysis to be valid is that the design η be derived from η_0 by an appropriate form of conditioning. We

use our theory to justify certain existing methods, question others, and finally suggest new methodological insights such as conditioning on approximate covariate balance.

Causal Inference for Multiple Non-Randomized Treatments using Fractional Factorial Designs

We explore a framework for addressing causal questions in an observational setting with multiple treatments. This setting involves attempting to approximate an experiment from observational data. With multiple treatments, this experiment would be a factorial design. However, certain treatment combinations may be so rare that we have no measured outcomes in the observed data corresponding to them. We propose to conceptualize a hypothetical fractional factorial experiment instead of a full factorial experiment and lay out a framework for analysis in this setting. We connect our design-based methods to standard regression methods. We finish by illustrating our approach using biomedical data from the 2003-2004 cycle of the National Health and Nutrition Examination Survey to estimate the effects of four common pesticides on body mass index.

Chapter 1

Insights on Variance Estimation for Blocked and Matched Pairs Designs

1.1 Introduction

Beginning with Neyman and Fisher, there is a long literature of analyzing randomized experiments by focusing on the assignment mechanism rather than some generative model of the data. One major family of experimental designs in this literature is blocked randomized experiments, where units are grouped to hopefully create homogenous collections, and then treatment assignment is randomized within each group (see Fisher, 1926). Ideally, this process gives a higher precision estimate of the overall average treatment effect, as compared to a completely randomized design.

We follow the potential outcome causal literature, (as in Imbens and Rubin, 2015; Rosenbaum, 2010), as opposed to the experimental design literature (as in Cochran and Cox, 1950; Wu and Hamada, 2000). Much of the prior work on randomized experiments within the potential outcomes framework has focused on two forms of blocking: blocking where there are several treated and control units in each block and blocking where there is exactly one treated and one control unit in each block (matched pairs). See, for example Imai et al. (2008) or Imbens (2011) for treatments of large blocks and Abadie and Imbens

(2008) or Imai (2008) for treatments of matched pairs. The literature has not, however, treated the cases where researchers have generated groups of varying size but where there is still only one treated and/or one control in some of the blocks, which we call the "hybrid design." Recent textbooks such as *Field Experiments: Design, Analysis and Interpretation* (Gerber and Green, 2012) and *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Imbens and Rubin, 2015) do not propose a clear answer for Neyman-style variance estimation in this case. While obtaining a point estimate for the overall average treatment effect is straightforward in this context, assessing the uncertainty of such an estimate is not. Currently one would instead have to turn to Fisher-style permutation tests, which typically rely on constant treatment effect assumptions, or regression-based approaches, which can be biased and usually require assumptions as to the residual error structure. We build on prior work to fill this gap by providing novel methods for conducting Neyman-style analyses for this more general hybrid design.

This gap is important as hybrid experiments with blocks of different sizes, and different numbers of treated and control units within the blocks, can easily arise in many modern social science experiments. For example, multisite trials in education often have several sites (e.g., districts) with only a few schools in each site. Many matching methods used in observational studies generate hybrid designs as well. For instance, Coarsened Exact Matching (CEM) (Iacus et al., 2012) can lead to many variable-sized blocks, some of which have singleton treatment or control units. "Full matching," which identifies collections of units that are similar on some baseline covariates (Hansen, 2004; Rosenbaum, 1991), creates variable-sized blocks, each with either one treated or one control unit. Our approach allows for a Neyman-style analysis in these contexts. See Section 1.7 for more on these applications.

There are several different models used for Neyman-style causal inference. The first, the finite sample model, takes the sample of units in the experiment as fixed, using the assignment mechanism as the sole source of randomness. Other so-called super-population or population models assume that the units in the experimental sample come from some larger population; this can induce additional uncertainty that needs to be accounted for.

With blocking, there is the further complication of how the blocks in the final experimental sample are formed. There can be fixed blocks in which every unit inherently belongs to one of a finite number of blocks; flexible blocks made by the experimenter once a sample is obtained; and structural blocks that capture natural groupings of units. There are also several possible sampling mechanisms possible beyond the classic simple random sampling of units typically assumed, such as sampling from strata corresponding to the blocks or sampling entire blocks rather than individual units. We believe these variants in how blocks are formed and sampled has caused the gap of the hybrid design: because much of the current literature uses different frameworks tailored to the specific special cases of either large blocks or matched pairs, it is not easily reconciled as the variance and variance estimators vary across these variants. As part of our work we carefully outline the common frameworks used and discuss how they are different from each other and how they connect to different types of experimental design. We also analyze the performance of uncertainty estimation for all cases.

As a consequence of our taxonomy of frameworks and block types, we also resolve some apparent contradictions found in the literature on the classical blocked and matched pairs designs. In particular, we derive expressions comparing blocking to complete randomization for multiple frameworks and show which frameworks give guarantees on the benefits of blocking and which do not. We also carefully separate out sampling variance from assignment variance to obtain more precise statements of the benefits of blocking than have been given in prior literature. We finally clarify how these comparisons depend on the block types and sampling mechanism, not the block sizes.

Recent work by Fogarty (2018) has also addressed some of these issues. In particular, Fogarty presents a method for estimating variance with small blocks of variable size, not just matched pairs. He also makes explicit the issue of differing results under different population and sampling frameworks by comparing multiple settings. In this chapter, we tackle the issue of creating a cohesive hybrid estimator for experiments with large and small blocks and also discuss the comparison of blocking and complete randomization. We do

not focus on the use of covariates to model treatment effect heterogeneity. The approach to causal inference used in this work has strong connections to the survey sampling literature, as treated in, e.g., Särndal et al. (2003) or Cochran (1977).

In Section 1.2 we set out our notation and discuss complete randomization and blocking. We begin with the finite sample framework because it is a building block for the infinite population frameworks. Section 1.3 provides methods for estimating uncertainty in the case of large blocks, small blocks, and the hybrid of the two, and gives their bias under the finite sample framework. We then, in Section 1.4, provide true variance formula and the performance characteristics of the variance estimators for several infinite population frameworks. In Section 1.5 we systematically compare blocking to complete randomization under the full range of frameworks and discuss how the findings differ. Section 1.6 contains finite sample simulation studies to illustrate estimator performance and Section 1.7 illustrates estimation in two data examples. For clarity in presentation, we have moved the derivations of provided formulae to the Appendix. To make our estimators easily usable, we refer readers to our R package, `blkvar` (Miratix and Pashley, 2020). Sample scripts demonstrating its use and replicating our simulations are also available.

1.2 Overall setup and notation

We use the Neyman-Rubin model of potential outcomes (Rubin, 1974; Splawa-Neyman et al., 1923/1990). We assume the Stable Unit Treatment Value Assumption of no differential forms of treatment and no interference between units (Rubin, 1980). We will discuss both completely randomized and block randomized experiments. Let there be n units in the sample. In a completely randomized experiment, the entire collection of the units in the sample is divided into a treatment group and a control group by taking a simple random sample of pn units as the treatment group and leaving the remainder as control. In a blocked randomized experiment, our sample is divided into K blocks, formed based on some pretreatment covariate(s), with n_k units in block k. Each block k is then treated as a mini-experiment, with a fixed number of $p_k n_k$ units being randomly assigned to treatment

and the rest to control, independently of the other blocks.

The sample average treatment effect (SATE) is the typical estimand in so-called finite sample inference, which takes our sample as fixed, leaving the assignment mechanism as the only source of randomness. Under blocking, the SATE within block k, for $k = 1, ..., K$, is

$$\tau_{k,\mathcal{S}} = \frac{1}{n_k} \sum_{i:b_i=k} \big(Y_i(t) - Y_i(c)\big),$$

where $Y_i(t)$ and $Y_i(c)$ are the potential outcomes for unit i under treatment and control, respectively, and where b_i indicates the block that unit i belongs to. The overall SATE (see Imbens and Rubin (2015), p. 86) is then

$$\tau_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i(t) - Y_i(c)\big).$$

In this work, we consider two estimators for the SATE (and later the population average treatment effect), one typically used for complete randomization and one for blocked randomization. Define the variable Z_i as $Z_i = t$ if unit i is assigned treatment and $Z_i = c$ if unit i is assigned control, for $i = 1, ..., n$. Let $\mathbb{I}_{Z_i=t}$ be the indicator that unit i received treatment, n_t be the total number of treated units, and n_c be the total number of control units. So, $n_t = \sum_{i=1}^{n} \mathbb{I}_{Z_i=t}$, $n_c = n - n_t$. Similarly, let $n_{t,k}$, $n_{c,k}$ indicate these values within block k. Define $Y_i^{obs} = Y_i(Z_i)$ as the outcome we observe for unit i given a specific treatment Z_i. The blocked randomization estimator is then a weighted average of simple difference estimators for each block

$$\widehat{\tau}_{(BK)} = \sum_{k=1}^{K} \frac{n_k}{n} \widehat{\tau}_k,$$

with the

$$\widehat{\tau}_k = \frac{1}{n_{t,k}} \sum_{i:b_i=k} \mathbb{I}_{Z_i=t} Y_i(t) - \frac{1}{n_{c,k}} \sum_{i:b_i=k} (1 - \mathbb{I}_{Z_i=t}) Y_i(c),$$

$k = 1, ..., K$, being simple difference estimators within each block. The complete randomization estimator, $\widehat{\tau}_{(CR)}$, is

$$\widehat{\tau}_{(CR)} = \frac{1}{n_t} \sum_{i=1}^{n} \mathbb{I}_{Z_i=t} Y_i(t) - \frac{1}{n_c} \sum_{i=1}^{n} (1 - \mathbb{I}_{Z_i=t}) Y_i(c).$$

We will often take the expectation over the randomization of units to treatment for a fixed sample. In particular, $\mathbb{E}\left[\widehat{M}|\mathcal{S},\mathbf{P}\right]$ is the expected value of some estimator $\widehat{M}$ for a given, fixed, finite sample $\mathcal{S}$ and for some assignment mechanism $\mathbf{P}$, which may be complete randomization or blocked randomization. To reduce clutter, we drop the $\mathbf{P}$ and simply write $\mathbb{E}\left[\widehat{M}|\mathcal{S}\right]$ if the estimator makes the assignment mechanism clear.

In general, our estimators are unbiased, with

$$\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{S}\right] = \mathbb{E}\left[\widehat{\tau}_{(BK)}|\mathcal{S}\right] = \tau_{\mathcal{S}}.$$

It is assessing variance, and the precision gains of blocking, that is more tricky. This assessment is the goal of this chapter, but first we need to introduce a few more useful concepts.

An important aspect of blocking is how the blocks are formed. Explicit articulation of block formation will be useful when we discuss asymptotic properties of our estimators and will also be used to differentiate the various population frameworks in Section 1.4. We identify three primary ways that blocks are formed:

(a) Fixed blocks: Occurs when the total number of blocks and the covariate distribution of blocks is fixed before looking at the sample. E.g., blocking that occurs on a single categorical covariate.

(b) Flexible blocks: Occurs when the covariate distribution and total number of blocks may not be known before looking at the sample's covariates. E.g. if there are many covariates or continuous covariates and matching or discretizing is used to form blocks.

(c) Structural blocks: Occurs when units have some natural grouping such that the blocks are self-contained. The members of each block are fixed and if a block is represented in the sample, typically all members of that block are in the sample. E.g., twins or classrooms.

Note that structural blocks are often thought of as clusters. With clusters, however, treatment assignment is commonly assigned at the cluster level, whereas we are focusing on treatment

assigned within cluster. We use "structural block" to clarify this difference.

1.3 Variance estimation

The prior section outlined the different experiments and treatment effect estimators in the finite sample. We next discuss how to estimate the estimators' variances, which are integral to obtaining standard errors and confidence intervals. We discuss estimators built from a Neyman-Rubin randomization perspective. See Appendix A.1 for a discussion of alternative variance estimators (such as from linear models) that make additional assumptions on the data structure. We first investigate bias of variance estimators under a finite sample framework and extend to other frameworks in Section 1.4.

We start by giving the true variance in the finite sample for each of the designs. To do so, we need some additional notation. The mean of the potential outcomes for the units in the sample under treatment z is

$$\bar{Y}(z) = \frac{1}{n}\sum_{i=1}^{n} Y_i(z).$$

The sample variance of potential outcomes under treatment z is

$$S^2(z) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(z) - \bar{Y}(z))^2.$$

The sample variance of the individual level treatment effects is

$$S^2(tc) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(Y_i(t) - Y_i(c) - \tau_{\mathcal{S}}\Big)^2.$$

$\bar{Y}_k(z)$, $S_k^2(z)$, and $S_k^2(tc)$ are defined analogously over the units in block k.

For the finite sample, the variances of $\widehat{\tau}_{(CR)}$ and $\widehat{\tau}_{(BK)}$ are well known (see Imbens and Rubin (2015)). For complete randomization,

$$\text{var}(\widehat{\tau}_{(CR)}|\mathcal{S}) = \frac{S^2(t)}{n_t} + \frac{S^2(c)}{n_c} - \frac{S^2(tc)}{n}. \tag{1.1}$$

Extending this to blocked randomization (see Imbens (2011)), the overall variance is

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\text{var}(\widehat{\tau}_k|\mathcal{S}) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\Big(\frac{S_k^2(t)}{n_{t,k}} + \frac{S_k^2(c)}{n_{c,k}} - \frac{S_k^2(tc)}{n_k}\Big). \tag{1.2}$$

For complete randomization we use the Neyman-style variance estimator of

$$\widehat{\sigma}^2_{(CR)} = \widehat{\text{var}}(\widehat{\tau}_{(CR)}) = \frac{s^2(c)}{n_c} + \frac{s^2(t)}{n_t}$$

with $s^2(z)$ the sample variance of the units which received treatment z. This gives a conservative estimate due to the dropping of the $S^2(tc)/n$ term. Some tightening is possible by exploiting features such as differences in the shape of the observed treatment and control outcome distributions; for examples see Aronow et al. (2014), Chapter 6 of Imbens and Rubin (2015), or Schochet (2016).

For blocked experiments, the type of variance estimator one would use depends on the sizes of blocks one has. In cases where we have at least two treated and two control units in each block, we can directly extend the completely randomized estimator strategy by using it within each block and weighting. Defining $s_k^2(z)$ analogously to $s^2(z)$ within block k, the typical variance estimator is

$$\widehat{\sigma}^2_{(BK)} = \widehat{\text{var}}(\widehat{\tau}_{(BK)}) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{s_k^2(c)}{n_{c,k}} + \frac{s_k^2(t)}{n_{t,k}}\right). \tag{1.3}$$

This is the estimator for blocking derived in Mukerjee et al. (2018). See Imbens (2011) for a more in depth discussion of blocking of this form. We call this the "big block" style of blocking, and the "big block" estimator.

For the "small blocks" case, where our blocks have only one treated unit or one control unit, we need to use an alternative approach as we cannot estimate the variance for a treatment arm with a single unit. Our approach is presented below. To give some background, the analytical problems that arise when estimating the variance in matched pairs experiments, especially when working in the finite sample framework, have been lamented by many statisticians (see, e.g., Imbens, 2011). The issues arise from the fact that there is no way to estimate the within pair variance with only one unit assigned to treatment

and one unit assigned to control in each pair. Previous work has found conservative estimators, however, which we build on. For instance, Imai (2008) showed that the standard matched pairs estimator is biased in the finite sample setting and put bounds on the true variance. The RCT-Yes R package and documentation (Schochet, 2016) also provides a conservative variance estimator for the matched pairs design (as well as estimators for blocked designs); this is discussed more in Appendix A.1.3.

For a hybrid experiment with both big and small blocks, we combine results to create an overall variance estimator.

1.3.1 Small block experiments with equal size blocks

When we have small blocks of the same size, we can directly use the usual variance estimator in the matched pairs literature (e.g., Imai, 2008) as a variance estimator for $\widehat{\tau}_{(BK)}$, no matter what the block sizes are, as also noted by Fogarty (2018). This gives a variance estimator of

$$\widehat{\sigma}^2_{(SMALL/s)} = \frac{1}{K(K-1)} \sum_{k=1}^{K} (\widehat{\tau}_k - \widehat{\tau}_{(BK)})^2. \tag{1.4}$$

This estimator directly estimates the variance of the overall block treatment effect estimator, rather than estimating the variance for each individual block and then weighting. We will see that, depending on the framework used, this estimator can give positively biased estimates if the true τ_k tends to differ across blocks.

1.3.2 Small block experiments with varying size blocks

For experiments with small blocks of varying sizes we offer two variance estimators. The first directly extends the standard matched pairs estimator by grouping the blocks by size into J groups and using Equation 1.4 for each group. We then weight and combine to get an overall variance estimator.

Stratified Small Block Variance Estimator:

$$\widehat{\sigma}^2_{(SMALL/m)} = \frac{1}{\left(\sum_{j=1}^{J} m_j K_j\right)^2} \sum_{j=1}^{J} (m_j K_j)^2 \widehat{\sigma}^2_{(SMALL),j}, \tag{1.5}$$

where K_j is the number of blocks of size m_j and

$$\widehat{\sigma}^2_{(SMALL),j} = \frac{1}{K_j(K_j-1)} \sum_{k:n_k=m_j} (\widehat{\tau}_k - \widehat{\tau}_{(SMALL),j})^2 \tag{1.6}$$

with $\widehat{\tau}_{(SMALL),j} = \sum_{k:n_k=m_j} \widehat{\tau}_k / K_j$. That is, grouping by the same size allows for using the equal size block estimator above. While straightforward, this is not ideal because it requires at least two blocks of each size in the overall experiment to estimate each $\widehat{\sigma}^2_{(SMALL),j}$. See Appendix A.5.1 for further detail.

The second approach allows the variance of all of the small blocks to be estimated at the same time, without requiring multiple blocks of the same size.

Unified Small Block Variance Estimator:

$$\widehat{\sigma}^2_{(SMALL/p)} = \sum_{k=1}^{K} \frac{n_k^2}{(n-2n_k)(n+\sum_{i=1}^{K} \frac{n_i^2}{n-2n_i})} (\widehat{\tau}_k - \widehat{\tau}_{(BK)})^2. \tag{1.7}$$

For $\widehat{\sigma}^2_{(SMALL/p)}$ to be defined and guaranteed conservative, no one block can make up half or more of the units. We derived this estimator using the basic form of the matched pairs variance estimator as a weighted sum of the squared differences between the estimated average block treatment effects and the estimated overall average treatment effect. The weights then come from a simple optimization (see Appendix A.6), and partially account for the different blocks having different levels of precision when estimating the variance of the block-level impacts. This estimator has similar finite sample properties to the standard estimator for blocks of the same size (Equation 1.4). In particular, it is also conservative and unbiased when the block average treatment effects are all the same. When block sizes are all the same, this reduces to the usual matched pairs type estimator.

1.3.3 Hybrid experiments

When doing variance estimation in a hybrid blocked design, we can split the blocks up into small blocks and big blocks. Grouping the big and small blocks together allows us to write the causal effect estimand as a combination of two estimands for our two different types of

block sizes. Let there be n_{sb} total units in small blocks in the sample. Then

$$\tau_{\mathcal{S}} = \frac{n - n_{sb}}{n}\tau_{(BIG),\mathcal{S}} + \frac{n_{sb}}{n}\tau_{(SMALL),\mathcal{S}}$$

where

$$\tau_{(BIG),\mathcal{S}} = \frac{1}{n - n_{sb}} \sum_{k:n_{t,k}\geq 2, n_{c,k}\geq 2} n_k\tau_k \quad \text{and} \quad \tau_{(SMALL),\mathcal{S}} = \frac{1}{n_{sb}} \sum_{k:n_{t,k}=1 \text{ or } n_{c,k}=1} n_k\tau_k.$$

The estimator for the overall treatment effect can also be written as

$$\widehat{\tau}_{(BK)} = \frac{n - n_{sb}}{n}\widehat{\tau}_{(BIG)} + \frac{n_{sb}}{n}\widehat{\tau}_{(SMALL)}.$$

For finite sample inference, we can similarly break down the variance, and estimator of the variance, of $\widehat{\tau}_{(BK)}$ because the block estimators are independent due to the block randomized treatment assignment.

Hybrid Variance Estimator:

$$\widehat{\text{var}}\left(\widehat{\tau}_{(BK)}\right) = \frac{(n - n_{sb})^2}{n^2}\widehat{\text{var}}\left(\widehat{\tau}_{(BIG)}\right) + \frac{n_{sb}^2}{n^2}\widehat{\text{var}}\left(\widehat{\tau}_{(SMALL)}\right).$$

Here we would use $\widehat{\sigma}^2_{(BK)}$ (Equation 1.3) for $\widehat{\text{var}}\left(\widehat{\tau}_{(BIG)}\right)$ and either $\widehat{\sigma}^2_{(SMALL/m)}$ (Equation 1.5) or $\widehat{\sigma}^2_{(SMALL/p)}$ (Equation 1.7) for $\widehat{\text{var}}\left(\widehat{\tau}_{(SMALL)}\right)$. Thus, when we have small blocks, we can estimate the variance for those small blocks separately and use the usual blocking estimator on the larger blocks. Alternatively, one could use $\widehat{\sigma}^2_{(SMALL/m)}$ or $\widehat{\sigma}^2_{(SMALL/p)}$ for all blocks, but we do not recommend this.

1.3.4 Finite sample bias of the variance estimators

In the finite setting all of the above estimators are conservative, and are only unbiased in specific circumstances. First, $\widehat{\sigma}^2_{(CR)}$ is known (Imbens and Rubin, 2015, p. 92; Splawa-Neyman et al., 1923/1990) to have bias

$$\mathbb{E}\left[\widehat{\sigma}^2_{(CR)}|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) = \frac{S^2(tc)}{n}.$$

If all of the blocks have at least two treated and two control units, we can extend this

result to $\widehat{\sigma}^2_{(BK)}$ (Equation 1.3), which has bias

$$\mathbb{E}\left[\widehat{\sigma}^2_{(BK)}|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K} \frac{n_k}{n^2} S_k^2(tc).$$

For the small blocks of varying sizes, we have two corollaries. See Appendix A.5.2 and A.6 for proofs. The first is

Corollary 1.3.4.1. *The bias of $\widehat{\sigma}^2_{(SMALL/m)}$ (Equation 1.5) under the finite framework is*

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{S}\right] - var\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right) = \sum_{j=1}^{J} \frac{K_j m_j^2}{n_{sb}^2(K_j-1)} \sum_{k:n_k=m_j} \left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.$$

The above extends prior results for $\widehat{\sigma}^2_{(SMALL/s)}$ for matched pairs (see Imai (2008), Imbens and Rubin (2015), p. 227, or, for a more general case, Fogarty (2018)). $\widehat{\sigma}^2_{(SMALL/m)}$ is conservative and unbiased when the average treatment effect is the same for all blocks of the same size (similar to the unbiased result from Imai (2008) for $\widehat{\sigma}^2_{(SMALL/s)}$).

For $\widehat{\sigma}^2_{(SMALL/p)}$ we have

Corollary 1.3.4.2. *The bias of $\widehat{\sigma}^2_{(SMALL/p)}$ (Equation 1.7) under the finite framework is*

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/p)}|\mathcal{S}\right] - var\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right) \\ = \sum_{k=1}^{K} \frac{n_k^2}{(n_{sb}-2n_k)(n_{sb}+\sum_{i=1}^{K}\frac{n_i^2}{n_{sb}-2n_i})}(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S}})^2.\end{aligned}$$

If the average treatment effect is the same across all small blocks then this estimator is unbiased, and if there is heterogeneity, it is conservative.

Remark. Both small block estimators are conservative, which raises the question of whether one is superior. The constant in front of each term of the bias of both estimators is of order n_k^2/n^2. Then we expect the bias of $\widehat{\sigma}^2_{(SMALL/m)}$ to be less than the bias of $\widehat{\sigma}^2_{(SMALL/p)}$ when the treatment effects of blocks of similar sizes are similar because the variance of impacts within blocks of a given size will be smaller than across all of the blocks. However, $\widehat{\sigma}^2_{(SMALL/m)}$ has the drawback that it can only be used when we have at least two blocks of each small size.

The improved potential performance of $\widehat{\sigma}^2_{(SMALL/m)}$ when there is homogeneity within block sizes does suggest that we could group blocks in some other way if we had prior

knowledge of which blocks were most similar. That is, $\widehat{\sigma}^2_{(SMALL/m)}$ relies on the blocks being equal size so the weights factor out of the sum to give the expression for the cross-block estimate of variation. But we could first subdivide our blocks based on some similarity measure and apply $\widehat{\sigma}^2_{(SMALL/p)}$ to each group, combining the parts with the hybrid weighting approach. This could make $\widehat{\sigma}^2_{(SMALL/p)}$ less conservative while maintaining its validity.

It is also worth considering the extent of conservatism of the estimators. For the case where all blocks are the same size, when we have blocks with m control units and 1 treated unit, as m increases the variance of the treatment effect estimator will decrease, as we are getting a more precise estimate for the control units. However, the form of the bias of $\widehat{\sigma}^2_{(SMALL/s)}$ remains the same. Therefore, with large m the bias of $\widehat{\sigma}^2_{(SMALL/s)}$ due to treatment heterogeneity becomes larger relative to the true variance. This intuition extends to the variable size case as well. In these cases alternative variance estimation strategies, such as discussed in Appendix A.1, may become more appealing.

It is important to note that the type of blocks will impact whether the bias of these estimators go to zero as sample size increases. For instance, one might argue for the use of $\widehat{\sigma}^2_{(SMALL/p)}$ instead of $\widehat{\sigma}^2_{(BK)}$ even if we have big blocks, because the condition for unbiasedness for $\widehat{\sigma}^2_{(SMALL/p)}$ (that all blocks have the same average treatment effect) could be considered less stringent than for $\widehat{\sigma}^2_{(BK)}$ (that there is zero treatment variation within each block). However, with fixed blocks, the number of units within each block increases as sample size increases and the bias of $\widehat{\sigma}^2_{(BK)}$ will go to zero, the standard result, but the bias of $\widehat{\sigma}^2_{(SMALL/p)}$ will not unless all of the blocks have the same average treatment effect. In this case, as the blocks grow to be big, we would use $\widehat{\sigma}^2_{(BK)}$.

In the hybrid setting the overall bias will be a weighted sum of the biases for the big and small block components. Therefore, because the overall weighting depends on the block sizes, having a poor estimator for the small blocks may not have a large effect on the overall bias if small blocks make up only a small proportion of the sample.

There is no way to unbiasedly estimate variance within small blocks without additional structure or covariates. If we think that the treatment effects of different strata are not too

far apart, then we suggest using one of the previous estimators. We at least know that the bias incurred is positive. However, if we have reason to believe that the treatment effects of different strata will be very far apart, a plug-in estimator, as discussed in Appendix A.1, may be more appropriate.

1.4 Infinite Population Frameworks

Up to this point we have examined blocking in a finite sample framework, conditioning on the units in the experiment in question. In the literature, however, blocking has often been examined under a variety of infinite population frameworks. In particular, the matched pairs literature uses a framework where the blocks themselves are sampled from an infinite population of blocks, whereas the big block literature typically assumes stratified random sampling from a finite number of infinite size strata. Using different population frameworks will give different answers to important questions of what the true variance of the treatment effect estimate is and what the bias of our variance estimators are. In this section, we first discuss the literature related to variance estimation for infinite populations, identifying the apparent tensions that exist. We then systematically discuss different frameworks, deriving the true variance of the treatment effect estimators under each of them. We also evaluate the bias of the variance estimators introduced in Section 1.3. We focus on infinite superpopulations; finite superpopulations substantially larger than the sample would give similar results. We explore work pertaining to the use of linear models, such as Cochran (1953) and Lin (2013), in Appendix A.1.1. An important note is that in some cases these sampling schemes are chosen for convenience and that the generalizability of the experiment to the population will depend upon the assumptions made in them being true. The sampling model may also be considered to serve as a conservative approach to finite sample inference (see Ding et al., 2017).

Related work

For matched pairs experiments, Imai (2008) showed that with a superpopulation of an infinite

number of structural blocks, specifically matched pairs, from which pairs are randomly sampled, the standard matched pairs variance estimator (Equation 1.4), is unbiased for the population average treatment effect (PATE). On the other hand, Imbens (2011) showed that the standard matched pairs variance estimator is biased in the setting where we have fixed blocks and units are drawn using stratified random sampling (see Section 1.4.3 for more on this setting). This is a clear example of how the population framework being used matters. We therefore advise practitioners to carefully consider what population and sampling structure they are assuming and to not simply assume a framework for convenience.

The general blocked design has been previously discussed in various forms. Imbens (2011) discussed blocking in the context of a superpopulation with a fixed number of strata from which units are sampled using a stratified sampling method. He formed unbiased estimators for the variance in this context, assuming that the blocks each have at least two units assigned to treatment and control. These results are similar to finite sample results discussed in Section 1.3 and will be discussed more in Section 1.4.3. Imai et al. (2008) analyzed estimation error and variance with the blocked design. Scosyrev (2014) also analyzed the blocked experiment in the finite sample and under two sampling frameworks, recognizing that the different settings resulted in different outcomes. Sävje (2015) analyzed flexible "threshold" blocking and made critical points about the importance of block structure and sampling design when analyzing blocked experiments, which we will echo and expand on.

1.4.1 Infinite populations in general

Inference for the population average treatment effect (PATE) typically takes the sample as a random sample from some larger population, as opposed to inference for the SATE discussed earlier which held the sample of potential outcomes as fixed. This makes estimation an implicit two-step process, estimating the treatment effect for the sample and extrapolating this estimate to the population. Frequently, in fact, the estimators themselves are the same as for finite sample inference even though the estimands are different.

Define the PATE as

$$\tau = \mathbb{E}[Y_i(t) - Y_i(c)|\mathcal{F}],$$

where $\mathcal{F}$ both indicates the block type and sampling framework. This is the same as the direct average of the unit-level treatment effects for all of the units in the population, as is commonly used (see Imbens and Rubin, 2015, p. 99), as long as our sampling mechanism is not biased. Here we will only consider frameworks where the sampling scheme provides a sample that, on average, has the same average treatment effect as the population but note that bias from the sampling mechanism can be fixed using weighting if the sampling mechanism is known (see Miratrix et al., 2018).

Under blocking, the PATE within block k is

$$\tau_k = \mathbb{E}[Y_i(t) - Y_i(c)|b_i = k, \mathcal{F}],$$

where, again, b_i indicates the block that unit i belongs to. It is possible that k indexes a (countably) infinite set of blocks in the case of some infinite population models.

Overall, using the law of total expectation and variance decompositions, we can generally obtain the properties of our estimators with respect to population estimands by first obtaining expressions for a finite sample and then averaging these expressions across the sampling distributions. In other words, we heavily exploit $\mathbb{E}\left[\widehat{M}|\mathcal{F}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{M}|\mathcal{S}\right]|\mathcal{F}\right]$, where $\mathcal{S}$ is a sample obtained from $\mathcal{F}$, our population and sampling framework. Under any unbiased framework $\mathcal{F}$, we have the typical result (e.g. see Imbens, 2011)

$$\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{F}\right] = \mathbb{E}\left[\widehat{\tau}_{(BK)}|\mathcal{F}\right] = \mathbb{E}\left[\tau_{\mathcal{S}}|\mathcal{F}\right] = \tau.$$

There are several different frameworks that one might assume. These can generally be characterized by two primary features: the block types, which also dictates the population strata structure, and the sampling scheme. Note that the term strata is used for the population here analogously to blocks in the sample. We may obtain a sample using simple random sampling and then form blocks based on covariates post-sampling and pre-randomization, i.e. flexible blocks. Or we may have fixed blocks (e.g. blood types)

and use stratified sampling where we sample units from each population stratum. Finally, we may have structural blocks and conceptualize a population of an infinite number of these blocks (e.g. schools in an "infinite" population of schools) from which we randomly select a fixed number of blocks. As we show next, the bias of the variance estimators can differ depending on the framework assumed. We refer to frameworks using their sampling method as a shorthand, leaving the block type and population structure implicit.

1.4.2 Simple random sampling, flexible blocks

In this framework, denoted SRS, units are sampled at random, without regard to block membership, from the population. This gives the classic result for complete randomization (see Imbens and Rubin, 2015, p. 101) of

$$\text{var}(\widehat{\tau}_{(CR)}|SRS) = \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t}$$

where $\sigma^2(z)$ is the population variance of the potential outcomes under treatment z. The classic variance estimator $\widehat{\sigma}^2_{(CR)}$ is unbiased in this setting.

For blocking with SRS, we focus on the use of flexible blocks, e.g. blocking using clustering on a continuous covariate or based on observed covariates in the sample obtained. Structural blocks do not make sense in this framework (e.g. one would always sample pairs of twins not individuals who are twins if we wish to run a twin study) and fixed blocks give rise to difficulties when the sample does not have units from all population strata. This sampling framework was examined for blocked experiments with fixed blocks in Scosyrev (2014).

For a blocked experiment, the variance in this framework, using the basic variance decomposition, is

$$\text{var}(\widehat{\tau}_{(BK)}|SRS) = \mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \text{var}\left(\tau_{\mathcal{S}}|SRS\right).$$

The expectation is across the sampling and blocking process.

Theorem 1.4.2.1. *The variance estimator*

$$\widehat{\sigma}^2_{SRS} = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{s^2_k(c)}{n_{c,k}} + \frac{s^2_k(t)}{n_{t,k}} \right) + \sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)} \right)^2 \tag{1.8}$$

is an unbiased estimator for $var(\widehat{\tau}_{(BK)}|SRS)$.

See Appendix A.7 for a derivation. The first term in the estimator looks similar to our usual big block estimator and captures part of the first term in our variance decomposition. The second term looks similar to our proposed small block estimator and accounts for the rest of the variation. This is very similar to the estimator found in Scosyrev (2014), however we make adjustments to achieve unbiasedness of the estimator whereas Scosyrev (2014) focuses on consistency. Scosyrev (2014) also works with fixed blocks where the number of blocks is assumed known before sampling and weights are used to match the sample to the population proportions, as opposed to flexible blocks which allow random numbers of blocks that are created post-sampling.

Remark. If we naïvely use $\widehat{\sigma}^2_{(BK)}$ (Equation 1.3) our bias will be

$$\mathbb{E}\left[\widehat{\sigma}^2_{(BK)}|SRS\right] - \text{var}(\widehat{\tau}_{(BK)}|SRS) = \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k}{n} S^2_k(tc) - S^2(tc) \middle| SRS\right].$$

This result follows from the derivations in Appendix A.7 and it implies that $\widehat{\sigma}^2_{(BK)}$ could be anti-conservative in this setting if there is generally treatment variation across samples (making $S^2(tc) > 0$), but units put within the same block are nearly identical in terms of impacts (making $S^2_k(tc) \approx 0$). This could happen when the experimenter is successfully making homogenous blocks.

Similarly, if we use either of the small block variance estimators, the bias will be the difference between the expected finite sample bias for those estimators (which depends on treatment effect heterogeneity between blocks) and $\mathbb{E}\left[S^2(tc)\middle|SRS\right]/n$, which corresponds to treatment effect heterogeneity across the whole population. Therefore whether these estimators are conservative or not depends upon the structure of the population and how the blocks are formed.

1.4.3 Stratified sampling, fixed blocks

In the "stratified sampling" framework, denoted $\mathcal{F}_1$, there are K fixed strata of infinite size in the population. Then n_k units are randomly sampled from strata k (i.e., stratified random sampling is used). Here we have fixed blocks. We assume that n_k is fixed and that n_k/n is the population proportion of units in stratum k, for simplicity. Otherwise, a weighting scheme, as mentioned in Section 1.4.1, would be needed to create an unbiased estimator of the direct average of treatment effects in the population. This is the framework used in Imbens (2011) and Miratrix et al. (2013), who show the following result under equal proportions treated within each block, which simplifies the weights.

As in the finite sample, overall variance is a weighted sum of within block variances:

$$\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \text{var}\Big(\hat{\tau}_k|\mathcal{F}_1\Big) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\Big(\frac{\sigma_k^2(c)}{n_{c,k}} + \frac{\sigma_k^2(t)}{n_{t,k}}\Big), \tag{1.9}$$

with $\sigma_k^2(z)$ the population variance of the potential outcomes under treatment z in strata k.

As noted in Imbens (2011), the variance estimator of big blocks, $\widehat{\sigma}^2_{(BK)}$ (Equation 1.3), is unbiased in this framework. The estimators for the variance of the small blocks, however, can have bias. We have two results pertaining to this.

First, as with the finite sample, we can extend results for $\widehat{\sigma}^2_{(SMALL/s)}$ (see Imbens, 2011) to $\widehat{\sigma}^2_{(SMALL/m)}$.

Corollary 1.4.3.1. *The bias of $\widehat{\sigma}^2_{(SMALL/m)}$ (Equation 1.5) under the stratified sampling framework is*

$$\mathbb{E}\Big[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{F}_1\Big] - var(\widehat{\tau}_{(SMALL)}|\mathcal{F}_1) = \sum_{j=1}^{J} \frac{K_j m_j^2}{n_{sb}^2(K_j-1)} \sum_{k:n_k=m_j} \Big(\tau_k - \tau_{(SMALL)}\Big)^2.$$

See Appendix A.5.2 for the derivation. As with finite sample inference, this shows that $\widehat{\sigma}^2_{(SMALL/m)}$ is a conservative estimator unless the average treatment effect is the same across all small blocks of the same size, in which case it is unbiased.

Second, for our new variance estimator we have the following result:

Corollary 1.4.3.2. *The bias of $\widehat{\sigma}^2_{(SMALL/p)}$ (Equation 1.7) under the stratified sampling framework*

is

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/p)}|\mathcal{F}_1\right] - var(\widehat{\tau}_{(SMALL)}|\mathcal{F}_1)$$
$$= \sum_{k=1}^{K} \frac{n_k^2}{(n_{sb} - 2n_k)(n_{sb} + \sum_{i=1}^{K} \frac{n_i^2}{n_{sb}-2n_i})} \left(\tau_k - \tau_{(SMALL)}\right)^2.$$

This shows that $\widehat{\sigma}^2_{(SMALL/p)}$ is also a conservative estimator (given no block makes up more than half the sample) and it is unbiased when the average treatment effect is the same across all small blocks. See Appendix A.6 for a derivation.

1.4.4 Random sampling of strata, structural blocks

In the "random sampling of strata" framework, denoted $\mathcal{F}_2$, there are an infinite number of strata of finite size, i.e. an infinite number of structural blocks. K strata are then randomly chosen to be in the sample and randomization is done within each of the sample blocks. This setting, with equal block sizes, is often used in the matched pairs literature, such as in Imai (2008).

Within this framework, which blocks are included in the sample is itself random. Therefore, the variance estimator needs to capture not only the within strata variance but also the variance due to which strata are chosen to be in the sample. Furthermore, if the block sizes vary, the total number of units is random which introduces additional complexities.

For the more general variable-size version of this framework, the variance of $\widehat{\tau}_{(BK)}$ is

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_2\right) = \mathbb{E}\left[\sum_{k:B_k=1} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)|\mathcal{F}_2\right] + \text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_2\right), \qquad (1.10)$$

where B_k is the indicator that stratum k is included in the sample, with $B_k = 1$ indicating sample membership and $B_k = 0$ otherwise.

When blocks are of the same size, we can simplify the expression with $\frac{n_k^2}{n^2} = \frac{1}{K^2}$, which is no longer random. If we have all blocks of the same size, then we can rewrite $\widehat{\sigma}^2_{(SMALL/s)}$

(Equation 1.4) using sample inclusion indicators as

$$\widehat{\sigma}^2_{(SMALL/s)} = \frac{1}{K(K-1)} \sum_k B_k(\widehat{\tau}_k - \widehat{\tau}_{(BK)})^2,$$

and this is an unbiased estimator for $\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_2)$. This is simply the variance of the estimated block effect in the sample. Imai (2008) showed that this estimator is unbiased in this setting with an infinite population of matched pairs. See Appendix A.5.2 for the proof of this result extended to other small block types of equal size.

Variance estimators when the strata vary in size are more complicated. In particular, under this framework there is a chance that there is only a single block of a given size, making the first variance estimator infeasible. If we condition on the number of strata drawn of each possible strata size, assuming that there are multiple strata of the each size in the sample, we obtain the following Corollary:

Corollary 1.4.4.1. *In the conditioned case, assuming it is defined, $\widehat{\sigma}^2_{(SMALL/m)}$ (Equation 1.5) is an unbiased estimator for $var(\widehat{\tau}_{(BK)}|\mathcal{F}_2)$.*

This result can be seen directly from the results in Appendix A.5.2.

Alternatively, if we are willing to assume that block size is independent of treatment effect, then we have the following more general result:

Theorem 1.4.4.1 (Unbiasedness of $\widehat{\sigma}^2_{(SMALL/p)}$ given independence)**.** *In the case the random sampling of strata setting where block sizes are independent of block average treatment effects, $\widehat{\sigma}^2_{(SMALL/p)}$ (Equation 1.7) is an unbiased estimator for $var(\widehat{\tau}_{(BK)}|\mathcal{F}_2)$.*

The proof is in Appendix A.6.2.

Remark. We may also consider an infinite number of strata of infinite size, as is commonly used in multisite randomized trials. This is the setting considered in Schochet (2016) and the RCT-YES software (Schochet, 2016) estimator discussed in Appendix A.1.3 could be used. The sampling scheme then has two steps: first sample the strata, then sample units from the strata. To discuss variance, we need to add a bit of notation. Let $\tau^*_{\mathcal{S}}$ denote the expectation of the treatment effect estimator given the blocks in the sample. That is, we fix

which strata are in the sample and take the expectation over the sampling of units from the infinite size strata. So conditioning on which strata are in the sample we are in a stratified sampling set up. Let this framework be denoted by $\mathcal{F}_3$. Then the variance of $\widehat{\tau}_{(BK)}$ is

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_3\right) = \mathbb{E}\left[\sum_{k:B_k=1} \frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(c)}{n_{c,k}} + \frac{\sigma_k^2(t)}{n_{t,k}}\right)|\mathcal{F}_3\right] + \text{var}\left(\tau_{\mathcal{S}}^*|\mathcal{F}_3\right).$$

It is straightforward to extend the results of Corollary 1.4.4.1 and Theorem 1.4.4.1 to this case.

1.4.5 Discussion

While the variance formulas that we presented above share a similar structure with each other and the finite sample forms, there are important differences. In the finite sample framework (Equation 1.2), there is a term regarding treatment effect variation that reduces the variance due to the correlation of potential outcomes. This term is retained in the random sampling of strata framework of Section 1.4.4 but not in the stratified sampling framework of Section 1.4.3. This difference in the true variance implies that different variance estimators may be more appropriate in different settings and that comparisons of blocking to complete randomization under these different assumptions will also diverge. In fact, this difference explains much of apparent discrepancy between the matched pairs literature and the blocking literature.

Relatedly, different variance estimators can have different amounts of bias depending on the framework being used. The small blocks estimators ($\widehat{\sigma}^2_{(SMALL/m)}$ and $\widehat{\sigma}^2_{(SMALL/p)}$) in the finite sample and the stratified sampling framework are unbiased if the average treatment effect is the same across all of the small blocks (or all of the small blocks of the same size for $\widehat{\sigma}^2_{(SMALL/m)}$) and otherwise are more conservative as the variance of the average treatment effects across blocks increases. For the infinite number of strata framework, under some assumptions all of our small block variance estimators are unbiased. We have no small block estimator that is guaranteed to be unbiased or conservative for the simple random sampling (flexible block) framework.

The big blocks estimator ($\widehat{\sigma}^2_{(BK)}$) in the finite sample is unbiased if the treatment effect is additive within each block and otherwise depends on the treatment effect heterogeneity within each block. In the stratified sampling framework, however, $\widehat{\sigma}^2_{(BK)}$ will be unbiased.

Overall, only the framework of Section 1.4.4 of sampling structural blocks, with the additional assumption given there, has unbiased variance estimators for a mixture of big and small blocks. This means that, without additional assumptions allowing for plug-in approaches, the hybrid estimators will always be conservative in the other settings discussed.

1.5 Comparing blocking to complete randomization

Much confusion in the literature surrounding the benefits of blocking can be attributed to researchers performing isolated investigations of estimators' properties under specific sampling frameworks. Although researchers typically focus on finite vs. infinite population as a distinction when assessing if blocking may be beneficial or harmful, the block types and sampling framework being used also matter. Blocking investigations have been made from many perspectives (e.g. Snedecor and Cochran, 1989) but we will discuss the prior work that has been done within the causal inference potential outcomes framework. Imai (2008) compared the true variance for the matched pairs design to the variance of the estimator for the completely randomized design under two sampling schemes, recognizing the important role that the sampling scheme plays. From this he concluded that "the relative efficiency of the matched-pair design depends on whether matching induces positive or negative correlations regarding potential outcomes within each pair" (Imai, 2008, p. 4865), a comment similar to one made on p. 101 of Snedecor and Cochran (1989). In contrast, Imbens (2011), assuming a stratified sampling superpopulation model, claimed that "In experiments with randomization at the unit-level, stratification is superior to complete randomization, and pairing is superior to stratification in terms of precision of estimation" (p. 1). Imai et al. (2008) similarly concluded that the variance under the blocked design was lower than under complete randomization for a superpopulation set-up. Although these conclusions are correct for the context the authors were working in, the use of different frameworks for

blocking and matched pairs can make the results seem inconsistent.

For our primary discussion, we assume that $p_k = p$ for all k, i.e. the proportion of treated units is constant across blocks and is the same as the proportion of treated units under complete randomization. We compare the variances of a blocked design vs. complete randomization under several population frameworks to clarify the similarities and differences under these different sampling methods. All derivations are in Appendix A.8. If we do not assume $p_k = p$ for all k, then units in the same treatment arm can be weighted differently and we do not have the same guarantees; see Appendix A.8.6. Section A.9 of the Appendix relatedly discusses the performance of the complete randomization variance estimator when treatment assignment is done according to blocked randomization, i.e., it gives the performance of the variance estimator that ignores blocking. Appendix A.4 examines the variances of the variance estimators, and compares the *variances* under blocking to complete randomization.

Finite framework

For the finite setting, we find a result similar to those presented in other papers, such as Imai et al. (2008) and Miratrix et al. (2013). The difference in variance of the treatment effect estimator between the completely randomized design and the blocked design, in the finite sample, is

$$\begin{aligned} &\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\ &= \frac{1}{n-1}\left[\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^{K}\frac{n_k}{n}\frac{n-n_k}{n}\text{var}(\widehat{\tau}_k|\mathcal{S})\right] \end{aligned} \tag{1.11}$$

where

$$\text{Var}_k\left(X_k\right) \equiv \sum_{k=1}^{K}\frac{n_k}{n}\left(X_k - \sum_{j=1}^{K}\frac{n_j}{n}X_j\right)^2. \tag{1.12}$$

Whether this quantity is positive or negative depends on whether some form of between block variation is larger than a form of within block variation. Finite sample numerical studies in Appendix A.3.2 show an example where even in the worst case for blocking, when all blocks have the same distribution of potential outcomes, the increase in variance

is not too great. Most prior work state that although the difference in the brackets can be negative, as the sample size grows this difference will go to a non-negative quantity. However, this statement depends on the type of blocks we have. In particular, if we have structural blocks such that as n grows, the number of blocks K also grows, the difference in the brackets of Equation 1.11 will not necessarily go to zero or become positive as $n \to \infty$. This has ties to the random sampling of strata framework.

Simple random sampling, flexible blocks

The difference of variances in this setting is just the expectation over Equation 1.11 with respect to sampling of units. If we have a fixed set of covariates and a fixed algorithm used to do blocking, then even if the covariates used to form blocks are independent from outcomes, blocking will not increase variance and in fact this difference will be zero, as shown in Appendix A.8.2. Hence, in the "worst case" of blocking on something irrelevant, blocking will still not increase variance on average. However, if the algorithm or the set of covariates used to block was allowed to adversarially change from sample to sample, then it could hypothetically choose the worst possible blocking for each sample, making the sample variance for blocking always higher than for complete randomization.

Stratified sampling, fixed blocks

Theorem 1.5.0.1 (Variance comparison under stratified sampling)**.** *The difference in variance between complete randomization and blocked randomization under the stratified sampling framework is*

$$var\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - var\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right) = \frac{1}{n-1} Var_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right) \geq 0$$

where $\mu_k(z)$ is the population mean of potential outcomes under treatment z in strata k and $Var_k(\cdot)$ is defined as in Equation (1.12).

Now, it is not possible for blocking to be harmful in the stratified sampling framework with equal proportion treated in all blocks.

Interestingly, when comparing blocking to completely randomized experiments in an infinite population setting, researchers have typically evaluated the completely randomized

design under the simpler sampling mechanism of simple random sampling and analyze the blocked design under the stratified sampling framework (e.g., see Imbens, 2011). For further discussion and mathematical formulation of that comparison, see Appendix A.8.4.

Random sampling of strata, structural blocks

The difference between the variances is again the expectation over Equation 1.11 with respect to sampling of blocks. Here the blocks themselves are sampled with block membership fixed, so the expectation can be thought of as over all blocks in the population. As in the finite framework, because the strata themselves are finite, it is possible that blocking could result in higher variance. In particular, it is possible to have systematically poor blocks if the block means do not vary. For instance, if we use elementary school classrooms as blocks, we may find that schools break up students into classes such that the classrooms all look similar to each other, in that they have similar proportions of high and low achieving students, but by the same token have higher within classroom variability. For the case with infinite strata of infinite size, see Appendix A.8.5.

Discussion

Different types of blocks and sampling framework can change the answer to the question "Is blocking always beneficial in terms of the precision of my estimators?" These findings, however, assume equal treatment proportions across blocks. With unequal proportions we do not have such guarantees (see Appendix A.8.6). In our simulations (see Appendix A.4.1) we in fact see degradation in the benefits of blocking on weakly predictive covariates when the proportion treated is only approximately equal, rather than precisely. This comes from the additional variation induced by different units being weighted differently, and is akin to the increase in variance from weighted survey sampling. This suggests that in many realistic scenarios, one might not want to block on covariates that are only weakly predictive.

1.6 Simulations

We compare different estimators of the variance for hybrid blocked experiments where there are a few big blocks and many small blocks in a finite sample context. Here, 50% of our

units are in small blocks with only one treated unit and the remainder are in big blocks with at least two treated units. In none of our blocks were there many treated units due to only having approximately 20% of the units treated. The 20% was approximate in order to create varying size small blocks to see the different performance of the hybrid estimators. We have 15 blocks with sizes ranging from 3 to 20.

The simulations presented here are for the finite sample framework, as it is both a common mode of inference as well as a core building block to the population frameworks. These results, however, are largely applicable to these other settings. For instance, the biases for the small blocks variance estimators have the same form for the finite sample and the stratified sampling frameworks.

We considered our two hybrid estimators, which correspond to estimating the variance of the small blocks two different ways. We also considered two regression estimators: the HC1 sandwich estimate (Hinkley, 1977) from a linear model with fixed effects and no interaction between treatment indicator and blocking factor, and the standard variance estimate (inverse Fisher information) from a weighted regression, weighting each unit by the inverse probability of being assigned to its given treatment status in its block, multiplied by the overall proportion of units in its treatment group (this is a variant of the approach in Gerber and Green (2012); see also Miratrix et al. (2020)).[1] Note that the HC1 estimator is the "robust" estimator used in Stata (StataCorp, 2017) for estimating standard deviations.

In our simulations, we varied how well blocking separated units based on the potential outcomes under control and on the treatment effect. The average potential outcome under control and the treatment effect for each block were both negatively correlated with block size, so that smaller blocks had larger control potential outcomes and larger treatment effects. The correlation of potential outcomes within blocks was also varied between $\rho = 0, 0.5,$ and 1. See Appendix A.3.1 for more on the data generating process.

[1]There are actually different weighting approaches one can use in regression adjustment; in particular one can use precision weighting or survey weighting. In additional explorations we examined survey weighting as implemented by `svyglm`, and found these other options generally performed more poorly, with some approaches resulting in substantial underestimation of variance and others having a great deal of inflation.

We compared all of the variance estimators to the actual variance of the corresponding blocking treatment estimator in Figure 1.1 by looking at the percent relative bias ($[\text{mean}(\hat{\sigma}^2_*) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})]/\text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})$).[2] The variation due to changing the between block difference in the mean of control potential outcomes was found to be minimal so we average these differences on the plots. The two hybrid estimators, the one using $\widehat{\sigma}^2_{(SMALL/m)}$ (Equation 1.5) (Hybrid_m) for the small blocks and the one using $\widehat{\sigma}^2_{(SMALL/p)}$ (Equation 1.7) (Hybrid_p) for the small blocks, outperform the linear model estimators, especially as the treatment effect variation across the blocks increases. We see that Hybrid_m also has lower bias than Hybrid_p as treatment heterogeneity increases. This is because the value of treatment effects are correlated with block size and $\widehat{\sigma}^2_{(SMALL/m)}$ groups variance estimation by block size. Weighted regression performance was generally similar to that of Hybrid_p. It was slightly anti-conservative for samples with low treatment heterogeneity when $\rho = 1$.

For discussion of the variance of the variance estimators, see Appendix A.4.2. The variance estimators' variances were found to be comparable, with weighted regression generally having the lowest variance.

When comparing the performance of estimators, there is an important note about the linear model estimator: the sandwich estimate for a linear model is associated with a different treatment effect estimator than the others. In particular, a linear model with fixed effects is estimating a precision weighted estimate of the treatment effect across the blocks. It is well known that as treatment heterogeneity increases, this estimator can become increasingly biased. See, Raudenbush and Schwartz (2020) for a longer discussion on this and related estimators. This is not an issue for the weighted regression which, similar to adding interactions between treatment and block dummy variables, will recover $\widehat{\tau}_{(BK)}$.

[2]We compare all estimators to the variance of $\widehat{\tau}_{(BK)}$ to put everything on the same scale, even though the sandwich estimate for a linear model is estimating variance for the linear model estimator.

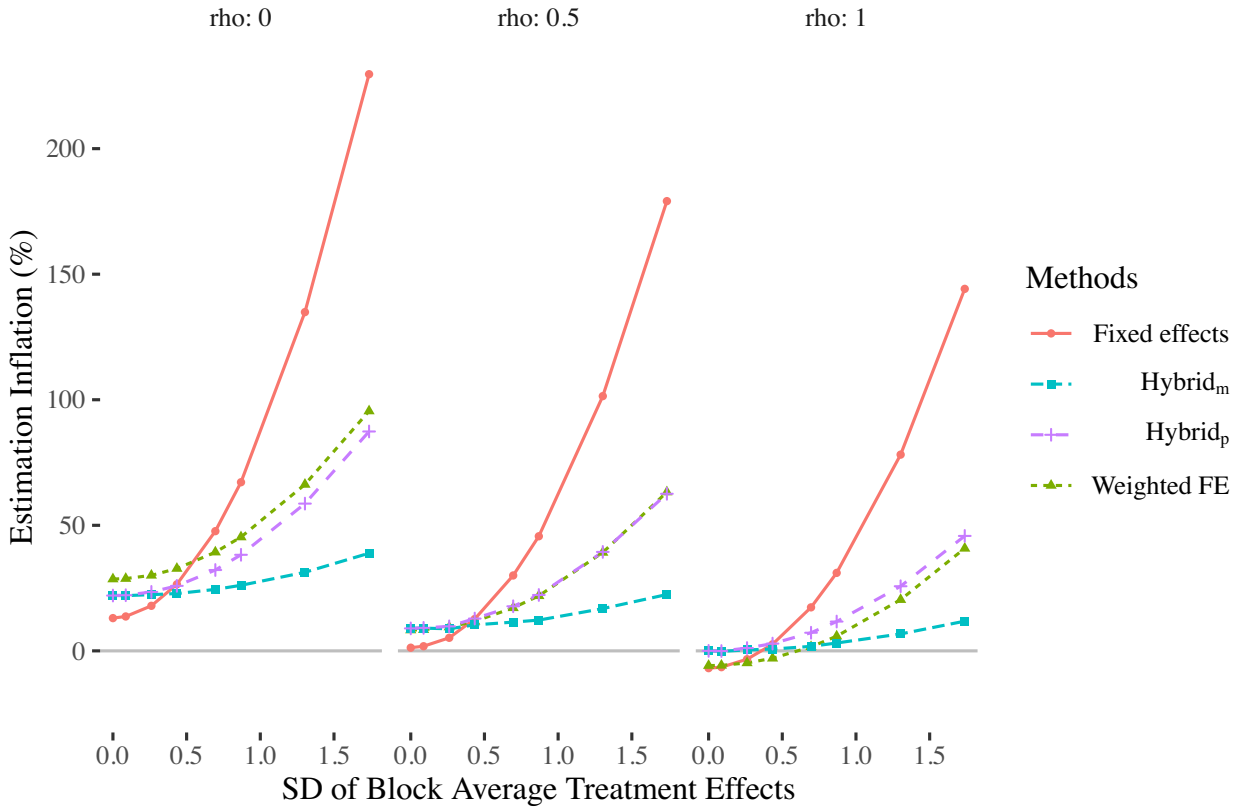

Figure 1.1: *Simulations to assess variance estimators' relative bias as a function of treatment variation across blocks. Each column represents a different value of ρ, with values denoted at the top of the graph. The x-axis shows the standard deviation of block treatment effects. Dots indicate average over changes in control means for specific finite samples. FE stands for fixed effects.*

1.7 Data Example

One area where analysts are often faced with many small blocks of varying sizes is found in the matching literature. In particular, full matching (see Hansen (2004), Rosenbaum (1991)) finds sets of similar units, with either one treated and several control or vice versa, that could be considered as-if randomized. After matching, a researcher could then analyze these data using permutation tests and associated sensitivity checks (see, e.g. Rosenbaum (2010)), but in this context generating confidence intervals or standard errors using permutation inference would typically rely on a constant treatment effect assumption across the blocks. One might alternatively wish for a Neyman-style randomization analysis such as would

	NHANES		Lalonde	
Estimator	Estimate	$\widehat{SE}$	Estimate	$\widehat{SE}$
Hybrid blocking with $\widehat{\sigma}^2_{(SMALL/m)}$	2.45	N/A	\$560	\$570
Hybrid blocking with $\widehat{\sigma}^2_{(SMALL/p)}$	2.45	0.20	\$560	\$606
Weighted regression	2.45	0.11	\$560	\$560
Fixed effects regression (HC1)	2.75	0.13	\$425	\$601

Table 1.1: *Results of NHANES (full matching), and Lalonde (CEM) for different estimation strategies.*

be typically done for large block experiments to obtain inference for the average effect in the presence of treatment variation. The average treatment effect estimate is easy to obtain; it is the uncertainty estimation that causes the trouble. Our small block variance estimators fills this gap. To illustrate, we analyze a data set from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 given in the `CrossScreening` package (Rosenbaum and Zhao, 2017) in `R` statistical software (R Core Team, 2017). This data set was also used by Zhao et al. (2018b) to analyze the effect of high fish consumption (defined as 12 or more servings of fish or shellfish in the previous month) versus low fish consumption (defined as 0 or 1 servings of fish or shellfish in the previous month) on a number of biomarkers.

Although Zhao et al. (2018b) analyzed numerous outcomes, we focus on a measure of mercury (LBXGM), converted to the $\log_2$ scale, as a simple illustration of our methods. We use unrestricted full matching to obtain a set of all small blocks of varying size. As in Zhao et al. (2018b), we matched on smoking, age, gender, race, income, and education. We used Bayesian logistic regression through the `brglm` package (Kosmidis, 2017) and `optmatch` (Hansen and Klopfer, 2006) in `R` (R Core Team, 2017). This resulted in 197 blocks with only one treated or one control unit in each. Sizes of blocks ranged from 2 to 47. This type of matching would fall into the category of flexible blocks, and we here focus on estimation of the SATE for the finite sample.

There were some block sizes that were unique, so the hybrid estimator with $\widehat{\sigma}^2_{(SMALL/m)}$ could not be used. Alternate forms of full matching could potentially avoid this concern: full matching can include additional restrictions, such as using only a portion of the control

group or exact matching on some important covariates, which could make the block sizes more homogenous (Hansen and Klopfer, 2006); for simplicity we do not explore this here. The blocking treatment effect estimate ($\widehat{\tau}_{(BK)}$) was 2.45 but using a fixed effects model with no interaction the treatment effect estimate was 2.75. Looking at Table 1.1, we see that our hybrid estimator using $\widehat{\sigma}^2_{(SMALL/p)}$ gave a much larger variance estimate (relative to the scale of the precision estimates) than the two linear model based variance estimators.

A second method for analyzing observational datasets where our variance estimators could be useful is coarsened exact matching (CEM). CEM coarsens covariates used to match and then exactly matches to these coarsened variables (Iacus et al., 2012). We follow the example from the vignette of the `cem` package (Iacus et al., 2016) in `R` using the most automated version of CEM on the classic LaLonde data set (LaLonde, 1986), available in the `cem` package. This data set consists of individuals who received or did not receive a job training program with the outcome of interest as earnings in 1978. We use the unmodified version of the LaLonde dataset, but otherwise follow the automated process for CEM laid out in the vignette to create blocks (we do not follow the analysis). This resulted in the creation of 69 blocks, some small and some big, with some ungrouped units being dropped. The blocking treatment effect estimate ($\widehat{\tau}_{(BK)}$) was \$560 but using a fixed effects model with no interaction the treatment effect estimate was \$425. From Table 1.1, the precision estimates from all methods were similar though, again, the hybrid estimator using $\widehat{\sigma}^2_{(SMALL/p)}$ was the largest and likely the most conservative.

1.8 Discussion

Blocking can be viewed under a wide variety of population frameworks ranging from a fixed, finite-sample model to one where we envision the units as being sampled from a larger population in pre-set groups. Because some findings regarding blocking change depending on what framework is used, the current literature can seem confusing and contradictory. Furthermore, because different types of blocking tend to use different frameworks, there is a lack of clarity on how one should proceed when faced with a randomized trial containing

blocks of all different sizes.

We have worked to clarify these subtleties and to fill this gap. We identified and compared the true variance of a blocking-based estimator under multiple settings, and created corresponding estimators of the impact estimator's variance. We also provide simple, model-free variance estimators for two types of experiments that have not received much attention: blocked experiments with variable-sized blocks containing singleton treatment or control units, and hybrid blocked experiments with large and small blocks combined. These contexts are quite common, frequently appearing in, for example, the matching literature. We analyzed the performance of both our new variance estimators and the classic variance estimators under different frameworks, identifying when they are unbiased or conservative. This investigation again illustrates how different sampling frameworks and block types can impact assessments of an estimator's performance.

We also carefully compared complete randomization to blocking, identifying that prior literature has often collapsed the sampling step and randomization step. Overall, we showed that blocking often, but not always, improves precision, and that guarantees about blocking depend on the framework adopted. Blocking will not reduce precision, compared to a complete randomization, when working in the stratified sampling framework with equal proportion of units treated across blocks, no matter how small the blocks are or how poorly they separate the units. Similarly, we found that in the simple random sampling setting, given a fixed algorithm for creating flexible blocks, and when blocking on covariates that are independent of potential outcomes, the blocking estimator will have equal variance to the estimator under complete randomization. In the other two main frameworks considered, however, blocking is not guaranteed to reduce variance. Without prior knowledge about the distribution of potential outcomes, it is impossible to know before an experiment is conducted whether blocking will improve the estimates in these latter two settings. This being said, the above assumes the blocks have equal proportions of units treated; if the proportions treated differ, we do not have guarantees that blocking will reduce variance, regardless of framework.

We do not discuss in this chaper two other issues which would be useful areas of further investigation: degrees of freedom and stability of variance estimators. The tradeoff between increased precision and reduced degrees of freedom by using blocked or matched pairs designs has been noted by others (e.g., Box et al., 2005, p. 93; Imbens, 2011; Snedecor and Cochran, 1989, p. 101) and is an important practical limitation to consider when using these designs. The variability of our variance estimators is an important aspect in determining the best design and analysis to use. Although we touch on the stability of our variance estimators in the Appendix, a more in depth exploration is needed. One might think, given the above, to implement blocking to realize some gains, but then analyze as a completely randomized experiment to avoid these concerns. Unfortunately, this does not result in a guaranteed valid or conservative analysis: even when proportion treated across blocks is constant, such a move is not necessarily conservative, depending on the framework adopted (see Appendix A.9). Future work should investigate how the real costs of degrees of freedom loss and instability in variance estimation depend on the experimental design within these frameworks.

Future work includes extending these results to other population settings and sampling methods, in particular finding small block estimators for the setting of constructing blocks post-sampling and pre-randomization. Variance estimation is also a missing and needed piece in post-stratification research, as noted in Miratrix et al. (2013). Although conditional answers for post-stratification would correspond to the estimators presented in this work, the unconditional case remains open.

Chapter 2

Conditional As-If Analyses in Randomized Experiments

2.1 Introduction

It is a long-standing idea in statistics that the design of an experiment should inform its analysis. Fisher placed the physical act of randomization at the center of his inferential theory, enshrining it as "the reasoned basis" for inference (Fisher, 1935). Building on these insights, Kempthorne (1955) proposed a randomization theory of inference from experiments, in which inference follows from the precise randomization mechanism used in the design. This approach has gained popularity in the causal inference literature because it relies on very few assumptions (Splawa-Neyman et al., 1990; Imbens and Rubin, 2015).

Yet the injunction to 'analyze as you randomize' is not always followed in practice, as noted by Senn (2004) who argues that in clinical trials the analysis does not always follow strictly from the randomization performed. For instance, a Bernoulli randomized experiment might be analyzed as if it were a completely randomized experiment, or we might analyze a completely randomized experiment as if it had been stratified.

This chapter studies such 'as-if' analyses in detail in the context of Neymanian inference, and makes three contributions. First we formalize the notion of 'as-if' analyses, motivating

their usefulness and proposing a rigorous validity criterion (Section 2.2). Our framework is grounded in the randomization-based approach to inference. In the two examples we described above, the analysis conditions on some aspect of the observed assignment; for instance, in the first example, the complete randomization is obtained by fixing the number of treated units to its observed value. The idea that inference should be conditional on quantities that affect the precision of estimation is not new in the experimental design literature (e.g., Cox, 1958, 2009) or the larger statistical inference literature (e.g., Särndal et al., 1989; Sundberg, 2003), and it has been reaffirmed recently in the causal inference literature (Branson and Miratrix, 2019; Hennessy et al., 2016). Our second contribution is to show that in our setting, conditioning leads to valid 'as-if' analyses. We also warn against a dangerous pitfall: some 'as-if' analyses look conditional on the surface, but are in fact neither conditional nor valid. This is the case, for instance, of analyzing a completely randomized experiment by conditioning on the covariate balance being no worse than that of the observed assignment (Section 2.3). Our third contribution is to show how our ideas can be used to suggest new methods (Section 2.4) and also show how they can be used to evaluate existing methods (Section 2.5). Throughout we will focus on building the foundational theory for the 'as-if' method of analysis.

2.2 As-if confidence procedures

2.2.1 Setup

Consider N units and let $Z_i \in \{0,1\}$ $(i = 1,\ldots,N)$ be a binary treatment indicator for unit i. We adopt the potential outcomes framework (Rubin, 1974; Splawa-Neyman et al., 1990), where, under the Stable Unit Treatment Value Assumption (Rubin, 1980), each unit has two potential outcomes, one under treatment, $Y_i(1)$, and one under control, $Y_i(0)$, and the observed response is $Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. We denote by $\boldsymbol{Z}$, $\boldsymbol{Y(1)}$ and $\boldsymbol{Y(0)}$ the vectors of binary treatment assignments, treatment potential outcomes and control potential outcomes for all N units. Let τ be our estimand of interest, which can be any

function of the potential outcome vectors $(\boldsymbol{Y}(\mathbf{0}), \boldsymbol{Y}(\mathbf{1}))$. An estimator $\hat{\tau}$ is a function of the assignment vector $\boldsymbol{Z}$ and the observed outcomes. For clarity, we will generally write $\hat{\tau} = \hat{\tau}(\boldsymbol{Z})$ to emphasize the dependence of $\hat{\tau}$ on $\boldsymbol{Z}$, but keep the dependence on the potential outcomes implicit. Denote by η_0 the design describing how treatment is allocated, so for any particular assignment $\boldsymbol{Z}'$, $\eta_0(\boldsymbol{Z}')$ gives the probability of observing $\boldsymbol{Z}'$ under design η_0. In randomization-based inference, we consider the potential outcomes $(\boldsymbol{Y}(\mathbf{0}), \boldsymbol{Y}(\mathbf{1}))$ as fixed and partially unknown quantities; the randomness comes exclusively from the assignment vector $\boldsymbol{Z}$ following the distribution η_0. The estimand τ is therefore fixed because it is a function of the potential outcomes only, while the observed outcomes and the estimator $\hat{\tau}(\boldsymbol{Z})$ are random because they are functions of the random assignment vector $\boldsymbol{Z}$.

Our focus is on the construction of confidence intervals for the estimand τ under the randomization-based perspective. We define a confidence procedure as a function C mapping any assignment $\boldsymbol{Z} \in \mathcal{Z}_{\eta_0}$ and associated vector of observed outcomes, $\boldsymbol{Y}^{obs}$, to an interval in $\mathbb{R}$, where $\mathcal{Z}_{\eta_0} = \{\boldsymbol{Z} \in \{0,1\}^N : \eta_0(\boldsymbol{Z}) > 0\}$ is the support of the randomization distribution. Standard confidence intervals are examples of confidence procedures, and are usually based on an approximate distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$. For careful choices of $\hat{\tau}$ and η_0, the random variable $\hat{\tau}(\boldsymbol{Z}) - \tau$, standardized by variance $\text{var}_{\eta_0}(\hat{\tau})$ induced by the design η_0, is asymptotically standard normal (see Li and Ding, 2017). We can then construct an interval

$$\left[\hat{\tau}(\boldsymbol{Z}) - 1.96\sqrt{\hat{V}_{\eta_0}(\boldsymbol{Z})}, \hat{\tau}(\boldsymbol{Z}) + 1.96\sqrt{\hat{V}_{\eta_0}(\boldsymbol{Z})}\right], \tag{2.1}$$

where $\hat{V}_{\eta_0}$ is an estimator of $\text{var}_{\eta_0}(\hat{\tau})$. Discussing the validity of these kinds of confidence procedures is difficult for two reasons. First, they are generally based on asymptotics, so validity in finite populations can only be approximate. Second, they use the square root of estimates of the variance which, in practice, tend to be biased. These two issues obscure the conceptual underpinning of 'as-if' analysis. We circumvent these issues by focusing instead on oracle confidence procedures, which are based on the true quantiles of the distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$ induced by the design η_0. Specifically, we consider $1 - 2\alpha$ level confidence

intervals ($0 < \alpha < 1$) of the form

$$C(\hat{\tau}(\boldsymbol{Z}); \eta_0) = [\hat{\tau}(\boldsymbol{Z}) - U_\alpha(\eta_0), \hat{\tau}(\boldsymbol{Z}) - L_\alpha(\eta_0)] \tag{2.2}$$

where $U_\alpha(\eta_0)$ and $L_\alpha(\eta_0)$ are the α upper and lower quantiles, respectively, of the distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$ under design η_0. Because they don't depend on $\boldsymbol{Z}$, the quantiles $U_\alpha(\eta_0)$ and $L_\alpha(\eta_0)$ are fixed. The confidence procedure in Equation (2.2) is said to be an oracle procedure because unlike the interval in Equation (2.1), it cannot be computed from observed data. Oracle procedures allow us to set aside the practical issues of approximation and estimability to focus on the essence of the problem. We discuss some of the practical issues that occur without oracles in supplementary material B.2.

2.2.2 As-if confidence procedures

Given data from an experiment, it is natural to consider the confidence procedure $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ constructed with the design η_0 that was actually used to randomize the treatment assignment. Consider, however, the oracle procedure

$$C(\hat{\tau}(\boldsymbol{Z}); \eta) = [\hat{\tau}(\boldsymbol{Z}) - U_\alpha(\eta), \hat{\tau}(\boldsymbol{Z}) - L_\alpha(\eta)],$$

based on the distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$ induced by some other design η that assigns positive probability to the observed assignment. In this case we say that the experiment is analyzed 'as-if' it were randomized according to η. We generalize this idea further by allowing the design η used in the oracle procedure to vary depending on the observed assignment. This can be formalized with the concept of a design map.

Definition 1 (Design map)**.** *Let $\mathcal{D}$ be the set of designs, or probability distributions, with support in $\{0,1\}^N$. A function $H : \boldsymbol{\mathcal{Z}}_{\eta_0} \to \mathcal{D}$ which maps each assignment $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0}$ to a design $H(\boldsymbol{Z}) \in \mathcal{D}$ is a design map.*

A confidence procedure $C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))$ can then be constructed using design map H as follows:

$$C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z})) = [\hat{\tau}(\boldsymbol{Z}) - U_\alpha(H(\boldsymbol{Z})), \hat{\tau}(\boldsymbol{Z}) - L_\alpha(H(\boldsymbol{Z}))].$$

This is an instance of 'as-if' analysis, in which the design used to analyze the data depends on the observed assignment. That is, while we traditionally have one rule for how to create a confidence interval, we may now have many rules, possibly as many as $|\boldsymbol{\mathcal{Z}}_{\eta_0}|$, specified via the design map. In the special case in which the design map H is constant, i.e $H(\boldsymbol{Z}) = \eta$ for all $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0}$, we write $C(\hat{\tau}(\boldsymbol{Z}); \eta)$ instead of $C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))$, with a slight abuse of notation. Note that the design map function itself is fixed before observing the treatment assignment.

Example 1. *Consider an experiment run according to a Bernoulli design with probability of treatment $\pi = 0.5$, where we remove assignments with no treated or control units, to ensure that our estimator is defined. That is, $\boldsymbol{\mathcal{Z}}_{\eta_0}$ is the set of all assignments such that at least one unit receives treatment and one unit receives control, to ensure the estimator is defined, and $\eta_0 = Unif(\boldsymbol{\mathcal{Z}}_{\eta_0})$. Let η_k ($k = 1, \ldots, N-1$) be the completely randomized design with k treated units. We use the design map $H(\boldsymbol{Z}) = \eta_{N_1(\boldsymbol{Z})}$ where $N_1(\boldsymbol{Z}) = \sum_i^N Z_i$. That is, we analyze the Bernoulli design as if it were completely randomized, with the observed number of treated units assigned to treatment. Consider the concrete case where $N = 10$. Suppose that we observe $\boldsymbol{Z}^{obs}$ with $N_1(\boldsymbol{Z}^{obs}) = 3$. The design $H(\boldsymbol{Z}^{obs}) = \eta_3$ corresponds to complete randomization with 3 units treated out of 10, and the confidence procedure $C(\boldsymbol{Z}^{obs}; H(\boldsymbol{Z}^{obs}))$ is constructed using the distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$ induced by η_3. We analyze as if a completely randomized design with 3 treated units had actually been run. If instead we observe $\boldsymbol{Z}^{obs}$ with $N_1(\boldsymbol{Z}^{obs}) = 4$, then the confidence procedure $C(\boldsymbol{Z}^{obs}; H(\boldsymbol{Z}^{obs}))$ would be constructed by considering the distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$ induced by η_4.*

Example 2. *Let our units have a single categorical covariate X_i. Then this covariate forms blocks; that is the units in the sample are partitioned based on their covariate value. Assume that the actual experiment was run using complete randomization with fixed number of treated units N_1 but discarding assignments where there is not at least one treated unit and one control unit in each block. That is, $\boldsymbol{\mathcal{Z}}_{\eta_0}$ is the set of all assignments with N_1 treated units such that at least one unit receives treatment and one unit receives control in each block and $\eta_0 = Unif(\boldsymbol{\mathcal{Z}}_{\eta_0})$. This restriction on the assignment space is to account for the associated blocked estimator being undefined. However, with moderate size blocks we can ignore this nuisance event due to low probability. For vector $\boldsymbol{k}$ whose jth entry is an integer strictly less than the size of the jth block and strictly greater than 1, let $\eta_{\boldsymbol{k}}$ be a*

block randomized design with the number of treated units in each block corresponding to the numbers given in vector $\boldsymbol{k}$*. We use the design map* $H(\boldsymbol{Z}) = \eta_{N_{1,block}(\boldsymbol{Z})}$ *where* $N_{1,block}(\boldsymbol{Z})$ *is the vector that gives the number of treated units within each block and* $\eta_{N_{1,block}(\boldsymbol{Z})}$ *is the blocked design corresponding to* $N_{1,block}(\boldsymbol{Z})$*. We use post-stratification (Holt and Smith, 1979; Miratrix et al., 2013) to analyze this completely randomized design as if it were block randomized.*

Example 3. *Assume that the actual experiment was run using complete randomization with exactly* N_1 *units treated. That is,* $\boldsymbol{\mathcal{Z}}_{\eta_0}$ *is the set of all assignments such that* $N_1 < N$ *units receive treatment and* $\eta_0 = Unif(\boldsymbol{\mathcal{Z}}_{\eta_0})$*. Let* X_i *be a continuous covariate for each unit* i*. Define a covariate balance measure* $\Delta_X(\boldsymbol{Z})$*, e.g.*

$$\Delta_X(\boldsymbol{Z}) = \frac{1}{N_1}\sum_i Z_i X_i - \frac{1}{N - N_1}\sum_i (1 - Z_i) X_i.$$

For an assignment $\boldsymbol{Z}$*, define* $\mathcal{A}(\boldsymbol{Z}) = \{\boldsymbol{Z}' : |\Delta_X(\boldsymbol{Z}')| \leq |\Delta_X(\boldsymbol{Z})|\}$*, the set of assignments with covariate balance better than or equal to the observed covariate balance. We use the design map* $H : \boldsymbol{Z}' \rightarrow pr_{\eta_0}\{\boldsymbol{Z} \mid \boldsymbol{Z} \in \mathcal{A}(\boldsymbol{Z}')\}$*. Then this design map leads to analyzing the completely randomized design as if it were rerandomized (see Morgan and Rubin, 2012, for more on rerandomization) with acceptable covariate balance cut-off equal to that of the observed covariate balance.*

Example 4. *Assume that the actual experiment was run using block randomization where* $N_{1,block}(\boldsymbol{Z})$ *is a fixed vector that gives the number of treated units within each block. That is,* $\boldsymbol{\mathcal{Z}}_{\eta_0}$ *is the set of all assignments such that number of treated units in each block is* $N_{1,block}(\boldsymbol{Z})$ *and* $\eta_0 = Unif(\boldsymbol{\mathcal{Z}}_{\eta_0})$*. Let* η *correspond to a completely randomized design, as laid out in Example 2, with* N_1*, the total number of units treated across all blocks, treated units. We use the design map* $H(\boldsymbol{Z}) = \eta$*. This corresponds to analyzing this block randomized design as if it were completely randomized .*

Throughout the chapter, we focus on settings in which the same estimator is used in the original analysis and in the 'as-if' analysis. In practice, the two analyses might employ different estimators. For instance, in Example 2, we might analyze the completely randomized experiment with a difference-in-means estimator, but use the standard blocking estimator to analyze the 'as-if' stratified experiment. We discuss this point further in the

supplementary material B.2 but in the rest of this article, we fix the estimator and focus on the impact of changing only the design.

2.2.3 Validity, relevance and conditioning

We have formalized the concept of an 'as-if' analysis, but we have not yet addressed an important question: why should we even consider such analyses instead of simply analyzing the way we randomize? Before we answer this question, we first introduce a minimum validity criterion for 'as-if' procedures.

Definition 2 (Valid confidence procedure)**.** *Fix $\gamma \in [0,1]$. Let $\eta_0 \in \mathcal{D}$ be the design used in the original experiment and let H be a design map on $\boldsymbol{\mathcal{Z}}_{\eta_0}$. The confidence procedure $C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))$ is said to be valid with respect to η_0, or η_0-valid, at level γ if $pr_{\eta_0}\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))\} \geq \gamma$. When a procedure is valid at all levels, we simply say that it is η_0-valid.*

This criterion is intuitive: a confidence procedure is valid if its coverage is as advertised over the original design. The following simple result formalizes the popular injunction to "analyze the way you randomize" (p. 317 Lachin, 1988):

Proposition 2.2.3.1. *The procedure $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ is η_0-valid.*

Given that the procedure $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ is η_0-valid, why should we consider alternative 'as-if' analyses, even valid ones? That is, having observed $\boldsymbol{Z}$, why should we use a design $H(\boldsymbol{Z})$ to perform the analysis, instead of the original η_0? A natural, but only partially correct, answer would be that the goal of an 'as-if' analysis is to increase the precision of our estimator and obtain smaller confidence intervals while maintaining validity. After all, this is the purpose of restricted randomization approaches when considered at the design stage. For instance, if we have reasons to believe that a certain factor might affect the responses of experimental units, stratifying on this factor will reduce the variance of our estimator. This analogy, however, is misleading. The primary goal of an 'as-if' analysis is not to increase the precision of the analysis but to increase its relevance. In fact, we argue heuristically in supplementary material B.4 that an 'as-if' analysis will generally not increase power.

Informally, an observable quantity is relevant if a reasonable person cannot tell if a confidence interval will be too long or too short given that quantity. The concept of relevance captures the idea that our inference should be an accurate reflection of our uncertainty given the information from the realized randomization; it should be more precise if our uncertainty is lower and less precise if our uncertainty is higher. Defining the concept of relevance formally is difficult. See Buehler (1959) and Robinson (1979) for a more formal treatment or supplementary material B.3 for a precise discussion in the context of betting games, following Buehler (1959). But for this chapter we will not attempt a formal definition and instead, following Liu and Meng (2016), we will illustrate its essence with a simple example.

Consider the Bernoulli design scenario of Example 1. From Equation (2.2), the oracle interval $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ constructed from the original design has the same length regardless of the assignment vector $\boldsymbol{Z}$ actually observed. Yet, intuitively, an observed assignment vector with 1 treated unit and 99 control units should lead to less precise inference, and therefore wider confidence intervals than a balanced assignment vector with 50 treated units and 50 control units. In a sense, the confidence interval $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ is too narrow if the observed assignment is severely imbalanced, too wide if it is well balanced, but right overall. Let $\boldsymbol{\mathcal{Z}}_1$ be the set of all assignments with a single treated unit and $\boldsymbol{\mathcal{Z}}_{50}$ the set of all assignments with 50 treated units. If the confidence interval $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ has level γ, we expect

$$\text{pr}_{\eta_0}\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_0) \mid \boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_1\} \leq \gamma; \quad \text{pr}_{\eta_0}\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_0) \mid \boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{50}\} \geq \gamma;$$

where, in general, the inequalities are strict. See the supplementary material B.3 for a proof in a concrete setting. More formally, we say that the procedure is valid marginally, but not conditional on the number of treated units. This example is illustrated in Figure 2.1, which shows that the coverage is below 0.95 if the proportion of treated units is not close to 0.5, and above 0.95 if the proportion is around 0.5. To remedy this, we should use wider confidence intervals in the first case, and narrower ones in the second. The right panel of Figure 2.1 shows that in this case, a large fraction of assignments have a proportion of

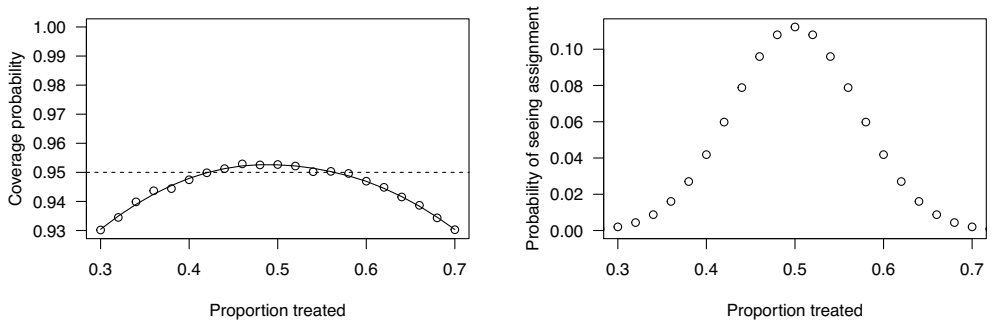

Figure 2.1: *Left: conditional coverage for a Bernoulli experiment with 100 units each having probability 0.5 of being treated. Right: distribution of the proportion of treated units for this Bernoulli experiment.*

treated units close to 0.5, therefore the confidence interval $C(\hat{\tau}(\boldsymbol{Z}); \eta_0)$ will be too large for many realizations of the design η_0. In summary, our confidence interval should be relevant to the assignment vector actually observed, and reflect the appropriate level of uncertainty. In the context of randomization-based inference, this takes the form of an 'as-if' analysis.

The concept of relevance and its connection to conditioning has a long history in statistics. Cox (1958) gives a dramatic example of a scenario in which two measuring instruments have widely different precisions, and one of them is chosen at random to measure a quantity of interest. He argues that the relevant measure of uncertainty is that of the instrument actually used, not that obtained by marginalizing over the choice of the instrument. In other words, the analysis should be conditional on the instrument actually chosen. This is an illustration of the conditionality principle (Birnbaum, 1962). In the context of randomization-based inference, this conditioning argument leads to valid 'as-if' analyses, as we show in Section 2.3. An important complication, explored in Section 2.3, is that conditional 'as-if' analyses are only a subset of possible 'as-if' analyses, and while the former are guaranteed to be valid, the latter enjoy no such guarantees.

2.3 Conditional as-if analyses

2.3.1 Conditional design maps

We define a conditional as-if analysis as an analysis conducted with a conditional design map as defined below.

Definition 3. *[Conditional design map] Consider an experiment with design η_0. Take any function $w : \boldsymbol{\mathcal{Z}}_{\eta_0} \to \Omega$, for some set Ω, and for $\omega \in \Omega$ define the design $\eta_\omega \in \mathcal{D}$ as $\eta_\omega(\boldsymbol{Z}') = pr_{\eta_0}\{\boldsymbol{Z}' \mid w(\boldsymbol{Z}') = \omega\}$. Then a function $H : \boldsymbol{\mathcal{Z}}_{\eta_0} \to \mathcal{D}$ such that $H(\boldsymbol{Z}) = \eta_{w(\boldsymbol{Z})}$ is called a conditional design map.*

It is easy to verify that a conditional design map also satisfies Definition 1. For $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0}$, $H(Z)$ is a design, not the probability of $\boldsymbol{Z}$ under some design. The probability of assignment $\boldsymbol{Z}' \in \boldsymbol{\mathcal{Z}}_{\eta_0}$ under design $H(\boldsymbol{Z})$ is $H(\boldsymbol{Z})(\boldsymbol{Z}') = \text{pr}_{\eta_0}\{\boldsymbol{Z}' \mid w(\boldsymbol{Z}') = w(\boldsymbol{Z})\}$, where $\boldsymbol{Z}$ is fixed and the probability is that induced by η_0. We introduced the shorthand $H(\boldsymbol{Z})(\boldsymbol{Z}') = \eta_{w(\boldsymbol{Z})}(\boldsymbol{Z}')$ in order to ease the notation.

For an alternative perspective on conditional design maps, notice that any function $w : \boldsymbol{\mathcal{Z}}_{\eta_0} \to \Omega$ induces a partition $\mathcal{P}_w = \{\boldsymbol{\mathcal{Z}}_\omega\}_{\omega \in \Omega}$ of the support $\boldsymbol{\mathcal{Z}}_{\eta_0}$, where $\boldsymbol{\mathcal{Z}}_\omega = \{\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0} : w(\boldsymbol{Z}) = \omega\}$. The corresponding conditional design map would then, for a given $\boldsymbol{Z}$, restrict and renormalize the original η_0 to the $\boldsymbol{\mathcal{Z}}_\omega$ containing $\boldsymbol{Z}$. An important note is that the mapping function, and therefore the partitioning of the assignment space, must be fixed before observing the treatment assignment.

Example 1 (cont.)**.** *The design map $H(\boldsymbol{Z})$ in Example 1 is a conditional design map, with $w(\boldsymbol{Z}) = N_1(\boldsymbol{Z})$. Here we partition the assignments by the number of treated units.*

Example 2 (cont.)**.** *The design map $H(\boldsymbol{Z})$ in Example 2 is a conditional design map, with $w(\boldsymbol{Z}) = N_{1,block}(\boldsymbol{Z})$. Here we partition the assignments by the vector of the number of treated units in each block .*

While Definition 3 implies Definition 1, the converse is not true: some design maps are not conditional. For instance, the design maps we consider in Examples 3 and Example 4

are not conditional, as will be discussed in Section 2.3.2. We can now state our main validity result.

Theorem 2.3.1.1. *Consider a design η_0 and a function $w : \boldsymbol{\mathcal{Z}}_{\eta_0} \to \Omega$. Then an oracle procedure built with the conditional design map $H(\boldsymbol{Z}) = \eta_{w(\boldsymbol{Z})}$ is η_0-valid.*

Proof of Theorem 2.3.1.1 is provided in supplementary material B.1. In fact, the intervals obtained are not just valid marginally; they are also conditionally valid within each $\boldsymbol{\mathcal{Z}}_\omega \in \mathcal{P}_w$, in the sense that

$$\mathrm{pr}_{\eta_0}\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z})) \mid w(\boldsymbol{Z}) = k\} = \gamma$$

for any $k \in \Omega$. Conditional validity implies unconditional validity because if we have valid inference for each partition of the assignment space then we will have validity over all partitions. Conditional validity further implies increased relevance, at least with respect to function w. We discuss this connection more in supplementary material B.3 in the context of betting games.

Corollary 2.3.1.1. *Let $\mathcal{P} = \{\boldsymbol{\mathcal{Z}}_{(1)}, \ldots, \boldsymbol{\mathcal{Z}}_{(K)}\}$ be a partition of the set of all possible assignments, $\boldsymbol{\mathcal{Z}}_{\eta_0}$. That is,*

$$w(\boldsymbol{Z}) = \sum_{k}^{K} k \mathbb{1}\{\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{(k)}\}.$$

Then $w(\boldsymbol{Z}) : \boldsymbol{\mathcal{Z}}_{\eta_0} \to \{1, \ldots, K\}$ indexes the partition that $\boldsymbol{Z}$ is in. Using $H(\boldsymbol{Z}) = \eta_{w(\boldsymbol{Z})}$ gives an η_0-valid procedure, as a consequence of Theorem 2.3.1.1.

Details are provided in supplementary material B.1. Corollary 2.3.1.1 states that any partition $\mathcal{P}$ of the support $\boldsymbol{\mathcal{Z}}_{\eta_0}$ induces a valid oracle confidence procedure; having observed assignment $\boldsymbol{Z} \sim \eta_0$, one simply needs to identify the unique element $\boldsymbol{\mathcal{Z}} \in \mathcal{P}$ containing $\boldsymbol{Z}$ and construct an oracle interval using the design obtained by restricting η_0 to the set $\boldsymbol{\mathcal{Z}}$.

Now that we have a formal result on the validity of conditional designs, we can discuss an added benefit of using these conditional design maps. Consider Example 1, and the corresponding discussion of relevance with Bernoulli designs in Section 2.2.3. This example provides insight into an additional potential benefit of increasing relevance in terms of replicability of studies. Under the original analysis for the Bernoulli design we would expect

that the estimates for the bad randomizations with an extreme proportion of treated units will be far from the truth. But if we do not adjust the precision of our estimators based on this information, we will not only have an estimate that is far from the truth but our confidence intervals will imply confidence in that poor estimate. Although our conditional analysis will not cover the truth the same proportion of the time as the original analysis, we would expect the length of our confidence intervals to reflect less certainty when we have a poor randomization. In terms of reproducibility, this means that we are less likely to end up being confident in an extreme result.

2.3.2 Non-conditional design maps

Theorem 2.3.1.1 states that a sufficient condition for an 'as-if' procedure to be valid is that it be a conditional 'as-if' procedure. Although this condition is not necessary, we will now show that some non-conditional 'as-if' analyses can have arbitrarily poor properties. Example 3 provides a sharp illustration of this phenomenon and, although it is an edge case, it helps build intuition for why some design maps are not valid.

Example 3 (cont.)**.** *The design map* $H(\boldsymbol{Z}) = pr_{\eta_0}\{\boldsymbol{Z}' \mid \boldsymbol{Z}' \in \mathcal{A}(\boldsymbol{Z})\}$ *introduced in Example 3 is not a conditional design map. This can be seen by noticing that the sets* $\{\mathcal{A}(\boldsymbol{Z})\}_{\boldsymbol{Z}\in\boldsymbol{\mathcal{Z}}_{\eta_0}}$ *where* $\mathcal{A}(\boldsymbol{Z}) = \{\boldsymbol{Z}' : |\Delta_X(\boldsymbol{Z}')| \leq |\Delta_X(\boldsymbol{Z})|\}$ *do not form a partition of* $\boldsymbol{\mathcal{Z}}_{\eta_0}$.

This example is particularly deceptive because the design map $H(\boldsymbol{Z})$ does involve a conditional distribution. And yet, it is not a conditional design map in the sense of Definition 3 because it does not partition the space of assignments; each assignment $\boldsymbol{Z}'$, except for the assignments with the very worst balance, will belong to multiple $\mathcal{A}(\boldsymbol{Z})$. Therefore Theorem 2.3.1.1 does not apply.

Consider the special case where covariate X_i are drawn from a continuous distribution and $Y_i(0) = Y_i(1) = X_i$ $(i = 1, \ldots, N)$ and we are interested in the average treatment effect, which is the difference in mean potential outcomes under treatment versus control. Suppose that assignments are balanced such that half the units are assigned to treatment and half are assigned to control. Then given any $\boldsymbol{Z}$, with probability one there are only

two assignments with exactly the same value for $|\Delta_X(\boldsymbol{Z})|$, $\boldsymbol{Z}$ and the assignment $1 - \boldsymbol{Z}$; see supplementary material B.1 for proof of this statement. By construction, then, our assignment is one of two worst case assignments in terms of balance for the set $\mathcal{A}(\boldsymbol{Z})$. Under the model $Y_i(0) = Y_i(1) = X_i$, the observed difference, $\hat{\tau}(\boldsymbol{Z}) - \tau = \hat{\tau}(\boldsymbol{Z})$, will be the most extreme in $\mathcal{A}(\boldsymbol{Z})$ and so would lie outside the oracle confidence interval if $2/|\mathcal{A}(\boldsymbol{Z})| \leq \gamma$, which it will generally do. Thus this design map would lead to poor coverage. In fact, we show in supplementary material B.1 that if we instead make out inequality strict and take $\mathcal{A}(\boldsymbol{Z}) = \{\boldsymbol{Z}' : |\Delta_X(\boldsymbol{Z}')| < |\Delta_X(\boldsymbol{Z})|\}$, the 'as-if' procedure of Example 3 has a coverage of 0. Intuitively, this is because the observed assignment $\boldsymbol{Z}$ always has the worst covariate balance of all assignments within the support $\mathcal{A}(\boldsymbol{Z})$. Although this is an extreme example, it illustrates the fact that 'as-if' analyses are not guaranteed to be valid if they are not conditional, and can indeed be extremely ill-behaved.

Example 4 (cont.)**.** *The design map introduced in Example 4, in which we analyze a blocked design as if it were completely randomized, is not a conditional design map. This can be seen by noticing that the complete randomization is does not partition the blocked design but rather the blocked design is a single partition of the completely randomized design.*

Thus we can show that Example 4 can also lead to invalid analysis. If this partition is a particularly bad in the sense of having wider conditional intervals over it, or higher variance of estimators, we will not have guaranteed validity using a completely randomized design for analysis. See Appendix A for further discussion on when a blocked design can result in higher variance of estimators than an unblocked design, particularly when blocking is structural, such as with clusters, and that units within blocks are heterogeneous but blocks look similar to each other.

2.3.3 How to build a better conditional analysis

The original goal of the 'as-if' analysis of Example 3 was to incorporate the observed covariate balance in the analysis to increase relevance. We have shown that the design map originally proposed was not a conditional design map. We now show how to construct

a conditional design map, and therefore a valid procedure, for this problem. The idea is to partition the support $\boldsymbol{\mathcal{Z}}_{\eta_0}$ into sets of assignments with similar covariate balance and then use the induced conditional design map, as prescribed by Corollary 2.3.1.1. Let $\{\Delta_X(\boldsymbol{Z}) : \boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0}\}$ be the set of all possible covariate imbalance values achievable by the design η_0, and $\mathcal{G} = \{\mathcal{G}_{(1)}, \ldots, \mathcal{G}_{(K)}\}$ be a partition of that set into K ordered elements. That is, for any k, k' with $k < k'$, $\delta < \delta'$ for all $\delta \in \mathcal{G}_{(k)}$, $\delta' \in \mathcal{G}_{(k')}$. This induces a partition $\mathcal{P} = \{\boldsymbol{\mathcal{Z}}_{(1)}, \ldots, \boldsymbol{\mathcal{Z}}_{(K)}\}$ of $\boldsymbol{\mathcal{Z}}_{\eta_0}$, where

$$\boldsymbol{\mathcal{Z}}_{(k)} = \{\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0} : \Delta_X(\boldsymbol{Z}') \in \mathcal{G}_{(k)}\}.$$

We can therefore apply the results of Corollary 2.3.1.1. This approach is similar in spirit to the conditional randomization tests proposed by Rosenbaum (1984); see also Branson and Miratrix (2019) and Hennessy et al. (2016). The resulting 'as-if' analysis improves on the original analysis under η_0 by increasing its relevance. Indeed, suppose that the observed assignment has covariate balance δ. Then the confidence interval constructed using η_0 will involve all of the assignments in $\boldsymbol{\mathcal{Z}}_{\eta_0}$, including some whose covariate balances differ sharply from δ. In contrast, the procedure we just introduced restricts the randomization distribution to a subset of assignments $\boldsymbol{\mathcal{Z}}_{(k)}$ containing only assignments with balance close to δ.

This does not, however, completely solve the original problem. Suppose, for instance, that by chance, $\Delta_X(\boldsymbol{Z}) = \max \mathcal{G}_{(k)}$. By definition, the randomization distribution of the 'as-if' analyses we introduced above will include the assignment $\boldsymbol{Z}'$ such that $\Delta_X(\boldsymbol{Z}') = \min \mathcal{G}_{(k)}$, but not $\boldsymbol{Z}^*$ such that $\Delta_X(\boldsymbol{Z}^*) = \min \mathcal{G}_{(k+1)}$ even though $\boldsymbol{Z}^*$ might be more relevant to $\boldsymbol{Z}$ than $\boldsymbol{Z}'$, in the sense that we may have $|\Delta_X(\boldsymbol{Z}) - \Delta_X(\boldsymbol{Z}^*)| \leq |\Delta_X(\boldsymbol{Z}) - \Delta_X(\boldsymbol{Z}')|$. This issue does not affect validity, but it raises concerns about relevance when the observed assignment is close to the boundary of a set $\boldsymbol{\mathcal{Z}}_{(k)}$. Informally, we would like to choose $\boldsymbol{\mathcal{Z}}_{(k)}$ in such a way that the observed assignment $\boldsymbol{Z}$ is at the center of the set, as measured by covariate balance. For instance, fixing $c > 0$, we would like to construct an 'as-if' procedure that randomizes within a set of the form $\mathcal{B}(\boldsymbol{Z}) = \{\boldsymbol{Z}' : \Delta_X(\boldsymbol{Z}') \in [\Delta_X(\boldsymbol{Z}) - c, \Delta_X(\boldsymbol{Z}) + c]\}$, rather than $\boldsymbol{\mathcal{Z}}_{(k)}$.

A naive approach would be to use the design mapping $H : \boldsymbol{Z} \to \text{pr}_{\eta_0}\{\boldsymbol{Z}' \mid \boldsymbol{Z}' \in \mathcal{B}(\boldsymbol{Z})\}$, but this is not a conditional design mapping. Branson and Miratrix (2019) discussed a similar approach in the context of randomization tests and also noted that it was not guaranteed to be valid.

Let's explore further why $H : \boldsymbol{Z} \to \text{pr}_{\eta_0}\{\boldsymbol{Z}' \mid \boldsymbol{Z}' \in \mathcal{B}(\boldsymbol{Z})\}$ does not have guaranteed validity. In this case, each assignment vector has an interval or window of acceptable covariate balances centered around it. Let assignment $\boldsymbol{Z}'$ have covariate balance within the acceptable covariate balance window for some other assignment, $\boldsymbol{Z}^*$. That is, $\boldsymbol{Z}' \in \mathcal{B}(\boldsymbol{Z}^*)$. If $\boldsymbol{Z}'$ and $\boldsymbol{Z}^*$ do not have the same covariate balance, they will have overlapping but non-identical windows. The confidence interval for $\boldsymbol{Z}^*$ is guaranteed to have $\gamma \cdot 100\%$ coverage over all assignments in $\mathcal{B}(\boldsymbol{Z}^*)$. However, there are no guarantees about which assignments in this set will result in a confidence interval covering the truth. In particular, over the smaller subset of assignments with exactly the same covariate balance as $\boldsymbol{Z}^*$, which lead to the same design over $\mathcal{B}(\boldsymbol{Z}^*)$, the coverage may be less than $\gamma \cdot 100\%$. Furthermore, although the interval built for $\boldsymbol{Z}'$ under the design for $\boldsymbol{Z}^*$ may cover the truth, we will never observe it because $\boldsymbol{Z}'$ uses a different design to construct intervals.

To build a better solution, we need more flexible tools. The following section will discuss how we can be more flexible, while still guaranteeing validity, by introducing some randomness.

2.4 Stochastic conditional as-if

The setting of Example 3 posed a problem of how to build valid procedures that allow the design mapping to vary based on the assignment. That is, we want to avoid making a strict partition of the assignment space but still guarantee validity. We can do this by introducing some randomness into our design map.

Definition 4. *[Stochastic conditional design map] Consider an experiment with design η_0. For observed assignment $\boldsymbol{Z} \sim \eta_0$, draw $w \sim m(w \mid \boldsymbol{Z})$ from a given distribution $m(\cdot \mid \boldsymbol{Z})$, indexed by*

$\boldsymbol{Z}$ with support on some set Ω, and consider the design

$$\eta_w(\boldsymbol{Z}') = pr_{\eta_0}\{\boldsymbol{Z}' \mid w\} \propto m(w \mid \boldsymbol{Z}')pr_{\eta_0}\{\boldsymbol{Z}'\}.$$

The mapping $H : \boldsymbol{Z} \to \mathcal{D}$, with $H(\boldsymbol{Z}) = \eta_w$ and $w \sim m(w \mid \boldsymbol{Z})$, is called a stochastic design mapping.

In the special case where the distribution $m(w \mid \boldsymbol{Z})$ degenerates into $\delta_{w=w(\boldsymbol{Z})} = \mathbb{1}(w = w(\boldsymbol{Z}))$, Definition 4 is equivalent to Definition 3. When m is non-degenerate, the stochastic design map H becomes a random function.

Before stating our theoretical result for stochastic design maps, we first examine how the added flexibility these maps afford can be put to use in the context of Example 3. For $c > 0$, let $\Omega = \mathbb{R}$ and define

$$m(w \mid \boldsymbol{Z}) = \text{Unif}\Big(w' : |\Delta_X(\boldsymbol{Z}) - w'| \leq c\Big) = \text{Unif}\Big(\Delta_X(\boldsymbol{Z}) - c, \Delta_X(\boldsymbol{Z}) + c\Big).$$

Note that $m(w \mid \boldsymbol{Z})$ selects a w^* around the observed imbalance. Having observed $\boldsymbol{Z} \sim \eta_0$ and drawn $w \sim m(w \mid \boldsymbol{Z})$, we then consider the design

$$H(\boldsymbol{Z}) = \text{pr}_{\eta_0}\{\boldsymbol{Z}' \mid w\} = \begin{cases} \frac{\text{pr}_{\eta_0}\{\boldsymbol{Z}'\}}{\nu(w)} & \text{if } \Delta_X(\boldsymbol{Z}') \in [w - c, w + c] \\ 0 & \text{otherwise} \end{cases}$$

where normalizing factor $\nu(w) = \text{pr}_{\eta_0}\{w\}/m(w \mid \boldsymbol{Z}) = 2c\text{pr}_{\eta_0}\{w\}$. In other words, we analyze the experiment by restricting the randomization to a set $\mathcal{A}(w) = \{\boldsymbol{Z}' : \Delta_X(\boldsymbol{Z}') \in [w - c, w + c]\}$. Now comparing $\mathcal{A}(w)$ to our original randomization set $\mathcal{B}(\boldsymbol{Z})$, we see that while $\mathcal{B}(\boldsymbol{Z})$ is a set of w imbalances centered on observed $\Delta_X(\boldsymbol{Z})$, $\mathcal{A}(w)$ is only centered on $\Delta_X(\boldsymbol{Z})$ on average over draws of w. The following theorem guarantees that the procedure is valid. The proof is in supplementary material B.1.

Theorem 2.4.0.1. *Consider a design η_0 and a variable w, with conditional distribution $m(w \mid \boldsymbol{Z})$. Then an oracle procedure built at level γ with the stochastic conditional design map $H(\boldsymbol{Z})$, which draws w^{obs} and maps to $\eta_{w^{obs}} = pr_{\eta_0}\{\boldsymbol{Z}' \mid w^{obs}\}$, is η_0-valid at level γ.*

The idea of a stochastic conditional design map mirrors that of conditioning mechanisms

introduced by Basse et al. (2019) in the context of randomization tests. Inference is also stochastic here in the sense that a single draw of w determines the final reference distribution to calculate the confidence interval. We discuss some practical challenges with this approach in supplementary material B.2.

The randomness intrinsic to this method is similar to the introduction of randomness into uniformly most powerful (UMP) tests (see Lehmann and Romano, 2005). One way around this stochastic nature, which we do not discuss here, would be in the use of fuzzy confidence intervals (see Geyer and Meeden, 2005) to report a distribution of results over all possible random w's, rather than a result based on one draw of w. In a fuzzy interval, similar to fuzzy sets, membership in the interval is not binary but rather allowed to take values in $[0, 1]$.

2.5 Discussion: Implications for Matching

In this section we illustrate how the 'as-if' framework and theory we have developed can be applied to evaluate a particular analysis method: analyzing data matched post-randomization as if it was pair randomized. Matching is a powerful tool for extracting quasi-randomized experiments from observational studies (Ho et al., 2007; Rubin, 2007; Stuart, 2010). To highlight the conceptual difficulty with post-matching analysis, we consider the idealized setting where treatment is assigned according to a known Bernoulli randomization mechanism, η_0, and matching is performed subsequently. Specifically, units are assigned to treatment independently with probability $p_i = logit^{-1}(\alpha_0 + \alpha_1 X_i)$, where $X_i = (X_{i,1}, X_{i,2}) \in \mathbb{R}^2$ is a unit-level covariate vector. In addition, for simplicity, we focus on pair matching, where treated units are paired to control units with similar covariate values. One way to analyze the pairs is as if the randomization had been a matched pairs experiment. Although this method of analysis has already received scrutiny in the literature (see, e.g., Schafer and Kang, 2008; Stuart, 2010), it is worth asking: is this a conditional design map with guaranteed validity?

If we can exactly match on X_i, then the situation is identical to that of Example 2; the as

if pair randomized design map is a conditional design map, and the procedure is therefore guaranteed to be valid. Exact matching is, however, often hard to achieve in practice. Instead, we generally rely on approximate matching, in which the covariate distance between the units within a pair is small, but not zero. Unfortunately, we will show that with approximate matching, the as if pair randomized design map is not a conditional design map. To make this formal, let R be a matching algorithm which, given an assignment and fixed covariates, returns a set of L pairs which we denote M, $M = \{(i_{1,1}, i_{1,2}), \ldots, (i_{L,1}, i_{L,2})\}$. Assuming a deterministic matching algorithm, $M^{obs} = R(\boldsymbol{Z}^{obs})$ is the matching obtained from observed assignment $\boldsymbol{Z}^{obs}$. Treating the matched data as a matched pairs experiment implies analyzing over all assignment vectors that permute the treatment assignment within the pairs. Let $\boldsymbol{Z}'$ be an assignment obtained by such a permutation within the pairs of M^{obs}. A necessary condition for pairwise randomization to be a conditional procedure is that $R(\boldsymbol{Z}') = M^{obs}$; that is, for any permutation of treatment within pairs, the matching algorithm R must return the original matches.

This condition for a valid conditional procedure is not guaranteed. To illustrate, consider the first three steps inside the light grey box in Figure 2.2, in which we consider a greedy matching algorithm. If we analyze the matched data as if it were a matched pairs design, then the permutation shown is allowable by the design. However, we see in the dotted rectangle of Figure 2.2 that if we had observed that permutation as the treatment assignment, we would have ended up matching the units differently, and therefore would have conducted a different analysis. This is essentially the issue we encountered with the non-conditioning so-called fix to the rerandomization example in Section 2.4; we have not created a partition of our space. The upshot is that when matching is not exact, analyzing the data as if it came from a paired-randomized study cannot be justified by a conditioning argument. A proper conditional analysis would need to take into account the matching algorithm. Specifically, let $\mathcal{P}(M)$ be the set of all treatment assignment vectors that are permutations of treatment within a set of matches M and let $\mathcal{V}(M)$ be the set of assignments that would

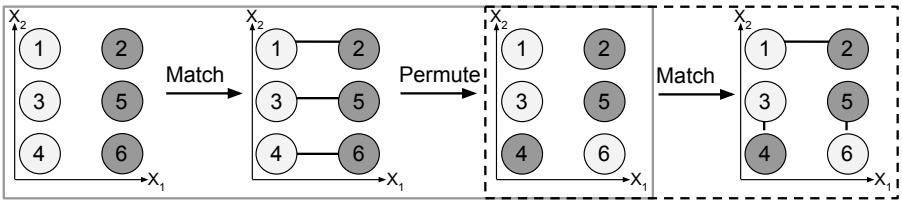

Figure 2.2: *Units are numbered in circles. Position in the graphs corresponds to covariate values. Shaded circles are treated. Lines indicate matches. The solid edge rectangle indicates original match and one permutation. The dashed edge rectangle indicates match based on that permutation.*

lead to matches M using algorithm R. Then

$$\text{pr}_{\eta_0}\left(\boldsymbol{Z}' \mid R(\boldsymbol{Z}') = R(\boldsymbol{Z})\right) = \text{pr}_{\eta_0}\left(\boldsymbol{Z}' \mid \boldsymbol{Z}' \in \mathcal{V}(R(\boldsymbol{Z}))\right).$$

Note that this is equal to $Unif\left(\mathcal{V}(R(\boldsymbol{Z}))\right)$ if assignments in $\mathcal{V}(R(\boldsymbol{Z}))$ have equal probability under the original design. If we condition on having observed a certain matched set, we would not use all within pair permutations of those matches, but rather would use all permutations that would have resulted in the same matches given our matching algorithm.

Still, this distinction appears to matter more in theory than in practice. Matching has been shown to be, in practice, a reliable tool to achieve covariate balance such that an observational study resembles an experiment and thus gives us hope of obtaining good inference.

Chapter 3

Causal Inference for Multiple Non-Randomized Treatments using Fractional Factorial Designs

3.1 Introduction

There has been a growth in the literature regarding the use of factorial designs in causal inference, though primarily focusing on randomized experiments (e.g. Branson et al., 2016; Dasgupta et al., 2015; Dong, 2015; Egami and Imai, 2019; Espinosa et al., 2016; Lu, 2016a,b; Lu and Deng, 2017; Mukerjee et al., 2018; Zhao et al., 2018a). These are settings in which we have multiple treatments applied contemporaneously to subjects and interest is in understanding not only the effect of a single treatment but also how treatments may interact. A factorial experiment involves random assignment of all possible treatment combinations to units and can be used to help understand these different effects. However, estimating the effects of multiple treatments is not exclusively relevant for experiments, it is also of interest when analyzing observational studies. In observational studies the mechanism by which units receive different treatment combinations is unknown and possibly related to unit-specific characteristics, which need to be accounted for. Regression models with interaction terms

are commonly used in observational studies to estimate the effects of multiple treatments (Bobb et al., 2015; Oulhote et al., 2017; Patel et al., 2010; Valeri et al., 2017). For instance, Bobb et al. (2015) considers a Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. However, the use of regression without a careful design phase, in which one tries to uncover or approximate some underlying experiment, can lead to incorrect conclusions (e.g., see Rubin, 2008).

We thus view the problem of estimating causal effects with multiple treatments in an observational study through an experimental design perspective. With a single binary treatment (or equivalently a single factor with two levels), conceptualizing observational studies into plausible treatment-control hypothetical randomized experiments has been a common strategy for estimating the causal effect of a treatment on an outcome (Bind and Rubin, 2019; Rosenbaum, 2002; Rubin, 2008; Stuart, 2010). This is achieved via conceptual and design stages in which one attempts to approximate a randomized experiment through methods that aim to balance treatment groups with respect to covariates (Bind and Rubin, 2019; Imbens and Rubin, 2015; Rubin, 2008). Covariate balance is important not only to limit overt biases (Rosenbaum, 2002, Chapter 3) but also to increase plausibility of assumptions that allow the analysis stage to be implemented with limited model extrapolation (Imbens and Rubin, 2015, Chapter 12). For observational data with multiple contemporaneously applied treatments, a factorial design is a natural choice of experimental design. There have been some extensions of matching techniques for a single binary treatment to multiple treatments. For instance, Lopez and Gutman (2017) review and extend current methods such as matching for obtaining causal estimates in observational studies with multiple treatments, although they focus on a single factor with multiple levels, rather than multiple treatments that may be applied contemporaneously. Nilsson (2013) considers matching using the generalized propensity score (GPS) (Hirano and Imbens, 2004) in a 2^2 factorial setting. However, further exploration of matching and other methods, such as weighting, that attempt to approximate an experiment in observational studies with multiple treatments is still needed. We discuss obtaining covariate balance further in Section 3.5.2.

An important issue with modeling multiple treatments in the observational setting, in addition to the usual concerns that we attempt to address in the design phase, is that we may have limited data or no data available for some treatment combinations. If a single treatment combination has no measurements, then the recreation of a full factorial design is not possible. Linear or additive regression models, often used in practice, will not be able to recover a full factorial design and the implicit assumptions these regression models are making about the missing treatment combinations in the analysis are not transparent, especially with respect to the implicit imputation of the missing potential outcomes. We further discuss regression estimates in such setting in Section 3.4. When there are only one or two observations for a certain treatment combination, utilizing a factorial design would rely heavily on those few individuals being representative. We propose instead to embed an observational study into a hypothetical fractional factorial experiment.

In this chapter, we discuss the estimation of the causal effects of multiple non-randomized treatments, in particular when we do not observe all treatment combinations. First, Section 3.2 reviews full factorial designs within the potential outcomes framework described in Dasgupta et al. (2015). Next, Section 3.3 reviews extensions of this framework to fractional factorial designs and expands upon current inference results. Section 3.4 explores the use of incomplete factorial designs. Then Section 3.5 examines how to embed an observational study into one of these experimental designs. Section 3.6 illustrates our method and challenges when working with observational data with multiple treatments with an application examining the effects of four pesticides on body mass index (BMI) using data from the National Health and Nutrition Examination Survey (NHANES), which is conducted through the Centers for Disease Control and Prevention (CDC). Finally, Section 3.7 concludes.

3.2 Full factorial designs

3.2.1 Set up

We work in the Rubin Causal Model (Holland, 1986), also known as the potential outcomes framework (Splawa-Neyman et al., 1990; Rubin, 1974). Additionally, we assume the Stable Unit Treatment Value Assumption (SUTVA), that is, there is no interference and no different forms of each treatment level (Rubin, 1980). Throughout this chapter, we focus on two-level factorial and fractional factorial designs. We closely follow the potential outcomes framework for 2^K factorial designs proposed by Dasgupta et al. (2015). We start by reviewing the notation and basic setup.

Let there be K two-level factors; that is, there are K treatments (e.g. medications), each having two levels (e.g. receiving a certain medication or not receiving that medication) that are assigned in combination. This creates $2^K = J$ total possible treatment combinations. Let $\boldsymbol{z}_j$ denote the j^{th} treatment combination. See Table 3.1 for an example where $K = 3$ which illustrates the notation, with treatment combinations listed in lexicographic order (the standard ordering, used here for consistency). Let $z_{j,k} \in \{-1, +1\}$ be the level of the k^{th} factor in the j^{th} treatment combination, so $\boldsymbol{z}_j = (z_{j,1}, ..., z_{j,K})$. Let there be n units in the sample. The potential outcome for unit i under treatment combination $\boldsymbol{z}_j$ is $Y_i(\boldsymbol{z}_j)$ and the sample average potential outcome under treatment $\boldsymbol{z}_j$ is $\bar{Y}(\boldsymbol{z}_j) = \sum_{i=1}^{n} Y_i(\boldsymbol{z}_j)/n$. $\bar{\boldsymbol{Y}} = (\bar{Y}(\boldsymbol{z}_1), \bar{Y}(\boldsymbol{z}_2), \ldots, \bar{Y}(\boldsymbol{z}_J))$ is the vector of mean potential outcomes for all 2^K treatment combinations.

Let $W_i(\boldsymbol{z}_j) = 1$ if unit i received treatment combination $\boldsymbol{z}_j$ and $W_i(\boldsymbol{z}_j) = 0$ otherwise. As in Dasgupta et al. (2015), $\sum_{j=1}^{J} W_i(\boldsymbol{z}_j) = 1$ and the observed potential outcome for unit i is

$$Y_i^{obs} = \sum_{j=1}^{J} W_i(\boldsymbol{z}_j) Y_i(\boldsymbol{z}_j).$$

Let there be a fixed number, n_j, of units randomly assigned to treatment combination $\boldsymbol{z}_j$. That is, we are assuming a (possibly unbalanced) completely randomized design. An

estimator of the observed average potential outcome for treatment z_j is

$$\bar{Y}^{obs}(z_j) = \frac{1}{n_j}\sum_{i=1}^{n} W_i(z_j)Y_i(z_j) = \frac{1}{n_j}\sum_{i:W_i(z_j)=1}^{n} Y_i^{obs}.$$

Denote $\bar{\boldsymbol{Y}}^{obs} = (\bar{Y}^{obs}(z_1), \bar{Y}^{obs}(z_2), \ldots, \bar{Y}^{obs}(z_J))$ as the vector of observed mean potential outcomes for all 2^K treatment combinations.

Treatment	Factor 1	Factor 2	Factor 3	Outcomes
z_1	-1	-1	-1	$\bar{Y}(z_1)$
z_2	-1	-1	+1	$\bar{Y}(z_2)$
z_3	-1	+1	-1	$\bar{Y}(z_3)$
z_4	-1	+1	+1	$\bar{Y}(z_4)$
z_5	+1	-1	-1	$\bar{Y}(z_5)$
z_6	+1	-1	+1	$\bar{Y}(z_6)$
z_7	+1	+1	-1	$\bar{Y}(z_7)$
z_8	+1	+1	+1	$\bar{Y}(z_8)$
	g_1	g_2	g_3	$\bar{\boldsymbol{Y}}$

Table 3.1: *Example of a 2^3 factorial design.*

Using the framework from Dasgupta et al. (2015), denote the contrast column j in the design matrix as g_j, as illustrated in Table 3.1. Following that paper, we can also define the contrast vector for the two-factor interaction between factors k and k' as

$$g_{k,k'} = g_k \circ g_{k'},$$

where $\circ$ indicates the Hadamard (element-wise) product. Similarly, the contrast vector for three-factor interactions is

$$g_{k,h,i} = g_k \circ g_h \circ g_i,$$

and all higher order interaction contrast vectors can be calculated analogously.

3.2.2 Estimands and estimators

Continuing to follow Dasgupta et al. (2015), define the finite population main causal effect for factor k and the finite population interaction effect for factors k and k' as

$$\tau(k) = \frac{1}{2^{K-1}} g_k^T \bar{\boldsymbol{Y}} \quad \text{and} \quad \tau(k,k') = \frac{1}{2^{K-1}} g_{k,k'}^T \bar{\boldsymbol{Y}},$$

respectively, where by finite population we mean that we are only interested in inference for the units we have in the experiment. Higher-level interaction terms are defined analogously. We can similarly define the individual-level effects as

$$\tau_i(k) = \boldsymbol{g}_k^T \boldsymbol{Y}_i \text{ and } \tau_i(k,k') = \frac{1}{2^{K-1}} \boldsymbol{g}_{k,k'}^T \boldsymbol{Y}_i,$$

where $\boldsymbol{Y}_i = (Y_i(\boldsymbol{z}_1), \ldots, Y_i(\boldsymbol{z}_J))^T$. We also define the average potential outcome across treatments as

$$\tau(0) = \frac{\sum_{i=1}^{J} \bar{Y}(\boldsymbol{z}_k)}{2^K} = \frac{1}{2^K} \boldsymbol{g}_0^T \bar{\boldsymbol{Y}},$$

where $\boldsymbol{g}_0$ is a vector of length 2^K of all $+1$'s.

There is a correspondence between the potential outcomes and the causal effects. Consider a full factorial model matrix, $\boldsymbol{G}$, that includes the mean and interactions and whose rows are comprised of $\boldsymbol{g}_k^T$, where
k $\in \{0; 1; ...; K; 1, 2; ...; K-1; K\}$. Based on the definition of our estimands in Section 3.2.2, the matrix $\boldsymbol{G}$ relates the mean potential outcomes and the factorial effects as follows:

$$2^{-(K-1)} \underbrace{\begin{bmatrix} \boldsymbol{g}_0^T \\ \boldsymbol{g}_1^T \\ \vdots \\ \boldsymbol{g}_K^T \\ \boldsymbol{g}_{1,2}^T \\ \vdots \\ \boldsymbol{g}_{K-1,K}^T \\ \vdots \\ \boldsymbol{g}_{1,2,\ldots,K-1,K}^T \end{bmatrix}}_{\boldsymbol{G}} \underbrace{\begin{bmatrix} \bar{Y}(\boldsymbol{z}_1) \\ \bar{Y}(\boldsymbol{z}_2) \\ \vdots \\ \bar{Y}(\boldsymbol{z}_J) \end{bmatrix}}_{\bar{\boldsymbol{Y}}} = \underbrace{\begin{bmatrix} 2\tau(0) \\ \tau(1) \\ \vdots \\ \tau(K) \\ \tau(1,2) \\ \vdots \\ \tau(K-1,K) \\ \vdots \\ \tau(1,2,...,K-1,K) \end{bmatrix}}_{\boldsymbol{\tau}}.$$

Note that because of orthogonality, the inverse of $\boldsymbol{G}$ is simply $\boldsymbol{G}^T$ rescaled (i.e., $\boldsymbol{G}^{-1} = \frac{1}{2^k}\boldsymbol{G}^T$), as argued by Dasgupta et al. (2015) in the context of imputing potential outcomes under Fisher's sharp null hypothesis. The mean potential outcomes can be rewritten in terms of the factorial effects as

$$\bar{\boldsymbol{Y}} = \frac{1}{2}\boldsymbol{G}^T \boldsymbol{\tau}.$$

Let the j^{th} row of $\boldsymbol{G}^T$ be denoted by $\boldsymbol{h}_j$. The first entry of $\boldsymbol{h}_j$ is $+1$, corresponding to the mean, the next K entries are equal to the entries of $\boldsymbol{z}_j$, and the remaining entries correspond

to interactions, with the order given by the order of the rows of $\boldsymbol{G}$. For instance, in the 2^3 factorial design shown in Table 3.1, $h_1 = (+1, -1, -1, -1, +1, +1, +1, -1)$ We have

$$\bar{Y}(\boldsymbol{z}_j) = \frac{1}{2}\boldsymbol{h}_j\boldsymbol{\tau}.$$

This representation will be useful in Sections 3.4 and 3.5.3 when considering what can be estimated when we do not observe all treatment combinations.

Let us now focus on estimation. The estimator for $\tau(k)$ and $\tau(k, k')$ are defined as

$$\widehat{\tau}(k) = \frac{1}{2^{K-1}}\boldsymbol{g}_k^T\bar{\boldsymbol{Y}}^{obs} \text{ and } \widehat{\tau}(k, k') = \frac{1}{2^{K-1}}\boldsymbol{g}_{k,k'}^T\bar{\boldsymbol{Y}}^{obs},$$

respectively. Estimators for higher-level interaction terms and $\tau(0)$ are defined analogously by replacing $\bar{\boldsymbol{Y}}$ by $\bar{\boldsymbol{Y}}^{obs}$.

In Dasgupta et al. (2015), the variance of the factorial effect estimators was derived under a balanced design; however, the variance expression was extended to unbalanced designs in Lu (2016b), as follows:

$$\text{Var}\left(\widehat{\tau}(k)\right) = \frac{1}{2^{2(K-1)}}\sum_{j=1}^{J}\frac{1}{n_j}S^2(\boldsymbol{z}_j) - \frac{1}{n}S_k^2, \tag{3.1}$$

where

$$S^2(\boldsymbol{z}_j) = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i(\boldsymbol{z}_j) - \bar{Y}(\boldsymbol{z}_j)\right)^2 \quad \text{and} \quad S_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tau_i(k) - \tau(k)\right)^2.$$

An expression for the covariance between two factorial effect estimators was provided by Dasgupta et al. (2015), which again can be extended to unbalanced factorial designs (Lu, 2016b):

$$\text{Cov}\left(\hat{\tau}(k), \hat{\tau}(k')\right) = \frac{1}{2^{2(K-1)}}\left[\sum_{j:g_{kj}=g_{k'j}}\frac{1}{n_j}S^2(\boldsymbol{z}_j) - \sum_{j:g_{kj}\neq g_{k'j}}\frac{1}{n_j}S^2(\boldsymbol{z}_j)\right] - \frac{1}{n}S_{k,k'}^2, \tag{3.2}$$

where $S_{k,k'}^2 = \sum_{i=1}^{n}\left(\tau_i(k) - \tau(k)\right)\left(\tau_i(k') - \tau(k')\right)/(n-1)$.

3.2.3 Statistical inference

Among the various types of statistical analyses that could be performed for a factorial design (e.g., Neymanian, frequentist linear regression, Fisherian, and Bayesian), in this chapter, we follow the Neymanian perspective. To do so, we first need good estimators for the variances and covariances of the estimated causal effects. A conservative Neyman-style estimator for the variance was proposed by Dasgupta et al. (2015) and extended to the unbalanced case by Lu (2016b):

$$\widehat{\text{Var}}\left(\widehat{\tau}(k)\right) = \frac{1}{2^{2(K-1)}} \sum_{j=1}^{J} \frac{1}{n_j} s^2(\boldsymbol{z}_j), \tag{3.3}$$

where

$$s^2(\boldsymbol{z}_j) = \frac{1}{n_j - 1} \sum_{i:W_i(\boldsymbol{z}_j)=1} \left(Y_i(\boldsymbol{z}_j) - \bar{Y}^{obs}(\boldsymbol{z}_j)\right)^2$$

is the estimated sampling variance of potential outcomes under treatment combination $\boldsymbol{z}_j$. Dasgupta et al. (2015) discussed other variance estimators and their performance under different assumptions (e.g. strict additivity, compound symmetry). In that paper, the authors also provided a Neyman-style estimator for the covariance by again substituting $s^2(\boldsymbol{z}_j)$ for $S^2(\boldsymbol{z}_j)$. Unfortunately, this estimator is not guaranteed to be conservative because $S^2_{k,k'}$ may be positive or negative. The same authors provided a Neyman confidence region for $\boldsymbol{\tau}$, the vector of all $J-1$ factorial effects. First, they define, in Equation 26,

$$T_n = \hat{\boldsymbol{\tau}}^T \hat{\Sigma}_{\hat{\boldsymbol{\tau}}}^{-1} \hat{\boldsymbol{\tau}},$$

where $\hat{\boldsymbol{\tau}}$ is the vector of Neyman estimators of $\boldsymbol{\tau}$ that we defined earlier and $\hat{\Sigma}_{\hat{\boldsymbol{\tau}}}$ is the estimator of the covariance matrix of $\hat{\boldsymbol{\tau}}$, $\Sigma_{\hat{\boldsymbol{\tau}}}$. Then the $100(1-\alpha)\%$ confidence region for $\boldsymbol{\tau}$ (Equation 27 of Dasgupta et al. (2015)), is

$$\{\hat{\boldsymbol{\tau}} : p_{\alpha/2} \leq T_n \leq p_{1-\alpha/2}\},$$

where $0 < \alpha < 1$ and p_α is the is the α quantile of the asymptotic distribution of T_n. In that paper, a Neyman-style confidence interval for individual effects was also provided as

$$\hat{\tau}(k) \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}\left(\widehat{\tau}(k)\right)},$$

where the authors rely on a normal approximation for the distribution of $\hat{\tau}(k)$, typically assumed to hold asymptotically. See Proposition 1 in Li et al. (2020) for conditions under which the asymptotic normality holds in this setting. These asymptotic properties are also explored in Li and Ding (2017).

Note that a model that linearly regresses an outcome against the factors coded using contrast values -1 and $+1$ and all interactions between factors (but no other covariates) will result in the same point estimates for the factorial effects as presented here, divided by 2 (Dasgupta et al., 2015; Lu, 2016b). For a balanced design or when treatment effects are assumed to be additive, so that the variances $S^2(\boldsymbol{z}_j)$ are all the same, the standard linear regression variance estimate, relying on homoskedasticity, will be the same as the Neymanian variance estimate (Dasgupta et al., 2015; Imbens and Rubin, 2015). However, this is not true for an unbalanced design. Samii and Aronow (2012) showed that the HC2 heteroskedasticity robust variance estimator (see MacKinnon and White, 1985, for more details) is the same as the Neymanian variance estimator presented here for a single treatment experiment. Lu (2016b) extended this finding to factorial designs, showing that this estimator is also the same as the Neymanian variance estimator in that case. Note that because the regression estimator is different by a factor of 2, this regression variance is different by a factor of 4.

Fisherian and Bayesian types of analyses are also possible. See Dasgupta et al. (2015) for a discussion of creating "Fisherian Fiducial" intervals (Fisher, 1930; Wang, 2000) for factorial effects. The basic idea is to invert our estimands so that we write our potential outcomes in terms of the estimands. Then, under a Fisher sharp null hypothesis, we can impute all missing potential outcomes and generate the randomization distribution. Dasgupta et al. (2015) also provide an in-depth discussion of Bayesian analyses for factorial designs.

3.3 Fractional factorial designs

3.3.1 Set up

There are situations in which a full 2^K factorial design experiment cannot be conducted or is not optimal. For instance, limited resources may mean there are not enough units to

randomly assign to each of the J ($= 2^K$) treatment combinations. Or the full factorial design might not be the most efficient allocation of resources, for instance if the experimenter believes that higher order interactions are not significant (Wu and Hamada, 2000). Instead, an experimenter might implement a $2^{K-p} = J'$ fractional factorial design in which only J' of the total J treatment combinations are used (see, e.g., Wu and Hamada, 2000). Here we will give a brief overview of this design, but we recommend Wu and Hamada (2000) to obtain a much more detailed review. We follow notation in Dasgupta et al. (2015) and Dasgupta and Rubin (2015).

To create a 2^{K-p} design, we can write out a full factorial design for the first $K - p$ factors, and fill in the p columns for the remaining factors using multiplicative combinations of subsets of the previous contrast columns. We use a generator to choose the factors whose treatment levels we multiply together to get the treatment levels for the other factors (Wu and Hamada, 2000). For example, if we have want to create a 2^{3-1} design using the generator $3 = 12$, then we would start by writing out a 2^2 design for factors 1 and 2, as in column 1 and 2 of Table 3.2. Then we generate the third column, corresponding to treatment levels for factor 3, by multiplying together the contrast vectors for factors 1 and 2, as shown in column 3 of Table 3.2. Note that which two factors are used initially is irrelevant in this case because of symmetry. That is, because $3 = 12$, we also have $1 = 23$ and $2 = 13$. However, this may not hold in general and one should choose which factors to use based on the final aliasing structure, which tells us which effects are confounded (discussed more below).

Treatment	Factor 1	Factor 2	Factor 3	Outcomes
z_1^*	-1	-1	+1	$\bar{Y}(z_1^*)$
z_2^*	-1	+1	-1	$\bar{Y}(z_2^*)$
z_3^*	+1	-1	-1	$\bar{Y}(z_3^*)$
z_4^*	+1	+1	+1	$\bar{Y}(z_4^*)$
	g_1^*	g_2^*	g_3^*	$\bar{\boldsymbol{Y}}_*$

Table 3.2: *Example of a 2^{3-1} factorial design.*

Under this design, the contrast columns, g_k^*, are now shortened versions or subsets of the contrast columns of a full 2^3 factorial design, g_k. The treatment combinations in the fractional design, z_j^*, are a subset of the full set of treatment combinations, z_j, for the 2^3 design. Again, note that, referring to Table 3.2, $g_3^* = g_1^* \circ g_2^*$. The generator $3 = 12$ means

that the main effect of factor 3 is aliased with the two-factor interaction 12. Two effects being aliased means that we cannot distinguish the effects – they are confounded or combined in our estimators. The full aliasing in this design is as follows: $I = 123$, $3 = 12$, $2 = 13$, $1 = 23$. Note that I corresponds to a vector of all +1's (g_0). The relation $I = 123$ is called the defining relation, which characterizes the aliasing and how to generate the rest of the columns. We see that the main effects, as defined in Section 3.2.2 are aliased with the two-factor interactions. If the two-factor interactions are negligible, the 2^{3-1} design is a parsimonious design to estimate the main effects. That is, we can create unbiased estimators (reviewed in Section 3.3.2) for the main effects. The resulting design is also orthogonal, that is all of the pairs columns are orthogonal, and balanced, which means that each g_k^* has an equal number +1's and -1's (Wu and Hamada, 2000). These properties simplify the aliasing structure.

We typically choose the generator or defining relation based on the maximum resolution criterion for the design. The resolution indicates the aliasing structure, especially which levels of effects the main effects are interacted with. Resolution is defined as the word length (i.e., the number of factors) in the shortest word of the defining relation (see Wu and Hamada, 2000, for more details). In our example of a 2^{3-1} design, we only have one defining relation, $I = 123$, and the word is length 3. This means that the main effects are aliased with two-factor interactions. We can imagine an alternate aliasing structure in which some main effects are aliased with other main effects. The effect hierarchy principle assumes that lower-order effects are more significant than higher-order interaction effects (Wu and Hamada, 2000). Therefore, generally one chooses a fractional design where main effects, and some other lower-order interaction effects, are only aliased with higher-order terms (Wu and Hamada, 2000). In particular, we may want the main effects and two-factor interactions to be clear. This assumption goes along with the assumption of effect sparsity, that the number of significant or important effects is small, as a justification of the use of fractional designs over the full factorial (Wu and Hamada, 2000).

3.3.2 Estimators

We now review estimators for the fractional design, which are similar to the full factorial case. We follow the framework laid out in Dasgupta et al. (2015) and Dasgupta and Rubin (2015). The estimator for $\tau(k)$ is defined as

$$\widehat{\tau}^*(k) = \frac{1}{2^{K-p-1}} \boldsymbol{g}_k^{*T} \bar{\boldsymbol{Y}}_*^{obs}.$$

The estimator for $\tau(k, k')$ is defined as

$$\widehat{\tau}^*(k, k') = \frac{1}{2^{K-p-1}} \boldsymbol{g}_{k,k'}^{*T} \bar{\boldsymbol{Y}}_*^{obs}.$$

Note that these estimators are no longer unbiased. Let $\mathcal{S}$ be the set of all effects aliased with factorial effect k as well as k itself. The number of factors aliased with factor k is $2^p - 1$ (Montgomery, 2017, Chapter 8). So, $\mathcal{S}$ has 2^p elements. Factor k may be aliased with the negative of a main effect or interaction. Let $S_{k,j}$ be the indicator for whether factor j is negatively aliased with factor k ($S_{k,j} = 1$) or positively aliased ($S_{k,j} = 0$). Then we find, using the orthogonality of the $\boldsymbol{g}_k$ vectors, that

$$\begin{aligned} E\left[\widehat{\tau}^*(k)\right] &= \frac{1}{2^{K-p-1}} \boldsymbol{g}_k^{*T} \bar{\boldsymbol{Y}}_* \\ &= \sum_{j \in \mathcal{S}} (-1)^{S_{k,j}} \tau(j) \\ &= \tau(k) + \sum_{j \in \mathcal{S} \setminus \{k\}} (-1)^{S_{k,j}} \tau(j). \end{aligned}$$

Hence by aliasing these effects, they get combined in our estimator.

We see that the estimator for $\tau(k)$ is unbiased if the effects aliased with factorial effect k are zero, and will be close to unbiased as long as the aliased effects are negligible, which may be justified by the effect hierarchy principle. When aliasing occurs such that main effects are aliased with low-order interaction terms, such as two-factor interactions, this assumption may be unrealistic. However, if we have a large number of factors and are only aliasing main effects with higher-order effects, this may be reasonable. For instance, in a 2^{6-1} fractional design, main effects can be aliased with five-factor interactions and two-factor interactions can be aliased with four-factor interactions, where both higher-order interactions may be assumed to be small.

Now we extend the variance and variance estimator expressions from Dasgupta et al. (2015) (see Section 3.2) to this setting. Recall $J' = 2^{K-p}$ and define $\tilde{\tau}(k) = E[\hat{\tau}^*(k)]$. Further define the analog of S_k^2,

$$\tilde{S}_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\tilde{\tau}_i(k) - \tilde{\tau}(k)\right)^2,$$

which is the variation in our newly defined aliased effects, $\tilde{\tau}_i(k)$. Let n_j^* be the number of units assigned to treatment combination $\boldsymbol{z}_j^*$. Then the variance of the estimator $\hat{\tau}^*(k)$ is

$$\text{Var}\left(\hat{\tau}^*(k)\right) = \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J'} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*) - \frac{1}{n} \tilde{S}_k^2. \tag{3.4}$$

See Appendix C.1.2 for more details on this derivation. Similarly, we can obtain the covariance between two fractional factorial effect estimators. Here we define an altered version of $S_{k,k'}^2$:

$$\tilde{S}_{k,k'}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\tilde{\tau}_i(k) - \tilde{\tau}(k)\right) \left(\tilde{\tau}_i(k') - \tilde{\tau}(k')\right). \tag{3.5}$$

Then, the covariance of $\hat{\tau}^*(k)$ and $\hat{\tau}^*(k')$ is

$$\text{Cov}\left(\hat{\tau}^*(k), \hat{\tau}^*(k')\right) = \frac{1}{2^{2(K-p-1)}} \left[\sum_{j: g_{kj}^* = g_{k'j}^*} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*) - \sum_{j: g_{kj}^* \neq g_{k'j}^*} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*) \right] - \frac{1}{n} \tilde{S}_{k,k'}^2. \tag{3.6}$$

See Appendix C.1.3 for more details on this derivation.

The variance expressions for fractional factorial designs are similar to the full factorial case, but are now defined over aliased or grouped effects.

3.3.3 Statistical inference

It is straightforward to extend the analysis for factorial designs in Dasgupta et al. (2015) and reviewed in Section 3.2 to fractional factorial designs. We can again use a conservative Neyman-style variance estimator:

$$\widehat{\text{Var}}\left(\hat{\tau}^*(k)\right) = \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J'} \frac{1}{n_j^*} s^2(\boldsymbol{z}_j^*). \tag{3.7}$$

This leads to

$$E\left[\widehat{\text{Var}}\left(\hat{\tau}^*(k)\right)\right] = \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J'} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*)$$

and thus $\widehat{\text{Var}}\left(\hat{\tau}^*(k)\right)$ is a conservative estimator and unbiased if and only if $\tilde{S}_k^2 = 0$, which would occur if the aliased effects $\tilde{\tau}_i(k)$ are constant. Note that this situation occurs when all effects aliased with factor k are constant additive effects.

We can build confidence regions and confidence intervals analogously to what was reviewed in Section 3.2.3. To be more rigorous in this section, we show how to use the results of Li and Ding (2017) to build Wald-type confidence regions. First, note that there is a new matrix $\boldsymbol{G}^*$ that relates the treatment combinations in the fractional experiment to the unique treatment effect estimands and estimators defined for this experiment. For the 2^{3-1} example we have, we have

$$2^{-(K-p)} \underbrace{\begin{bmatrix} \boldsymbol{g}_0^{*T} \\ \boldsymbol{g}_1^{*T} \\ \boldsymbol{g}_2^{*T} \\ \boldsymbol{g}_3^{*T} \end{bmatrix}}_{\boldsymbol{G}^*} \underbrace{\begin{bmatrix} \bar{Y}(\boldsymbol{z}_1^*) \\ \bar{Y}(\boldsymbol{z}_2^*) \\ \bar{Y}(\boldsymbol{z}_3^*) \\ \bar{Y}(\boldsymbol{z}_4^*) \end{bmatrix}}_{\bar{\boldsymbol{Y}}^*} = \underbrace{\begin{bmatrix} 2\tilde{\tau}(0) \\ \tilde{\tau}(1) \\ \tilde{\tau}(2) \\ \tilde{\tau}(3) \end{bmatrix}}_{\tilde{\boldsymbol{\tau}}},$$

noting that, for instance $\boldsymbol{g}_0^* = \boldsymbol{g}_{123}^*$ and $\tilde{\tau}(2) = \tilde{\tau}(13)$ by the aliasing structure. Define $\hat{\boldsymbol{\tau}}^*$ as the 2^{K-p} vector of unique treatment effect estimators that corresponds to estimands in $\tilde{\boldsymbol{\tau}}$ (shown in previous example).

Now we can define asymptotic results. Assume that all $S^2(\boldsymbol{z}_j^*)$ and $\tilde{S}_{k,k'}^2$ have limiting values, that the n_j^*/n have positive limiting values, and $\max_{1\leq j\leq J'} \max_{1\leq i\leq n} \left(Y_i(z_j^*) - \bar{Y}(z_j^*)\right)^2 / n \to 0$. Then, according to Theorem 5 of Li and Ding (2017), $n\text{var}(\hat{\boldsymbol{\tau}}^*)$ has a limiting value, which, following their notation, we denote by $\boldsymbol{V}$ and

$$\sqrt{n}\left(\hat{\boldsymbol{\tau}}^* - \tilde{\boldsymbol{\tau}}\right) \xrightarrow{d} \text{N}(\boldsymbol{0}, \boldsymbol{V}).$$

Further let $\hat{\boldsymbol{V}} = \sum_{j=1}^{J'} n_j^* \boldsymbol{G}_j^{*2} s(\boldsymbol{z}_j^*) \boldsymbol{G}_j^{*T}$, where $\boldsymbol{G}_j^*$ is the jth column of $\boldsymbol{G}^*$. In addition to the previous assumptions, we require that $s^2(\boldsymbol{z}_j^*) - S^2(\boldsymbol{z}_j^*)$ for all $1 \leq j \leq J'$ and that the limit of $\sum_{j=1}^{J'} n_j^* \boldsymbol{G}_j^{*2} S(\boldsymbol{z}_j^*) \boldsymbol{G}_j^{*T}$ is nonsingular. Then under Proposition 3 of Li and Ding (2017) and

following their notation, the Wald-type confidence region,

$$\{\boldsymbol{\mu} : (\hat{\boldsymbol{\tau}}^* - \boldsymbol{\mu})^T\} \hat{\boldsymbol{V}} (\hat{\boldsymbol{\tau}}^* - \boldsymbol{\mu}) \leq q_{J',1-\alpha},$$

where $q_{J',1-\alpha}$ corresponds to the $1-\alpha$ quantile of the $\chi^2_{J'}$ distribution, has at least $1-\alpha$ asymptotic coverage.

In these asymptotic results we assume that the number of treatments is constant as $n \to \infty$, but Li and Ding (2017) discuss the case where the number of treatment combinations grows as well.

As in the full factorial setting, linear regression yields the same point estimates (divided by 2) and the HC2 variance estimator yields the same variance estimate (divided by 4) as the Neymanian estimates. For proof, see Appendix C.2.

3.4 Incomplete factorial designs

3.4.1 Design and estimators

In this section, we discuss alternative experimental designs to the fractional factorial design, in particular the incomplete factorial design which uses a subset of data from a full factorial design but a different subset than the fractional design (Byar et al., 1993). We discuss aspects of incomplete factorial designs, as defined and discussed in Byar et al. (1993), but with a design-based and potential outcome perspective. In particular, the estimators we discuss here will not use all of the available data. We therefore can consider each estimator to be associated with a particular "design" that corresponds to an experiment that randomizes units only to treatment combinations used in the estimator. Different estimators may give non-zero weight to different treatment combinations, and so the hypothetical designs may be estimand-specific.

Let us assume we ran an experiment with K binary treatments where, whether due to human error (e.g., neglecting to randomize units to a particular treatment combination) or lack of resources, some of treatment combinations from the full 2^K factorial design were not included in the experiment. How should we analyze this? We could reduce the data to a fractional factorial design, which requires removing multiple treatment groups. For

instance, we may have a factorial structure as in Table 3.3 but no outcome measurements for treatment combination z_7.

Treatment	1	2	3	Observed Outcomes
z_1	-1	-1	-1	$\bar{Y}^{obs}(z_1)$
z_2	-1	-1	+1	$\bar{Y}^{obs}(z_2)$
z_3	-1	+1	-1	$\bar{Y}^{obs}(z_3)$
z_4	-1	+1	+1	$\bar{Y}^{obs}(z_4)$
z_5	+1	-1	-1	$\bar{Y}^{obs}(z_5)$
z_6	+1	-1	+1	$\bar{Y}^{obs}(z_6)$
z_7	+1	+1	-1	?
z_8	+1	+1	+1	$\bar{Y}^{obs}(z_8)$
	g_1	g_2	g_3	$\bar{\boldsymbol{Y}}^{obs}$

Table 3.3: *Example of a 2^3 factorial design with no observations for one treatment combination.*

Let us focus on estimation of $\tau(1)$ first. If we recreate a fractional factorial design that aliases the main effects with the two-way interactions, then we estimate $\tau(1) + \tau(23)$, which would involve using outcome data from units assigned to treatment combinations z_2, z_3, z_5, and z_8, but not from units assigned to treatment combinations z_1, z_4, and z_6.

Instead, we might consider building an estimator using all of the treatment combinations except for z_3, which has the same levels for factors 2 and 3 but the opposite level for factor 1 as combination z_7. Thus, in some sense, removing z_3 "balances" the remaining treatment combinations and is essentially the "naïve estimator" discussed in Byar et al. (1993). This strategy creates a different hypothetical experimental design with a different aliasing structure. In this case, we would be estimating

$$\begin{aligned}\dot{\tau}(1) &= \frac{\bar{Y}(z_5) + \bar{Y}(z_6) + \bar{Y}(z_8)}{3} - \frac{\bar{Y}(z_1) + \bar{Y}(z_2) + \bar{Y}(z_4)}{3} \\ &= \frac{\tau(1|F_2 = -1, F_3 = -1) + \tau(1|F_2 = -1, F_3 = +1) + \tau(1|F_2 = +1, F_3 = +1)}{3},\end{aligned}$$

where we let F_k denote the factor level of the k^{th} factor so that
$\tau(k|F_j = x, F_i = y)$ is the main effect of factor k conditional on level x of factor j and level y of factor i, as in Dasgupta et al. (2015). That is, we estimate the average of the conditional effects of factor 1 given the combinations (-1,-1), (-1,+1), and (+1,+1) for factors 2 and 3.

To find the aliasing structure, we can refer to the matrix $\boldsymbol{G}^T$ in Section 3.2.2 to rewrite

the estimand above as

$$\begin{aligned}\dot{\tau}(1) &= \frac{1}{3}\frac{1}{2}\left(\boldsymbol{h}_8 + \boldsymbol{h}_6 + \boldsymbol{h}_5 - \boldsymbol{h}_4 - \boldsymbol{h}_2 - \boldsymbol{h}_1\right)\boldsymbol{\tau} \\ &= \tau(1) + \frac{-\tau(13) + \tau(23) + \tau(123)}{3}.\end{aligned}$$

Now we have aliased the main effect for factor 1 with the two-factor interactions for factors 1 and 3 and factors 2 and 3, as well as the three-factor interaction, all divided by three. This is partial aliasing because the factors are neither fully aliased nor completely clear of each other (Wu and Hamada, 2000, Chapter 7). Whether this is preferable to aliasing the main effect with just the two-factor interaction between factors 2 and 3 depends entirely on subject-matter knowledge. For instance, if we know that the three-factor interaction is negligible, then we might expect this estimator to have lower bias, as we are dividing both two-factor interactions by 3. However, even if we had knowledge that the two-factor interactions were of the same sign, we would not know the direction of the bias in this case without knowing the relative magnitudes of the two-factor interactions. When estimating the main effect for factor 2, we would naturally approximate a different design. In that case, using the same logic as before, we would remove treatment combination $\boldsymbol{z}_5$.

As an alternative design, we may alias our main effects with the highest-order interaction possible, allowing for a different design for each main effect estimator. So when estimating the main effect of factor 1, we would use a design that aliases factor 1 with the three-factor interaction. This leads to a design using treatment combinations $\boldsymbol{z}_1$, $\boldsymbol{z}_4$, $\boldsymbol{z}_5$, and $\boldsymbol{z}_8$ for which we can build the following estimand:

$$\begin{aligned}\dot{\tau}(1) &= \frac{\bar{Y}(\boldsymbol{z}_5) + \bar{Y}(\boldsymbol{z}_8)}{2} - \frac{\bar{Y}(\boldsymbol{z}_1) + \bar{Y}(\boldsymbol{z}_4)}{2} \\ &= \frac{1}{2}\frac{1}{2}\left(\boldsymbol{h}_8 + \boldsymbol{h}_5 - \boldsymbol{h}_4 - \boldsymbol{h}_1\right)\boldsymbol{\tau} \\ &= \tau(1) + \tau(123).\end{aligned}$$

Based on the hierarchy principle, an estimator with this aliasing structure should be a superior estimator to the original fractional factorial estimator mentioned because the three-factor interaction is more likely to be negligible than the two-factor interactions.

Denote by $\dot{\boldsymbol{g}}_k$ the analog of $\boldsymbol{g}_k$ with zeros where outcome data are missing or excluded

in a given estimator for factor k. If there is a single treatment combination with no outcome measured, we can choose the aliasing such that factor k is aliased with the K-factor interaction. When more rows are missing, the pattern of missingness will dictate what aliasing structure is possible. For example, if we are missing two treatment combinations but they are missing from the same design that aliases factor k with the negative of the K-factor interaction, then we can still recreate the design that aliases factor k with the positive of the K-factor interaction. But if we are missing a row from each of these designs, neither option is possible and we must choose a different aliasing structure. If this method is continued for each factor, one ends up with a set of $\dot{\boldsymbol{g}}_k$, each with zeros for different treatment combinations.

Section 4.5 of Wu and Hamada (2000) gives a general strategy to design experiments while attempting to reduce aliasing for certain main effects. There are also other designs we can construct, such as nonregular designs that have partial aliasing (see Wu and Hamada, 2000, Chapter 7 for more details on nonregular designs).

3.4.2 Estimation and inference

More generally, denote our estimator for k^{th} factor under one of these alternative incomplete factorial designs as $\hat{\bar{\tau}}(k) = \dot{\boldsymbol{g}}_k^T \bar{\boldsymbol{Y}}^{obs}$. It is straightforward to extend the variance expression $\text{Var}\left(\hat{\bar{\tau}}(k)\right)$ and variance estimator $\widehat{\text{Var}}\left(\hat{\bar{\tau}}(k)\right)$, as well as the covariance of $\hat{\bar{\tau}}(k)$ and $\hat{\bar{\tau}}(k')$, from Section 3.3.3.[1] See Appendix C.3 for the specification and derivation. However, terms for some treatment combinations will be set to zero in these expressions because they are present in one estimator but not the other.

If we run a regression with all interactions on a design with missing treatment combinations, it is ambiguous what design the resulting estimators correspond to and, as discussed above, multiple designs may be plausible to approximate the observational study. If we specify a regression of the outcome on the fully interacted treatments, but some treatment combinations are not present in the data, the regression will not be able to estimate all interactions. Not including a set of interactions implies that the linear model assumes that

[1] In fact, it is easy to similarly extend these results to any linear combinations of interest on the 2^K treatment combinations.

those interaction effects are zero, and we can assess how reasonable this assumption is. However, the specific aliasing structure between the effects included in the model and those dropped by the regression is not obvious from the usual computer output alone, though can be discerned from the design matrix. Byar et al. (1993) and Byar et al. (1995) give more discussion of these types of estimators and Appendix C.3.3 has further discussion.

3.5 Embedding observational studies in fractional factorial designs

3.5.1 General issues

To address causality in a non-randomized study, it has been argued that one needs to conceptualize a hypothetical randomized experiment that could correspond to the observational data (Bind and Rubin, 2019; Rosenbaum, 2002; Rubin, 2008; Stuart, 2010). The hypothetical randomization is plausible if the treatment groups are "similar" with respect to confounding variables (Rubin, 2007, 2008). In a setting with multiple treatments of interest, we recreate a hypothetical factorial randomized experiment. However, in observational studies, there may be no units that received certain treatment combinations. Therefore, our strategy is to recreate a hypothetical fractional factorial experiment instead of a full factorial experiment.

The choice of which fractional factorial design to recreate should be based on some criteria such as the maximum resolution criteria. Typically, it is desirable to alias main effects with the highest order interactions, which usually means a small p, i.e., a small fraction of total design is removed. In practice, we may not be able to control the aliasing structure. We must choose an aliasing structure for the 2^{K-p} fractional design such that the treatment combination with no observations is not used. However, this strategy usually results in the removal of units assigned to treatment combinations that were present in the observational data set but not used in the design. That is, if only one treatment combination is not present in the data set, we could use a 2^{K-1} design, but then we are not using $2^{K-1} - 1$ treatment combinations for which we have data.

After a design is chosen, a strategy to balance covariates should be used to ensure that the units across treatment combinations are similar with respect to background covariates. We assume strong ignorability, that is, that conditional on measured covariates, the assignment

mechanism is individualistic, probabilistic, and unconfounded (Rosenbaum and Rubin, 1983). Then we can obtain unbiased causal estimates by analyzing the data as if it arose from a hypothetical randomized experiment.

Note that these assumptions apply to all treatment combinations in the final experimental design used. We must also assume that the treatment assignment is, at least hypothetically, manipulable such that all potential outcomes are well-defined. This in turn ensures that our estimands of interest are defined. If this assumption on manipulation does not hold, we would need to consider a causal estimand that does not depend on unmeasurable or undefined potential outcome. In particular, although our fractional factorial estimators do not use all treatment combinations, the estimand and aliasing structure both do depend on those potential outcomes for the unobserved treatment combinations being well defined.[2] Thus, it is important that subject-matter knowledge guides us in deciding which covariates are informative about which units have well-defined potential outcomes under all treatment combinations. That is, only some sub-populations may have all potential outcomes in the experiment be well-defined, similar to some analyses for noncompliance where the estimand is only well-defined for compliers. For other groups where not all treatment combinations could theoretically be received, there still may be interesting estimands based on the well-defined potential outcomes, but we do not explore those specifically in this work. A similar argument must hold for the unconfoundedness assumption. The unconfoundedness assumption is often assumed when reasonable covariate balance is achieved (Imbens and Rubin, 2015).

Achieving covariate balance in multiple treatment groups in non-randomized studies can be non-trivial. We discuss this issue further in the context of hypothetical fractional factorial designs in Section 3.5.2. Therefore, a first step might be to only obtain covariate balance between two treatment groups, a task commonly done in the causal inference literature, and compare outcomes for these groups. For instance, we could estimate the difference in outcomes between the units that were assigned level $+1$ for all factors and the units that

[2]Under some assumptions, this estimand may be the same as the original. For instance, if we have a 2^2 experiment and there is no interaction between factor 1 and factor 2, then $\tau(1) = \bar{Y}(+1,+1) - \bar{Y}(-1,+1) = \bar{Y}(+1,-1) - \bar{Y}(-1,-1)$. Hence, even if we only observed one level of factor 2 but both levels of factor 1, we could recover $\tau(1)$. However, $\tau(1)$ would not be of scientific interest if all potential outcomes are not defined.

were assigned -1 for all factors. Under certain assumptions, testing the difference in these groups can act as a global test for whether any effects of interest are significant. We discuss the test and assumptions that are needed for this simple comparison to be meaningful in Section 3.5.3.

Once we have decided to recreate a particular fractional factorial design and have obtained a data set with covariate balance, estimation and inference can follow in a similar way to Section 3.3 or Section 3.4. We discuss some of the relative benefits of using fractional factorial vs. incomplete designs in Section 3.5.4

3.5.2 Covariate balance

An important stage when estimating the causal effect of non-randomized treatments is the design phase (Rubin, 2007, 2008). At this stage, we attempt to obtain a subset of units for which we can assume unconfoundedness. That is, units for which $P(W_i|Y_i(z), \boldsymbol{X}_i) = P(W_i|\boldsymbol{X}_i)$ where $\boldsymbol{X}_i$ is a m-dimensional vector of covariates (Imbens and Rubin, 2015). Matching strategies are often used to ensure no evidence of covariate imbalance between treatment groups, as reviewed by Stuart (2010). Note that matching often involves removing units and so the statistical analysis is generally performed on a subset of the original study population. This implies that the population of units we are doing inference with respect to may change after trimming units.

In settings with multiple treatments, matching can be difficult. There have been extensions of propensity score balancing to multiple treatments, most notably the generalized propensity score (GPS) (Hirano and Imbens, 2004) and more recently the covariate balancing propensity score (CBPS) (Imai and Ratkovic, 2014) was introduced and shown to extend to multiple treatments. Lopez and Gutman (2017) review techniques, including matching, for observational studies with multiple treatments and although they do not explore factorial (full or fractional) designs, it may be straightforward to extend their methods to this design. Nilsson (2013) discusses matching in the 2^2 design. Recent work by Bennett et al. (2020) uses template matching, which matches units to a "template" population of units, for multiple treatments, though again not with factorial designs. Therefore, methods for creating covariate balance specifically for factorial type designs require further exploration. In our

data illustration in Section 3.6, we employ sequential trimming and checks on covariate balance, as discussed below.

Testing for covariate imbalance across multiple treatment groups can be done using multivariate analyses of variance (MANOVA), which uses the covariance between variables to test for mean differences across treatment groups, as used in Branson et al. (2016). Recall that the factorial design has J treatment combinations. Define the following H and E matrices (Coombs and Algina, 1996):

$$\boldsymbol{H} = \sum_{k=1}^{J} n_k(\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})^T$$

$$\boldsymbol{E} = \sum_{k=1}^{J}(n_k - 1)S_k, \text{ where } S_k = \frac{1}{n_k - 1}\sum_{j=1}^{n_k}(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)^T,$$

where $\overline{\boldsymbol{X}}_k$ is the m-dimensional vector of mean covariate values for treatment group k, $\overline{\boldsymbol{X}}$ is the average m-dimensional vector of mean covariate values for all units, that is $\overline{\boldsymbol{X}} = \sum_{k=1}^{J} \frac{n_k}{n}\overline{\boldsymbol{X}}_k$, and $\boldsymbol{X}_{kj}$ is the m-dimensional vector of covariates for the jth unit in treatment group k. Denote by θ_k the ordered eigenvalues of $\boldsymbol{HE}^{-1}$, where $k \in \{1, ..., s\}$ and $s = \min(m, K-1)$. Standard MANOVA statistics which can be used to test covariate balance are typically functions of the eigenvalues of $\boldsymbol{HE}^{-1}$ (Coombs and Algina, 1996). We chose the Wilks' statistic (Wilks, 1932),

$$\text{Wilks} = \frac{|\boldsymbol{E}|}{|\boldsymbol{H}| + |\boldsymbol{E}|} = \prod_{k=1}^{K}\frac{1}{1+\theta_k},$$

where θ_k corresponds to the k^{th} eigenvalue of $\boldsymbol{HE}^{-1}$.

As discussed in Imai et al. (2008), a potential drawback of testing for evidence against covariates imbalance is that as we drop units we lose power to detect deviations from the null hypothesis of no difference in covariates between the treatment groups. Another diagnostic for covariate balance is checking covariate overlap via plots and other visual summaries of the data. So called "Love plots", which show standardized differences in covariate means between two treatment groups before and after adjustment (Ahmed et al.,

2006), are difficult to generalize directly because of the multitude of treatment groups and comparisons. However, plots of standardized means or of distributions may be helpful to detect imbalance.

3.5.3 Initial test for significance of effects

As discussed in the previous section, achieving covariate balance for more than two treatment groups can be a challenge. Therefore, instead of attempting to achieve balance among all treatment groups, a simple first step might be to examine two carefully chosen treatment groups and attempt to balance these two groups only. Obtaining balance between two treatment groups has been well-studied in causal inference, see Stuart (2010) for a review of common matching methods for such settings. Once significant covariate imbalance can be ruled out, we can test whether the mean difference between these two groups is significantly different. But what can we learn from this comparison about our factorial effects?

We have from Section 3.2.2 that

$$\bar{Y}(\boldsymbol{z}_j) = \frac{1}{2}\boldsymbol{h}_j\boldsymbol{\tau}.$$

So when we subtract two observed means, assuming that the observed means are unbiased estimates of the true means (i.e. we have randomization or strong ignorability), we are estimating

$$\bar{Y}(\boldsymbol{z}_j) - \bar{Y}(\boldsymbol{z}_{j'}) = \frac{1}{2}(\boldsymbol{h}_j - \boldsymbol{h}_{j'})\boldsymbol{\tau},$$

which is the sum of terms that are signed differently in $\boldsymbol{h}_j$ and $\boldsymbol{h}_{j'}$.

As an example for a 2^3 design, we have the following matrix for $\boldsymbol{G}^T$:

$$\boldsymbol{G}^T = \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \\ \boldsymbol{h}_3 \\ \boldsymbol{h}_4 \\ \boldsymbol{h}_5 \\ \boldsymbol{h}_6 \\ \boldsymbol{h}_7 \\ \boldsymbol{h}_8 \end{bmatrix} = \begin{pmatrix} [r]+1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & +1 & -1 & +1 & -1 & +1 & -1 & -1 \\ +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \end{pmatrix}.$$

Now consider taking the difference between $\bar{Y}(\boldsymbol{z}_8) - \bar{Y}(\boldsymbol{z}_1) = \bar{Y}(+1,+1,+1) - \bar{Y}(-1,-1,-1)$,

which yields

$$\begin{aligned}\bar{Y}(\boldsymbol{z}_8) - \bar{Y}(\boldsymbol{z}_1) &= \frac{1}{2}(\boldsymbol{h}_8 - \boldsymbol{h}_1)\boldsymbol{\tau} \\ &= \tau(1) + \tau(2) + \tau(3) + \tau(123).\end{aligned}$$

Hence, testing whether the difference between $\bar{Y}(\boldsymbol{z}_8)$ and $\bar{Y}(\boldsymbol{z}_1)$ is zero is the same as testing whether $\tau(1) + \tau(2) + \tau(3) + \tau(123)$ is zero. If it is reasonable to assume that all main effects are of the same sign based on subject matter knowledge and that the three-factor interaction is also of the same sign or negligible, then this global test of whether there are any treatment effects is relevant to the estimand of interest. According to the effect hierarchy principle (Wu and Hamada, 2000), the main effects should dominate the three-factors interaction. Hence, even if the interaction differs in sign, we would expect to see an effect under this assumption if there is one. If the global test is not rejected, then we would move on to recreating the entire factorial design. If it is unclear whether the signs of the effects are the same (also referred as antagonistic effects), then this global test would be inappropriate because effects could still be different from zero but their sum could be (close to) zero.

If we are particularly interested in the causal effects involving the first factor, we can create a test specifically for those effects. For instance in the the 2^3 setting, we might use the estimand $\bar{Y}(\boldsymbol{z}_8) - \bar{Y}(\boldsymbol{z}_4)$ as a proxy for the effect of the first factor, as follows:

$$\begin{aligned}\bar{Y}(\boldsymbol{z}_8) - \bar{Y}(\boldsymbol{z}_4) &= \bar{Y}(+1,+1,+1) - \bar{Y}(-1,+1,+1) \\ &= \frac{1}{2}(\boldsymbol{h}_8 - \boldsymbol{h}_4)\boldsymbol{\tau} \\ &= \tau(1) + \tau(12) + \tau(13) + \tau(123). \qquad (3.8)\end{aligned}$$

In Equation 3.8, if all terms have the same sign and we find that the difference is significantly different than zero, then we would conclude that factor 1 has an effect, either on its own or through interactions with the other factors. A different choice of levels for the other factors, for instance comparing the mean potential outcome when one factor is high vs. low when all other factors are at the low level, would result in different signs for the interactions. The choice of which levels to compare should be based on subject matter knowledge and reasonable assumptions.

To reiterate, throughout this section we focused on the meaning of several estimands, which are easily estimated under a randomized experiment. In an observational study, we would first need to obtain balance for any treatment groups we would be using in our estimator.

3.5.4 Comparing designs

So far our discussion of designs has focused on randomized experiments. However, there may be added complications in the observational setting that make one design more desirable than another. For instance, consider using a different design for each estimator as in Section 3.4. Here we are able to use more of the data than a fractional factorial design, but this can also incur a cost. If we have no outcome measurement for only one treatment combination, we would use all $2^K - 1$ treatment combinations if we did an analysis for all effects and used a different design for each. This approach would require us to either first obtain balance among all $2^K - 1$ treatment groups or to obtain balance among the treatment groups within each design separately. The former option may be difficult; as the number of treatment combinations grows, obtaining covariate balance across all treatment groups becomes increasingly difficult and may result in smaller and smaller sample sizes, especially if trimming is used. The latter option will make joint inferences more challenging because different units would be used in each analysis.

Therefore, although these incomplete factorial designs may improve the bias of our estimators, the fractional factorial design in which we are using the same 2^{K-p} rows may be more attractive in terms of obtaining covariate balance. The fractional factorial design also has the benefit of being a classical experimental design with an aliasing structure that is easy to understand.

3.6 Data illustration

3.6.1 Data description

Here we give an illustration of the implementation of our methods using data on pesticide exposure and body mass index (BMI). We use the 2003-2004 cycle of the National Health

and Nutrition Examination Survey (NHANES) collected by the Centers for Disease Control and Prevention (CDC). We access the data via the `R` (R Core Team, 2017) package `RHANES` (Susmann, 2016). We focus on four organochlorine pesticides, measured via a blood serum test and then dichotomized based on whether they were above (+1) or below (-1) the detection limit, as given in the NHANES dataset, as factors. Organochlorine pesticides are persistent in the environment and adverse health effects have been reported by the CDC (2009), making them an interesting group of pesticides to study. To keep this illustration simple, we chose to use only four pesticides and those were chosen primarily based on data availability and exposure rates. That is, we did not use those pesticides that were so common (or rare) that virtually everyone (or no one) in the data set was above the detection limit (or below the detection limit). The following are the the four pesticides that were chosen: beta-Hexachlorocyclohexane (beta-Hex), heptachlor epoxide (Hept Epox), mirex, and p.p'-DDT. Previous findings of an association between pesticide exposures and body mass index (BMI) (Buser et al., 2014; Ranjbar et al., 2015) led us to choose BMI as the outcome of interest. BMI is the ratio between weight and height-squared.

We removed 271 units with missing values of pesticide and BMI observations, noting that because this is an illustration and not intended to draw causal conclusions we drop those units for simplicity. We also decided to study a non-farmer population as farmers are more likely to be exposed to pesticides than the general population and may also differ on other unobserved covariates that may affect health outcomes. This is our first step to achieving covariate balance, leaving a dataset with 1,259 observations (see Figure C.1 in the Appendix for more).

3.6.2 Design stage

To show the process of how estimands change as we adjust our design stage, we consider different "designs." First we consider analyzing the data as if it came from a 2^4 factorial hypothetical experiment, without adjusting the sample to balance for covariates. Second, we instead analyze the data as if it came from a fractional factorial 2^{4-1} hypothetical experiment to assess how estimates change when going from a factorial design to a fractional factorial design. Finally, we trim units to obtain a fractional factorial 2^{4-1} hypothetical experiment

with covariate overlap and with no evidence of covariate imbalance with respect to gender, age, and smoking status, to see how estimates change when a true design phase with the aim of obtaining covariate balance is implemented.

Note that we trim units to obtain better balance on gender, age, and smoking status as these are our first tier of most important covariates. Then additional adjustment for our second tier of covariates, race and ethnicity and income, are done via linear regression, as described in the following section. We also found that even with trimming, gender and smoking status were imbalanced across the treatment groups so these were also adjusted for in the linear regression.

There are a few notes on the definitions of these covariates. From now on we will refer to the "race and ethnicity" covariate as simply ethnicity, as a short hand and to make clear that this is one categorical variable in the NHANES dataset. The income variable is defined as annual household income. For categorizing individuals as smokers vs non-smokers, we use the question "Have you smoked at least 100 cigarettes in your life time?" and we categorized "Yes" and "Don't know" as smokers. Note that only one observation with value "Don't know" was recorded.

3.6.3 Statistical analysis

We analyzed the three datasets described in the design stage using: 1) a multiple linear regression that regresses BMI on treatment factors; 2) a multiple linear regression that regresses BMI on treatment factors, as well as the following covariates, as factors: ethnicity, income, gender, and smoking status; 3) Fisher-randomization tests of the sharp null hypothesis of no treatment effects.

Recall from Sections 3.3.3 and 3.2.3 that regression estimates when including all factors and interactions, but not covariates, are our standard Neymanian estimates divided by two. Figures C.15 and C.16 in Appendix C.4.3 show that further adjustment for ethnicity and income are needed, even after balancing gender, age, and smoking status. Additional data descriptions and the full statistical analyses are available in Appendix C.4. Note that due to the right-skewed nature of the weight variable, BMI exhibits some degree of right-skewness. Therefore we use log-transformed BMI as the outcome in these analyses.

Full factorial design

Table 3.4 provides the counts of observations for each treatment combination (z_j). There were 1,259 units in this dataset, although, when adjusting for covariates in the regression, units with missing covariate values were removed resulting in 1,183 units. We see that factor combination 10 (z_{10}) has only one observation. Relying on only one observation for a treatment combination will lead to unstable estimates. Hence, Section 3.6.3 aims to avoid this issue by embedding the observational study in a fractional factorial hypothetical experiment. Nonetheless, we perform the analysis of the 2^4 factorial hypothetical experiment in this section and will compare the results to the analysis of the fractional factorial hypothetical experiment. Because there is only one unit assigned to z_{10}, it is not possible to estimate Neymanian variances here. See Appendix C.4.1 for full analysis results.

	Factor Levels				Number of Obs.
	beta-Hex	**Hept Epox**	**Mirex**	**p,p'-DDT**	
z_1	+1	+1	+1	+1	426
z_2	-1	+1	+1	+1	12
z_3	+1	-1	+1	+1	70
z_4	-1	-1	+1	+1	51
z_5	+1	+1	-1	+1	291
z_6	-1	+1	-1	+1	25
z_7	+1	-1	-1	+1	94
z_8	-1	-1	-1	+1	54
z_9	+1	+1	+1	-1	21
z_{10}	-1	+1	+1	-1	1
z_{11}	+1	-1	+1	-1	19
z_{12}	-1	-1	+1	-1	19
z_{13}	+1	+1	-1	-1	42
z_{14}	-1	+1	-1	-1	19
z_{15}	+1	-1	-1	-1	37
z_{16}	-1	-1	-1	-1	78
	g_1	g_2	g_3	g_4	1259

Table 3.4: *Counts of observations for each treatment combination of the pesticides with farmers removed for the factorial design. Red rows treatment combinations that we will use when recreating a fractional factorial design.*

Fractional (2^{4-1}) factorial design

In the 2^{4-1} fractional factorial design, instead of using $I = 1234$, we chose $I = -1234$ to exclude row 10 in Table 3.4 that has only a single observation. The dataset in this hypothetical experiment consists of 523 observations. However, when adjusting for covariates in the regression, units with missing covariate values were removed resulting in 488 units. In this design, aliasing is as follows: $I = -1234$, $4 = -123$, $3 = -124$, $2 = -134$, $1 = -234$, $12 = -34$, $13 = -24$, $14 = -23$. The main effects are aliased with the negative of the three-factor interactions and the two-factor interactions are aliased with each other with reversed signs. In order to identify main effects, we will assume that the three-factor interaction is negligible. In practice, researchers should assess whether this aliasing assumption is realistic. See Appendix C.4.2 for full analysis results.

Fractional factorial design with no evidence of covariate imbalance

We examined the distributions across treatments of the following covariates: gender (recorded as male vs. female), smoking status (smoker vs. non-smoker, as defined earlier), and age at the time of survey (in years). Figure 3.1 shows covariate imbalance with respect to gender, smoking, and age in the fractional factorial design.

We used a rejection approach that sequentially pruned the observations of the fractional factorial dataset until we found no evidence of imbalance across exposure groups with respect to gender, smoking status, and age. To test for covariate balance across treatments, we perform a MANOVA using the Wilks' statistic (Wilks, 1932), as defined in Section 3.5.2. Figure 3.2 shows the covariate distribution for gender, smoking status and age after trimming. We see that gender and smoking status are still imbalanced even after trimming, and hence were adjusted for in the linear regression that includes covariates (this is true for all "designs").

The first dataset that resulted in no evidence of covariate imbalance consisted of 169 observations, and the new treatment counts are presented in Table 3.5. After removing units with missing covariate values for the regression that adjusts for income and ethnicity, we had 158 units. Note that while gender, smoking status, and age are our first tiers of covariates, ethnicity and income constitutes the second tiers. In practice, balancing a large

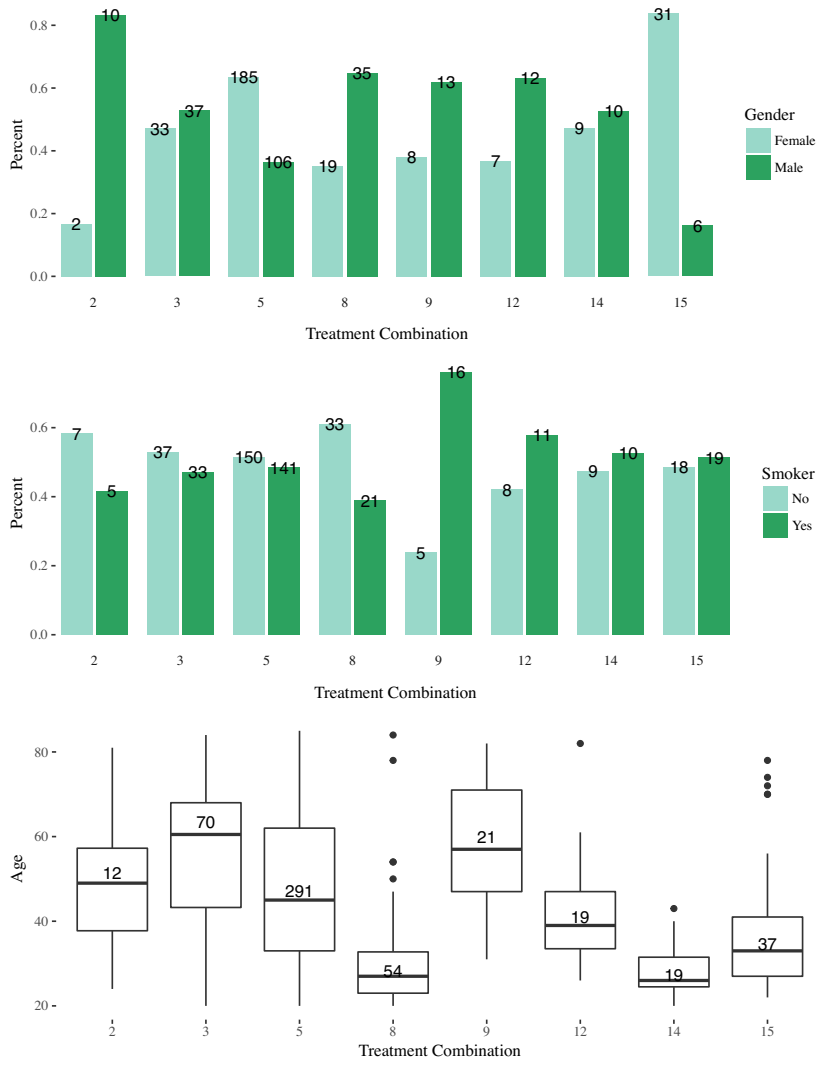

Figure 3.1: *Comparing covariates across treatment combinations in the* 2^{4-1} *fractional factorial design. Text labels give number of observations per group. For age, individuals with ">= 85 years of age" were set to 85 on the graph. Note that all individuals older than 85 were dropped in the covariate balance stage.*

set of covariates is not always possible. Therefore, choosing the most important ones should be done using subject matter knowledge. We see here that the number of units has been drastically reduced in our attempt to achieve covariate balance, a major challenge in this setting. In fact, one of the treatment groups only has three observations, a very small number. See Appendix C.4.3 for full analysis results.

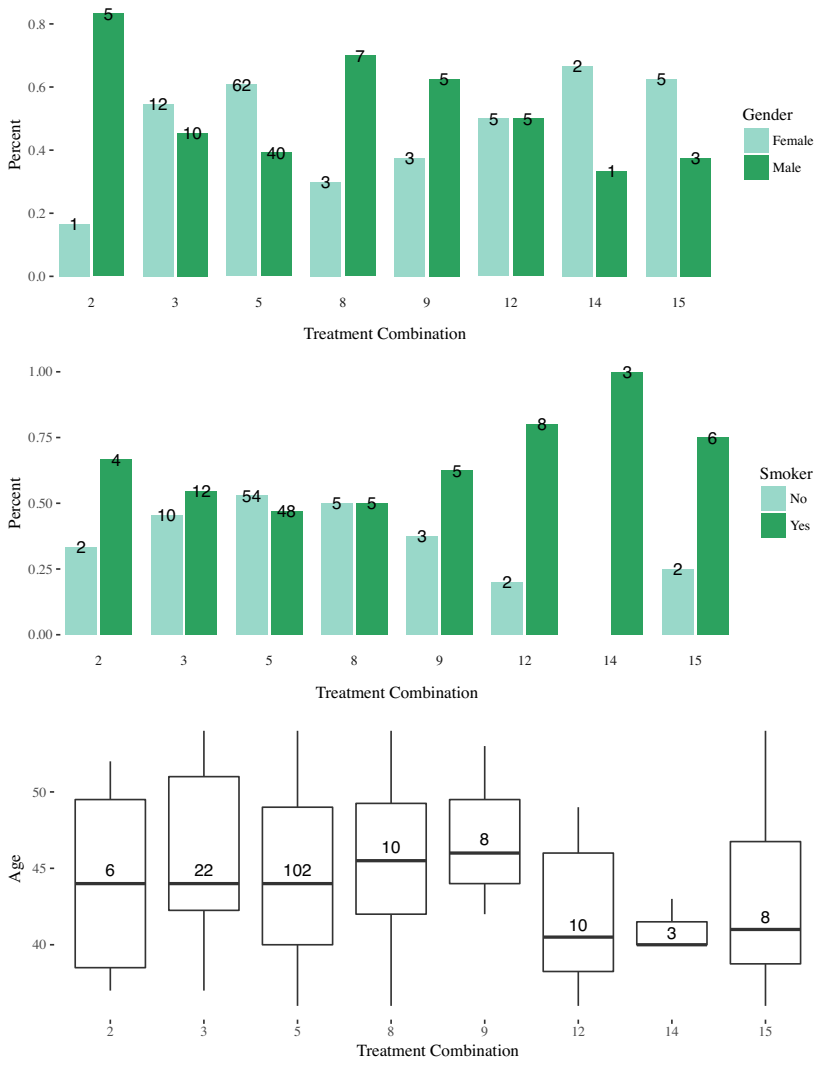

Figure 3.2: *Comparing covariates across treatment combinations in the* 2^{4-1} *fractional factorial design after trimming. Text labels give number of observations per group.*

3.6.4 Results comparison across different conceptualized experiments and statistical approaches

Figure 3.3 shows a comparison of the regression estimates of the main effects across designs and statistical analyses. To compare the different methods we present univariate analyses, that is we utilize individual tests for each main effect rather than joint tests, for better illustration of the different methods. Recall that the standard Fisherian and Neymanian point estimates are the unadjusted regression estimates multiplied by two. The bars show

	Factor Levels				Number of Obs.	
	beta-Hex	**Hept Epox**	**Mirex**	**p,p'-DDT**	**Original**	**Trimmed**
z_2	-1	+1	+1	+1	12	6
z_3	+1	-1	+1	+1	70	22
z_5	+1	+1	-1	+1	291	102
z_8	-1	-1	-1	+1	54	10
z_9	+1	+1	+1	-1	21	8
z_{12}	-1	-1	+1	-1	19	10
z_{14}	-1	+1	-1	-1	19	3
z_{15}	+1	-1	-1	-1	37	8
	g_1^*	g_2^*	g_3^*	g_4^*	523	169

Table 3.5: *Counts of observations for each treatment combination of the pesticides with farmers removed for the fractional factorial design, before and after trimming.*

two standard errors above and below the estimate calculated by the usual ordinary least squares as the Neymanian variance estimates are not available for the full factorial design. In practice, adjustment for multiple comparisons should be considered. All methods and designs seem to agree on the positive effect of heptachlor epoxide and the negative effect of mirex on BMI. Although the full factorial estimates generally agree with the estimates of the two fractional factorial designs, differences in estimates of beta-Hexachlorocyclohexane (Beta-Hex) and p,p'-DDT may be due to the aliasing of the three-factor interactions with the main effects. However, it is also plausible that we have reduced our data in the fractional design to a group with different average main effects than in the full data set.

Figure 3.4 shows a comparison of the significance of these estimates. The Fisher p-value is the p-value for the test of no effects of any pesticides, based on effect estimates for a given pesticide, which is suggested as a screening stage in Espinosa et al. (2016). We obtained low p-values for the main effect of mirex on BMI across all methods and designs. However, the p-values disagree for the other pesticides, especially the p-values testing the effects of p,p'-DDT and beta-Hexachlorocyclohexane. Note that the HC2/Neyman p-value is the significance based on the HC2 variance estimate (or Neyman variance estimate as we have shown this estimator to be equivalent in settings with no covariates) and the Normal approximation. This p-value was only calculated for the regression analysis without covariates and was unavailable for the full factorial model due to limited data.

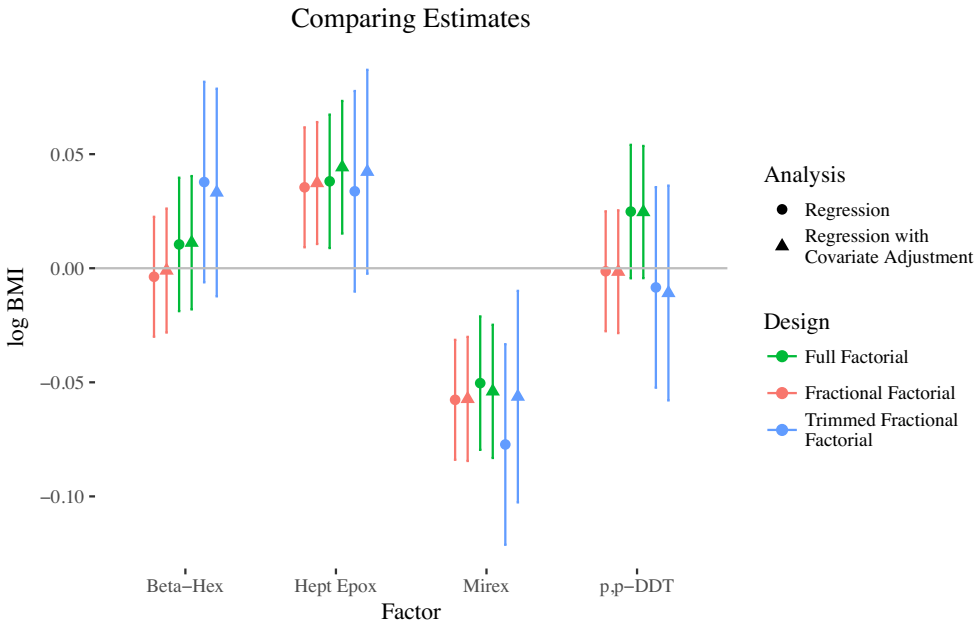

Figure 3.3: *Plots of estimates of factorial effects, on the log BMI scale. Bars indicate two standard errors (using standard OLS standard estimates) above and below point estimate.*

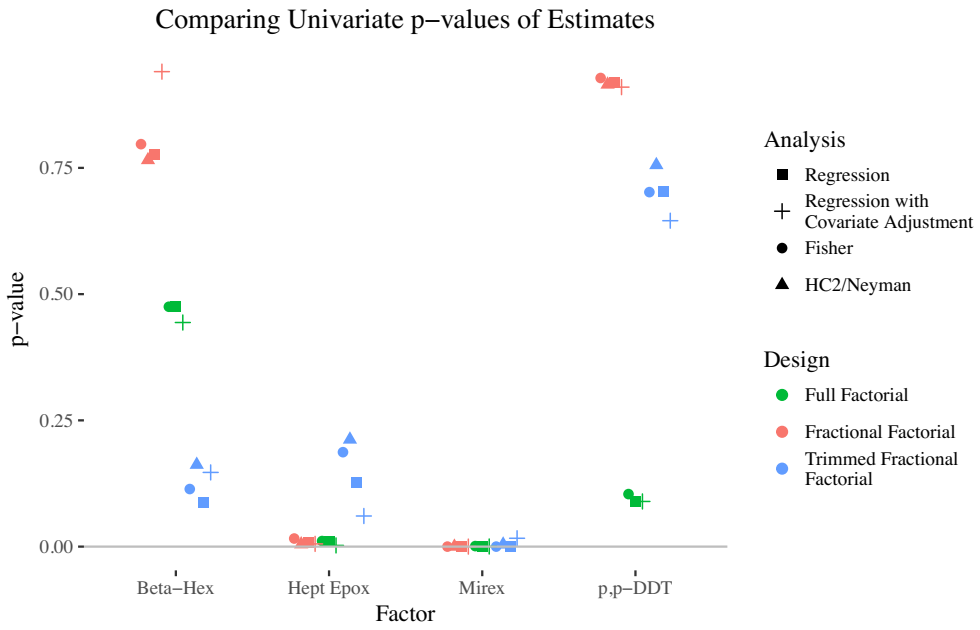

Figure 3.4: *Plots of p-values of factorial effect estimates, which are compared in Figure 3.3.*

3.6.5 Discussion of data illustration

We have performed a data illustration to show the benefits and challenges of using our method and working with observational data with multiple treatments in general. Here we outline some of the problems that we ran into, that could be improved in future analyses, and also some of our findings. First, it is important to note that simplifications were made in the statistical analyses to focus on illustrating how researchers can capitalize on using fractional factorial designs to estimate the main and interactive effects of multiple treatments in observational studies. For instance, we are aware that multiply imputing the missing data would have been more appropriate to provide valid estimates and inferences in the original study population. However, for simplicity we focused on the complete-case observations. We additionally did not adjust for all important hypothetical covariates, such as diet. Another consideration is that we log-transformed BMI to address the fact that BMI is a ratio and so its distribution tends to have heavy tails. However, heavy-tailed distributions could have been considered.

There were also some major challenges in working with this data and our method, mostly related to sample size. Trimming helps mitigate bias but greatly reduced our sample size, potentially leading to decreased power and precision. There was also a great reduction in sample size by using the fractional factorial design which entailed dropping treatment combinations. We could have used an incomplete factorial type design, but this could have made the covariate balancing even more difficult, as discussed previously.

For p,p'-DDT, the full factorial design resulted in a lower p-value than the other designs, which could be an issue of aliased interactions watering down the effect in the fractional design. It could also have occurred because the populations are different in these two designs. For beta-Hexachlorocyclohexane, the covariate balance adjusted fractional factorial design differs, leading to a lower p-value for the association of beta-Hexachlorocyclohexane with BMI compared to the other designs. We are more inclined to trust the balanced design as this should have reduced bias. This result may indicate that there was some confounding that made the effect appear less significant before trimming. In fact, the stark contrast between the unbalanced and balanced fractional design suggests that confounding may be to blame. Alternatively, by trimming we may have reduced our sample to a subpopulation

where beta-Hexachlorocyclohexane has a larger effect on BMI than the rest of the population.

3.7 Discussion

In this chapter, we have proposed to embed observational studies with multiple treatments in fractional factorial hypothetical experiments. This type of design is useful in settings with many treatments, especially when some treatment combinations have few or no observations and the aliasing assumptions are plausible. Once we recreate a factorial or fractional factorial experiment in the design phase, we can use standard methods, extended as in Sections 3.2 and 3.3, to estimate causal effects of interest. We first reviewed the basic setup for factorial and fractional factorial designs. Our work includes extensions of some of the known factorial design results variance and regression results to the fractional factorial setting. We also explored the use of incomplete factorial designs in the design-based potential outcome framework. A main contribution of this chapter consists of extending these ideas for observational studies. This includes discussion of tests that can be performed before the full analysis. We have also discussed covariate balance complications that may arise when dealing with multiple non-randomized treatments in practice.

We illustrated these methods on a data set with pesticide exposure and BMI. This helped to exemplify the uses of our new methodology as well as identify challenges that occur when working with observational data with multiple treatments. It is important to note that using the small subset of the NHANES dataset, we do not intend to provide policy recommendations on pesticide use. In the general population, organochlorine pesticide exposure primarily occurs through diet (excluding those with farm-related jobs), particularly eating foods such as dairy products and fatty fish (Centers for Disease Control and Prevention (CDC), 2009). Without further adjustments for diet, we are not be able to disentangle the causal effect of diet and pesticides. For instance, in our study individuals are likely to have been exposed to mirex largely through fish consumption (Agency for Toxic Substances and Disease Registry [ATSDR], 1995). Further studies could investigate BMI differences in a group of fish consumers with high level of mirex and a "similar" group of fish consumers with low level of mirex, where similar is with respect of important confounding variables. Indeed, it could be that eating fish cause individuals to have both

high levels of mirex and also lower BMI.

We have given a short overview of factorial and fractional factorial designs, as well as some other designs, in the potential outcomes framework. However, there are many aspects of these designs and classic analysis techniques that we did not cover. For instance, there are many nonregular design types, such as Plackett-Burman designs, that we could have explored more. Practitioners may also use variable selection and the principle of effect heredity to select their model for estimating factorial effects, including via Bayesian variable selection. See Wu and Hamada (2000) for more details on these methods from a more classical experimental design perspective.

We see many avenues of future exploration connected to our approach. For instance, coupling fractional factorial designs with a Bayesian framework would provide more statistical tools and would potentially offer different methodology for dealing with missing data. Additionally, development of balancing techniques for factorial designs with many treatment combinations should be an area of future exploration. A particular challenge is that as we increase the number of treatment combinations and therefore treatment groups, matching becomes more and more difficult due to the increased dimensionality and weighting methods may produce unstable estimators. One direction could also be to choose the design of the observational study based upon the ability to balance different treatment combinations. Random allocation designs (Dempster, 1960, 1961), in which randomness of the design is incorporated, could also be utilized in this framework. Finally, we could explore other causal estimands that may be of interest in observational studies with multiple treatments, such as those proposed in Egami and Imai (2019).

References

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, 91/92:175–187.

Agency for Toxic Substances and Disease Registry (ATSDR). (1995). Public Health Statement for Mirex and Chlordecone. *Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.* https://www.atsdr.cdc.gov/phs/phs.asp?id=1189&tid=276.

Ahmed, A., Husain, A., Love, T. E., Gambassi, G., Dell'Italia, L. J., Francis, G. S., Gheorghiade, M., Allman, R. M., Meleth, S., and Bourge, R. C. (2006). Heart failure, chronic diuretic use, and increase in mortality and hospitalization: An observational study using propensity score methods. *European Heart Journal*, 27(12):1431–1439.

Aronow, P. M., Green, D. P., Lee, D. K., et al. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871.

Basse, G. W., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.

Basu, D. (1964). Recovery of ancillary information. *Sankhyā A*, 26(1):3–16.

Bennett, M., Vielma, J. P., and Zubizarreta, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *Journal of Computational and Graphical Statistics*, (just-accepted):1–42.

Bind, M.-A. C. and Rubin, D. B. (2019). Bridging observational studies and randomized experiments by embedding the former in the latter. *Statistical Methods in Medical Research*, 28(7):1958–1978.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508.

Box, G. E., Hunter, J. S., and Hunter, W. G. (2005). Statistics for experimenters. In *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, NJ.

Branson, Z., Dasgupta, T., and Rubin, D. B. (2016). Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City Department of Education High School Study. *The Annals of Applied Statistics*, 10(4):1958–1976.

Branson, Z. and Miratrix, L. W. (2019). Randomization tests that condition on non-categorical covariate balance. *Journal of Causal Inference*, 7(1).

Buehler, R. J. (1959). Some validity criteria for statistical inferences. *The Annals of Mathematical Statistics*, 30(4):845–863.

Buser, M. C., Murray, H. E., and Scinicariello, F. (2014). Association of urinary phenols with increased body weight measures and obesity in children and adolescents. *The Journal of Pediatrics*, 165(4):744–749.

Byar, D. P., Freedman, L. S., and Herzberg, A. M. (1995). Identifying which sets of parameters are simultaneously estimable in an incomplete factorial design. *Journal of the Royal Statistical Society: Series D*, 44(4):451–456.

Byar, D. P., Herzberg, A. M., and Tan, W.-Y. (1993). Incomplete factorial designs for randomized clinical trials. *Statistics in Medicine*, 12(17):1629–1641.

Centers for Disease Control and Prevention (CDC) (2009). Fourth Report on Human Exposure to Environmental Chemicals. *Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention*. https://www.cdc.gov/exposurereport/.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS) (2003-2004). National Health and Nutrition Examination Survey Data. *Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention*. https://www.cdc.gov/nchs/nhanes/. Accessed Oct. 1, 2018.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS) (2013-2014). National Health and Nutrition Examination Survey Data. *Hyattsville, MD: U.S. Department of Health and Human Services, CDC*. https://www.cdc.gov/nchs/nhanes/.

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):684–691.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, New York, 3d edition.

Cochran, W. G. and Cox, G. M. (1950). *Experimental Designs*. John Wiley & Sons, New York, NY.

Coombs, W. T. and Algina, J. (1996). New test statistics for manova/descriptive discriminant analysis. *Educational and Psychological Measurement*, 56(3):382–402.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

Cox, D. R. (2009). Randomization in the design of experiments. *International Statistical Review*, 77(3):415–429.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B*, 77(4):727–753.

Dasgupta, T. and Rubin, D. B. (2015). Harvard university STAT 240: Matched Sampling and Study Design lecture notes and draft textbook, Fall 2015.

Dempster, A. (1960). Random allocation designs I: On general classes of estimation methods. *The Annals of Mathematical Statistics*, 31(4):885–905.

Dempster, A. (1961). Random allocation designs II: Approximate theory for simple random allocation. *The Annals of Mathematical Statistics*, 32(2):387–405.

Ding, P., Li, X., and Miratrix, L. W. (2017). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2).

Dong, N. (2015). Using propensity score methods to approximate factorial experimental designs to analyze the relationship between two variables and an outcome. *American Journal of Evaluation*, 36(1):42–66.

Egami, N. and Imai, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, 114(526):529–540.

Espinosa, V., Dasgupta, T., and Rubin, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics*, 58(1):62–73.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of Ministry of Agriculture*, 33:503–513.

Fisher, R. A. (1930). Inverse probability. In *Proceedings of the Cambridge Philosophical Society*, volume 26, pages 528–535. Cambridge University Press.

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

Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B*, 80(5):1035–1056.

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

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

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis and Interpretation*. Norton, New York.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values. *Statistical Science*, 20(4):358–366.

Ghosh, M., Reid, N., and Fraser, D. A. S. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20(4):1309–1332.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., and Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4(1):61–80.

Higgins, M. J., Sävje, F., and Sekhon, J. S. (2015). Blocking estimators and inference under the Neyman-Rubin model. *arXiv preprint arXiv:1510.01103*.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In Gelman, A. and Meng, X., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, chapter 7, pages 73–84. Hoboken, N.J.: Wiley.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

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

Holt, D. and Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society: Series A*, 142(1):33–46.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.

Iacus, S. M., King, G., and Porro, G. (2016). *cem: Coarsened Exact Matching*. R package version 1.1.17.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27(24):4857–4873.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A*, 171(2):481–502.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76(1):243–263.

Imbens, G. W. (2011). Experimental design for unit and cluster randomid trials. *Conf. International Initiative for Impact Evaluation, Cuernavaca*.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967.

Kosmidis, I. (2017). *brglm: Bias Reduction in Binary-Response Generalized Linear Models*. R package version 0.6.1.

Lachin, J. M. (1988). Properties of simple randomization in clinical trials. *Controlled Clinical Trials*, 9(4):312–326.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer texts in statistics. Springer, New York, 3. edition.

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.

Li, X., Ding, P., and Rubin, D. B. (2020). Rerandomization in 2^K factorial experiments. *The Annals of Statistics*, 48(1):43–63.

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

Liu, K. and Meng, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Application*, 3:79–111.

Lohr, S. L. (2009). *Sampling: Design and Analysis*. Cengage Learning, Boston, 2nd edition.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science*, 32(3):432–454.

Lu, J. (2016a). Covariate adjustment in randomization-based causal inference for 2^K factorial designs. *Statistics & Probability Letters*, 119:11–20.

Lu, J. (2016b). On randomization-based and regression-based inferences for 2^K factorial designs. *Statistics & Probability Letters*, 112:72–78.

Lu, J. and Deng, A. (2017). On randomization-based causal inference for matched-pair factorial designs. *Statistics & Probability Letters*, 125:99–103.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Miratix, L. W. and Pashley, N. E. (2020). blkvar. `https://rdrr.io/github/lmiratrix/blkvar/`.

Miratrix, L., Weiss, M., and Henderson, B. (2020). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. Working paper.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth weighting? How to think about and use weights in survey experiments. *Political Analysis*, 26(3):275–291.

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(2):369–396.

Montgomery, D. C. (2017). *Design and analysis of experiments*. Wiley, New York, NY.

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

Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113(522):868–881.

Nilsson, M. (2013). Causal inference in a 2^2 factorial design using generalized propensity score. Master's thesis, Uppsala University, Sweden.

Oulhote, Y., Bind, M.-A. C., Coull, B., Patel, C. J., and Grandjean, P. (2017). Combining ensemble learning techniques and G-computation to investigate chemical mixtures in environmental epidemiology studies. *bioRxiv*.

Pashley, N. E. and Bind, M.-A. C. (2019). Causal inference for multiple non-randomized treatments using fractional factorial designs. *arXiv preprint arXiv:1905.07596*.

Pashley, N. E. and Miratrix, L. W. (2017). Insights on variance estimation for blocked and matched pairs designs. *arXiv preprint arXiv:1710.10342*.

Patel, C. J., Bhattacharya, J., and Butte, A. J. (2010). An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*, 5(5):e10746.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranjbar, M., Rotondi, M. A., Ardern, C. I., and Kuk, J. L. (2015). The influence of urinary concentrations of organophosphate metabolites on the relationship between BMI and cardiometabolic health risk. *Journal of Obesity*, 2015. Article ID 687914.

Raudenbush, S. W. and Schwartz, D. (2020). Randomized experiments in education, with implications for multilevel causal inference. *Annual Review of Statistics and Its Application*, 7:177–208.

Robinson, G. K. (1979). Conditional properties of statistical procedures. *The Annals of Statistics*, 7(4):742–755.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B*, 53(3):597–610.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York, NY, 2nd edition.

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

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Zhao, Q. (2017). *CrossScreening: Cross-Screening in Observational Studies that Test Many Hypotheses*. R package version 0.1.1.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

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

Rubin, D. B. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

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

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3):527–537.

Sävje, F. (2015). The performance and efficiency of threshold blocking. *arXiv preprint arXiv:1506.02824*.

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13(4):279–313.

Schochet, P. Z. (2016). Statistical theory for the RCT-YES software: Design-based causal inference for RCTs, Second Edition. Technical Report (NCEE 2015-4011), Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Scosyrev, E. (2014). Causal inference in block-randomized experiments: Analysis based on Neyman's stochastic causal model. Unpublished.

Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine*, 23(24):3729–3753.

Snedecor, G. and Cochran, W. (1989). *Statistical Methods*. Iowa State University Press, 8th edition.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472.

StataCorp (2017). Stata Statistical Software: Release 15. *StataCorp. College Station, TX: StataCorp*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1.

Sundberg, R. (2003). Conditional statistical inference and quantification of relevance. *Journal of the Royal Statistical Society: Series B*, 65(1):299–315.

Susmann, H. (2016). *RNHANES: Facilitates Analysis of CDC NHANES Data*. R package version 1.1.0.

Valeri, L., Mazumdar, M. M., Bobb, J. F., Claus Henn, B., Rodrigues, E., Sharif, O. I. A., Kile, M. L., Quamruzzaman, Q., Afroz, S., Golam, M., Amarasiriwardena, C., Bellinger, D. C., Christiani, D. C., Coull, B. A., and Wright, R. O. (2017). The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20-40 months of age: Evidence from rural Bangladesh. *Environmental Health Perspectives*, 125(6). CID:067015.

Wang, Y. H. (2000). Fiducial intervals: What are they? *The American Statistician*, 54(2):105–111.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494.

Wu, C. F. J. and Hamada, M. S. (2000). *Experiments : Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, New York, NY.

Zhao, A., Ding, P., Mukerjee, R., and Dasgupta, T. (2018a). Randomization-based causal inference from split-plot designs. *The Annals of Statistics*, 46(5):1876–1903.

Zhao, Q., Small, D. S., and Rosenbaum, P. R. (2018b). Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association*, 113(523):1070–1084.

Appendix A

Appendix to Chapter 1

A.1 Alternative strategies for variance estimation

In Chapter 1, we examined strategies for variance estimation that put no structure on how the individual blocks may differ from each other. At root, the focus is on estimating the residual variance of units around their block means, and aggregating appropriately.

This section discusses alternatives along with when they might be more or less appropriate. The first two subsections describe estimators that require model assumptions. Therefore these estimators may perform well in certain circumstances, but involve assumptions that we do not make in our analysis. The third subsection describes an estimator used for matched pairs that is proposed in the RCT-YES software documentation. The RCT-YES estimator assumes a specific population model that we do not consider in Chapter 1.

A.1.1 Linear regression

Perhaps the most common method of estimation for randomized trials is to simply fit a linear model to the data with a treatment indicator and a dummy variable for each block. If there is no interaction between the treatment and block dummies, this approach will produce a precision-weighted estimate of the treatment effect, with an overall implicit estimand of a weighted average of the average impacts within each block, weighted by the estimated block precision under a homoscedasticity assumption. If there is impact variation correlated with this precision, then this precision-weighted estimand could be different

than the overall ATE, resulting in a biased estimator. Furthermore, if blocks have different proportions of treated units and different sizes, this weighting might not correspond to any easily interpretable quantity. As pointed out by Freedman, this regression model is also not justified by randomization, which results in complications with the corresponding standard errors (Freedman, 2008a,b). Unfortunately, however, this approach is likely the most common in the field.

The above issues are, in part, repairable. Lin (2013) shows that the estimator from a linear model including interactions between the treatment indicator and block dummies is unbiased. In fact, this estimator is equivalent to the blocked estimator presented in Chapter 1. The question is then how to estimate the standard errors from within the ordinary least squares framework. Lin advocates a Huber-White sandwich estimator for the general covariate case, but these have problems when the blocks have single treated or single control units. In particular, several variants of these estimators, such as HC2 and HC3, will not even be defined due to the characteristics of the corresponding design matrix. The HC0 estimator can still be heavily biased if there is systematic heteroscedasticity across the blocks. Gerber and Green (2012) (p. 116-117) advocate a weighted estimator, but this can also fail in the presence of blocks with singleton treated or control units.

A.1.2 Pooling variance estimates

As we have seen, the most straightforward way to get an overall variance is to obtain variance estimates for all of the block specific estimates, and then combine them in a weighted average. For small blocks, obtaining these block specific estimates is difficult because we do not have enough units to estimate variances of the treatment arms. In this case we have two general options: When treatment effects are considered to be homogenous, we can use the variance of the estimates across the blocks knowing that the variability is dictated purely by block mean variability and not treatment effect variability. An alternative strategy is to use the variance estimates in other blocks to estimate the variances in the intractable blocks. We discuss this next.

One such estimator is to, given a means of assessing how similar blocks are in terms of variance, simply use the variance of the closest big block for each small block. This typically

requires some assumptions that, based on covariate values of the blocks, the variances of the potential outcomes are the same or similar. Similarly, Abadie and Imbens created an estimator of the variance for matched pairs that involves pairing the closest matched pairs and creating a pooled variance estimator for the two blocks together (Abadie and Imbens, 2008). They found that their estimator was asymptotically unbiased given certain conditions, such as the closeness of pairs increasing as the sample size grew. Although the asymptotic results derived in their paper are not necessarily appropriate here, this could be a reasonable plug-in estimator under the assumptions that (i) the covariate(s) that create the strata are related to the potential outcomes and variance and that (ii) the small strata are more similar to each other than the larger strata.

Covariates could also be exploited using linear regression to create variance estimates; see, for example, Fogarty (2018). Or, if we believe that the variance of the estimator in each block is related to the block size, we could fit a linear regression for the big blocks, of variance versus their size, and then extrapolate to the small blocks. Alternatively, if nothing is known and there are very few small blocks, an average (or the largest) of the big block variance estimates might be used. This type of plug-in is used in simulations in Section 1.6.

Any of these plug-in estimators could be used, instead of plugging in for the overall variance, to simply fill in the missing component for small blocks. For instance, if all of the small blocks are such that they have multiple controls but only one treated unit, we can calculate $s_k^2(c)$ as usual but approximate $s_k^2(t)$ based on one of the previously mentioned methods.

There are many other plug-ins that might be appropriate, based on what assumptions the researcher is able to make. The choice of plug-in estimator should be chosen prior to running the experiment and should be based on the researchers assumptions and knowledge at that time. Trying several plug-in estimators and using the smallest will create bias.

A.1.3 The RCT-YES estimator

One might also consider an estimator suggested in the RCT-YES manuscript (Schochet, 2016, p. 83). The form of this estimator, using block sizes as weights, is

$$\widehat{\sigma}^2_{RCT} = \frac{1}{K\,(K-1)\,(\frac{n}{K})^2} \sum_{k=1}^{K} \left(n_k \widehat{\tau}_k - \frac{n}{K} \widehat{\tau}_{(BK)} \right)^2 .$$

As discussed and proven in the RCT-YES documentation, this estimator is consistent under an infinite population of an infinite number of strata of infinite size, where we sample strata and then units within strata. This is the random sampling of strata setting in the remark of Section 1.4.4. This estimator differs from our estimators $\widehat{\sigma}^2_{(SMALL/m)}$ and $\widehat{\sigma}^2_{(SMALL/p)}$ by putting the weights inside the square. Unfortunately, moving the weighting inside the squares can cause large bias in the finite setting and the stratified sampling framework. In fact, in the simulations comparing variance estimator performance in the finite sample, presented in Section 1.6, the RCT-YES bias and variance was high enough that it was not comparable to the other estimators presented. This estimator is targeting a superpopulation quantity, thus the standard errors are larger in part to capture the additional variation of the strata being a random sample.

We discussed the performance of the original RCT-YES estimator with Dr. Schochet (personal correspondence, April 2018), and he proposed an alternate estimator. Again using the block sizes as weights, this estimator has the form

$$\widehat{\sigma}^2_{RCT,2} = \frac{1}{K\,(K-1)\,(\frac{n}{K})^2} \sum_{k=1}^{K} n_k^2 \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)} \right)^2$$

and is rooted in survey sampling methods (Cochran, 1977). This estimator is more stable because the weights are not inside the parentheses. This estimator is still motivated by a superpopulation sampling framework, and takes the variability of the blocks into account. Finite sample performance using this estimator on all of the blocks does not perform well, unless all blocks have the same τ_k, which aligns with what we expect from explorations of our small block estimators. If used in combination as a hybrid estimator, its performance is very similar to that of the hybrid using $\widehat{\sigma}^2_{(SMALL/p)}$.

A.2 Consequences of ignoring blocking

In this section we explore what happens when a blocked randomization was implemented but then the experiment was analyzed as if complete randomization was used. There is a misconception that implementing a blocked design and then ignoring the blocking when calculating the variance estimator will result in an estimator that is conservative for the variance of the $\widehat{\tau}_{(BK)}$.

Theorem A.2.0.1 (Using completely randomized variance estimator for blocked experiment)**.** *In the finite sample setting, analyzing a blocked experiment as if it were completely randomized could give anti-conservative estimators for variance.*

That is, it is possible to have $\mathbb{E}\left[\widehat{\sigma}^2_{(CR)}|\mathcal{S}, \mathbf{P}_{blk}\right] \leq \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}, \mathbf{P}_{blk}\right)$, where $\mathbf{P}_{blk}$ is a blocked randomization assignment mechanism. See Appendix A.9.1 for a derivation that proves this result (assuming $p_k = p$ for all k and with a positive correlation of potential outcomes).

However, in the stratified sampling framework, ignoring blocking when a blocked design was run will always result in a conservative estimator for the variance of $\widehat{\tau}_{(BK)}$.

Corollary A.2.0.1 (Using completely randomized variance estimator for blocked experiment)**.** *Analyzing a blocked experiment as if it were completely randomized will not give anti-conservative estimators for variance if we are analyzing for a superpopulation with fixed blocks and stratified random sampling.*

See Appendix A.9.2 for more on this result (assuming $p_k = p$ for all k).

This discussion is related to the misconception that the complete randomization variance estimator is always more stable than the blocking variance estimator, which is discussed in Appendix A.4.1.

If we do not have $p_k = p$ for all k, on the other hand, then it is possible to have $\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{S}, \mathbf{P}_{blk}\right] \neq \tau_{\mathcal{S}}$. This means $\widehat{\tau}_{(CR)}$ could be biased, even under a constant treatment effect assumption, because some units will be weighted more heavily than others in both the treatment and control arms. In this case, variance comparison is less relevant.

A.3 Details on the numerical studies

A.3.1 Data generating process for simulations

The data generating process used both for the simulations comparing the variance estimators and also those comparing blocking to complete randomization gives us a single finite data set. We then repeatedly randomize units to treatment, according to blocked randomization, multiple times to assess finite sample behavior. Note that we found that the simulation values agreed with biases calculated for our variance estimators via the bias formulas presented in Chapter 1, in Section 1.3.4. Potential outcomes for the units in each block were drawn from a bivariate normal distribution, with the means and covariance matrix as follows (shown for a unit in block k):

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \sim MVN\left(\begin{pmatrix} \alpha_k \\ \alpha_k + \beta_k \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

The correlation of potential outcomes, ρ, was varied among 0, 0.5, and 1. We controlled how differentiated the blocks were, and how heterogeneous the treatment effects across blocks were, by varying α_k and β_k. We set α_k as $\alpha_k = \Phi^{-1}\left(1 - \frac{k}{K+1}\right) a$. Similarly, $\beta_k = 5 + \Phi^{-1}\left(1 - \frac{k}{k+1}\right) b$. The larger the a, the more the mean control potential outcomes for the blocks were spread apart. The larger the b, the more heterogeneous the treatment impacts. The parameters a and b were varied among the values (0,0.1,0.3,0.5,0.8,1,1.5,2). We keep the number and sizes of blocks fixed. The blocks were ordered by size with the smallest block as block one. As a consequence, the smaller blocks have both larger means under control and larger average treatment effects.

Simulations were run over assignment of units to treatments under a blocked design, which was done 5000 times for each combination of factors.

A.3.2 Blocking vs. complete randomization

Here we have numerical examples to explore the potential benefits and costs of blocking in the finite sample framework. These numerical studies look at the actual variances of the treatment effect estimators, not the costs and differences of estimating these variances. The data generating mechanism was the same as in the simulations of Section 1.6. However,

we kept the correlation of potential outcomes at $\rho = 0.5$ in the interest of keeping the plots uncluttered. Other values of ρ gave similar results. We examined a series of scenarios ranging from a collection of blocks where there is little variation from block to block (causing blocking to be less beneficial) to scenarios where the blocks are well separated and blocking is critical for controlling variation. In our first numerical study we treat 20% of all of the blocks, which enforces specific block sizes of 5, 10, 15, and 20. 50% of the units were in small blocks. In the second numerical study we kept the sizes of the blocks the same as in the first numerical study but allowed the proportion treated to vary from block to block, from 0.1 to 0.4. We kept the overall proportion treated the same so that the completely randomized design (which randomized the same total number of units to treatment as the blocked design) was the same in both numerical studies. To do so, the big blocks all had less than 20% treated. The first numerical study corresponds to the mathematical argument presented in Section 1.5 and examines how much blocking can hurt in a variety of scenarios. The second numerical study is to illustrate what changes when proportions treated are not the same across blocks, as discussed in Appendix A.8.6.

We see on the x-axis of Figure A.1 an R^2-like measure of how predictive blocks are of outcome, calculated for each finite data set investigated. R^2 was varied by manipulating the spread of block means under control and the spread of block treatment effects. The y-axis is the ratio of variances of the different treatment effect estimators. Generally, as expected, we see large gains in blocking for moderate to large R^2, and a slight penalty to blocking when the R^2 is relatively small. In this example with varying proportions treated, the gains of blocking is muted. The difference between equal and unequal proportions is potentially augmented by the fact that, to hold the overall proportion treated constant, the four largest blocks, of sizes 20, 15 and 10, (which have a larger overall impact on variance) had roughly proportion 0.1 treated.

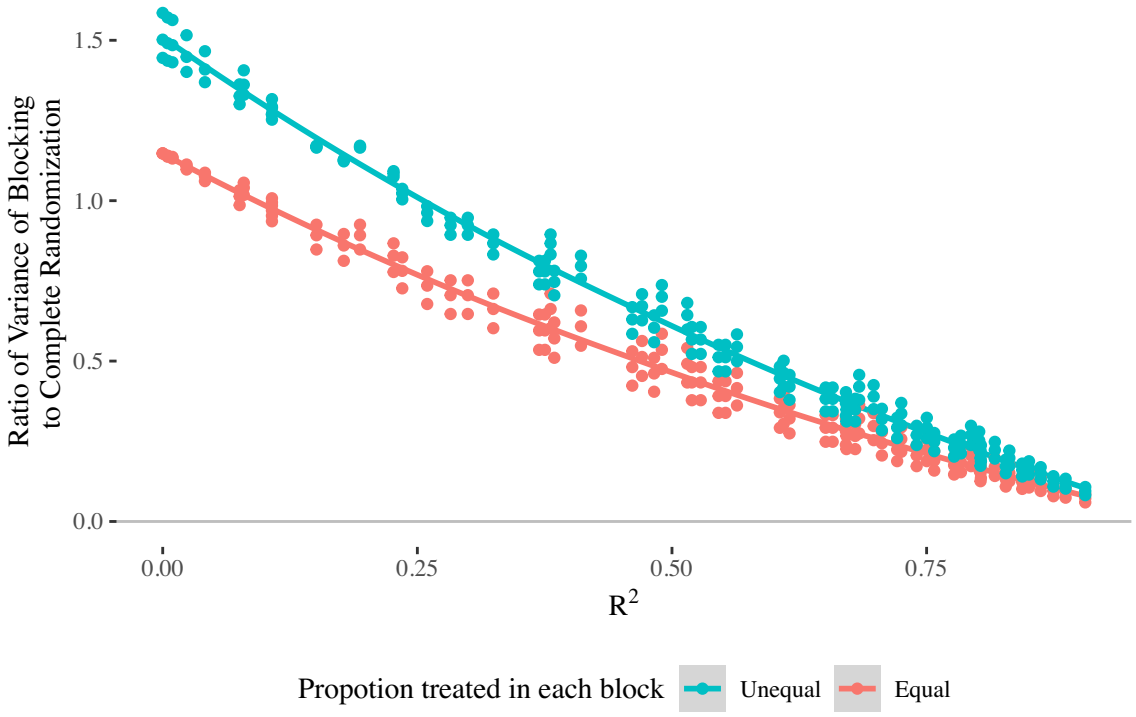

Figure A.1: *Numerical study to assess completely randomized versus blocked design, when* $p_k = 0.2$ *(equal proportions) or unequal proportions across blocks. The y axis is* $\frac{Var(\hat{\tau}_{(BK)}|\mathcal{S})}{Var(\hat{\tau}_{(CR)}|\mathcal{S})}$.

A.4 The variance of the variance estimators

A.4.1 Blocking versus complete randomization

Discussions of blocking versus complete randomization often include a discussion on the performance of the variance estimators in terms of their own variance. There is a misconception that the variance of the blocking variance estimator will be larger than that of the complete randomization variance estimator. For instance, the standard big block variance estimator ($\widehat{\sigma}^2_{(BK)}$) under the stratified sampling framework was assessed in Imbens (2011). In particular, in this piece there is a discussion about how the true variance of the blocking estimator may be lower than under complete randomization but that the estimate of that variance may be more variable under blocking than under complete randomization. We consider that discussion here. First, assume that $n_{z,k} \geq 2$ for all blocks k and all treatment assignments z. Then, as we have seen, $\widehat{\sigma}^2_{(BK)}$ and $\widehat{\sigma}^2_{(CR)}$ are unbiased estimators of $\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1)$ and $\text{var}(\widehat{\tau}_{(CR)}|SRS)$, respectively. This implies, based on results from Section A.8.4, that on average the variance estimator under the blocked design is less than that under the completely randomized design, as noted by Imbens (2011).

In Imbens (2011), p. 11, a further claim is made that

$$\text{var}(\widehat{\sigma}^2_{(BK)}) \geq \text{var}(\widehat{\sigma}^2_{(CR)})$$

It is not entirely clear whether the variance of $\widehat{\sigma}^2_{(CR)}$ is with respect to SRS or $\mathcal{F}_1$ but let us assume that it is with respect to $\mathcal{F}_1$ (our example extends easily to considering complete randomization with respect to SRS instead). This statement is only true if $\text{var}(s^2(c)|\mathcal{F}_1) \leq \text{var}(s^2_k(c)|\mathcal{F}_1)$ and $\text{var}(s^2(t)|\mathcal{F}_1) \leq \text{var}(s^2_k(t)|\mathcal{F}_1)$ for most $k = 1, ..., K$. Imbens (2011) gave an example in which this might be true. In his example, there is no variance in the potential outcomes under control ($\sigma^2_k(c) = 0$ for all $k = 1, ...K$) and the distribution of the potential outcomes under treatment is the same in all of the strata ($\sigma^2_k(t) = \sigma^2(t)$ for all $k = 1, .., K$). Then, Imbens argues, because $s^2(t)$ is a less noisy estimator of $\sigma^2(t)$ than any of the $s^2_k(t)$, the variance of the variance estimator would be smaller under the completely randomized design. This notion of the variance estimator for complete randomization being less noisy as it is using more data may be true in many situations. However, this result does not hold

in general. For instance, consider a population with four strata. Within each stratum, there is zero treatment effect and all units are identical. Between strata, however, the potential outcomes differ. For convenience, say in stratum one $Y_i(c) = Y_i(t) = 1$, in stratum two $Y_i(c) = Y_i(t) = 2$, in stratum three $Y_i(c) = Y_i(t) = 3$, and in stratum four $Y_i(c) = Y_i(t) = 4$. Now assume that four units are sampled from each of the strata. In a blocked design, our variance estimate would always be 0. But in a completely randomized design, the variance estimate would change based on which units were assigned to treatment and control. Thus, the blocking variance estimator would have 0 variance whereas the completely randomized variance estimator would have non-zero variance.

So, when blocking is "good", we expect the true variance of our treatment effect estimator to be lower under blocking than complete randomization and the variance of our variance estimator to also be lower. However, when the blocking is "bad" we would expect blocking to not be beneficial in terms of variance of our estimator and in this case the variance of our variance estimator could also be higher than under complete randomization.

A.4.2 Variance simulations

The previous discussion relates how variances of variance estimators differ under blocking and complete randomization. This raises the question of how the blocking variance estimators compare amongst themselves in terms of variance. To assess this, we examine the variance of our variance estimators from our simulation study in Section 1.6 with data generating process given in Section A.3.1.

Results are on Figure A.2. We see that, in terms of variance, the estimators are generally comparable. We expect more instability from estimators that utilize only information from the estimated average treatment effects, not from the variation of the individual units. We see that the weighted regression estimator has the lowest variability.

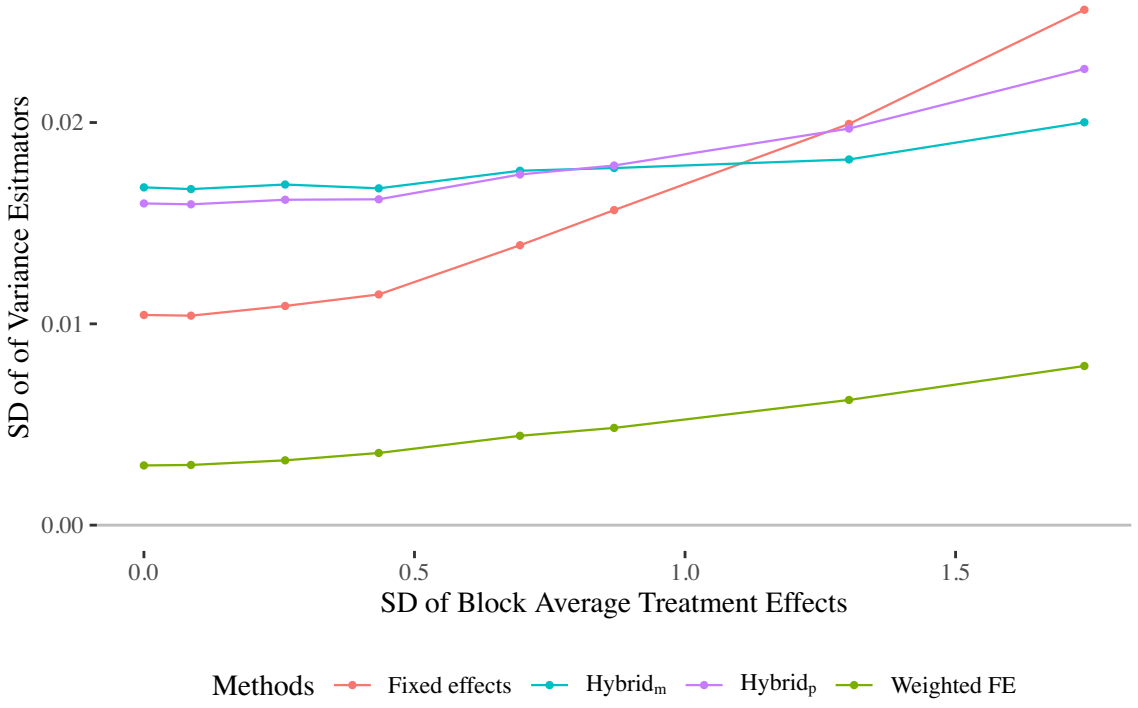

Figure A.2: *Simulations to assess variance estimators' variance. The x-axis shows the standard deviation of block average treatment effects. Points show the average of values of ρ and the standard deviation of block average control potential outcomes in the simulation. FE stands for fixed effects.*

A.5 Creation and bias of $\widehat{\sigma}^2_{(SMALL/m)}$

A.5.1 Creation of $\widehat{\sigma}^2_{(SMALL/m)}$, Equation (1.5)

To formally state the method described by Equation (1.5) from Section 1.3.2, we first express our estimands and estimators in terms of weighted averages of estimates within collections of same-size blocks. In the following, let there be J unique block sizes in the sample. Let m_j be the jth block size and let K_j be the number of blocks in the population of size m_j. So then $n = \sum_{j=1}^{J} m_j K_j$. In particular, the sample average treatment effect for all units in blocks of size m_j is

$$\tau_{(SMALL),\mathcal{S},j} = \frac{1}{K_j} \sum_{k:n_k=m_j} \tau_{k,\mathcal{S}}.$$

Let $N_j = m_j K_j$ be the total number of units in the small blocks of size m_j. Then the overall sample average treatment effect in terms of these $\tau_{(SMALL),\mathcal{S},j}$ is

$$\tau_{(SMALL),\mathcal{S}} = \frac{1}{\sum_{i=1}^{J} N_j} \sum_{j=1}^{J} N_j \tau_{(SMALL),\mathcal{S},j}. \tag{A.1}$$

$\tau_{(SMALL),\mathcal{S}}$ is the same as $\tau_{\mathcal{S}}$ as before; we add the subscript "small" here to clarify the notation when we discuss hybrid experiments. Note that these definitions are analogous for the infinite population, which are indicated by removing the $\mathcal{S}$ subscript.

The treatment effect estimators can be written in analogous form to the above. We have unbiased estimators $\widehat{\tau}_{(SMALL),j} = \frac{1}{K_j} \sum_{k:n_k=m_j} \widehat{\tau}_k$ for the average treatment effects in blocks of size m_j. Simply plug them into Equation A.1 to obtain an overall treatment effect estimator.

As discussed in Section 1.3.2, within each piece j, use a variance estimator with the same form as Equation 1.4:

$$\widehat{\sigma}^2_{(SMALL),j} = \frac{1}{K_j(K_j-1)} \sum_{k:n_k=m_j} (\widehat{\tau}_k - \widehat{\tau}_{(SMALL),j})^2.$$

Then combine to create an overall variance estimator:

$$\widehat{\sigma}^2_{(SMALL/m)} = \frac{1}{\left(\sum_{j=1}^{J} N_j\right)^2} \sum_{j=1}^{J} N_j^2 \widehat{\sigma}^2_{(SMALL),j}.$$

A.5.2 Bias of $\widehat{\sigma}^2_{(SMALL/m)}$

Proof of Corollary 1.3.4.1 and Corollary 1.4.3.1

Proof. For this section assume that we are in the finite sample framework. The results for the stratified sampling from an infinite population framework follow directly by changing the expectations and notation.

First we will focus on $\widehat{\sigma}^2_{(SMALL),j}$ which is the variance estimator for $\widehat{\tau}_{(SMALL),j}$. Note that

$$\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) = \text{var}\left(\frac{1}{K_j}\sum_{k:n_k=m_j}\widehat{\tau}_k\Big|\mathcal{S}\right) = \frac{1}{K_j^2}\sum_{k:n_k=m_j}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right).$$

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right] \\
&= \mathbb{E}\left[\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}(\widehat{\tau}_k - \widehat{\tau}_{(SMALL),j})^2\Big|\mathcal{S}\right] \\
&= \frac{1}{K_j(K_j-1)}\mathbb{E}\left[\sum_{k:n_k=m_j}\left(\widehat{\tau}_k^2 - 2\widehat{\tau}_k\widehat{\tau}_{(SMALL),j} + \widehat{\tau}^2_{(SMALL),j}\right)\Big|\mathcal{S}\right] \\
&= \frac{1}{K_j(K_j-1)}\left(\sum_{k:n_k=m_j}\left[\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \tau^2_{k,\mathcal{S}}\right] - K_j\left[\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) + \tau^2_{(SMALL),\mathcal{S},j}\right]\right) \\
&= \frac{1}{K_j(K_j-1)}\left(\sum_{k:n_k=m_j}\left[\frac{K_j-1}{K_j}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \tau^2_{k,\mathcal{S}}\right] - K_j\tau_{(SMALL),\mathcal{S},j}\right) \\
&= \frac{1}{K_j^2}\sum_{k:n_k=m_j}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2 \\
&= \text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) + \frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.
\end{aligned}$$

So the bias is

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) = \frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.$$

Now we move our attention to $\widehat{\sigma}^2_{(SMALL/m)}$ which is a variance estimator for $\widehat{\tau}_{(SMALL)}$.

We have

$$\begin{aligned}\text{var}\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right) &= \text{var}\left(\frac{1}{\sum_{i=1}^{J} m_i K_i}\sum_{j=1}^{J} m_j K_j \widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) \\ &= \frac{1}{\left(\sum_{i=1}^{J} m_i K_i\right)^2}\sum_{j=1}^{J}\left(m_j K_j\right)^2 \text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right).\end{aligned}$$

So then

$$\begin{aligned}&\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{S}\right] \\ &= \frac{1}{\left(\sum_{i=1}^{J} m_i K_i\right)^2}\sum_{j=1}^{J}\left(m_j K_j\right)^2 \mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right] \\ &= \text{var}\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right) + \sum_{j=1}^{J}\frac{m_j^2 K_j}{\left(\sum_{i=1}^{J} m_i K_i\right)^2 (K_j - 1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.\end{aligned}$$

So the bias is

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right) = \sum_{j=1}^{J}\frac{m_j^2 K_j}{\left(\sum_{i=1}^{J} m_i K_i\right)^2 (K_j - 1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.$$

□

Proof of Corollary 1.4.4.1

Proof. Assume that we are in the random sampling of strata framework of Section 1.4.4. We are focusing in on just the set of strata of size m_j. We can either consider that there only exist strata of this size or we can imagine a sampling mechanism that draws these strata independently from strata of other sizes (e.g. stratified sampling by strata size), which is the same as conditioning on the number of strata of each size in the sample. Also,

$$\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{F}_2\right) = \mathbb{E}\left[\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)|\mathcal{F}_2\right] + \text{var}\left(\mathbb{E}\left[\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right]|\mathcal{F}_2\right).$$

From Appendix A.5.2, we can see that

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right] = \text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) + \frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.$$

From standard results from sampling theory (see (Lohr, 2009, Chapter 2)) we have

$$\mathbb{E}\left[\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2|\mathcal{F}_2\right]=\frac{\sigma_\tau^2}{K_j}=\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{F}_2\right).$$

Hence, we end up with

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{F}_2\right]=\mathbb{E}\left[\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)|\mathcal{F}_2\right]+\text{var}\left(\mathbb{E}\left[\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right]|\mathcal{F}_2\right).$$

Thus, this is an unbiased variance estimator in this setting.

If we have varying size but condition on the number of strata of each size that we sample, then we use the fact that stratified sampling causes the estimators for each strata size to be independent.

The proof of this result for an infinite number of infinite size strata is direct by replacing the conditioning on $\mathcal{S}$ by conditioning on $\boldsymbol{B}$ and using results from the stratified sampling framework. □

A.6 Creation and bias of $\widehat{\sigma}^2_{(SMALL/p)}$

A.6.1 Proof of Corollary 1.3.4.2 and Corollary 1.4.3.2

Proof. For this section assume that we are in the finite sample framework. The results for the stratified sampling framework follow directly by changing what we are taking the expectation with respect to and exchanging notation. We explain how $\widehat{\sigma}^2_{(SMALL/p)}$ was derived which also provides the bias of the estimator in these two frameworks.

To begin we consider, for an experiment with all small blocks, a variance estimator of the form

$$X\equiv\sum_{k=1}^{K}a_k\left(\widehat{\tau}_k-\widehat{\tau}_{(BK)}\right)^2$$

for some collection of a_k. We then wish to find non-negative a_k's that would make this estimator as close to unbiased as possible. In particular, we aim to create an estimator with similar bias to $\widehat{\sigma}^2_{(SMALL/m)}$ but that allows for blocks of varying size. That is, we are looking to create a similarly conservative estimator that is unbiased when the average treatment effect is constant across blocks. This also means that we are creating the minimally biased

conservative estimator of this form, without further assumptions.

The expected value of an estimator of this form is

$$
\begin{aligned}
&\mathbb{E}\left[X|\mathcal{S}\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K} a_k \left(\widehat{\tau}_k - \tau_{k,\mathcal{S}} + \tau_{k,\mathcal{S}} - \tau_{\mathcal{S}} + \tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)^2 |\mathcal{S}\right] \\
=&\mathbb{E}\Big[\sum_{k=1}^{K} a_k \Big(\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)^2 + \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 + \left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)^2 \\
&+ 2\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right) + 2\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right) + 2\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right) \Big) |\mathcal{S}\Big] \\
=&\sum_{k=1}^{K} a_k \left(\mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \mathbb{E}\left[\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 |\mathcal{S}\right] + \underbrace{\mathbb{E}\left[\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)^2 |\mathcal{S}\right]}_{\mathbf{A}} + 2\underbrace{\mathbb{E}\left[\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)|\mathcal{S}\right]}_{\mathbf{B}} \right)
\end{aligned}
$$

For **A**:

$$
\mathbb{E}\left[\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)^2 |\mathcal{S}\right] = \mathrm{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right)
$$

For **B**:

$$
\begin{aligned}
\mathbb{E}\left[\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)|\mathcal{S}\right] &= \mathbb{E}\left[\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right) \sum_{j=1}^{K} \frac{n_j}{n}\left(\tau_{j,\mathcal{S}} - \widehat{\tau}_j\right) \Big| \mathcal{S}\right] \\
&= \mathbb{E}\left[-\frac{n_k}{n}\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)^2 + \left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right) \sum_{j\neq k} \frac{n_j}{n}\left(\tau_{j,\mathcal{S}} - \widehat{\tau}_{j,\mathcal{S}}\right) \Big| \mathcal{S}\right] \\
&= -\frac{n_k}{n} \mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right)
\end{aligned}
$$

Due to the assignment mechanism, $\widehat{\tau}_j$ will be independent of $\widehat{\tau}_k$ so the cross terms are all zero in the above equation.

Putting **A** and **B** together, we get

$$
\begin{aligned}
\mathbb{E}\left[X|\mathcal{S}\right] = &\sum_{k=1}^{K} a_k \mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 + \sum_{k=1}^{K} a_k \sum_{j=1}^{K} \frac{n_j^2}{n^2} \mathrm{var}\left(\widehat{\tau}_j|\mathcal{S}\right) \\
&- 2\sum_{k=1}^{K} a_k \frac{n_k}{n} \mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right) \\
= &\sum_{k=1}^{K} \left(a_k - 2a_k \frac{n_k}{n} + \frac{n_k^2}{n^2} \sum_{j=1}^{K} a_j\right) \mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2
\end{aligned}
$$

We now select a_k to make the above as close to the true variance as possible. The second term will be small if the $\tau_{k,\mathcal{S}}$ do not vary much. But this is unknown and thus we cannot select universal a_k to control it. We would like

$$a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j = \frac{n_k^2}{n^2}$$

so that the first term is the true variance. Then the second term, the bias, would be similar to that of the standard matched pairs variance estimator.

If we solve the a_k as above we will obtain a conservative estimator that is unbiased when we have equal average treatment effect for all blocks, for the stratified sampling or finite framework. To show this, consider the bias:

$$\begin{aligned}
&\mathbb{E}\left[X|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&\qquad = \sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j - \frac{n_k^2}{n^2}\right)\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2. \qquad \text{(A.2)}
\end{aligned}$$

We know $\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) \geq 0$ for all k. We also have $\sum_{k=1}^{K} a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \geq 0$ so at a minimum it is 0. This implies that to always be conservative, the first term in the above expression must always be at least 0. Hence, to minimize the bias but remain conservative, we set $a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j - \frac{n_k^2}{n^2} = 0$. Note that $\tau_{k,\mathcal{S}}$ and $\tau_{\mathcal{S}}$ are unknown and so we cannot optimize with respect to them.

Denote $C = \sum_{k=1}^{K} a_k$. Then we want to solve

$$\begin{aligned}
a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}C &= \frac{n_k^2}{n^2} \\
a_k(1 - 2\frac{n_k}{n}) &= \frac{n_k^2}{n^2}(1 - C) \\
a_k &= \frac{n_k^2}{n}\frac{1 - C}{n - 2n_k}
\end{aligned}$$

But then

$$C = \sum_{k=1}^{K} a_k = \sum_{k=1}^{K} \frac{n_k^2(1-C)}{n(n-2n_k)}$$

$$\left(1 + \frac{1}{n}\sum_{k=1}^{K} \frac{n_k^2}{n-2n_k}\right) C = \frac{1}{n}\sum_{k=1}^{K} \frac{n_k^2}{n-2n_k}$$

$$C = \frac{1}{n}\sum_{k=1}^{K} \frac{n_k^2}{n-2n_k}\left(\frac{1}{1+\frac{1}{n}\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j}}\right) = \frac{\sum_{k=1}^{K}\frac{n_k^2}{n-2n_k}}{n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j}}$$

Then we have

$$a_k = \frac{n_k^2}{n}\left(\frac{1-C}{n-2n_k}\right) = \frac{n_k^2}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}.$$

So then

$$\sum_{k=1}^{K} \frac{n_k^2}{(n-2n_k)\left(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j}\right)}\left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2$$

has bias

$$\sum_{k=1}^{K} \frac{n_k^2}{(n-2n_k)\left(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j}\right)}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2.$$

Bigger strata get weighted more heavily.

As a check, in the case where the n_k are all the same, so that $n = Kn_k$, the above boils down to $a_k = \frac{1}{K(K-1)}$ and $C = \frac{1}{K-1}$, giving us the classic matched pairs variance estimator. □

A.6.2 Proof of Theorem 1.4.4.1: Unbiasedness of $\widehat{\sigma}^2_{(SMALL/p)}$ given independence

Proof. We start from the Equation A.2. As in Section A.6, let

$$X \equiv \sum_{k=1}^{K} a_k \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2.$$

Then

$$\mathbb{E}\left[X|\mathcal{S}\right] = \sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K}a_j\right)\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K}a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2.$$

Previously we were concerned with getting the first term correct. But in the Random Sampling of Strata setting, the second term is trickier. This is especially the case if we have large blocks and thus can estimate the first term. So first we focus on the second term. Keeping the variance decomposition in mind, we ultimately want the expectation of this second term to look like $\text{var}(\tau_{\mathcal{S}})$.

As a reminder, the variance decomposition is

$$\begin{aligned}\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_2) &= \mathbb{E}\left[\text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_2\right] + \text{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|\mathcal{F}_2\right)\\ &= \mathbb{E}\left[\text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_2\right] + \text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_2\right).\end{aligned}$$

We assume, for simplicity, that the block sizes are independent from the treatment effects. This implies that $\mathbb{E}[\tau_{k,\mathcal{S}}|\mathcal{F}_2] = \tau$.

The expected value of the second term in this setting is

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2\Big|\mathcal{F}_2\right]\\ =&\mathbb{E}\left[\sum_{k=1}^{K}a_k\tau_{k,\mathcal{S}}^2 - 2\tau_{\mathcal{S}}\sum_{k=1}^{K}a_k\tau_{k,\mathcal{S}} + \tau_{\mathcal{S}}^2\sum_{k=1}^{K}a_k\Big|\mathcal{F}_2\right]\\ =&\mathbb{E}\left[\sum_{k=1}^{K}a_k\tau_{k,\mathcal{S}}^2 - 2\left(\sum_{k=1}^{K}a_k\frac{n_k}{n}\tau_{k,\mathcal{S}}^2 + \sum_{k=1}^{K}\sum_{j\neq k}a_k\frac{n_j}{n}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\right)\right.\\ &\left. + \left(\sum_{k=1}^{K}\frac{n_k^2}{n^2}\tau_{k,\mathcal{S}}^2 + \sum_{k=1}^{K}\sum_{j\neq k}\frac{n_kn_j}{n^2}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\right)\sum_{k=1}^{K}a_k\Big|\mathcal{F}_2\right]\\ =&\mathbb{E}\left[\sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{i=1}^{K}a_i\right)\tau_{k,\mathcal{S}}^2 - \sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n} - \frac{n_kn_j}{n^2}\sum_{i=1}^{K}a_i\right)\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\Big|\mathcal{F}_2\right]\\ =&\mathbb{E}\left[\sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{i=1}^{K}a_i\right)\tau_{k,\mathcal{S}}^2\Big|\mathcal{F}_2\right] - \mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n} - \frac{n_kn_j}{n^2}\sum_{i=1}^{K}a_i\right)\Big|\mathcal{F}_2\right]\tau^2.\end{aligned}$$

Now consider the true variance, which we are trying to estimate.

$$\begin{aligned}
\text{var}(\tau_{\mathcal{S}}|\mathcal{F}_2) &= \text{var}\left(\sum_{k=1}^{K}\frac{n_k}{n}\tau_{k,\mathcal{S}}\Big|\mathcal{F}_2\right) \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{K}\frac{n_k}{n}\tau_{k,\mathcal{S}}-\tau\right)^2\Big|\mathcal{F}_2\right] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{K}\frac{n_k}{n}\tau_{k,\mathcal{S}}\right)^2\Big|\mathcal{F}_2\right]-\tau^2 \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\tau_{k,\mathcal{S}}^2\Big|\mathcal{F}_2\right]+\mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\frac{n_k n_j}{n^2}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\Big|\mathcal{F}_2\right]-\tau^2 \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\tau_{k,\mathcal{S}}^2\Big|\mathcal{F}_2\right]-\mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\Big|\mathcal{F}_2\right]\tau^2
\end{aligned}$$

We have the last equality because

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\frac{n_k n_j}{n^2}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\Big|\mathcal{F}_2\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\frac{n_k n_j}{n}\Big|\mathcal{F}_2\right]\tau^2 \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}(1-\frac{n_k}{n})\Big|\mathcal{F}_2\right]\tau^2 \\
&= \mathbb{E}\left[1-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\Big|\mathcal{F}_2\right]\tau^2.
\end{aligned}$$

Matching this up with the expectation of our estimator, we want

$$\frac{n_k^2}{n^2} = a_k - 2a_k\frac{n_k}{n}+\frac{n_k^2}{n^2}\sum_{i=1}^{K}a_i$$

and

$$\sum_{k=1}^{K}\frac{n_k^2}{n^2} = \sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n}-\frac{n_k n_j}{n^2}\sum_{i=1}^{K}a_i\right).$$

The first equation we solved for before in Section A.6. So we will get

$$a_k = \frac{n_k^2}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}.$$

Let's see if this weight works for the second term.

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n}-\frac{n_k n_j}{n^2}\sum_{i=1}^{K}a_i\right)\\
&=\sum_{k=1}^{K}2a_k(1-\frac{n_k}{n})-(1-\sum_{k=1}^{K}\frac{n_k^2}{n^2})\sum_{i=1}^{K}a_i\\
&=\sum_{k=1}^{K}\frac{2n_k^2(1-\frac{n_k}{n})}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}-(1-\sum_{k=1}^{K}\frac{n_k^2}{n^2})\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2(1-2\frac{n_k}{n})}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}+\sum_{k=1}^{K}\frac{n_k^2}{n^2}\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2(n-2n_k)}{n(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}+\sum_{k=1}^{K}\frac{n_k^2}{n^2}\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2}{n(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}+\sum_{k=1}^{K}\frac{n_k^2}{n^2}\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2(n+\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)})}{n^2(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2}{n^2}
\end{aligned}
$$

Which is exactly what we wanted. So this weight works. And because it is the same as the weight for the finite sample (where we wanted to get the first term in the variance decomposition correct), this also takes care of the first term in the variance decomposition.

The proof of this result for an infinite number of infinite size strata is direct by replacing the conditioning on $\mathcal{S}$ by conditioning on $\boldsymbol{B}$ and using results from the stratified sampling framework. □

A.7 Creation of $\widehat{\sigma}^2_{SRS}$, Equation (1.8)

First, we start with the basic variance decomposition to examine what we are trying to estimate.

$$
\begin{aligned}
\operatorname{var}(\widehat{\tau}_{(BK)}|SRS) &= \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|SRS\right] + \operatorname{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|SRS\right) \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \operatorname{var}\left(\tau_{\mathcal{S}}|SRS\right) \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \frac{\sigma^2(tc)}{n} \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \mathbb{E}\left[\frac{S^2(tc)}{n}\Big|SRS\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right)\Big|SRS\right]}_{\mathbf{A}} + \underbrace{\mathbb{E}\left[\frac{S^2(tc)}{n} - \sum_{k=1}^{K} \frac{n_k}{n}\frac{S_k^2(tc)}{n}\Big|SRS\right]}_{\mathbf{B}}
\end{aligned}
\tag{A.3}
$$

Let's examine $S^2(tc)$ so we can simplify term **B**.

$$
S^2(tc) = \sum_{k=1}^{K} \frac{n_k - 1}{n-1} S_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2
$$

So term **B** can simplify as follows:

$$
\begin{aligned}
S^2(tc) - \sum_{k=1}^{K} \frac{n_k}{n} S_k^2(tc) &= \sum_{k=1}^{K} \frac{n_k - 1}{n-1} S_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 - \sum_{k=1}^{K} \frac{n_k}{n} S_k^2(tc) \\
&= \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 - \sum_{k=1}^{K} \frac{n - n_k}{n(n-1)} S_k^2(tc).
\end{aligned}
$$

Now recall from Appendix A.6 that

$$
\mathbb{E}\left[\sum_{k=1}^{K} a_k \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 |\mathcal{S}\right] = \sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j\right)\operatorname{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2.
$$

Letting $a_k = n_k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} n_k \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 \Big| \mathcal{S}\right] \\
&= \sum_{k=1}^{K} \left(n_k - \frac{n_k^2}{n}\right) \text{var}\left(\widehat{\tau}_k | \mathcal{S}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \text{var}\left(\widehat{\tau}_k | \mathcal{S}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 - \sum_{k=1}^{K} \frac{(n-n_k)}{n} S_k^2(tc) \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + (n-1)\left(S^2(tc) - \sum_{k=1}^{K} \frac{n_k}{n} S_k^2(tc)\right).
\end{aligned}$$

This means that

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 \Big| \mathcal{S}\right] = \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n^2(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + \frac{S^2(tc)}{n} - \sum_{k=1}^{K} \frac{n_k}{n} \frac{S_k^2(tc)}{n}.$$

So we have a way to estimate term **B** of Equation A.3, which means we just need to add in a correction to get term **A**.

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{n_k^2}{n^2} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) - \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n^2(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) \\
&= \sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right)
\end{aligned}$$

Putting it all together, we have

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 \Big| SRS\right] = \text{var}(\widehat{\tau}_{(BK)} | SRS).$$

So

$$\widehat{\sigma}_{SRS}^2 = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{s_k^2(c)}{n_{c,k}} + \frac{s_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2$$

is an unbiased variance estimator.

A.8 Derivations of blocking versus complete randomization differences

Unless otherwise noted, the following section assumes that $p_k = p$ for all $k = 1, \ldots, K$.

A.8.1 Finite sample, Equation (1.11)

We start by deriving the result of Equation (1.11),

$$\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&\quad = \sum_{k=1}^{K} \frac{1}{n(n-1)} \Bigg[n_k \left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t) \right) \right)^2 \\
&\quad - \frac{n-n_k}{n(n_k-1)} \sum_{i:b_i=k} \left(\sqrt{\frac{p}{1-p}}Y_i(c) + \sqrt{\frac{1-p}{p}}Y_i(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) \right) \right)^2 \Bigg].
\end{aligned}$$

We have from usual results (Imbens and Rubin, 2015, p. 89), in addition to independence of treatment assignment within blocks and the assumption that $p_k = p$ for $k = 1, ..., K$, that

$$\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) = \frac{S^2(c)}{n_c} + \frac{S^2(t)}{n_t} - \frac{S^2(tc)}{n} = \frac{S^2(c)}{(1-p)n} + \frac{S^2(t)}{pn} - \frac{S^2(tc)}{n}$$

and

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k} \right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \left(\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k} \right).$$

To take the difference, we need to write the complete randomization variance in terms of the block components. First we look at the expansion of $S^2(z)$:

$$\begin{aligned}
S^2(z) &= \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i(z) - \bar{Y}(z)\right)^2 \\
&= \frac{1}{n-1} \sum_{k=1}^{K} \sum_{i:b_i=k} \left(Y_i(z) - \bar{Y}_k(z) + \bar{Y}_k(z) - \bar{Y}(z)\right)^2 \\
&= \frac{1}{n-1} \sum_{k=1}^{K} \left[(n_k - 1)S_k^2(z) + n_k \left(\bar{Y}_k(z) - \bar{Y}(z)\right)^2 \right] \\
&= \sum_{k=1}^{K} \frac{n_k - 1}{n-1} S_k^2(z) + \sum_{k=1}^{K} \frac{n_k}{n-1} \left(\bar{Y}_k(z) - \bar{Y}(z)\right)^2.
\end{aligned}$$

Now do the same for $S^2(tc)$:

$$S^2(tc) = \sum_{k=1}^{K} \frac{n_k - 1}{n-1} S_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n-1} \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2.$$

$$\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&= \frac{S^2(c)}{(1-p)n} + \frac{S^2(t)}{pn} - \frac{S^2(tc)}{n} - \left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\right)\right] \\
&= \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(c) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n} + \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(t) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn} \\
&- \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2}{n} \\
&- \left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\right)\right] \\
&= \underbrace{\frac{\sum_{k=1}^{K} n_k\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n(n-1)} + \frac{\sum_{k=1}^{K} n_k\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn(n-1)} - \frac{\sum_{k=1}^{K} n_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2}{n(n-1)}}_{\mathbf{A}} \\
&\underbrace{- \sum_{k=1}^{K}\left[\left(\frac{n_k}{(1-p)n^2} - \frac{n_k-1}{(1-p)n(n-1)}\right) S_k^2(c) + \left(\frac{n_k}{pn^2} - \frac{n_k-1}{pn(n-1)}\right) S_k^2(t) - \left(\frac{n_k}{n^2} - \frac{n_k-1}{n(n-1)}\right) S_k^2(tc)\right]}_{\mathbf{B}}
\end{aligned}$$

We have now split our calculation into the between block variation and within block variation pieces.

A is the between block variation:

We need to expand $\sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2$:

$$\begin{aligned}
\sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 &= \sum_{k=1}^{K} n_k \left(\bar{Y}_k(t) - \bar{Y}_k(c) - \left(\bar{Y}(t) - \bar{Y}(c)\right)\right)^2 \\
&= \sum_{k=1}^{K} n_k \left[\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2 + \left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2 - 2\left(\bar{Y}_k(t) - \bar{Y}(t)\right)\left(\bar{Y}(c) - \bar{Y}(c)\right)\right]
\end{aligned}$$

So then

$$\begin{aligned}
\mathbf{A} &= \frac{\sum_{k=1}^{K} n_k \left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n(n-1)} + \frac{\sum_{k=1}^{K} n_k \left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn(n-1)} \\
&- \frac{\sum_{k=1}^{K} n_k \left[\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2 + \left(\bar{Y}(c) - \bar{Y}(c)\right)^2 - 2\left(\bar{Y}_k(t) - \bar{Y}(t)\right)\left(\bar{Y}(c) - \bar{Y}(c)\right)\right]}{n(n-1)} \\
&= \frac{\sum_{k=1}^{K} pn_k \left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n(n-1)} + \frac{\sum_{k=1}^{K}(1-p)n_k \left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn(n-1)} + 2\frac{\left(\bar{Y}_k(t) - \bar{Y}(t)\right)\left(\bar{Y}(c) - \bar{Y}(c)\right)}{n(n-1)} \\
&= \frac{1}{n-1}\sum_{k=1}^{K} \frac{n_k}{n}\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2 \\
&= \frac{1}{(n-1)}\mathrm{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right).
\end{aligned}$$

B is the within block variation:

$$\begin{aligned}
\mathbf{B} &= \sum_{k=1}^{K}\Big[\left(\frac{n-n_k}{(1-p)n^2(n-1)}\right)S_k^2(c) + \left(\frac{n-n_k}{pn^2(n-1)}\right)S_k^2(t) - \left(\frac{n-n_k}{n^2(n-1)}\right)S_k^2(tc)\Big] \\
&= \frac{1}{n^2(n-1)}\sum_{k=1}^{K}(n-n_k)n_k\Big[\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\Big] \\
&= \frac{1}{n^2(n-1)}\sum_{k=1}^{K}(n-n_k)n_k\mathrm{var}(\widehat{\tau}_k|\mathcal{S})
\end{aligned}$$

Then we can write the difference as

$$\begin{aligned}
&\mathrm{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \mathrm{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&\quad= \frac{1}{n-1}\left[\mathrm{var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^{K}\frac{(n-n_k)n_k}{n^2}\mathrm{var}(\widehat{\tau}_k|\mathcal{S})\right] \\
&\quad= \sum_{k=1}^{K}\frac{n_k}{n(n-1)}\Big[\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2 \\
&\qquad - \frac{n-n_k}{n}\mathrm{var}(\widehat{\tau}_k|\mathcal{S})\Big].
\end{aligned}$$

To write this another way, note that

$$\begin{aligned}
S_k^2(tc) &= \frac{1}{n_k - 1}\sum_{i:b_i=k}\left(\tau_i - \tau_{fs,k}\right)^2 \\
&= \frac{1}{n_k - 1}\sum_{i:b_i=k}\left[\left(Y_i(t) - \bar{Y}_k(t)\right)^2 + \left(Y_i(c) - \bar{Y}_k(c)\right)^2 - 2\left(Y_i(t) - \bar{Y}_k(t)\right)\left(Y_i(c) - \bar{Y}_k(c)\right)\right].
\end{aligned}$$

So then

$$
\begin{aligned}
&\frac{S_k^2(c)}{1-p} + \frac{S_k^2(t)}{p} - S_k^2(tc) \\
&= \frac{\frac{1}{n_k-1}\sum_{i:b_i=k}\left(Y_i(c) - \bar{Y}_k(c)\right)^2}{1-p} + \frac{\frac{1}{n_k-1}\sum_{i:b_i=k}\left(Y_i(t) - \bar{Y}_k(t)\right)^2}{p} \\
&- \frac{1}{n_k-1}\sum_{i:b_i=k}\left[\left(Y_i(t) - \bar{Y}_k(t)\right)^2 + \left(Y_i(c) - \bar{Y}_k(c)\right)^2 - 2\left(Y_i(t) - \bar{Y}_k(t)\right)\left(Y_i(c) - \bar{Y}_k(c)\right)\right] \\
&= \sum_{i:b_i=k}\frac{1}{n_k-1}\left[\frac{p\left(Y_i(c) - \bar{Y}_k(c)\right)^2}{1-p} + \frac{(1-p)\left(Y_i(t) - \bar{Y}_k(t)\right)^2}{p} + 2\left(Y_i(t) - \bar{Y}_k(t)\right)\left(Y_i(c) - \bar{Y}_k(c)\right)\right] \\
&= \sum_{i:b_i=k}\frac{1}{n_k-1}\left(\sqrt{\frac{p}{1-p}}\left(Y_i(c) - \bar{Y}_k(c)\right) + \sqrt{\frac{1-p}{p}}\left(Y_i(t) - \bar{Y}_k(t)\right)\right)^2 \\
&= \sum_{i:b_i=k}\frac{1}{n_k-1}\left(\sqrt{\frac{p}{1-p}}Y_i(c) + \sqrt{\frac{1-p}{p}}Y_i(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right)\right)^2.
\end{aligned}
$$

So we get

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&\qquad = \sum_{k=1}^{K}\frac{1}{n(n-1)}\Bigg[n_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2 \\
&\qquad\quad - \frac{n-n_k}{n(n_k-1)}\sum_{i:b_i=k}\left(\sqrt{\frac{p}{1-p}}Y_i(c) + \sqrt{\frac{1-p}{p}}Y_i(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right)\right)^2\Bigg].
\end{aligned}
$$

This allows us to see the similarity in the two terms.

A.8.2 Simple random sampling

The simple random sampling framework has two steps: First, obtain a random sample of units for the experiment. Then, we form blocks. We assume there is a procedure for every sample that clusters on the observed covariate matrix $\boldsymbol{X}$ to form blocks for the given sample. The number and size of blocks can be sample dependent. Now think about the the "worst" case, that this set of covariates, $\boldsymbol{X}$, are actually independent of the potential outcomes. Then

we have, assuming equal proportions treated in each block,

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|SRS\right) - \text{var}\left(\widehat{\tau}_{(BK)}|SRS\right) \\
&= \mathbb{E}\left[\text{var}(\widehat{\tau}_{(CR)}|\mathcal{S})|SRS\right] + \text{var}\left(\mathbb{E}[\widehat{\tau}_{(CR)}|\mathcal{S}]|SRS\right) - \mathbb{E}\left[\text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\text{SRS}\right] - \text{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|SRS\right) \\
&= \mathbb{E}\left[\frac{S^2(c)}{n_c} + \frac{S^2(t)}{n_t} - \frac{S^2(tc)}{n}\Big|SRS\right] - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] \\
&= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{S_k^2(c)}{n_c} + \frac{S_k^2(t)}{n_t} - \frac{S_k^2(tc)}{n}\right)\Big|SRS\right] \\
&= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{S_k^2(c)}{n_c} + \frac{S_k^2(t)}{n_t} - \frac{S_k^2(tc)}{n}\right)\Big|SRS, \boldsymbol{X}\right]\Big|SRS\right] \\
&= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n}\right)\Big|SRS\right] \\
&= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \left(\frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n}\right) \\
&= 0.
\end{aligned}
$$

Note that conditioning on $\boldsymbol{X}$ fixes the number of units in each block and number of blocks, but potential outcomes are independent of this value of the covariates so we can push the expectation through.

A.8.3 Proof of Theorem 1.5.0.1: Variance comparison under stratified sampling

Proof. First use decomposition of variance and then simplify using results from the derivation in Appendix A.8.1. Let $\mu(z)$ and $\sigma^2(z)$ be the population mean and variance, respectively, of the potential outcomes of all units under treatment z. Let $\mu_k(z)$ and $\sigma_k^2(z)$ be the population mean and variance of the potential outcomes of units in stratum k under treatment z.

$$
\begin{aligned}
&\operatorname{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) - \operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \\
=&\mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(CR)}|\mathcal{S})|\mathcal{F}_1\right] + \operatorname{var}\left(\mathbb{E}[\widehat{\tau}_{(CR)}|\mathcal{S}]|\mathcal{F}_1\right) - \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_1\right] + \operatorname{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|\mathcal{F}_1\right) \\
=&\mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(CR)}|\mathcal{S})|\mathcal{F}_1\right] - \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_1\right] \\
=&\mathbb{E}\Big[\underbrace{\frac{\sum_{k=1}^K n_k\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)(n-1)n} + \frac{\sum_{k=1}^K n_k\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{p(n-1)n}}_{\mathbf{A}} - \underbrace{\frac{\sum_{k=1}^K n_k(\tau_{\mathcal{S}} - \tau_{k,\mathcal{S}})^2}{n(n-1)}}_{\mathbf{B}}|\mathcal{F}_1\Big] \\
&- \sum_{k=1}^K \frac{n-n_k}{n^2(n-1)}\left(\frac{\sigma_k^2(c)}{(1-p)} + \frac{\sigma_k^2(t)}{p} - \sigma_k^2(tc)\right)
\end{aligned} \tag{A.4}
$$

We start with expectation of the numerators in term **A**:

$$
\begin{aligned}
\mathbb{E}\left[\left(\bar{Y}_k(z) - \bar{Y}(z)\right)|\mathcal{F}_1\right]^2 &= \mathbb{E}\left[\bar{Y}_k(z)^2 - 2\bar{Y}_k(z)\bar{Y}(z) + \bar{Y}(z)^2|\mathcal{F}_1\right] \\
&= \underbrace{\mathbb{E}\left[\bar{Y}_k(z)^2|\mathcal{F}_1\right]}_{\mathbf{A.1}} - 2\mathbb{E}\left[\bar{Y}_k(z)\bar{Y}(z)|\mathcal{F}_1\right] + \underbrace{\mathbb{E}\left[\bar{Y}(z)^2|\mathcal{F}_1\right]}_{\mathbf{A.2}}.
\end{aligned}
$$

A.1:

$$
\mathbb{E}\left[\bar{Y}_k(z)^2|\mathcal{F}_1\right] = \operatorname{var}\left(\bar{Y}_k(z)|\mathcal{F}_1\right) + \mathbb{E}\left[\bar{Y}_k(z)|\mathcal{F}_1\right]^2 = \frac{\sigma_k^2(z)}{n_k} + \mu_k(z)^2
$$

A.2:

$$
\mathbb{E}\left[\bar{Y}(z)^2|\mathcal{F}_1\right] = \operatorname{var}\left(\bar{Y}(z)|\mathcal{F}_1\right) + \mathbb{E}\left[\bar{Y}(z)|\mathcal{F}_1\right]^2 = \sum_{k=1}^K \frac{n_k}{n^2}\sigma_k^2(z) + \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(z)\right)^2
$$

Putting **A.1** and **A.2** together and combining like terms:

$$
\begin{aligned}
\sum_{k=1}^k \frac{n_k}{n}\mathbb{E}\left[\left(\bar{Y}_k(z) - \bar{Y}(z)\right)^2|\mathcal{F}_1\right] &= \sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\left[\bar{Y}_k(z)^2|\mathcal{F}_1\right] - \mathbb{E}\left[\bar{Y}(z)^2|\mathcal{F}_1\right] \\
&= \sum_{k=1}^K \frac{n-n_k}{n^2}\sigma_k^2(z) + \sum_{k=1}^K \frac{n_k}{n}\mu_k(z)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(z)\right)^2.
\end{aligned}
$$

The expectation of **B** follows in a very similar manner, except now we have treated and control units in the calculation.

B becomes

$$\sum_{k=1}^{K} \frac{n_k}{n} \mathbb{E}\left[(\tau_{\mathcal{S}} - \tau_{k,\mathcal{S}})^2 \,|\mathcal{F}_1 \right]$$
$$= \sum_{k=1}^{K} \frac{n - n_k}{n^2} \sigma_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n} \left(\mu_k(t)^2 - \mu_k(c) \right)^2 - \left[\left(\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(t) \right) - \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(c) \right) \right]^2.$$

Putting **A** and **B** into Equation A.4 and simplifying:

$$\begin{aligned}
&\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \\
&= \frac{\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(c)^2 - \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(c) \right)^2}{(1-p)(n-1)} + \frac{\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(t)^2 - \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(t) \right)^2}{p(n-1)} \\
&- \frac{\sum_{k=1}^{K} \frac{n_k}{n} \left(\mu_k(t) - \mu_k(c) \right)^2 - \left(\sum_{k=1}^{K} \frac{n_k}{n} \left(\mu_k(t) - \mu_k(c) \right) \right)^2}{n-1} \\
&= \frac{p}{1-p} \frac{\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(c)^2 - \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(c) \right)^2}{n-1} + \frac{1-p}{p} \frac{\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(t)^2 - \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(t) \right)^2}{n-1} \\
&+ 2 \frac{\sum_{k=1}^{K} \frac{n_k}{n} \mu_k(t) \mu_k(c) - \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu(t,k) \right) \left(\sum_{k=1}^{K} \frac{n_k}{n} \mu(c,k) \right)}{n-1} \\
&= \frac{1}{n-1} \left[\frac{p}{1-p} \text{Var}_k \left(\mu_k(c) \right) + \frac{1-p}{p} \text{Var}_k \left(\mu_k(t) \right) + 2 Cov_k \left(\mu_k(c), \mu_k(t) \right) \right] \\
&= \frac{1}{n-1} \text{Var}_k \left(\sqrt{\frac{p}{1-p}} \mu_k(t) + \sqrt{\frac{1-p}{p}} \mu_k(c) \right) \geq 0.
\end{aligned}$$

The variance in the last line is the variance over the blocks, as defined in Equation 1.12. Therefore we have that $\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \geq 0$ so we are always doing better with blocking in this setting. □

A.8.4 Stratified sampling vs SRS comparisons

Corollary A.8.4.1. *The difference between* $var(\widehat{\tau}_{(CR)}|SRS) - var\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right)$ *may be positive or negative.*

Proof. Compare the previous result to when we assume SRS for complete randomization

and $\mathcal{F}_1$ for blocked randomization.

$$\begin{aligned}
&\text{var}(\widehat{\tau}_{(CR)}|SRS) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \\
&\quad = \frac{1}{n_c}\left(\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(c)-\mu(c)\right)^2\right) + \frac{1}{n_t}\left(\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(t)-\mu(t)\right)^2\right) \\
&\quad \geq 0
\end{aligned}$$

A form of this result can be found in Imbens (2011).

We can then take the difference, $\text{var}(\widehat{\tau}_{(CR)}|SRS) - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right)$, between the results under the two frameworks to see whether using different sampling frameworks over or under estimates the benefits of blocking. First consider the form of the variance under the two different models.

$$\begin{aligned}
\text{var}(\widehat{\tau}_{(CR)}|SRS) &= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} \\
&= \frac{\sum_{k=1}^{K}\frac{n_k}{n}\sigma_k^2(c) + \sum\frac{n_k}{n}\left(\mu_k(c)-\mu(c)\right)^2}{n_c} + \frac{\sum_{k=1}^{K}\frac{n_k}{n}\sigma_k^2(t) + \sum\frac{n_k}{n}\left(\mu_k(t)-\mu(t)\right)^2}{n_t}
\end{aligned}$$

$$\begin{aligned}
\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) &= \mathbb{E}\left[\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right)|\mathcal{F}_1\right] + \text{var}\left(\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{S}\right]|\mathcal{F}_1\right) \\
&= \underbrace{\mathbb{E}\left[\frac{S^2(c)}{n_c} + \frac{S^2(t)}{n_t} - \frac{S^2(tc)}{n}|\mathcal{F}_1\right]}_{\mathbf{A}} + \underbrace{\text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_1\right)}_{\mathbf{B}}
\end{aligned}$$

For **A** we can use similar techniques from previous proofs to break the pieces up by block:

$$\begin{aligned}
\mathbb{E}\left[\frac{S^2(tc)}{n}|\mathcal{F}_1\right] &= \sum_{k=1}^{K}\left[\frac{n_k}{n^2}\sigma_k^2(tc) + \frac{n_k}{n(n-1)}\left(\tau_k-\tau\right)^2\right] \\
\mathbb{E}\left[\frac{S^2(z)}{n}|\mathcal{F}_1\right] &= \frac{1}{n_z}\sum_{k=1}^{K}\left[\frac{n_k}{n}\sigma_k^2(z) + \frac{n_k}{n-1}\left(\mu_k(z)-\mu(z)\right)^2\right]
\end{aligned}$$

For **B** we can split up by block and then use classical results from sampling theory on variation of sample means under SRS (Lohr, 2009, Chapter 2).

$$\text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_1\right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \frac{\sigma_k^2(tc)}{n_k}$$

Putting **A** and **B** together and collecting similar terms, we have

$$\begin{aligned}\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) = &\sum_{k=1}^{K} \left[\frac{n_k}{n}\frac{\sigma_k^2(c)}{n_c} + \frac{n_k}{n}\frac{\sigma_k^2(t)}{n_t}\right] + \sum_{k=1}^{K}\left[\frac{n_k}{(n-1)n_c}\left(\mu_k(c)-\mu(c)\right)^2 + \frac{n_k}{(n-1)n_t}\left(\mu_k(t)-\mu(t)\right)^2\right] \\ &- \sum_{k=1}^{K}\frac{n_k}{n(n-1)}\left(\tau_k-\tau\right)^2.\end{aligned}$$

So then we can do the subtraction:

$$\begin{aligned}&\text{var}(\widehat{\tau}_{(CR)}|SRS) - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) \\ &= \frac{1}{n(n-1)}\sum_{k=1}^{K} n_k\left[\left(\tau_k-\tau\right)^2 - \frac{\left(\mu_k(c)-\mu(c)\right)^2}{n_c} - \frac{\left(\mu_k(t)-\mu(t)\right)^2}{n_t}\right] \\ &= \sum_{k=1}^{K}\frac{n_k}{n(n-1)}\Big[\frac{\left(n_c-1\right)\left(\mu_k(c)-\mu(c)\right)^2}{n_c} + \frac{\left(n_t-1\right)\left(\mu_k(t)-\mu(t)\right)^2}{n_t} \\ &\qquad - 2\left(\mu_k(c)-\mu(c)\right)\left(\mu_k(t)-\mu(t)\right)\Big].\end{aligned}$$

We see that this difference could be positive or negative. In particular, if the blocks all have the same average treatment effect but different control and treatment means, then this expression will be negative, indicating that comparing the variance of blocking under $\mathcal{F}_1$ to complete randomization under SRS is an underestimate of the benefits of blocking. On the other hand, if n_c and n_t are large compared to the variation of block average control and treatment outcomes, then we would expect the negative terms in the first expression to be small, resulting in the difference being positive. This would mean that there is an overestimate of the benefits of blocking when comparing the variance of blocking under $\mathcal{F}_1$ to complete randomization under SRS.

If we have n_c and n_t large so $\frac{n_c-1}{n_c} \approx 1 \approx \frac{n_t-1}{n_t}$ then we have

$$\text{var}\left(\widehat{\tau}_{(CR)}|SRS\right) - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) = \frac{1}{n(n-1)}\sum_{k=1}^{K} n_k\left(\tau_k-\tau\right)^2.$$

□

A.8.5 Variance comparison under sampling of infinite size blocks

We onsider an infinite number of blocks of infinite size. Then the difference between the variance under the completely randomized design and the blocked design under this framework is

$$\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_3\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_3\right) \\
&\quad = \mathbb{E}\left[\frac{1}{n-1}\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right)|\mathcal{F}_3\right] \\
&\quad \geq 0.
\end{aligned}$$

Hence, making the population from which units in the blocks are sampled infinite guarantees that blocking is always beneficial or at least not harmful in terms of variance compared to complete randomization.

A.8.6 Unequal treatment proportions

Our comparison of complete randomization to blocking in Chapter 1 only applies to the small slice of possible experiments in which the treatment proportion is equal across all blocks. In practice, however, the proportion treated, p_k, may be unequal across blocks, and in this case the above results are not guaranteed to hold. In particular, with blocks of variable size, it can be difficult to have the same proportion treated within each block due to the discrete nature of units.

With varying p_k, the units within each block are weighted differently than they would be in a complete randomization when calculating a treatment effect estimate. That is, in a complete randomization, the treated units are all weighted proportional to $1/p$ but here the treated units in each block get weighted instead by $1/p_k$, meaning units with low probability of treatment will "count more" towards the overall treatment mean and their variability will have greater relevance for the overall variance of the estimator. Sävje (2015) also noted the effect of variable proportions treated on variance and Higgins et al. (2015) explored estimators for blocked designs with possibly unequal treatment proportions, but also multiple treatments. The costs here are similar to the costs of variable selection probabilities in survey sampling (see Särndal et al., 2003).

When different blocks have different proportions of units treated, it is possible to systematically have blocks and treatment groups with more variance to also have more weight, which could cause blocking to be harmful even in the stratified sampling setting.

Theorem A.8.6.1 (Variance comparison with unequal treatment proportions)**.**

$$
\begin{aligned}
&var\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - var\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right) \\
&= \frac{1}{n-1} Var_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right) + \sum_{k=1}^{K} \frac{(p-p_k)n_k}{n^2}\left[\frac{\sigma_k^2(c)}{(1-p_k)(1-p)} - \frac{\sigma_k^2(t)}{p_k p}\right].
\end{aligned}
$$

Proof. Let the proportion of units treated in a complete randomization be p and in block randomization be p_k for block k. Again, we need to break the complete randomization variance into block components. For the finite sample, the complete randomization variance, from Appendix A.8.1 is

$$
\begin{aligned}
\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) = &\frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(c) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n} \\
&+ \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(t) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn} \\
&- \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2}{n}.
\end{aligned}
$$

For block randomization, the variance is

$$
\begin{aligned}
\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) &= \sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right) \\
&= \sum_{k=1}^{K} \frac{n_k}{n^2}\left(\frac{S_k^2(c)}{1-p_k} + \frac{S_k^2(t)}{p_k} - S_k^2(tc)\right).
\end{aligned}
$$

Then the difference is

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&= \sum_{k=1}^{K}\left[\frac{(1-p_k)(n_k-1)n-(1-p)n_k(n-1)}{(1-p_k)(1-p)n^2(n-1)}S_k^2(c) + \frac{p_k(n_k-1)n-pn_k(n-1)}{p_kpn^2(n-1)}S_k^2(t)\right] \\
&+ \sum_{k=1}^{K}\left[\frac{n-n_k}{(n-1)n^2}S_k^2(tc)\right] + \frac{\sum_{k=1}^{K} n_k\left[\frac{(\bar{Y}_k(c)-\bar{Y}(c))^2}{1-p} + \frac{(\bar{Y}_k(t)-\bar{Y}(t))^2}{p} - (\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}})^2\right]}{(n-1)n} \\
&= \frac{1}{n^2(n-1)}\sum_{k=1}^{K}\left(\frac{(1-p_k)(n_k-1)n-(1-p)n_k(n-1)}{(1-p_k)(1-p)}S_k^2(c) + \frac{p_k(n_k-1)n-pn_k(n-1)}{p_kp}S_k^2(t)\right. \\
&\qquad \left. + (n-n_k)S_k^2(tc)\right) + \frac{\sum_{k=1}^{K} n_k\left[\frac{(\bar{Y}_k(c)-\bar{Y}(c))^2}{1-p} + \frac{(\bar{Y}_k(t)-\bar{Y}(t))^2}{p} - (\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}})^2\right]}{(n-1)n}.
\end{aligned}
$$

For the stratified random sampling ($\mathcal{F}_1$), we use results from the proof in Appendix A.8.3 and the above derivation, and combine like terms to get

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right) \\
&= \frac{1}{n-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\sqrt{\frac{1-p}{p}}\mu_k(c) + \sqrt{\frac{p}{1-p}}\mu_k(t) - \left[\sqrt{\frac{1-p}{p}}\mu(c) + \sqrt{\frac{p}{1-p}}\mu(t)\right]\right)^2 \\
&+ \sum_{k=1}^{K}\left[\frac{(p-p_k)n_k}{(1-p_k)(1-p)n^2}\sigma_k^2(c) + \frac{(p_k-p)n_k}{p_kpn^2}\sigma_k^2(t)\right].
\end{aligned}
$$

More details on the derivation can be given upon request.

This can also be written as

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right) \\
&= \frac{1}{n-1}\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right) + \sum_{k=1}^{K}\frac{(p-p_k)n_k}{n^2}\left[\frac{\sigma_k^2(c)}{(1-p_k)(1-p)} - \frac{\sigma_k^2(t)}{p_kp}\right].
\end{aligned}
\tag{A.5}
$$

The first term in Equation A.5 is equal to the result in Proposition 1.5.0.1 and the second term, capturing the additional variability due to varying proportions, is zero when $p_k = p$ for all k. In general, the second term of Equation A.5 can be positive or negative.

In particular, if by some bad luck, $pp_k\sigma_k^2(c) > (1-p)(1-p_k)\sigma_k^2(t)$ for all blocks where $p_k > p$ and $pp_k\sigma_k^2(c) < (1-p)(1-p_k)\sigma_k^2(t)$ for all blocks where $p_k < p$, this term will be negative. If, in this case, the population mean potential outcomes for all blocks are approximately equal, the entire expression will be negative. The two terms are of the same order with respect to sample size n, making comparison easier. If $p_k \approx p$ then the second term should not be too large. In small blocks, where one unit more in treatment or control can dramatically change the proportion of treated units, we would expect unequal proportions to have more of an impact. □

A.9 Proofs of consequences of ignoring blocking

A.9.1 Proof of Theorem A.2.0.1

Proof. We perform a block randomization but then use the variance estimator from a complete randomization, $\frac{s^2(c)}{n_c} + \frac{s^2(t)}{n_t}$. We will condition on $\mathbf{P}_{blk}$, the assignment mechanism being blocked randomization, throughout to make this clear. For the finite sample framework, the true variance would still be

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}, \mathbf{P}_{blk}\right) = \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right).$$

Again, we assume that $p_k = p$ for all $k = 1, \ldots, K$. Then

$$\bar{Y}^{obs}(z) = \sum_{k=1}^{K} \frac{n_{z,k}}{n_z}\bar{Y}_k^{obs}(z) = \sum_{k=1}^{K} \frac{n_k}{n}\bar{Y}_k^{obs}(z).$$

We have

$$
\begin{aligned}
s^2(z) =& \frac{1}{n_z - 1} \sum_{i:Z_i=z} \left(Y_i(z) - \bar{Y}^{obs}(z) \right)^2 \\
=& \frac{1}{n_z - 1} \sum_{k=1}^{K} \sum_{i:Z_i=z, b_i=k} \left(Y_i(z) - \bar{Y}_k(z) + \bar{Y}_k(z) - \bar{Y}^{obs}(z) \right)^2 \\
=& \frac{1}{n_z - 1} \sum_{k=1}^{K} \sum_{i:Z_i=z, b_i=k} \Big[\left(Y_i(z) - \bar{Y}_k(z) \right)^2 + 2 \left(Y_i(z) - \bar{Y}_k(z) \right) \left(\bar{Y}_k(z) - \bar{Y}^{obs}(z) \right) \\
& + \left(\bar{Y}_k(z) - \bar{Y}^{obs}(z) \right)^2 \Big] \\
=& \frac{1}{n_z - 1} \Big[\underbrace{\sum_{k=1}^{K} \sum_{i:Z_i=z, b_i=k} \left(Y_i(z) - \bar{Y}_k(z) \right)^2}_{\mathbf{A}} + \underbrace{2 \sum_{k=1}^{K} n_{z,k} \left(\bar{Y}_k^{obs}(z) - \bar{Y}_k(z) \right) \left(\bar{Y}_k(z) - \bar{Y}^{obs}(z) \right)}_{\mathbf{B}} \\
& + \underbrace{\sum_{k=1}^{K} n_{z,k} \left(\bar{Y}_k(z) - \bar{Y}^{obs}(z) \right)^2}_{\mathbf{C}} \Big].
\end{aligned}
$$

Now expand and take the expectation of each term separately. We start with **A** and note that $\mathbb{E}\left[\mathbb{I}_{Z_i=z} | \mathbf{P}_{blk}\right]$ is the same for all units because the proportion treated is assumed to be the same in all blocks.

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K} \sum_{i:Z_i=z, b_i=k} \left(Y_i(z) - \bar{Y}_k(z) \right)^2 | \mathcal{S}, \mathbf{P}_{blk} \right] &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{i:b_i=k} \mathbb{I}_{Z_i=z} \left(Y_i(z) - \bar{Y}_k(z) \right)^2 | \mathcal{S}, \mathbf{P}_{blk} \right] \\
&= \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{I}_{Z_i=z} | \mathbf{P}_{blk} \right] (n_k - 1) S_k^2(z)
\end{aligned}
$$

Now **B**,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k^{obs}(z)-\bar{Y}_k(z)\right)\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k^{obs}(z)\bar{Y}_k(z)-\bar{Y}_k(z)^2+\bar{Y}_k(z)\bar{Y}^{obs}(z)-\bar{Y}_k^{obs}(z)\bar{Y}^{obs}(z)\right)|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k(z)\bar{Y}^{obs}(z)-\bar{Y}_k^{obs}(z)\bar{Y}^{obs}(z)\right)|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=n_z\left(\bar{Y}(z)^2-\mathbb{E}\left[\bar{Y}^{obs}(z)^2|\mathcal{S},\mathbf{P}_{blk}\right]\right)\\
&=n_z\left(-\text{var}\left(\bar{Y}^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right)\right)\\
&=-n_z\sum_{k=1}^{K}\frac{n_k^2}{n^2}\text{var}\left(\bar{Y}_k^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right)\\
&=-n_z\sum_{k=1}^{K}\frac{n_{z,k}^2}{n_z^2}\frac{n_k-n_{z,k}}{n_k}\frac{S_k^2(z)}{n_{z,k}}\\
&=-\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk},b_i=k\right]\right)S_k^2(z).
\end{aligned}
$$

Now $\mathbf{C}$,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)^2|\mathcal{S}\right]\\
&=\sum_{k=1}^{K} n_{z,k}\bar{Y}_k(z)^2-2n_z\bar{Y}(z)^2+n_z\mathbb{E}\left[\bar{Y}^{obs}(z)^2|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\sum_{k=1}^{K} n_{z,k}\bar{Y}_k(z)^2-2n_z\bar{Y}(z)^2+n_z\text{var}\left(\bar{Y}^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right)+n_z\mathbb{E}\left[\bar{Y}^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right]^2\\
&=\sum_{k=1}^{K} n_{z,k}\bar{Y}_k(z)^2-n_z\bar{Y}(z)^2+\sum_{k=1}^{K}\frac{n_{z,k}(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right])}{n_z}S_k^2(z).
\end{aligned}
$$

Putting it all back together,

$$
\begin{aligned}
&\mathbb{E}\left[s^2(z)|\mathcal{S},\mathbf{P}_{blk}\right] \\
=&\frac{1}{n_z-1}\Big[\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{Z_i=z}\right](n_k-1)S_k^2(z)-2\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z|\mathbf{P}_{blk}}\right]\right)S_k^2(z)+\sum_{k=1}^{K}n_{z,k}\bar{Y}_k(z)^2-n_z\bar{Y}(z)^2 \\
&+\sum_{k=1}^{K}\frac{n_{z,k}(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right])}{n_z}S_k^2(z)\Big] \\
=&\frac{1}{n_z-1}\Big[\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n_k-1)S_k^2(z)-\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right]\right)S_k^2(z) \\
&+\sum_{k=1}^{K}n_{z,k}\bar{Y}_k(z)^2-n_z\bar{Y}(z)^2\Big] \\
=&\sum_{k=1}^{K}\left(\frac{n_k}{n}-\frac{\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n-n_k)}{n(n_z-1)}\right)S_k^2(z)+\frac{1}{n_z-1}\sum_{k=1}^{K}n_{z,k}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2.
\end{aligned}
$$

So then the bias is

$$
\begin{aligned}
&\mathbb{E}\left[\frac{s^2(c)}{n_c}+\frac{s^2(t)}{n_t}|\mathcal{S},\mathbf{P}_{blk}\right]-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}-\frac{S_k^2(tc)}{n_k}\right) \\
&=\sum_{k=1}^{K}\left(\frac{n_k}{(1-p)n^2}-\frac{(1-p)(n-n_k)}{(1-p)n^2(n_c-1)}\right)S_k^2(c)+\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_{c,k}}{n_c}\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2 \\
&+\sum_{k=1}^{K}\left(\frac{n_k}{pn^2}-\frac{p(n-n_k)}{pn^2(n_t-1)}\right)S_k^2(t)+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_{t,k}}{n_t}\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2 \\
&-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k}+\frac{S_k^2(t)}{pn_k}-\frac{S_k^2(tc)}{n_k}\right) \\
&=\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2 \\
&-\left(\sum_{k=1}^{K}\frac{n-n_k}{n^2(n_c-1)}S_k^2(c)+\sum_{k=1}^{K}\frac{n-n_k}{n^2(n_t-1)}S_k^2(t)-\sum_{k=1}^{K}\frac{n_k}{n^2}S_k^2(tc)\right)
\end{aligned}
$$

If there is no variation between the blocks (i.e. $\bar{Y}_k(z)=\bar{Y}(z)$ for all k) but there is within block variability (i.e. $S_k^2(z)\neq 0$) then this difference will be negative. The reduction is blunted, however, by the degree of treatment variation there is within block (making $S_k^2(tc)$ offset the negative term from the $S_k^2(z)$). □

A.9.2 Proof of Corollary A.2.0.1

Proof. We start from the result of Appendix A.9.1 and utilize work done in Appendix A.8.3 to simplify things.

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\frac{s^2(z)}{n_z}|\mathcal{S},\mathbf{P}_{blk}\right]|\mathcal{F}_1\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(\frac{n_k}{nn_z}-\frac{\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n-n_k)}{n_z n(n_z-1)}\right)S_k^2(z)+\frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2|\mathcal{F}_1\right] \\
=&\sum_{k=1}^{K}\left(\frac{n_k}{nn_z}-\frac{\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n-n_k)}{n_z n(n_z-1)}\right)\sigma_k^2(z)+\mathbb{E}\left[\frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2|\mathcal{F}_1\right] \\
=&\sum_{k=1}^{K}\left(\frac{n_k}{nn_z}-\frac{n-n_k}{n^2(n_z-1)}\right)\sigma_k^2(z)+\frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n-n_k}{n^2}\sigma_k^2(z)+\frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(z)-\mu(z)\right)^2 \\
=&\sum_{k=1}^{K}\frac{n_k}{nn_z}\sigma_k^2(z)+\frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(z)-\mu(z)\right)^2
\end{aligned}
$$

Now to get the bias we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{s^2(c)}{n_c}+\frac{s^2(t)}{n_t}|\mathbf{P}_{blk},\mathcal{F}_1\right]-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(c)}{n_{c,k}}+\frac{\sigma_k^2(t)}{n_{t,k}}\right) \\
=&\sum_{k=1}^{K}\frac{n_k}{nn_c}\sigma_k^2(c)+\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(c)-\mu(c)\right)^2+\sum_{k=1}^{K}\frac{n_k}{nn_t}\sigma_k^2(t)+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(t)-\mu(t)\right)^2 \\
&-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(c)}{n_{c,k}}+\frac{\sigma_k^2(t)}{n_{t,k}}\right) \\
=&\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(c)-\mu(c)\right)^2+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(t)-\mu(t)\right)^2.
\end{aligned}
$$

□

Appendix B

Appendix to Chapter 2

B.1 Proofs

of Theorem 2.3.1.1. Let $\mathcal{W}$ be the set of values that $w(\boldsymbol{Z})$ can take for $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0}$. Let $H(Z) = \eta_{w(\boldsymbol{Z})}$. Then

$$\begin{aligned}
P_{\eta_0}\left\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))\right\} &= \sum_{z \in \boldsymbol{\mathcal{Z}}_{\eta_0}} P_{\eta_0}(\boldsymbol{Z} = z)\mathbb{1}\{\tau \in C(\hat{\tau}(z); H(z))\} \\
&= \sum_{z \in \boldsymbol{\mathcal{Z}}_{\eta_0}} \sum_{x \in \mathcal{W}} P_{\eta_0}(\boldsymbol{Z} = z | w(\boldsymbol{Z}) = x) P_{\eta_0}(w(\boldsymbol{Z}) = x)\mathbb{1}\{\tau \in C(\hat{\tau}(z); H(z))\} \\
&= \sum_{x \in \mathcal{W}} P_{\eta_0}(w(\boldsymbol{Z}) = x) \sum_{z: w(z) = x} P_{\eta_0}(\boldsymbol{Z} = z | w(\boldsymbol{Z}) = x)\mathbb{1}\{\tau \in C(\hat{\tau}(z); \eta_x)\} \\
&= \sum_{x \in \mathcal{W}} P_{\eta_0}(w(\boldsymbol{Z}) = x) \sum_{z: w(z) = x} P_{\eta_x}(\boldsymbol{Z} = z)\mathbb{1}\{\tau \in C(\hat{\tau}(z); \eta_x)\} \\
&\geq \sum_{x \in \mathcal{W}} P_{\eta_0}(w(\boldsymbol{Z}) = x)\gamma \\
&\geq \gamma
\end{aligned}$$

which concludes the proof. The second to last step comes from Proposition 2.2.3.1:

$$\sum_{z: w(z) = x} P_{\eta_x}(\boldsymbol{Z} = z)\mathbb{1}\{\tau \in C(\hat{\tau}(z); \eta_x)\} = P_{\eta_x}\left(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_x)\right) = \gamma.$$

□

of Corollary 2.3.1.1. We have $\mathcal{P} = \{\boldsymbol{\mathcal{Z}}_{(1)}, \ldots, \boldsymbol{\mathcal{Z}}_{(K)}\}$ a partition of the set of all possible

assignments, $\boldsymbol{\mathcal{Z}}_{\eta_0}$ and

$$w(\boldsymbol{Z}) = \sum_{k}^{K} k \mathbb{1}\{\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{(k)}\}.$$

Then

$$P_{\eta_0}(w(\boldsymbol{Z}) = k) = \eta_0(\boldsymbol{\mathcal{Z}}_{(k)}).$$

We have

$$P_{\eta_0}(\boldsymbol{Z} \mid w(\boldsymbol{Z}) = k) = \begin{cases} \frac{\eta_0(\boldsymbol{Z})}{\eta_0(\boldsymbol{\mathcal{Z}}_{(k)})} & \text{if } \boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{(k)} \\ 0 & \text{otherwise} \end{cases}$$

$$= \eta_k(\boldsymbol{Z}).$$

Hence $H : \boldsymbol{Z} \to \eta_{w(\boldsymbol{Z})}$ leads to an η_0-valid procedure, as a consequence of Theorem 2.3.1.1.

□

of Probability 1 for Rerandomized As-if Design. We have that X_i is continuous for each i. For example, X_i could be normally distributed. So for any random finite population and any fixed vector a with ith entry a_i,

$$P(\sum_{i=1}^{n} a_i X_i = 0) = 0.$$

The difference in covariate means (balance) is

$$\Delta_X(\boldsymbol{Z}) = \frac{1}{N_1} \sum_{i=1}^{n} Z_i X_i - \frac{1}{N - N_1} \sum_{i=1}^{n} (1 - Z_i) X_i = \sum_{i=1}^{n} \left(\frac{Z_i}{N_1} - \frac{1 - Z_i}{N - N_1} \right) X_i.$$

The difference in the imbalance between assignment $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ is

$$\Delta_X(\boldsymbol{Z}) - \Delta_X(\boldsymbol{Z}') = \sum_{i=1}^{n} \left(\frac{Z_i - Z_i'}{N_1} - \frac{Z_i' - Z_i}{N - N_1} \right) X_i = \left(\frac{1}{N_1} - \frac{1}{N - N_1} \right) \sum_{i=1}^{n} (Z_i - Z_i') X_i.$$

Let $a_i = \left(\frac{1}{N_1} - \frac{1}{N - N_1} \right) (Z_i - Z_i')$. Then we have

$$P(\Delta_X(\boldsymbol{Z}) - \Delta_X(\boldsymbol{Z}') = 0) = 0.$$

Hence,

$$P(\Delta_X(\boldsymbol{Z}) - \Delta_X(\boldsymbol{Z}') = 0 \quad \forall \boldsymbol{Z}' \neq \boldsymbol{Z}) \leq \sum_{\boldsymbol{Z}' \neq \boldsymbol{Z}} P(\Delta_X(\boldsymbol{Z}) - \Delta_X(\boldsymbol{Z}') = 0) = 0.$$

This implies that for any random finite population, the probability of any given assignment having the same covariate balance measure as another assignment is zero. Or, in other words, with probability one each covariate balance is unique. □

of Special Case of 0 Coverage for Rerandomized As-if Design. We now consider the special case of Example 3, with strict inequality for the rerandomization. That is, we have an original design of complete randomization and design map $H : \boldsymbol{Z} \to P(\boldsymbol{Z}'|\boldsymbol{Z}' \in \mathcal{A}(\boldsymbol{Z}))$ where $\mathcal{A}(\boldsymbol{Z}) = \{\boldsymbol{Z}' : |\Delta_X(\boldsymbol{Z}')| < |\Delta_X(\boldsymbol{Z})|\}$ is the set of assignments with covariate balance strictly better than the observed covariate balance. Consider the special case where $Y_i(0) = Y_i(1) = X_i$ $(i = 1, ..., N)$, so $\tau = 0$. In that case, we have $\hat{\tau}(\boldsymbol{Z}) = \Delta_X(\boldsymbol{Z})$, and so for $\boldsymbol{Z} \sim \eta_0$,

$$\forall \boldsymbol{Z}' \in \mathcal{A}(\boldsymbol{Z}), \quad |\hat{\tau}(\boldsymbol{Z})| > |\hat{\tau}(\boldsymbol{Z}')|.$$

In particular, $\forall \boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}_{\eta_0}$, $\hat{\tau}(\boldsymbol{Z}) \notin [L_\alpha(H), U_\alpha(H)]$, and thus

$$\begin{aligned}
P_{\eta_0}(\tau \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))) &= P_{\eta_0}(0 \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))) \\
&= P_{\eta_0}(\hat{\tau} - U_\alpha(H) \leq 0 \leq \hat{\tau} - L_\alpha(H)) \\
&= P_{\eta_0}(U_\alpha(H) \geq \hat{\tau} \geq L_\alpha(H)) \\
&= P_{\eta_0}(\hat{\tau} \in [L_\alpha(H), U_\alpha(H)]) \\
&= 0.
\end{aligned}$$

Hence, analyzing the experiment as if it came from a stricter rerandomized design, whose non-inclusive maximal imbalance is the that of the observed assignment, leads to a coverage of 0. □

of Theorem 2.4.0.1. We have

$$\begin{aligned} P_{\eta_0,m}\left\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))\right\} &= \int p(w)\left(\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}_{\eta_0}} P_{\eta_0}(\boldsymbol{Z}=\boldsymbol{z} \mid w)\mathbb{1}\{\tau \in C(\hat{\tau}(\boldsymbol{z}); H(\boldsymbol{z}))\}\right) dw \\ &= \int p(w)\left(\sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}_{\eta_0}} \eta_w(\boldsymbol{z})\mathbb{1}\{\tau \in C(\hat{\tau}(\boldsymbol{z}); \eta_w)\}\right) dw \\ &= \int p(w) P_{\eta_w}\left(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_w)\right) dw. \end{aligned}$$

Again, the key is to notice that by Proposition 2.2.3.1, $P_{\eta_w}\left(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_w)\right) \geq \gamma$. Hence,

$$\begin{aligned} P_{\eta_0,m}\left\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); H(\boldsymbol{Z}))\right\} &= \int p(w) P_{\eta_w}\left(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_w)\right) dw \\ &\geq \int p(w)(\gamma)\, dw \\ &= \gamma, \end{aligned}$$

which concludes the proof. □

B.2 Estimators, Designs, and Practical Considerations

Our use of oracle procedures has allowed us to highlight the key mechanisms of conditioning, but it has obscured a number of practical issues with constructing as-if analyses in real-world applications, which we explore here.

First, an important note is that the estimators we should use may change based on the design we use. For instance, in post-stratification we would use the blocking estimator, which averages block treatment effect estimates, rather than the completely randomized estimator. These two estimators are identical when the proportion treated within each block is the same but otherwise the completely randomized estimator will be biased for the post-stratified design. Although this is not an issue when we have an oracle, because it will still correctly identify the distribution of $\hat{\tau}(\boldsymbol{Z}) - \tau$ for any estimator $\hat{\tau}(\boldsymbol{Z})$, in practice we would want to use unbiased estimators. Our validity result still holds even if we change the estimator for each different blocked design. This validity holds because we will have

conditional validity for each blocked design and therefore will be valid over the entire completely randomized design. See proof of Theorem 2.3.1.1, which shows that we just need conditional validity to get overall validity. But in practice, a full comparison of a conditional and non-conditional analysis may require a comparison of relevant estimators as well.

An obvious challenge in the absence of an oracle is that we may not have good analytical control over the randomization distribution of $\hat{\tau} - \tau$ in the sense of being able to derive and estimate quantities such as variance. For instance, we may have good control over the distribution of $\hat{\tau} - \tau$ under η_0 but not under conditional design $\eta_{w(\mathbf{Z})}$, so that even though Theorem 2.3.1.1 guarantees the theoretical validity of this conditional 'as-if' analysis, there is no way to implement it in practice. Therefore, we will often want to choose an estimator that has good properties, such as unbiasedness or low variance, under the conditional distribution $\eta_{w(\mathbf{Z})}$, as we discuss in supplementary material B.2.

A more subtle issue is that in practice to construct confidence intervals we will typically need to estimate variance and we often only have conservative estimators of the variance and estimators of the variance will often be even more conservative for conditional analyses. There may also be degrees of freedom losses under a conditional analysis, which may occur in post-stratification for example. Thus, we may see situations in which a conditional interval is smaller than the marginal interval under the oracle, but the estimated conditional interval is larger than the marginal conditional interval. Therefore, in practice the theoretical benefits of conditioning must be weighed against the drawbacks of excessively conservative variance estimation.

Other questions of practical importance when discussing conditioning are what and how much to condition on. The theory indicates that more is always better; the more variables we condition on, the more relevant our analysis become. But, as we condition on more information, the partition induced on the set of all assignments becomes increasingly fine. In the extreme case, each element of the partition may contain a single assignment. This poses a philosophical problem from the point of view of randomization-based inference because there is no randomness left! This is not a problem under an oracle, which would give us a single point at the true τ in this case, but this precludes any analysis in practice.

If we cannot condition on everything, we must choose what to condition on, given

multiple options. This is analogous to the well known fact in the frequentist literature that many ancillary statistics may exist in a given inference problem, and that two ancillary statistics may cease to be ancillary if taken jointly (Basu, 1964; Ghosh et al., 2010). Similarly, one must choose how fine to make the conditioning. For instance, in our stochastic design map example of Section 2.4, one must choose the bandwidth c of covariate closeness to use. The simple but vague answer is that one should condition on the quantity, or few quantities, that affect relevance the most, while remaining mindful of the analytical issues described in the previous paragraph. See Liu and Meng (2016) for a more detailed discussion on the trade-offs of relevance and robustness.

B.3 A simple example of relevance

B.3.1 Relevance and betting

First we review how the concepts of validity and relevance as formalized by Buehler (1959) and Robinson (1979) is made visual by the concept of betting. We follow broadly the setup of Buehler (1959), with some modifications to accommodate our notation. Consider a betting game between two players:

1. Player 1 chooses a distribution η^*, and a confidence level β. Intuitively, this corresponds to the claim $P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)) = \beta$.

2. Player 2 chooses a way to partition $\boldsymbol{\mathcal{Z}}$ into sets A^+ and A^- to use betting strategy $S(A^+, A^-)$. Under this strategy, if $\boldsymbol{Z}$ is in A^+ then Player 2 bets that τ is covered by the confidence interval based on Player 1's distribution. If $\boldsymbol{Z}$ is in A^- then Player 2 bets that τ is not covered by the confidence interval. Formally this strategy is defined as follows:

   - If $\boldsymbol{Z} \in A^+$, bet that $\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)$, risking β to win $1 - \beta$.
   - If $\boldsymbol{Z} \in A^-$, bet that $\tau \notin C(\hat{\tau}(\boldsymbol{Z}); \eta^*)$, risking $1 - \beta$ to win β.

The return of this game, for Player 2, is

$$R = \begin{cases} \mathbb{1}\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)\}(1-\beta) - \mathbb{1}\{\tau \notin C(\hat{\tau}(\boldsymbol{Z}); \eta^*)\}\beta & \text{if} \quad Z \in A^+ \\ \mathbb{1}\{\tau \notin C(\hat{\tau}(\boldsymbol{Z}); \eta^*)\}\beta - \mathbb{1}\{\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)\}(1-\beta) & \text{if} \quad Z \in A^-. \end{cases}$$

and the expected return is obtained by integrating over the design η. Define

$$\beta^+ = P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*) \mid Z \in A^+)$$

and

$$\beta^- = P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*) \mid Z \in A^-).$$

We then have

$$\begin{aligned} E(R) &= \{\beta^+(1-\beta) - (1-\beta^+)\beta\}P(Z \in A^+) + \{(1-\beta^-)\beta - \beta^-(1-\beta)\}P(Z \in A^-) \\ &= \{\beta^+ - \beta\}P(Z \in A^+) + \{\beta - \beta^-\}P(Z \in A^-). \end{aligned}$$

We can cast the validity criteria defined in Section 2.2.3 in terms of this betting (Buehler (1959) uses the term *weak exactness*). Consider the strategy $S(\boldsymbol{\mathcal{Z}}, \emptyset)$, where $\boldsymbol{\mathcal{Z}}$ is the set of all assignments. Clearly, $P(Z \in A^-) = P(Z \in \emptyset) = 0$ and $P(Z \in A^+) = P(Z \in \boldsymbol{\mathcal{Z}}) = 1$. Moreover, $\beta^+ = P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*) \mid Z \in A^+) = P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*))$, so

$$E(R) = \beta^+ - \beta = P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)) - \beta.$$

It is then easy to see that the expected return for this strategy is null if and only if the confidence interval of Player 1 has the advertised coverage β,

$$E(R) = 0 \quad \Leftrightarrow P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)) = \beta,$$

which corresponds to our frequentist strict validity criterion. We state the following proposition:

Proposition B.3.1.1. *The following assertions are equivalent:*

1. *The procedure $C(\hat{\tau}(\boldsymbol{Z}); \eta^*)$ is strictly valid (in the frequentist sense).*
2. *The expected return of strategy $S(\boldsymbol{\mathcal{Z}}, \emptyset)$ is zero.*

3. *The expected return of strategy $S(\emptyset, \boldsymbol{\mathcal{Z}})$ is zero.*

In order to better understand Proposition B.3.1.1, suppose that the choice of design η^* leads to an interval with coverage below the advertised level β. That is, such that

$$P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)) < \beta.$$

Now consider the strategy $S(\emptyset, \boldsymbol{\mathcal{Z}})$, we have

$$E(R) = \beta - P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*)) > 0$$

and so Player 2 can make money by betting against Player 1. It is easy to verify that if the interval has coverage greater than advertised, the strategy $S(\boldsymbol{\mathcal{Z}}, \emptyset)$ has positive return. The general idea here is that if Player 1 truly believes his confidence assertion, he should be willing to play against the strategies $S(\emptyset, \boldsymbol{\mathcal{Z}})$ or $S(\boldsymbol{\mathcal{Z}}, \emptyset)$.

One insight from the betting perspective is that the fact that the strategies $S(\boldsymbol{\mathcal{Z}}, \emptyset)$ and $S(\emptyset, \boldsymbol{\mathcal{Z}})$ have zero expected return may not be stringent enough a criterion for procedures. Suppose that the design η^* is such that the procedure $C(\hat{\tau}(\boldsymbol{Z}); \eta^*)$ has coverage probability β, as advertised. Now suppose that there is a set of assignments A such that

$$P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta^*) \mid Z \in A) = \beta^* > \beta.$$

The key question is the following: if the observed assignment $Z^{obs} \in A$, should Player 1 report the confidence β^* or β? This is equivalent to asking, if you have information based on your observed assignment that you should expect better or worse coverage using the standard method, should you use that information to determine your actual confidence level? The betting framework offers one perspective on the problem. Consider the strategy $S(A, \emptyset)$. The expected return is

$$E(R) = (\beta^* - \beta) P(Z \in A) > 0.$$

So the strategy $S(A, \emptyset)$ leads to a positive expected gain, and Player 2 can make money off of Player 1 by exploiting this strategy. Following the nomenclature in Buehler (1959), we give a definition of relevance.

Definition 5. *A subset of assignment $A \subset \boldsymbol{\mathcal{Z}}$ is said to be* relevant *for the pair $\{C(\hat{\tau}(\boldsymbol{Z}); \eta^*_Z), \beta\}$*

if either $S(A, \emptyset)$ or $S(\emptyset, A)$ have non-zero expected revenue. More generally a strategy $S(A^+, A^-)$ is said to be relevant *if it has non-zero expected revenue.*

This suggests that the existence of relevant strategies against $\{C(\hat{\tau}(\boldsymbol{Z}); \eta_Z^*), \beta\}$ is problematic for the procedure, even if the procedure is valid. In fact, the notion of relevance relates to that of conditional validity.

Proposition B.3.1.2. *The set A is not relevant for $\{C(\hat{\tau}(\boldsymbol{Z}); \eta_Z^*), \beta\}$ if and only if the procedure is valid conditionally on A. That, is*

$$E_{S(A,\emptyset)}(R) = 0 \quad \Leftrightarrow \quad E_{S(\emptyset,A)}(R) = 0 \quad \Leftrightarrow \quad P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta_Z^*) \mid Z \in A) = \beta.$$

B.3.2 Simple example

Consider the usual causal inference setup with N units, assuming SUTVA. The design η is Bernoulli, excluding $\boldsymbol{Z} = \vec{1}$ and $\boldsymbol{Z} = \vec{0}$. We consider the difference in means estimator $\hat{\tau}$. We assume a constant additive treatment effect such that for all i, $Y_i(1) - Y_i(0) = \tau$. Now, conditional on $N_1 = k$, this is a completely randomized experiment, and $E(\hat{\tau} \mid N_1 = k) = \tau$. The well known result for the variance of a completely randomized design yields

$$\begin{aligned}\text{Var}(\hat{\tau} \mid N_1 = k) &= V^* \left(\frac{1}{k} + \frac{1}{N-k} \right) \\ &\equiv v(k)\end{aligned}$$

where V^* is the sample variance of the potential outcomes under one treatment (which is the same as under the other treatment due to additive treatment effect). The estimator is also unbiased unconditionally $E(\hat{\tau}) = \tau$ but the unconditional variance becomes

$$\begin{aligned}\text{Var}(\hat{\tau}) &= E[Var(\hat{\tau} \mid N_1)] + \text{Var}(E[\hat{\tau} \mid N_1]) \\ &= E\left[V^* \left(\frac{1}{N_1} + \frac{1}{N_0} \right) \right] \\ &= V^* E\left(\frac{1}{N_1} + \frac{1}{N - N_1} \right) \\ &\equiv V\end{aligned}$$

where the expectation is with respect to the distribution of N_1 induced by the design η. We assume that we've reached the asymptotic regime, and so $\hat{\tau} \sim \mathcal{N}(\tau, V)$, where randomness is induced by design η (see Li and Ding (2017) Section 6 for an argument for why the CLT holds here). Consider the following confidence intervals:

$$C(\hat{\tau}(\boldsymbol{Z}); \eta) = [\hat{\tau} - 1.96\sqrt{V}, \hat{\tau} + 1.96\sqrt{V}].$$

By construction, we have $P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta)) = \beta = 0.95$. There exist winning betting strategies against this interval. Define $\mathcal{K} = \{k : v(k) < V\}$. Note that we have:

$$v(k) < V \quad \Leftrightarrow \quad \frac{1}{k} + \frac{1}{N-k} < E\left(\frac{1}{N_1} + \frac{1}{N - N_1}\right)$$

which doesn't depend on V^*. So $\mathcal{K}$ doesn't depend on V^* either. Define $A = \{Z : N_1 \in \mathcal{K}\}$, and $A_k = \{Z : N_1 = k\}$. Now notice that

$$P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta) | N_1 = k) = P(-1.96 \leq \frac{\hat{\tau} - \tau}{\sqrt{v(k)}} \sqrt{\frac{v(k)}{V}} \leq 1.96 \mid N_1 = k)$$

but $\frac{\hat{\tau} - \tau}{\sqrt{v(k)}} | N_1 = k \sim \mathcal{N}(0, 1)$. And so

$$\begin{aligned} P(\tau \in C(\hat{\tau}(\boldsymbol{Z}); \eta) | N_1 = k) &= P\left(-1.96\sqrt{\frac{V}{v(k)}} \leq \frac{\hat{\tau} - \tau}{\sqrt{v(k)}} \leq 1.96\sqrt{\frac{V}{v(k)}} \mid N_1 = k\right) \\ &= \Phi\left(1.96\sqrt{\frac{V}{v(k)}}\right) - \Phi\left(-1.96\sqrt{\frac{V}{v(k)}}\right) \\ &= 1 - 2\Phi\left(-1.96\sqrt{\frac{V}{v(k)}}\right) \\ &\equiv \beta_k \end{aligned}$$

Now the key is to notice the following:

Proposition B.3.2.1. *For all $k \in \mathcal{K}$, we have*

$$\beta_k > \beta$$

and similarly, for all $k \notin \mathcal{K}$, we have

$$\beta_k \leq \beta$$

where the strictness in the first equation comes for the the strictness in the definition of $\mathcal{K}$.

With this in place it is easy to verify that $S(A, A^c)$ is a relevant strategy. Indeed, if $N_1 = k \in \mathcal{K}$, we have

$$\begin{aligned} E(R \mid Z \in A_k) &= P(\tau \in C(\hat{\tau}(\mathbf{Z}); \eta) \mid Z \in A_k)(1-\beta) - P(\tau \notin C(\hat{\tau}(\mathbf{Z}); \eta) \mid Z \in A_k)\beta \\ &= \beta_k - \beta \\ &> 0 \end{aligned}$$

and so

$$\begin{aligned} E(R \mid Z \in A) &= \sum_{k \in \mathcal{K}} E(R \mid Z \in A_k) P(Z \in A_k | Z \in A) \\ &= \sum_{k \in \mathcal{K}} (\beta_k - \beta) P(Z \in A_k \mid Z \in A) \\ &> 0. \end{aligned}$$

Similar derivations show that $E(R \mid Z \in A^c) > 0$. And so

$$\begin{aligned} E(R) &= E(R \mid Z \in A) P(Z \in A) + E(R \mid Z \in A^c) P(Z \in A^c) \\ &> 0 \end{aligned}$$

which means that the strategy $S(A, A^c)$ is relevant for the procedure $\{C(\hat{\tau}(\mathbf{Z}); \eta), \beta\}$.

B.4 Heuristic argument on relevance and power

In general, if we keep our estimators the same, a conditional analysis will not result in higher precision or smaller confidence intervals than an unconditional one. This can be seen through a simple heuristic argument concerning the variance of our estimators. Although the mode of inference in this chapter does not rely strictly on the variance of our estimators, considering the variance is a proxy for considering the length of the confidence intervals. Consider restricting randomization by conditioning on some information about the assignment given by random variable w. We have the following well known variance decomposition: $\text{var}_{\eta_0}(\hat{\tau}) = E_{\eta_0}[\text{var}_{\eta_0}(\hat{\tau} \mid w)] + \text{var}_{\eta_0}(E_{\eta_0}[\hat{\tau} \mid w])$. It is easy to see that if $E_{\eta_0}[\hat{\tau} \mid w] = \tau$ then $\text{var}_{\eta_0}(\hat{\tau}) = E_{\eta_0}[\text{var}_{\eta_0}(\hat{\tau} \mid w)]$. This implies that on average, over the distribution of w, the conditional, or restricted, variance is equal to the unconditional, or

unrestricted, variance. This argument was made by Sundberg (2003) under a predictive view. Sundberg (2003) argued for the use of conditioning for lowering the mean-squared error of the predicted squared error for variance estimation. Of course, if we allow different estimators to be used for the conditional and unconditional designs, we do not have such guarantees and in fact may see a power gain through changing the estimator. For instance, in post-stratification typically we use a stratified adjusted blocking estimator rather than the simple difference in means estimator.

Appendix C

Appendix to Chapter 3

C.1 Variance derivations

C.1.1 Variance and covariance of observed mean potential outcomes

This section gives results for the building blocks necessary to obtain results such as Equation 3.1 and Equation 3.4. We start with the variance and covariance of the treatment indicators, which give results related to those in Lemma 1 and 2 of Dasgupta et al. (2015). For $i \neq k$,

$$\begin{aligned}
\text{Cov}\left(W_i(z_j), W_k(z_j)\right) &= E\left[W_i(z_j)W_k(z_j)\right] - E\left[W_i(z_j)\right]E\left[W_k(z_j)\right] \\
&= P(W_k(z_h) = 1 | W_i(z_j) = 1)P(W_i(z_j) = 1) - \frac{n_j^2}{n^2} \\
&= \frac{n_j}{n}\frac{n_j - 1}{n - 1} - \frac{n_j^2}{n^2} \\
&= \frac{n_j(n_j - n)}{n^2(n-1)}.
\end{aligned}$$

For $i \neq k$ and $j \neq h$,

$$\begin{aligned}
\text{Cov}\left(W_i(z_j), W_k(z_h)\right) &= E\left[W_i(z_j)W_k(z_h)\right] - E\left[W_i(z_j)\right]E\left[W_k(z_h)\right] \\
&= P(W_k(z_h) = 1 | W_i(z_j) = 1)P(W_i(z_j) = 1) - \frac{n_j n_h}{n^2} \\
&= \frac{n_h}{n-1}\frac{n_j}{n} - \frac{n_j n_h}{n^2} \\
&= \frac{n_j n_h}{n^2(n-1)}.
\end{aligned}$$

These results can be used directly to get the variance of $\bar{Y}^{obs}(z_j)$ and covariance of $\bar{Y}^{obs}(z_j)$ and $\bar{Y}^{obs}(z_h)$. See Lu (2016b) for proof.

C.1.2 Variance for fractional factorial design

This section gives details on the derivation of the variance of our estimators of factorial effects under a fractional factorial design, as given in Equation 3.4. The proofs here are similar to those given in Dasgupta et al. (2015) and Lu (2016b). We first breakdown $\tilde{S}_k^2$ in the same way that Dasgupta et al. (2015) and Lu (2016b) broke down S_k^2, to show that we obtain similar results for the fractional case as the full factorial. Let g_{kj}^* be the jth element of vector $\boldsymbol{g}_k^*$.

$$\begin{aligned}
\tilde{S}_k^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{\tau}_i(k) - \tilde{\tau}(k)\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} 2^{-2(K-p-1)}\left(\boldsymbol{g}_k^{*T}\boldsymbol{Y}_{*i} - \boldsymbol{g}_k^{*T}\bar{\boldsymbol{Y}}_*\right)^2 \\
&= \frac{1}{n-1}2^{-2(K-p-1)}\sum_{i=1}^{n}\left(\sum_{j=1}^{J'} g_{kj}^*\left(Y_i(z_j^*) - \bar{Y}(z_j^*)\right)\right)^2 \\
&= 2^{-2(K-p-1)}\left[\sum_{j=1}^{J'} g_{kj}^{*2}S^2(z_j^*) + \sum_j\sum_{h\neq j} g_{kj}^* g_{kh}^* S^2(z_j^*, z_h^*)\right]
\end{aligned}$$

Now we find the variance of $\widehat{\tau}^*(k)$ in a proof analogous to that given for balanced

factorial designs in Dasgupta et al. (2015) and for unbalanced factorial designs in Lu (2016b).

$$\begin{aligned}
\text{Var}\left(\widehat{\tau}^*(k)\right) &= \frac{1}{2^{2(K-p-1)}} \boldsymbol{g}_k^{*T} \text{Var}\left(\bar{\boldsymbol{Y}}_*^{obs}\right) \boldsymbol{g}_k^* \\
&= \frac{1}{2^{2(K-p-1)}} \left[\sum_{j=1}^{J'} g_{kj}^{*2} \text{Var}\left(\bar{Y}^{obs}(\boldsymbol{z}_j^*)\right) + \sum_j \sum_{h \neq j} g_{kj}^* g_{kh}^* \text{Cov}\left(\bar{Y}^{obs}(\boldsymbol{z}_j^*), \bar{Y}^{obs}(\boldsymbol{z}_h^*)\right) \right] \\
&= \frac{1}{2^{2(K-p-1)}} \left[\sum_{j=1}^{J'} \frac{n - n_j^*}{n n_j^*} S^2(\boldsymbol{z}_j^*) - \frac{1}{n} \sum_j \sum_{h \neq j} g_{kj}^* g_{kh}^* S^2(\boldsymbol{z}_j^*, \boldsymbol{z}_h^*) \right] \\
&= \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J'} \left(\frac{n - n_j^*}{n n_j^*} + \frac{1}{n} \right) S^2(\boldsymbol{z}_j^*) - \frac{1}{n} \tilde{S}_k^2 \\
&= \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J'} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*) - \frac{1}{n} \tilde{S}_k^2
\end{aligned}$$

C.1.3 Covariance for fractional factorial design

In this section we find the covariance between two factorial effect estimates from a fractional factorial design. We again closely follow proofs from Dasgupta et al. (2015) and Lu (2016b), showing that the same processes work in the fractional factorial case. First we breakdown $\tilde{S}_{k,k'}^2$.

$$\begin{aligned}
\tilde{S}_{k,k'}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\tilde{\tau}_i(k) - \tilde{\tau}(k) \right) \left(\tilde{\tau}_i(k') - \tilde{\tau}(k') \right) \\
&= \frac{1}{n-1} 2^{-2(K-p-1)} \sum_{i=1}^{n} \left(\boldsymbol{g}_k^{*T} \boldsymbol{Y}_{*i} - \boldsymbol{g}_k^{*T} \bar{\boldsymbol{Y}}_* \right) \left(\boldsymbol{g}_{k'}^{*T} \boldsymbol{Y}_{*i} - \boldsymbol{g}_{k'}^{*T} \bar{\boldsymbol{Y}}_* \right) \\
&= \frac{1}{n-1} 2^{-2(K-p-1)} \sum_{i=1}^{n} \left(\sum_{j=1}^{J'} g_{kj}^* (Y_i(\boldsymbol{z}_j^*) - \bar{Y}(\boldsymbol{z}_j^*)) \right) \left(\sum_{j=1}^{J'} g_{k'j}^* (Y_i(\boldsymbol{z}_j^*) - \bar{Y}(\boldsymbol{z}_j^*)) \right) \\
&= 2^{-2(K-p-1)} \left(\sum_{j=1}^{J'} g_{kj}^* g_{k'j}^* S^2(\boldsymbol{z}_j^*) + \sum_j \sum_{h \neq j} g_{kj}^* g_{k'h}^* S^2(\boldsymbol{z}_j^*, \boldsymbol{z}_h^*) \right)
\end{aligned}$$

$$
\begin{aligned}
\text{Cov}\left(\widehat{\tau}^*(k), \widehat{\tau}^*(k')\right) &= 2^{-2(K-p-1)} \boldsymbol{g}_k^{*T} \text{Var}\left(\bar{\boldsymbol{Y}}_*^{obs}\right) \boldsymbol{g}_{k'}^* \\
&= \frac{1}{2^{2(K-p-1)}} \Big[\sum_{j=1}^{J'} g_{kj}^* g_{k'j} \text{Var}\left(\bar{Y}^{obs}(\boldsymbol{z}_j^*)\right) \\
&+ \sum_j \sum_{h \neq j} g_{kj}^* g_{k'h}^* \text{Cov}\left(\bar{Y}^{obs}(\boldsymbol{z}_j^*), \bar{Y}^{obs}(\boldsymbol{z}_h^*)\right) \Big] \\
&= \frac{1}{2^{2(K-p-1)}} \left[\sum_{j=1}^{J'} \frac{n - n_j^*}{n n_j^*} g_{kj}^* g_{k'j} S^2(\boldsymbol{z}_j^*) - \frac{1}{n} \sum_j \sum_{h \neq j} g_{kj}^* g_{k'h}^* S^2(\boldsymbol{z}_j^*, \boldsymbol{z}_h^*) \right] \\
&= \frac{1}{2^{2(K-p-1)}} \left[\sum_{j=1}^{J'} \left(\frac{n - n_j^*}{n n_j^*} + \frac{1}{n} \right) g_{kj}^* g_{k'h}^* S^2(\boldsymbol{z}_j^*) \right] - \frac{1}{n} \tilde{S}_{k,k'}^2 \\
&= \frac{1}{2^{2(K-p-1)}} \left[\sum_{j=1}^{J'} \frac{1}{n_j^*} g_{kj}^* g_{k'h}^* S^2(\boldsymbol{z}_j^*) \right] - \frac{1}{n} \tilde{S}_{k,k'}^2 \\
&= \frac{1}{2^{2(K-p-1)}} \left[\sum_{j: g_{kj}^* = g_{k'j}^*} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*) - \sum_{j: g_{kj}^* \neq g_{k'j}^*} \frac{1}{n_j^*} S^2(\boldsymbol{z}_j^*) \right] - \frac{1}{n} \tilde{S}_{k,k'}^2
\end{aligned}
$$

C.2 Relating linear regression estimators to Neyman estimators in the fractional factorial design

This section gives a brief overview of a proof that the linear regression point estimates are the same as the Neymanian estimates for the fractional factorial design. For proofs of these results for the full factorial design, see Lu (2016b). The linear regression coefficient estimate is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}^{obs},$$

where $\boldsymbol{X}$ is a $n \times 2^{K-p}$ matrix whose columns correspond to first an intercept, then include the main effects, then the second order interactions not aliased with the main effects, and so on such that no two columns are aliased. For instance, for the 2^{3-1} design laid out in Section 3.3, $\boldsymbol{X}$ would be a design matrix with a first column of 1's and the rest of the columns corresponding to levels of the first, second, and third factor. No interactions would be included in this example because each of the two factor interactions is aliased with a main effect and the three factor interaction is aliased with the intercept, which are all

already included in the model. Thus the design looks a bit like the columns $\boldsymbol{g}_1^*$, $\boldsymbol{g}_2^*$, and $\boldsymbol{g}_3^*$ of Table 3.2, but with repeated rows for each of the units assigned to the same treatment combination. We need $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$. Denote $\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

$$(\boldsymbol{BX})_{ij} = \boldsymbol{b}_{i\cdot}\boldsymbol{f}_j$$

where $\boldsymbol{b}_{i\cdot}$ is the ith row of $\boldsymbol{B}$ and $\boldsymbol{f}_j$ is the jth column of $\boldsymbol{X}^T$. Using notation from Section 3.5.3, let $\boldsymbol{h}_j^*$ be the jth row of $\boldsymbol{G}^{*T}$ which is an expanded version of $\boldsymbol{z}_j^*$ which includes elements for interactions and the intercept. Then f_{ji}, the ith element of $\boldsymbol{f}_j$ is h_{kj}^*, the jth element of $\boldsymbol{h}_k^*$, where k is the treatment combination for the ith individual. It must be true that $(\boldsymbol{BX})_{ii} = 1$ and $(\boldsymbol{BX})_{ij} = 0$ for $i \neq j$. Consider letting the ith row of $\boldsymbol{B}$ be $\boldsymbol{b}_{i\cdot} = \frac{1}{2^{K-p}}(\tilde{\boldsymbol{n}}^{-1} \circ \boldsymbol{f}_i)^T$ where $\tilde{\boldsymbol{n}}^{-1}$ is the vector whose ith entry is $\frac{1}{n_j^*}$ where j is number of the treatment combination that the ith unit is assigned to ($j = \sum_{k=1}^{J'} kW_i(z_k)$). Then we have

$$\begin{aligned}
(\boldsymbol{BX})_{kk} &= \boldsymbol{b}_{k\cdot}\boldsymbol{f}_k \\
&= \frac{1}{2^{K-p}}(\tilde{\boldsymbol{n}}^{-1} \circ \boldsymbol{f}_k)^T\boldsymbol{f}_k \\
&= \frac{1}{2^{K-p}}(\tilde{\boldsymbol{n}}^{-1})^T\boldsymbol{f}_k \circ \boldsymbol{f}_k \\
&= \frac{1}{2^{K-p}}\sum_{j=1}^{J'}\sum_{i:W_i(\boldsymbol{z}_j^*)=1}\frac{1}{n_j^*} \\
&= \frac{1}{J'}\sum_j 1 \\
&= 1.
\end{aligned}$$

For $k \neq j$

$$\begin{aligned}
(\boldsymbol{BX})_{kj} &= \boldsymbol{b}_{k\cdot}\boldsymbol{f}_j \\
&= \frac{1}{2^{K-p}}(\tilde{\boldsymbol{n}}^{-1} \circ \boldsymbol{f}_k)^T \boldsymbol{f}_j \\
&= \frac{1}{2^{K-p}}(\tilde{\boldsymbol{n}}^{-1})^T \boldsymbol{f}_k \circ \boldsymbol{f}_j \\
&= \frac{1}{2^{K-p}} \sum_{s=1}^{J'} \frac{1}{n_s^*} \sum_{i:W_i(\boldsymbol{h}_s^*)=1} h_{sk}^* h_{sj}^* \\
&= \frac{1}{2^{K-p}} \left[\sum_{s:h_{sk}^*=h_{sj}^*} 1 - \sum_{s:h_{sk}^* \neq h_{sj}^*} 1 \right] \\
&= 0.
\end{aligned}$$

So the kth row of $\boldsymbol{B}$ is $\frac{1}{2^{K-p}}(\tilde{\boldsymbol{n}}^{-1} \circ \boldsymbol{f}_k)^T$. This means that

$$\begin{aligned}
\hat{\beta}_k &= \left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^{obs} \right)_k \\
&= (\frac{1}{2^{K-p}}\tilde{\boldsymbol{n}}^{-1} \circ \boldsymbol{f}_k)^T \boldsymbol{Y}^{obs} \\
&= \frac{1}{2^{K-p}} \sum_{j=1}^{J'} \frac{1}{n_j^*} \sum_{i:W_i(\boldsymbol{z}_j^*)=1} h_{jk}^* Y_i(\boldsymbol{h}_j^*) \\
&= \frac{1}{2^{K-p}} \sum_{j=1}^{J'} h_{jk}^* \boldsymbol{Y}^{obs}(\boldsymbol{z}_j^*) \\
&= \frac{1}{2^{K-p}} \boldsymbol{g}_k^{*T} \bar{\boldsymbol{Y}}^{obs} \\
&= \frac{1}{2}\widehat{\tau}^*(k).
\end{aligned}$$

So, indeed, the linear regression estimates are one half of the factorial effects.

Now let us consider the HC2 variance estimator. It has the form (MacKinnon and White, 1985)

$$\widehat{Var}_{HC2}(\widehat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\widehat{\Omega}\boldsymbol{X} \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

where $\widehat{\Omega} = \text{diag}\left(\frac{\hat{e}_i}{1-h_{ii}}\right)$ with $\hat{e}_i$ being the residual for observation i and h_{ii} being the ii value of the hat matrix, $\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. For discussion of this estimators in the single treatment case and the factorial case, see Samii and Aronow (2012) and Lu (2016b), respectively. We use similar ideas to those papers in the arguments below.

If unit i was assigned to treatment $\boldsymbol{z}_k^*$ then the ith column of $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is $\boldsymbol{b}_{\cdot i} = \frac{1}{2^{K-p}}\frac{1}{n_k^*}\boldsymbol{h}_k^{*T}$. We have

$$\begin{aligned}\left(\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)_{ii} &= \frac{1}{2^{K-p}}\frac{1}{n_k^*}\boldsymbol{h}_k^*\boldsymbol{h}_k^{*T} \\ &= \frac{1}{2^{K-p}}\frac{1}{n_k^*}2^{K-p} \\ &= \frac{1}{n_k^*}.\end{aligned}$$

So then $1 - h_{ii} = 1 - \frac{1}{n_k^*} = \frac{n_k^* - 1}{n_k^*}$. This in turn means that

$$\frac{\hat{e}_i}{1-h_{ii}} = \frac{n_k^*\left(Y_i^{obs} - \bar{Y}^{obs}(\boldsymbol{z}_k^*)\right)^2}{n_k^* - 1}$$

Now we can solve for the whole expression of $\widehat{Var}_{HC2}(\widehat{\boldsymbol{\beta}})$. We focus on the diagonal entries.

$$\begin{aligned}\left(\widehat{Var}_{HC2}(\widehat{\boldsymbol{\beta}})\right)_{kk} &= \left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\text{diag}\left(\frac{\hat{e}_j}{1-h_{jj}}\right)\right)_{i\cdot}\left(\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)_{\cdot i} \\ &= (\frac{1}{2^{K-p}}\tilde{\boldsymbol{n}}^{-1}\circ\boldsymbol{f}_k)^T\text{diag}\left(\frac{\hat{e}_j}{1-h_{jj}}\right)(\frac{1}{2^{K-p}}\tilde{\boldsymbol{n}}^{-1}\circ\boldsymbol{f}_k) \\ &= (\frac{1}{2^{K-p}}\boldsymbol{s}_1\circ\boldsymbol{f}_k)^T(\frac{1}{2^{K-p}}\tilde{\boldsymbol{n}}^{-1}\circ\boldsymbol{f}_k) \\ &= (\frac{1}{2^{K-p}}\boldsymbol{s}_1)^T(\frac{1}{2^{K-p}}\tilde{\boldsymbol{n}}^{-1}\circ\boldsymbol{f}_k\circ\boldsymbol{f}_k) \\ &= \frac{1}{2^{2(K-p)}}(\boldsymbol{s}_1)^T\tilde{\boldsymbol{n}}^{-1} \\ &= \frac{1}{2^{2(K-p)}}\sum_{j=1}^{J'}\frac{1}{n_j^*}\sum_{i:W_i(\boldsymbol{z}_j^*)=1}\frac{\left(Y_i(\boldsymbol{z}_j^*) - Y_i(\boldsymbol{z}_j^*)\right)^2}{n_j^* - 1} \\ &= \frac{1}{2^{2(K-p)}}\sum_j\frac{1}{n_j^*}s^2(\boldsymbol{z}_j^*)\end{aligned}$$

where $\boldsymbol{s}_1$ is the vector of whose ith entry, given entry i is assigned treatment combination $\boldsymbol{z}_k^*$, is $\frac{\left(Y_i(\boldsymbol{z}_k^*) - Y_i(\boldsymbol{z}_k^*)\right)^2}{n_k^* - 1}$. Thus, we have that $\left(\widehat{Var}_{HC2}(\widehat{\boldsymbol{\beta}})\right)_{kk}$ is $1/4$ times the Neyman style variance estimator.

C.3 Incomplete factorial Designs

C.3.1 Inference

This section discusses alternative incomplete factorial designs that one might use. See Byar et al. (1993) for more discussion on incomplete factorial designs. This discussion also applies more generally to recreating fractional designs in observational settings where we are using a subset of treatment combinations present in the data. We follow the same outline as proofs from Section C.1.2. In these incomplete factorial designs, for each main effect or other estimand of interest, we define a new experimental design tailored to estimating that effect. When this is done, we then analyze the data as if the treatment groups within the new design are the only possible treatment groups.

First we need to introduce some notation. Let $\dot{\boldsymbol{g}}_j$ be the same as $\boldsymbol{g}_j$ but with zero elements corresponding to treatment combinations that are not included in the particular design in use. Let 2^m treatment groups be used in the design, with half assigned to the $+1$ level of factor one and half assigned to the -1 level of factor one. Let $\dot{\tau}(k) = 2^{-(m-1)}\dot{\boldsymbol{g}}_k^T\bar{\boldsymbol{Y}}$. Then we can do a breakdown $\dot{S}_k^2$, defined in the first line below:

$$\begin{aligned}
\dot{S}_k^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\dot{\tau}_i(k) - \dot{\tau}(k)\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} 2^{-2(m-1)}\left(\dot{\boldsymbol{g}}_k^T\boldsymbol{Y}_i - \dot{\boldsymbol{g}}_k^T\bar{\boldsymbol{Y}}\right)^2 \\
&= \frac{1}{n-1}2^{-2(m-1)}\sum_{i=1}^{n}\left(\sum_{j=1}^{J}\dot{g}_{kj}\left(Y_i(\boldsymbol{z}_j) - \bar{Y}(\boldsymbol{z}_j)\right)\right)^2 \\
&= 2^{-2(m-1)}\left[\sum_{j=1}^{J}\dot{g}_{kj}^2 S^2(\boldsymbol{z}_j) + \sum_{j}\sum_{h\neq j}\dot{g}_{kj}\dot{g}_{kh}S^2(\boldsymbol{z}_j,\boldsymbol{z}_h)\right].
\end{aligned}$$

Then we have

$$\begin{aligned}
\text{Var}\left(\widehat{\tau}(k)\right) &= \frac{1}{2^{2(m-1)}}\dot{\boldsymbol{g}}_k^T \text{Var}\left(\bar{\boldsymbol{Y}}^{obs}\right)\dot{\boldsymbol{g}}_k \\
&= \frac{1}{2^{2(m-1)}}\left[\sum_{j=1}^{J}\dot{g}_{kj}^2\text{Var}\left(\bar{Y}^{obs}(\boldsymbol{z}_j)\right) + \sum_j\sum_{h\neq j}\dot{g}_{kj}\dot{g}_{kh}\text{Cov}\left(\bar{Y}^{obs}(\boldsymbol{z}_j),\bar{Y}^{obs}(\boldsymbol{z}_h)\right)\right] \\
&= \frac{1}{2^{2(m-1)}}\left[\sum_{j=1}^{J}\dot{g}_{kj}^2\frac{n-n_j}{nn_j}S^2(\boldsymbol{z}_j) - \frac{1}{n}\sum_j\sum_{h\neq j}\dot{g}_{kj}\dot{g}_{kh}S^2(\boldsymbol{z}_j,\boldsymbol{z}_h)\right] \\
&= \frac{1}{2^{2(m-1)}}\sum_{j=1}^{J}\dot{g}_{kj}^2\left(\frac{n-n_j}{nn_j}+\frac{1}{n}\right)S^2(\boldsymbol{z}_j) - \frac{1}{n}\dot{S}_k^2 \\
&= \frac{1}{2^{2(m-1)}}\sum_{j=1}^{J}\dot{g}_{kj}^2\frac{1}{n_j}S^2(\boldsymbol{z}_j) - \frac{1}{n}\dot{S}_k^2.
\end{aligned}$$

An important note is that in terms of estimation of this variance, whether we analyze the data as if the treatment levels in this design are the only possible treatment combinations or if we keep the assumption that units can be assigned to any possible treatment combinations (which aids in the interpretation and inference), the variance estimator will be the same. This is because we can only estimate the first term in this expression which only involves the specific treatment levels in this design.

C.3.2 Variance of estimators for incomplete factorial designs

Consider comparing two designs. The first involves J^* treatment groups and the second involves $\tilde{J} < J^*$ treatment groups. Let's assume that for all treatment groups $n_j = c$ and $S^2(\boldsymbol{z}_j) = S^2$. That is, all treatment groups are the same size and effects are additive so that the variance of potential outcomes in each treatment group is the same.

Then for the first design we have variance of

$$\frac{1}{J^{*2}}\sum_j^{J^*}\frac{1}{n_j}S^2(\boldsymbol{z}_j) - \frac{1}{n}S_k^2 = \frac{1}{cJ^*}S^2.$$

For the second design we have

$$\frac{1}{\tilde{J}^2}\sum_j^{\tilde{J}}\frac{1}{n_j}S^2(\boldsymbol{z}_j) - \frac{1}{n}S_k^2 = \frac{1}{c\tilde{J}}S^2.$$

So in this setting the design with more treatment groups will have lower variance, which is

intuitive. However, in general if we do not have the additive treatment effect assumption, it is possible that the design with more treatment groups includes treatment groups that are much more variable and therefore the variance of the estimator is actually larger.

C.3.3 Regression with missing levels

This section discusses what will result from a standard regression interacting all factors when not all treatment combinations are observed in the data.

If the dataset is missing m treatment combinations, then the regression will be able to estimate the first $2^K - m$ effects (including interactions) that are not aliased and the rest will be removed due to collinearity of the matrix columns. Then for each effect, there will be some aliasing structure imposed but the same "design" will not necessarily be used for each factor.

To explore this scenario more, let's take the specific example of three factors where we only observe five of the eight treatment combinations. For simplicity let there be one observation for each treatment combination and let the model matrix be as follows:

$$\boldsymbol{X} = \begin{pmatrix} 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 \end{pmatrix}.$$

The first column corresponds to the intercept, the second through fourth columns give the levels for the three factors, and the fifth column corresponds to the interaction between the first and second factor. Note that the first four rows correspond to a 2^{3-1} design, so it is possible to recreate that design here.

We have

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ -0.25 & -0.25 & 0.25 & 0.25 & 0 \\ -0.25 & 0.25 & -0.25 & 0.25 & 0 \\ 0.5 & 0 & 0 & 0 & -0.5 \\ -0.25 & -0.25 & -0.25 & 0.25 & 0.5 \end{pmatrix}.$$

Recalling that $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}^{obs}$, the first three columns correspond to estimates we would get using the fractional factorial design using the defining relation $I = 123$. The last two estimates, for factor 3 and the interaction between factors 1 and 2, have a different aliasing structure. In particular, the aliasing on factor 3 is similar to aliasing structures in Section 3.5.3 and we can find that factor 3 will be aliased with the two-factor interactions 13 and 23 as well as the three-factor interaction.

C.4 Data illustration

This section gives some additional descriptors for the data.

Figure C.1 shows that the proportion of farmers is different across the eight treatments of the hypothetical fractional factorial experiment defined in Section 3.6.3.

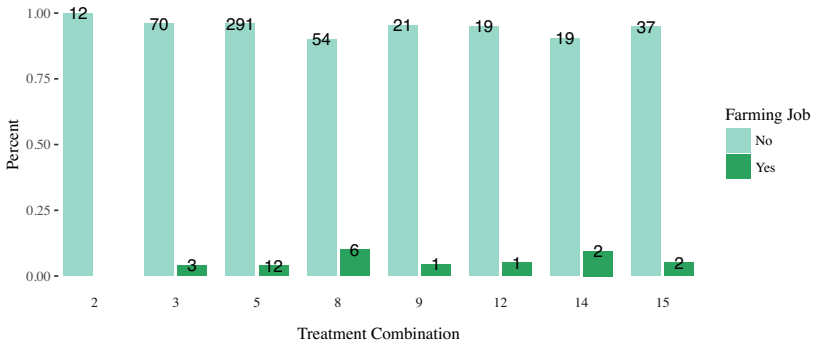

Figure C.1: *Comparing number of farmers across factor levels in the 2^{4-1} fractional factorial design. Text labels give number of observations per group.*

Figure C.2 shows the correlations between the levels of the different pesticides.

For each design, we explore regression (both saturated, i.e. with all interactions among

factors, and unsaturated, i.e. without interactions among factors), regression with covariates (both saturated and unsaturated for the factors), and Fisher tests for significance of effects. We also include the HC2 standard error estimate for the saturated model without covariates, where it is equivalent to the Neyman estimator, when possible. Note that all regression outputs are using log BMI as output. An important note is that individuals with missing values, either for factor levels, treatment levels, or covariates where they are used, were removed. This action likely changes the types of individuals within the analysis and therefore the generalizability of the results. However, as this is intended as simply an illustration of the methods and not as a full analysis to draw substantive conclusions from, this simple model suffices to allow us to continue with the analysis.

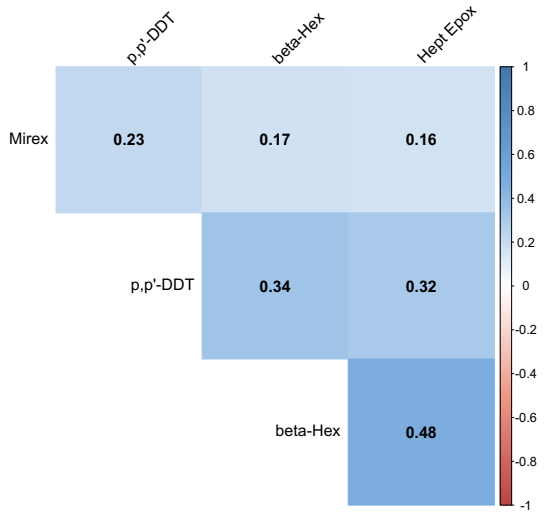

Figure C.2: *Plot of correlations between pesticide levels.*

C.4.1 Full factorial design

Full factorial: Regression analysis

We start by ignoring our limited data and use a full 2^4 factorial approach. Table C.1 shows an analysis with all factors and no interactions. Table C.2 shows the saturated model with all interactions. Note that the individual who received treatment combination $(-1, 1, 1, -1)$ has leverage 1 because they are the only individual with that treatment combination. Because of this data limitation, estimating variance using the HC2/Neyman variance estimator is not possible for the saturated case. Also note that in the saturated model, the variance estimates given in the linear model summary are all the same. This will be true of Neyman variance estimates too, since they are the same for each factorial effect estimators. Note the changes in significance of the estimator and even a change in sign for the estimate of beta-Hex going from the model with just main effects to the saturated model.

Table C.1: *All-pesticide model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.289	0.008	430.426	0.000
beta-Hex	-0.001	0.008	-0.067	0.946
Hept Epox	0.062	0.007	9.201	0.000
Mirex	-0.027	0.006	-4.709	0.000
p,p'-DDT	0.020	0.008	2.595	0.010

Full factorial: Regression analysis adjusting for covariates

This section gives the analysis of the full factorial design, adjusting for the covariates of income, ethnicity, gender and smoking status as linear factors in the model. For simplicity, we remove all individuals who had missing values for income or ethnicity, and assume that those values are missing at random. This resulted in 75 units being removed. In practice one should instead use multiple imputation to account for the missing values. One individual refused to give income (the only such individual) and so was removed. This did not affect the analysis. The unit assigned to the unique treatment combination, necessarily had leverage one in the saturated model. This baseline level for income in the analysis was "$0 to $4,999." Baseline for ethnicity is "Mexican American."

Table C.2: *Saturated model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.276	0.015	224.028	0.000
beta-Hex	0.010	0.015	0.713	0.476
Hept Epox	0.038	0.015	2.605	0.009
Mirex	-0.050	0.015	-3.445	0.001
p,p′-DDT	0.025	0.015	1.699	0.090
beta-Hex:Hept Epox	0.012	0.015	0.830	0.407
beta-Hex:Mirex	0.007	0.015	0.456	0.649
Hept Epox:Mirex	-0.037	0.015	-2.515	0.012
beta-Hex:p,p′-DDT	-0.008	0.015	-0.569	0.569
Hept Epox:p,p′-DDT	0.011	0.015	0.779	0.436
Mirex:p,p′-DDT	0.015	0.015	1.016	0.310
beta-Hex:Hept Epox:Mirex	0.026	0.015	1.790	0.074
beta-Hex:Hept Epox:p,p′-DDT	0.007	0.015	0.500	0.617
beta-Hex:Mirex:p,p′-DDT	0.003	0.015	0.179	0.858
Hept Epox:Mirex:p,p′-DDT	0.014	0.015	0.969	0.333
beta-Hex:Hept Epox:Mirex:p,p′-DDT	-0.004	0.015	-0.257	0.797

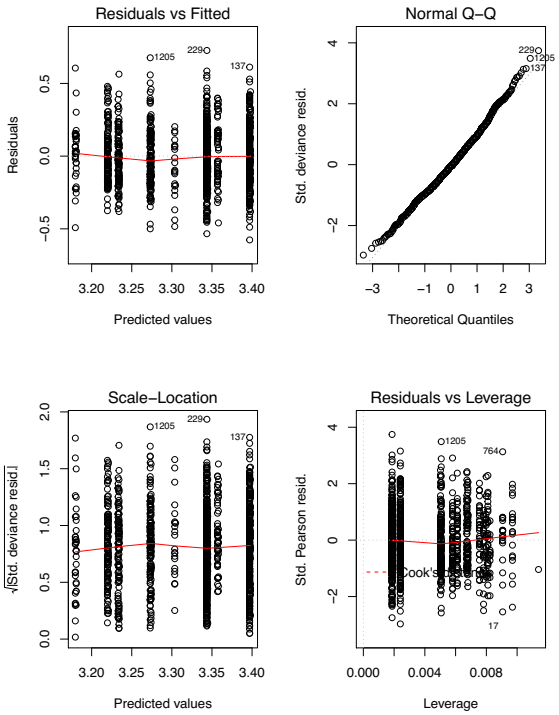

Figure C.3: *Basic diagnostics plot for the full model given in Table C.1.*

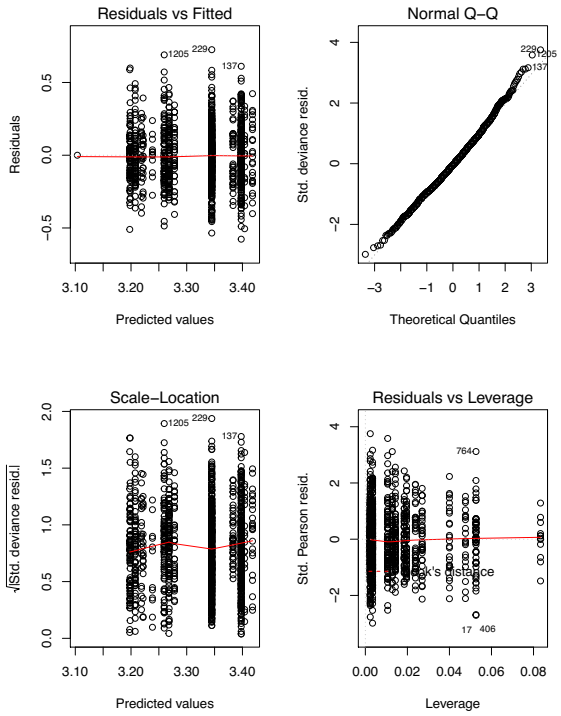

Figure C.4: *Basic diagnostics plot for the saturated model given in Table C.2. Note that the individual with unique treatment combination had leverage 1.*

Full factorial: Fisherian analysis

We assume the sharp null hypothesis of zero individual factorial effects, for all factors and interactions. This means that imputed missing potential outcomes for different assignments are just the observed potential outcomes. We do the imputation, or effectively rearrange the assignment vector, 1000 times.

As we can see from the plots and confirmed by calculation, only heptachlor epoxide (Hept Epox) and mirex appear to be significantly different from zero at the 0.05 level. We only examine the main effects here but could further consider interactions.

Table C.3: *All-pesticide model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.302	0.047	70.165	0.000
beta-Hex	0.001	0.008	0.070	0.944
Hept Epox	0.064	0.007	9.413	0.000
Mirex	-0.031	0.006	-5.240	0.000
p,p'-DDT	0.019	0.008	2.452	0.014
Income:$ 5,000 to $ 9,999	-0.005	0.051	-0.108	0.914
Income:$10,000 to $14,999	-0.016	0.049	-0.332	0.740
Income:$15,000 to $19,999	0.022	0.049	0.445	0.657
Income:$20,000 to $24,999	0.011	0.048	0.228	0.820
Income:$25,000 to $34,999	-0.014	0.047	-0.287	0.774
Income:$35,000 to $44,999	-0.002	0.048	-0.037	0.971
Income:$45,000 to $54,999	0.000	0.049	0.002	0.999
Income:$55,000 to $64,999	0.000	0.050	0.009	0.993
Income:$65,000 to $74,999	0.011	0.052	0.206	0.837
Income:$75,000 and Over	-0.004	0.046	-0.091	0.928
Income:Don't know	-0.168	0.142	-1.188	0.235
Income:Over $20,000	0.033	0.081	0.404	0.686
Ethnicity:Non-Hispanic Black	0.068	0.019	3.612	0.000
Ethnicity:Non-Hispanic White	-0.023	0.015	-1.514	0.130
Ethnicity:Other Hispanic	-0.022	0.033	-0.678	0.498
Ethnicity:Other Race - Including Multi-Racial	-0.055	0.029	-1.912	0.056
Gender:Male	-0.005	0.012	-0.405	0.686
Smoker: Yes	-0.011	0.011	-0.930	0.353

Table C.4: *Saturated model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.291	0.049	67.583	0.000
beta-Hex	0.011	0.015	0.766	0.444
Hept Epox	0.044	0.015	3.047	0.002
Mirex	-0.054	0.015	-3.698	0.000
p,p′-DDT	0.025	0.014	1.700	0.089
beta-Hex:Hept Epox	0.011	0.014	0.734	0.463
beta-Hex:Mirex	0.003	0.014	0.218	0.828
Hept Epox:Mirex	-0.039	0.015	-2.651	0.008
beta-Hex:p,p′-DDT	-0.011	0.015	-0.724	0.469
Hept Epox:p,p′-DDT	0.009	0.015	0.639	0.523
Mirex:p,p′-DDT	0.015	0.014	1.014	0.311
beta-Hex:Hept Epox:Mirex	0.028	0.014	1.927	0.054
beta-Hex:Hept Epox:p,p′-DDT	0.004	0.015	0.270	0.787
beta-Hex:Mirex:p,p′-DDT	0.005	0.015	0.371	0.711
Hept Epox:Mirex:p,p′-DDT	0.011	0.015	0.773	0.440
beta-Hex:Hept Epox:Mirex:p,p′-DDT	–0.002	0.014	-0.114	0.909
Income:$ 5,000 to $ 9,999	0.003	0.050	0.060	0.952
Income:$10,000 to $14,999	-0.016	0.048	-0.329	0.743
Income:$15,000 to $19,999	0.028	0.049	0.576	0.565
Income:$20,000 to $24,999	0.007	0.048	0.136	0.891
Income:$25,000 to $34,999	-0.011	0.047	-0.232	0.816
Income:$35,000 to $44,999	0.001	0.048	0.024	0.980
Income:$45,000 to $54,999	-0.001	0.049	-0.011	0.991
Income:$55,000 to $64,999	0.001	0.050	0.010	0.992
Income:$65,000 to $74,999	0.010	0.052	0.196	0.844
Income:$75,000 and Over	-0.003	0.046	-0.073	0.942
Income:Don’t know	-0.159	0.141	-1.127	0.260
Income:Over $20,000	0.039	0.081	0.488	0.626
Ethnicity:Non-Hispanic Black	0.067	0.019	3.571	0.000
Ethnicity:Non-Hispanic White	-0.023	0.015	-1.505	0.133
Ethnicity:Other Hispanic	-0.026	0.033	-0.792	0.429
Ethnicity:Other Race - Including Multi-Racial	-0.057	0.029	-1.971	0.049
Gender:Male	-0.002	0.012	-0.133	0.894
Smoker:Yes	-0.014	0.011	-1.215	0.225

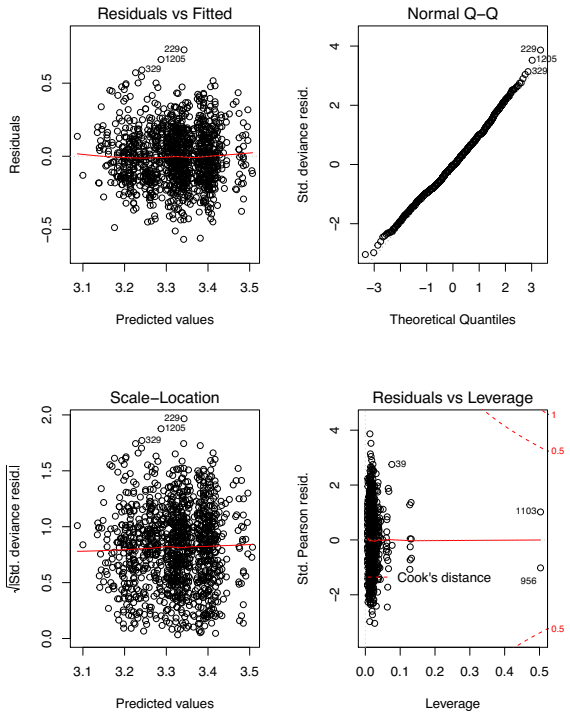

Figure C.5: *Basic diagnostics plot for the full model given in Table C.3.*

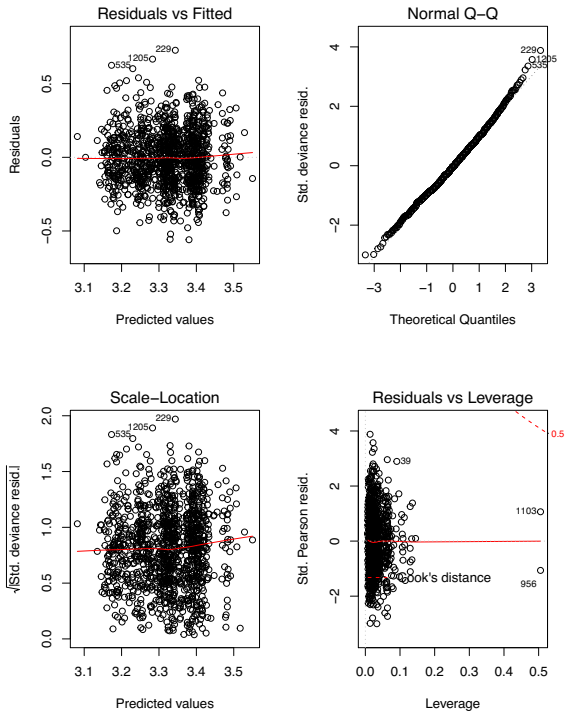

Figure C.6: *Basic diagnostics plot for the saturated model given in Table C.4. Note that the individual with unique treatment combination had leverage 1.*

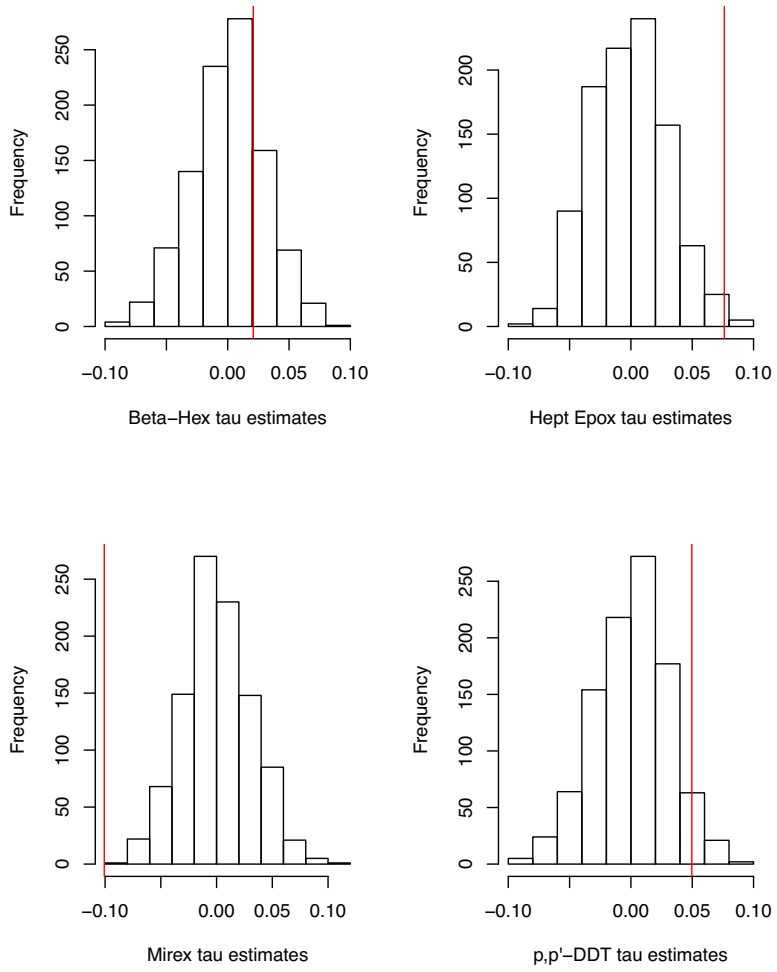

Figure C.7: *Plots of simulated treatment effects estimated. Observed treatment effect estimates plotted in red.*

C.4.2 Fractional factorial design

Fractional factorial: Regression analysis

This section gives the analysis of the fractional factorial design as is, using regression. Note that we remove farmers again, leaving 523 observations. Table C.5 shows an analysis with all factors and no interactions. Table C.6 shows the saturated model with all interactions. These results generally align with the full factorial analysis, in terms of sign and significance of terms. Figures C.8 and C.9 show basic diagnostic plots for the model with all main effects and the saturated model. Note again that it makes sense that the standard errors estimates are the same for all estimates in the saturated model because the same groups are being used to calculate them.

Table C.5: *All-pesticide model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.283	0.013	257.366	0.000
beta-Hex	0.007	0.012	0.570	0.569
Hept Epox	0.046	0.011	4.232	0.000
Mirex	-0.053	0.011	-4.673	0.000
p,p'-DDT	0.007	0.011	0.578	0.563

Table C.6: *Saturated model*

	Estimate	Std. Error	HC2	t value	Pr(>|t|)
(Intercept)	3.279	0.013	0.013	249.551	0.000
beta-Hex	-0.004	0.013	0.013	-0.284	0.776
Hept Epox	0.035	0.013	0.013	2.699	0.007
Mirex	-0.058	0.013	0.013	-4.389	0.000
p,p'-DDT	-0.001	0.013	0.013	-0.102	0.919
beta-Hex:Hept Epox	-0.003	0.013	0.013	-0.207	0.836
beta-Hex:Mirex	-0.005	0.013	0.013	-0.359	0.720
Hept Epox:Mirex	-0.028	0.013	0.013	-2.165	0.031

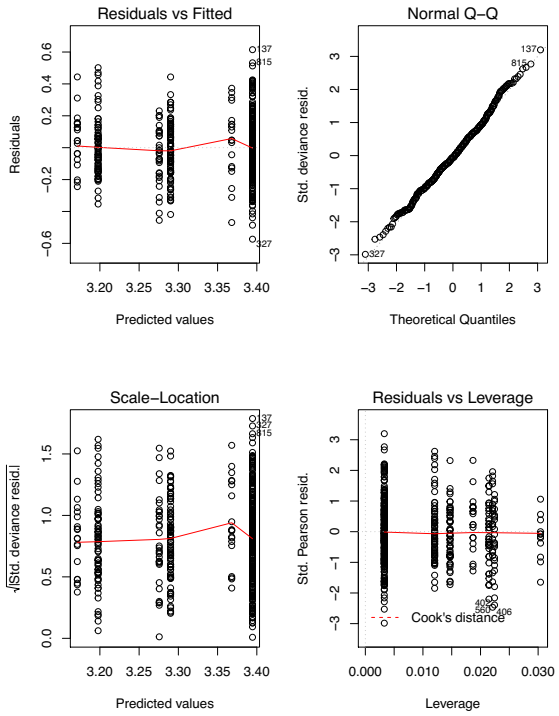

Figure C.8: *Basic diagnostics plot for the saturated model given in Table C.5.*

Fractional factorial: Regression analysis adjusting for covariates

This section gives the analysis of the fractional factorial design, adjusting for the covariates of income, ethnicity, gender and smoking status as linear factors in the model. We again removed units who were missing values for income or race, reducing the sample size by 34. One unit replied "Don't know" for income so this unit was removed. This did not affect the analysis.

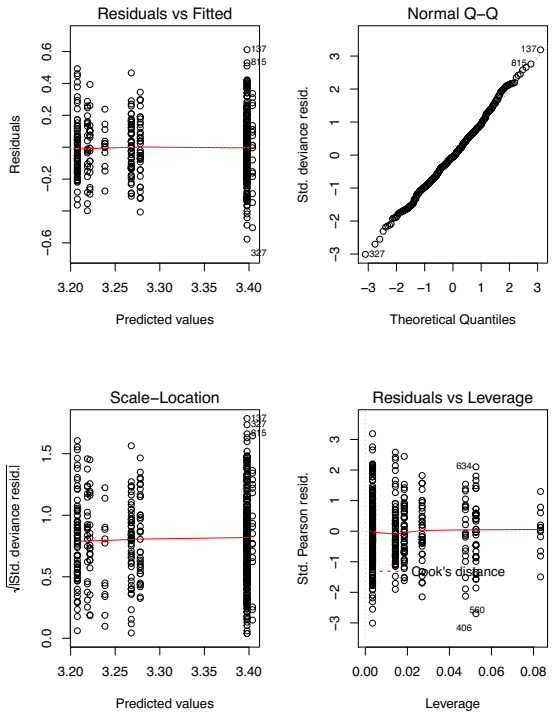

Figure C.9: *Basic diagnostics plot for the saturated model given in Table C.6.*

Fractional factorial: Fisherian analysis

Once again, only heptachlor epoxide (Hept Ex) and mirex appear to be significantly different than zero at the 0.05 level.

Table C.7: *All-pesticide model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.332	0.074	44.845	0.000
beta-Hex	0.008	0.012	0.682	0.495
Hept Epox	0.047	0.011	4.304	0.000
Mirex	-0.052	0.012	-4.464	0.000
p,p′-DDT	0.006	0.011	0.487	0.626
Income:$ 5,000 to $ 9,999	-0.079	0.083	-0.949	0.343
Income:$10,000 to $14,999	-0.066	0.077	-0.867	0.387
Income:$15,000 to $19,999	-0.060	0.078	-0.775	0.439
Income:$20,000 to $24,999	-0.031	0.077	-0.398	0.691
Income:$25,000 to $34,999	-0.027	0.076	-0.353	0.724
Income:$35,000 to $44,999	-0.014	0.076	-0.181	0.856
Income:$45,000 to $54,999	-0.036	0.076	-0.470	0.639
Income:$55,000 to $64,999	-0.001	0.080	-0.017	0.986
Income:$65,000 to $74,999	-0.017	0.080	-0.216	0.829
Income:$75,000 and Over	-0.048	0.073	-0.660	0.510
Income:Over $20,000	-0.022	0.118	-0.190	0.849
Ethnicity:Non-Hispanic Black	0.064	0.030	2.149	0.032
Ethnicity:Non-Hispanic White	-0.015	0.023	-0.667	0.505
Ethnicity:Other Hispanic	-0.025	0.051	-0.502	0.616
Ethnicity:Other Race - Including Multi-Racial	-0.040	0.043	-0.930	0.353
Gender:Male	-0.000	0.019	-0.014	0.989
Smoker:Yes	-0.016	0.018	-0.888	0.375

Table C.8: *Saturated model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.331	0.074	44.901	0.000
beta-Hex	-0.001	0.014	-0.075	0.941
Hept Epox	0.037	0.013	2.794	0.005
Mirex	-0.057	0.014	-4.212	0.000
p,p′-DDT	-0.002	0.013	-0.113	0.910
beta-Hex:Hept Epox	-0.005	0.013	-0.406	0.68
beta-Hex:Mirex	-0.007	0.013	-0.545	0.586
Hept Epox:Mirex	-0.027	0.013	-2.027	0.043
Income:$ 5,000 to $ 9,999	-0.086	0.083	-1.035	0.301
Income:$10,000 to $14,999	-0.076	0.077	-0.995	0.320
Income:$15,000 to $19,999	-0.064	0.077	-0.831	0.406
Income:$20,000 to $24,999	-0.045	0.077	-0.583	0.560
Income:$25,000 to $34,999	-0.031	0.076	-0.407	0.684
Income:$35,000 to $44,999	-0.023	0.076	-0.308	0.758
Income:$45,000 to $54,999	-0.046	0.076	-0.606	0.545
Income:$55,000 to $64,999	-0.013	0.080	-0.162	0.872
Income:$65,000 to $74,999	-0.027	0.080	-0.336	0.737
Income:$75,000 and Over	-0.056	0.073	-0.759	0.448
Income:Over $20,000	-0.028	0.117	-0.238	0.812
Ethnicity:Non-Hispanic Black	0.071	0.030	2.360	0.019
Ethnicity:Non-Hispanic White	-0.010	0.023	-0.451	0.652
Ethnicity:Other Hispanic	-0.025	0.051	-0.494	0.621
Ethnicity:Other Race - Including Multi-Racial	-0.041	0.042	-0.965	0.335
Gender:Male	0.003	0.019	0.139	0.889
Smoker:Yes	-0.016	0.018	-0.882	0.378

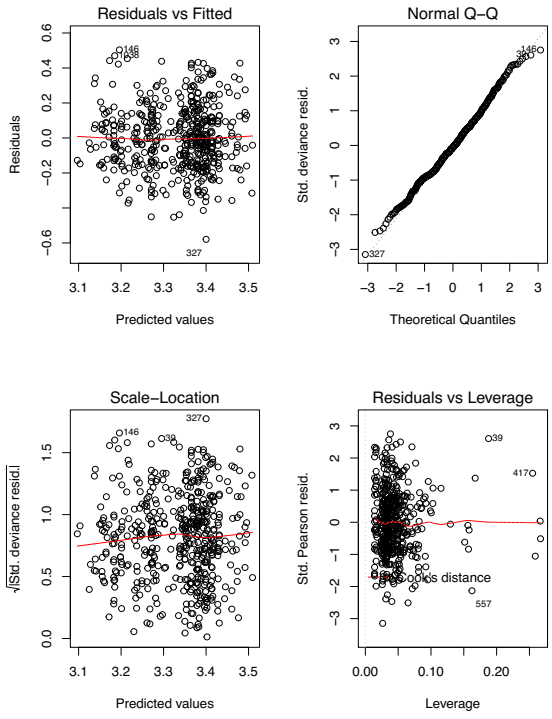

Figure C.10: *Basic diagnostics plot for the saturated model given in Table C.7.*

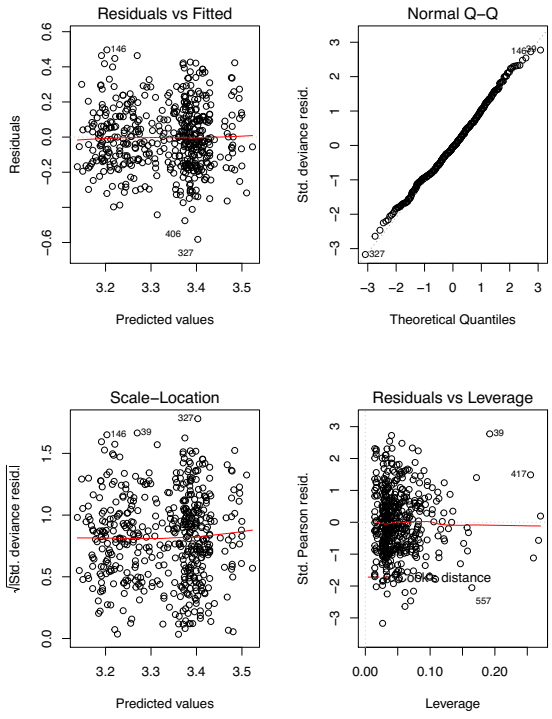

Figure C.11: *Basic diagnostics plot for the saturated model given in Table C.8.*

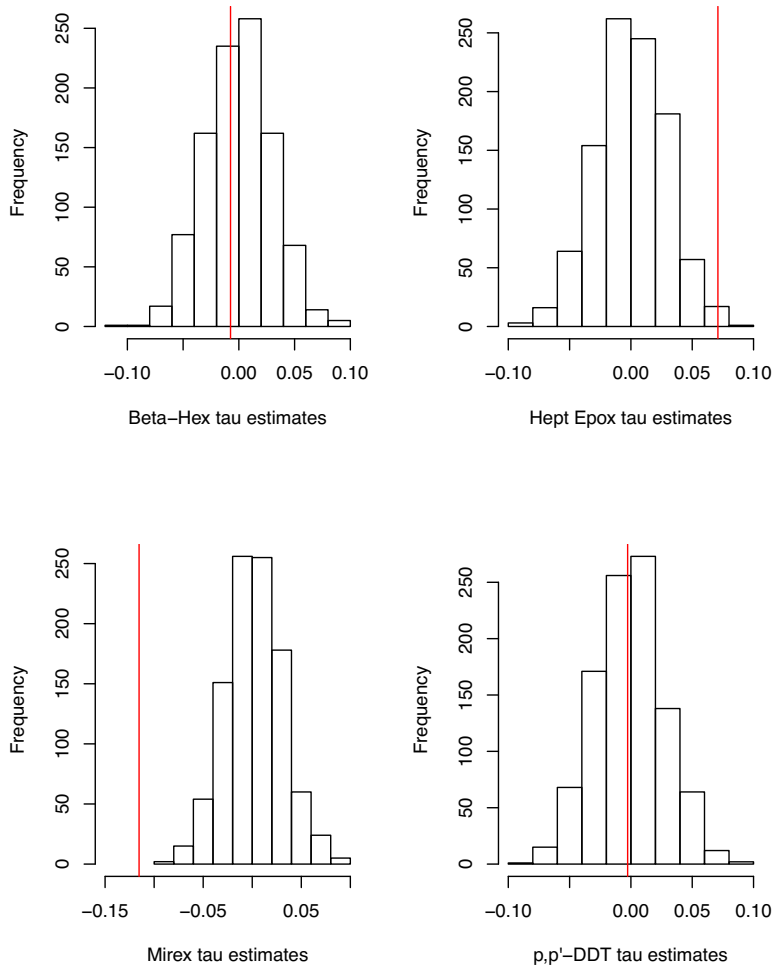

Figure C.12: *Plots of simulated treatment effects estimated. Observed treatment effect estimates plotted in red.*

C.4.3 Fractional factorial with covariate adjustment

169 units are retained after trimming.

Fractional factorial with covariate adjustment: Regression analysis

This section gives the analysis of the fractional factorial design after trimming to attain balance, using regression. Table C.9 shows an analysis with all factors and no interactions. Table C.10 shows the saturated model with all interactions. Figures C.13 and C.14 show basic diagnostic plots for the model with all main effects and the saturated model.

Table C.9: *All-pesticide model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.282	0.020	167.246	0.000
beta-Hex	0.041	0.020	2.053	0.042
Hept Epox	0.032	0.018	1.779	0.077
Mirex	-0.075	0.018	-4.127	0.000
p,p'-DDT	-0.006	0.020	-0.327	0.744

Table C.10: *Saturated model*

	Estimate	Std. Error	HC2 std	t value	Pr(>|t|)
(Intercept)	3.282	0.022	0.027	149.359	0.000
beta-Hex	0.038	0.022	0.027	1.719	0.088
Hept Epox	0.034	0.022	0.027	1.534	0.127
Mirex	-0.077	0.022	0.027	-3.515	0.001
p,p'-DDT	-0.008	0.022	0.027	-0.383	0.703
beta-Hex:Hept Epox	-0.007	0.022	0.027	-0.299	0.765
beta-Hex:Mirex	-0.003	0.022	0.027	-0.115	0.908
Hept Epox:Mirex	-0.006	0.022	0.027	-0.264	0.792

Fractional factorial with covariate balance: Regression analysis adjusting for covariates

This section gives the analysis on the trimmed data set adjusting for ethnicity, income, gender, and smoking status as linear factors in the regression. Figures C.15 and C.16 shows the balance across treatment groups for income and ethnicity after trimming, indicating further need to adjust. We see from Figure 3.2 that gender and smoking also require further adjustment, even after trimming.

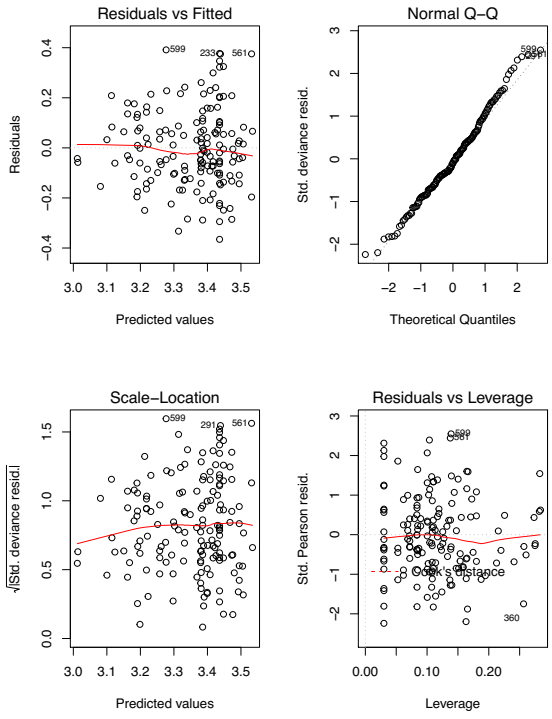

Figure C.13: *Basic diagnostics plot for the saturated model given in Table C.9.*

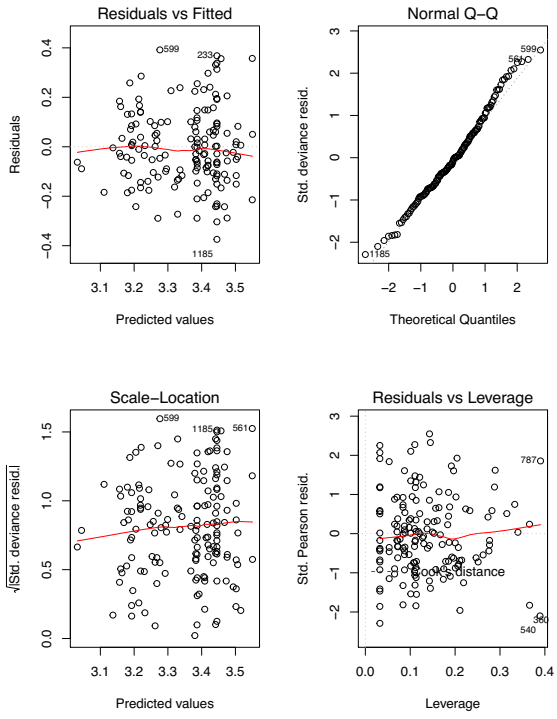

Figure C.14: *Basic diagnostics plot for the saturated model given in Table C.10.*

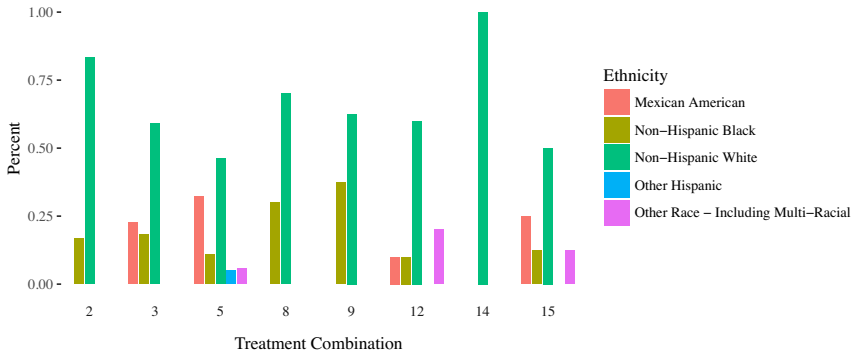

Figure C.15: *Balance of ethnicity in the different treatment groups after matching.*

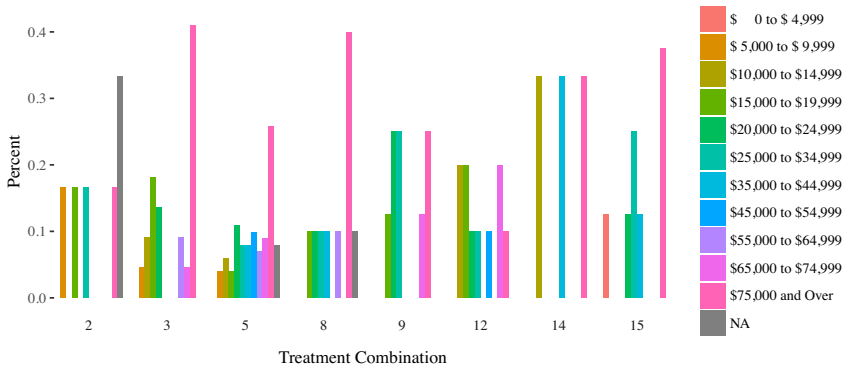

Figure C.16: *Balance of income in the different treatment groups after matching.*

Units who had missing values for income or ethnicity were removed, which reduced the sample size by 9. We found that two units had unique income values of "Over $20,000" and "0 to $4,999". These units were removed. This changes the baseline level for income in the analysis from "$0 to $4,999" to "$5,000 to $9,999."

Table C.11: *All-pesticide model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.292	0.086	38.110	0.000
beta-Hex	0.037	0.021	1.759	0.081
Hept Epox	0.047	0.019	2.509	0.013
Mirex	-0.060	0.020	-3.086	0.002
p,p'-DDT	-0.008	0.021	-0.383	0.702
Income:$10,000 to $14,999	-0.031	0.086	-0.360	0.720
Income:$15,000 to $19,999	-0.003	0.083	-0.038	0.970
Income:$20,000 to $24,999	-0.001	0.082	-0.014	0.989
Income:$25,000 to $34,999	-0.014	0.085	-0.160	0.873
Income:$35,000 to $44,999	0.069	0.088	0.777	0.439
Income:$45,000 to $54,999	0.039	0.088	0.445	0.657
Income:$55,000 to $64,999	0.059	0.087	0.675	0.501
Income:$65,000 to $74,999	-0.048	0.083	-0.577	0.565
Income:$75,000 and Over	0.015	0.075	0.202	0.840
Ethnicity:Non-Hispanic Black	0.059	0.047	1.251	0.213
Ethnicity:Non-Hispanic White	0.016	0.036	0.435	0.664
Ethnicity:Other Hispanic	0.056	0.082	0.677	0.499
Ethnicity:Other Race - Including Multi-Racial	-0.082	0.074	-1.109	0.269
Gender:Male	-0.058	0.030	-1.952	0.053
Smoker:Yes	-0.003	0.029	-0.087	0.930

Fractional factorial with covariate adjustment: Fisherian analysis.

In this analysis mirex is the only pesticide that appears to have a significant effect on BMI at the 0.05 level.

Table C.12: *Saturated model*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.278	0.088	37.234	0.000
beta-Hex	0.033	0.023	1.459	0.1470
Hept Epox	0.042	0.022	1.892	0.061
Mirex	-0.056	0.023	-2.428	0.017
p,p′-DDT	-0.011	0.024	-0.461	0.645
beta-Hex:Hept Epox	-0.006	0.023	-0.276	0.783
beta-Hex:Mirex	-0.019	0.023	-0.844	0.400
Hept Epox:Mirex	-0.014	0.023	-0.600	0.549
Income:$10,000 to $14,999	-0.035	0.088	-0.393	0.695
Income:$15,000 to $19,999	0.001	0.084	0.015	0.988
Income:$20,000 to $24,999	0.004	0.083	0.044	0.965
Income:$25,000 to $34,999	-0.008	0.086	-0.088	0.930
Income:$35,000 to $44,999	0.073	0.089	0.822	0.413
Income:$45,000 to $54,999	0.037	0.089	0.411	0.682
Income:$55,000 to $64,999	0.063	0.088	0.720	0.473
Income:$65,000 to $74,999	-0.049	0.084	-0.587	0.558
Income:$75,000 and Over	0.019	0.076	0.253	0.800
Ethnicity:Non-Hispanic Black	0.072	0.049	1.492	0.138
Ethnicity:Non-Hispanic White	0.022	0.037	0.598	0.551
Ethnicity:Other Hispanic	0.055	0.083	0.665	0.507
Ethnicity:Other Race - Including Multi-Racial	-0.090	0.075	-1.198	0.233
Gender:Male	-0.055	0.030	-1.830	0.069
Smoker:Yes	-0.005	0.030	-0.155	0.877

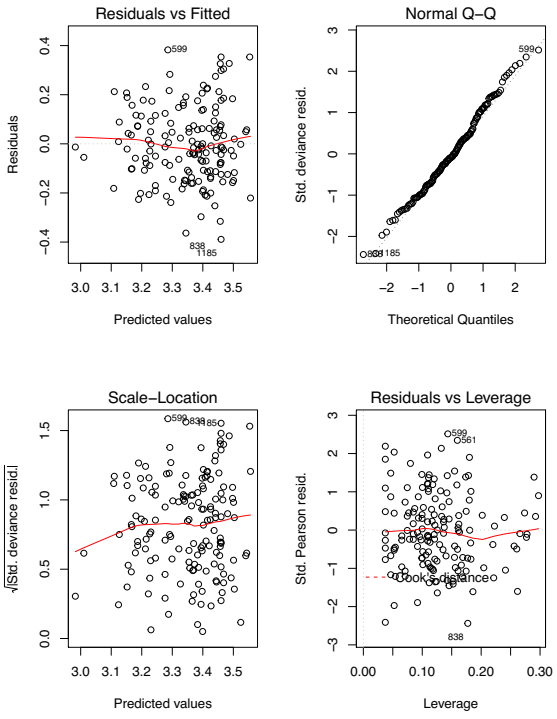

Figure C.17: *Basic diagnostics plot for the model given in Table C.11.*

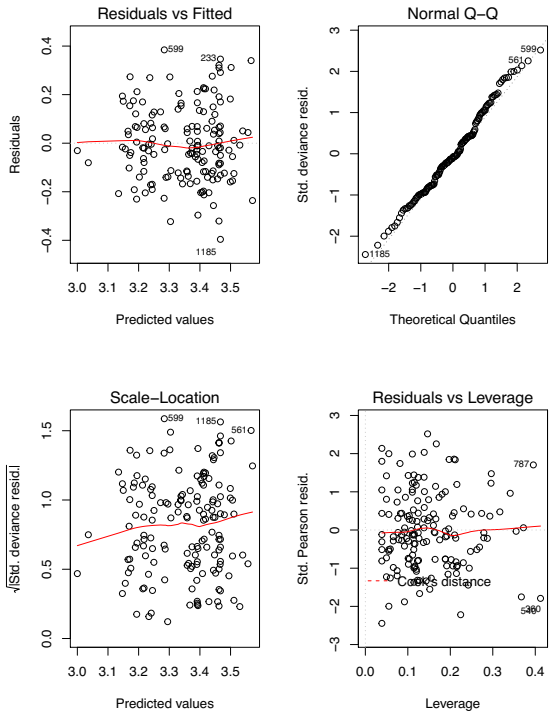

Figure C.18: *Basic diagnostics plot for the saturated model given in Table C.12.*

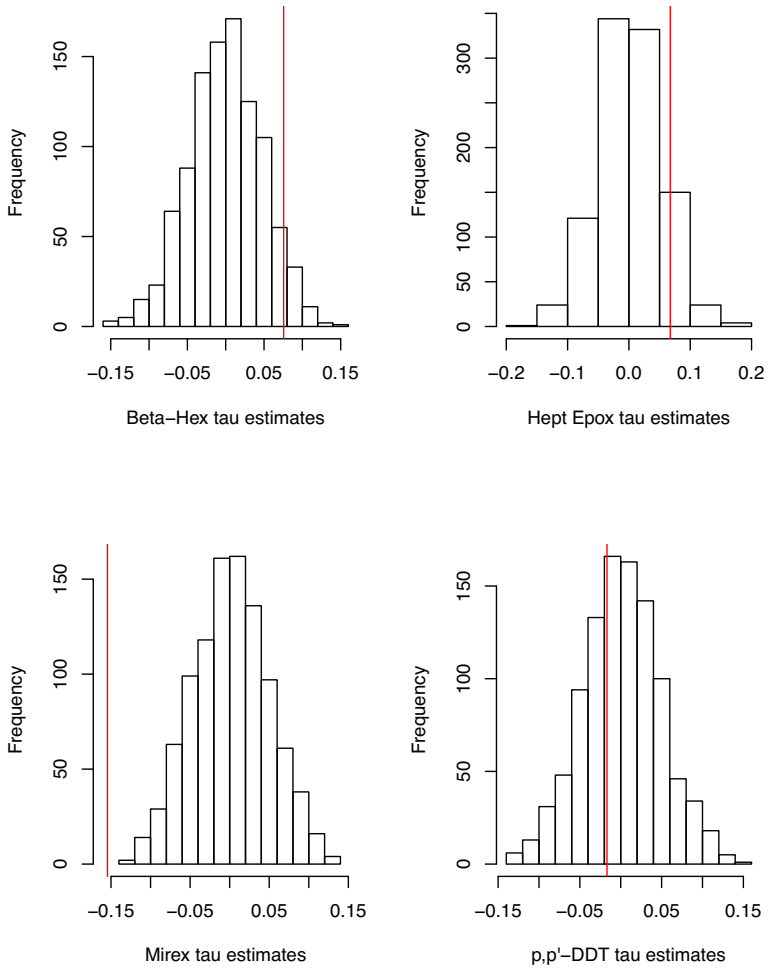

Figure C.19: *Plots of simulated treatment effects estimated. Observed treatment effect estimates plotted in red.*