A Primer on the Analysis of Randomized Experiments and a Survey of some Recent Advances *

Yuehao Bai
Department of Economics
University of Southern California
yuehao.bai@usc.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Max Tabord-Meehan
Department of Economics
University of Chicago
maxtm@uchicago.edu

May 6, 2024

1 Introduction

The past two decades have witnessed a surge of new research in the analysis of randomized experiments. The emergence of this literature may seem surprising given the widespread use and long history of experiments as the "gold standard" in program evaluation, but this body of work has revealed many subtle aspects of randomized experiments that may have been previously unappreciated. This article provides an overview of some of these topics, primarily focused on stratification, regression adjustment, and cluster randomization, although we also provide short discussions on a broad range of other topics at the end of the article.

The majority of our discussion is presented within the context of a framework in which units are sampled from a suitable "super-population"; i.e., we assume that the units are i.i.d. draws from some probability distribution. Importantly, these results differ from results that are derived within a complementary framework in which units are sampled from a fixed, finite population (without replacement). The special case in which all the units from the finite population are sampled is sometimes referred to as *design-based* inference. After defining some common notation in Section 2, we begin our review in Section 3 by providing a comparison of these two sampling frameworks. Specifically, we introduce both frameworks in the context of an analysis of

*The third author acknowledges support from NSF grant SES-2149408.

the difference-in-means estimator for the average treatment effect in a completely randomized experiment. There, we argue that the super-population framework approximates a finite-population framework in which a negligible fraction of the finite population is sampled; we argue further, however, that results derived in a super-population framework may provide useful methods for inference in a finite population even outside of this limiting case.

We then turn our attention in Section 4 to the impact of stratification on subsequent inferences drawn from randomized experiments. More specifically, instead of assigning treatment status according to complete randomization, it is common to stratify first according to some baseline covariates and then assign treatment status within each stratum so as to ensure that the treatment group and control group in the experiment are "balanced" according to these covariates. Examples of such schemes include stratified block randomization and matched pair designs, both of which are commonly used within economics and the sciences more generally: see Rosenberger and Lachin (2015) for a textbook treatment focused on clinical trials and Bruhn and McKenzie (2009) for a review focused on development economics. We first illustrate in Section 4.1 how stratification can improve the precision of the usual difference-in-means estimator of the average treatment effect by reducing what we refer to as the ex-post (as opposed to ex-ante) bias of the estimator. Here, the ex-post bias of the estimator is used to describe its behavior conditional on treatment status rather than unconditionally. Section 4.2 develops the implications of this increase in precision for inferences about the average treatment effect more formally for the special case of stratified block randomization with finitely many "large" strata; Section 4.2.1 describes how this analysis changes when strata are "small," for instance in the case of matched pair designs.

In Section 5 we consider the use of baseline covariates (including ones that are possibly not used in the assignment of treatment status) to further improve the precision of estimators of the average treatment effect through regression adjustment. We explain how naïve regression adjustment may increase rather than decrease the precision of estimators of the average treatment effect, and how a more careful use of the covariates can ensure an improvement in (asymptotic) precision. Our discussion includes linear adjustments (Section 5.1) as well as more general forms of adjustment (Section 5.2).

Finally, Section 6 extends our discussion to cluster randomized experiments, i.e., randomized experiments in which the unit of randomization is a cluster. Such designs are increasingly common in economics. Indeed, Muralidharan and Niehaus (2017) find in a survey of leading economics journals between 2001 and 2016 that more than 75 percent of randomized experiments were cluster randomized. In Section 6.1 we draw a distinction between different ways in which one may define the average effect of the treatment in such settings. In particular, we discuss two possible parameters of interest, which we call the equally-weighted average treatment effect and the size-weighted average treatment effect, that differ in how the average effect within a cluster is aggregated across individuals. In Section 6.2 we discuss inference in cluster randomized experiments.

Owing to space constraints, our discussion is, of course, necessarily incomplete. We therefore provide,

at the end of each section, a guide to some of the related literature on similar topics. For completeness, in Section 7 we also briefly discuss a number of topics that were regrettably omitted from the main discussion: this includes the analysis of treatment effect heterogeneity, re-randomization, multiple testing, imperfect compliance/attrition, experiments with interference, randomization inference, policy learning, and response-adaptive designs. Athey and Imbens (2017) also provide a review of some of these topics (primarily from the design based perspective).

Before proceeding, we emphasize that our survey is limited to the *analysis* of randomized experiments, and as a result we will only briefly comment on some aspects of experimental *design* as well as practical issues surrounding *implementation*, in passing; classical textbook treatments on experimental design are provided in Cox and Reid (2000), Pukelsheim (2006), Atkinson et al. (2007), and Wu and Hamada (2011). Other important contributions to the theory of experimental design (including theoretical justifications for *why* an experimenter may want to randomize) are provided in Savage (1951), Blackwell and Girshick (1954), Kiefer (1959), Li (1983), Kallus (2018, 2021), Section 5.10 in Lehmann and Romano (2022), and Bai (2023). Important references discussing other practical issues (particularly for field experiments conducted in economics) include Duflo et al. (2007), Glennerster and Takavarasha (2013), Karlan and Appel (2017), and List (2023).

2 Setup and Notation: The Potential Outcomes Framework

In this section, we present some notation which will be common to the majority of the article. Each individual i in the experiment is assigned a binary treatment $D_i \in \{0, 1\}$ (we focus on settings with binary treatments, but provide references on related extensions to multiple treatments throughout the article). Let $Y_i(1)$ denote the potential (or counterfactual) outcome for individual i if they are treated, and $Y_i(0)$ denote the potential outcome if they are untreated. Note that we never observe $Y_i(1)$ and $Y_i(0)$ simultaneously for the same individual, but rather we observe the outcome Y_i given by

$$Y_i = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \ . \end{cases}$$

We summarize the previous relationship succinctly by

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) \ .$$

For each individual, we may also observe a vector of baseline covariates denoted by X_i. The experimental sample is thus given by $\{(Y_i, D_i, X_i) : 1 \leq i \leq n\}$. For any variable indexed by i, for example D_i, we denote by $D^{(n)}$ the vector $(D_1, D_2, \ldots, D_n)$. Note that we will always model the experimental assignments $D^{(n)}$ as random; on the other hand, as we will explain below, depending on the sampling framework employed in the

analysis, $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq n\}$ may either be modelled as random vectors or fixed quantities.

Much of our discussion will center on the properties of the standard difference-in-means estimator: let $n_1 = \sum_{1 \leq i \leq n} D_i$ denote the number of treated units in the sample and $n_0 = n - n_1$ denote the number of control units. The difference-in-means estimator is then given by

$$\hat{\Delta}_n = \frac{1}{n_1} \sum_{1 \leq i \leq n} Y_i D_i - \frac{1}{n_0} \sum_{1 \leq i \leq n} Y_i (1 - D_i) \ .$$

Note that this estimator can equivalently be described as the estimator of the coefficient on D_i when estimating the following linear regression by least squares:

$$\texttt{regress } Y_i \texttt{ on } \text{constant} + D_i \ . \tag{1}$$

We will illustrate many of the basic concepts via the analysis of a *completely randomized* experiment. In a completely randomized experiment, the treatment assignment $D^{(n)}$ is implemented such that, for some fixed fraction of units, say π, exactly $n_1 = \lfloor \pi n \rfloor$ units are assigned to treatment and $n_0 = n - n_1$ units are assigned to control, with all such possible assignments being equally likely. Formally, let $(d_1, \ldots, d_n) \in \{0,1\}^n$, then $D^{(n)}$ is independent of $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq n\}$ with distribution given by

$$P\{D^{(n)} = (d_1, \ldots, d_n)\} = \begin{cases} \binom{n}{n_1}^{-1} & \text{if } \sum_{1 \leq i \leq n} d_i = n_1 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

We emphasize that an important feature of our discussion here and throughout the rest of the article is that we do not assume that $D^{(n)}$ are independently distributed; this will play a crucial role in our subsequent analyses when studying, for example, stratified randomization in Section 4.

3 What is Random? Finite versus Super-population Analyses of a Completely Randomized Experiment

In this section, we introduce the two main paradigms for the analysis of randomized experiments: the finite-population and super-population approaches to inference. In a finite-population analysis, we begin with a collection $\{(y_j(1), y_j(0), x_j) : 1 \leq j \leq N\}$ of *fixed* quantities that constitute the entire population of interest. A sample of size $n \leq N$, given by $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq n\}$, is then drawn *without* replacement from the population, and the experiment is performed on these n individuals. The most common case considered in the literature is when $n = N$, so that the experiment is performed on the entire population. We will refer to this case as *design-based* inference, since the only source of uncertainty arises from the randomness in $D^{(n)}$. This perspective is often considered attractive in settings where it is difficult to conceptualize an

appropriate sampling frame; Reichardt and Gollob (1999) provide further discussion.

In contrast, in a super-population analysis, the sample $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq n\}$ is modeled as being i.i.d. according to some probability distribution. Informally, we may view the distribution from which the potential outcomes are drawn as summarizing an essentially infinite "super-population."

To compare and contrast the super- and finite population paradigms, we present an analysis of the difference-in-means estimator $\hat{\Delta}_n$ for the average treatment effect under both approaches, following the original work of Neyman (1923), in a completely randomized experiment (as defined in Section 2). Our primary takeaway will be that, in a sense to be made formal below, we can view the super-population paradigm as an approximation to the finite-population paradigm in a regime where the sample size n is a vanishing fraction of the total population size N. Consequently, we will show that in the super-population framework it is possible to construct *consistent* variance estimators of $\text{Var}[\hat{\Delta}_n]$, whereas this will often be impossible in the finite-population framework. Instead, in the finite population framework we will explain how to construct *conservative* variance estimators of $\text{Var}[\hat{\Delta}_n]$. Moreover, we will argue that consistent variance estimators derived via a super-population analysis are often reasonable conservative variance estimators when viewed through the lens of a finite-population analysis.

3.1 Finite population analysis of $\hat{\Delta}_n$

First we consider a finite-population analysis. Recall that in this case we begin with $\{(y_j(1), y_j(0), x_j) : 1 \leq j \leq N\}$ which are *fixed*, non-random quantities that describe the outcomes (and covariates) of the *entire* population of N units. Accordingly, in this case, the average treatment effect is defined as

$$\Delta_N^{\text{fp}} = \frac{1}{N} \sum_{1 \leq j \leq N} (y_j(1) - y_j(0)) \ .$$

We emphasize that in this framework the parameter of interest Δ_N^{fp} is determined entirely by the values of the potential outcomes from the N units in the population.

To analyze $\hat{\Delta}_n$ from a finite population perspective, it is often useful to re-frame the problem as a problem of survey sampling from a finite population; re-framing the problem in this way allows us to employ classical results from survey sampling (see, for instance, Cochran, 1977; Lehmann and Romano, 2022). In Appendix B.1 we illustrate how this re-framing can be useful in deriving some "finite-N" properties of $\hat{\Delta}_n$ (i.e., properties that hold for every finite population size N) in a completely randomized experiment. In particular, there we show

$$E[\hat{\Delta}_n] = \Delta_N^{\text{fp}} \ ,$$

i.e., that $\hat{\Delta}_n$ is an unbiased estimator for Δ_N^{fp}, and

$$\text{Var}[\hat{\Delta}_n] = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\Delta^2}{N} \ , \tag{2}$$

where

$$S_d^2 = \frac{1}{N-1} \sum_{1 \leq j \leq N} \left(y_j(d) - \bar{y}_N(d)\right)^2$$

$$S_\Delta^2 = \frac{1}{N-1} \sum_{1 \leq j \leq N} \left(y_j(1) - y_j(0) - \Delta_N^{\text{fp}}\right)^2 \ ,$$

and $\bar{y}_N(d) = \frac{1}{N} \sum_{1 \leq j \leq N} y_N(d)$.

This expression for the variance of $\hat{\Delta}_n$ is fundamental to understanding why inference on Δ_N^{fp} is generally conservative in the finite population paradigm: note that S_d^2, $d \in \{0,1\}$ are simply the variances for the potential outcomes in the finite population, and we explain below how to construct estimators of these quantities. To rule out degenerate situations, we henceforth assume that at least one of S_d^2, $d \in \{0,1\}$ are nonzero. The quantity S_Δ^2, however, is the variance of the unit-level treatment effects $(y_j(1) - y_j(0) : 1 \leq j \leq N)$ in the finite population, and these are by definition *never* observed for a given unit. In settings where the experiment size n is a non-trivial fraction of the population size N (for instance, in design-based inference when $n = N$), this feature of the variance introduces an unavoidable roadblock for estimating $\text{Var}[\hat{\Delta}_n]$ consistently,[1] and thus estimators of $\text{Var}[\hat{\Delta}_n]$ will necessarily be conservative unless treatment effects are constant (i.e., that $y_j(1) - y_j(0) = \Delta_N^{\text{fp}}$ for every $1 \leq j \leq N$). In cases where n is a vanishing fraction of N, however, we see that S_Δ^2/N becomes a negligible component of $\text{Var}[\hat{\Delta}_n]$ as N gets large, in the sense that in order to estimate $\text{Var}[\hat{\Delta}_n]$ consistently it is sufficient to estimate only its first two terms consistently. As we show in Section 3.2, this feature of $\text{Var}[\hat{\Delta}_n]$ when n is a vanishing fraction of N will exactly mirror our findings in the super-population analysis.

Before discussing estimation of $\text{Var}[\hat{\Delta}_n]$ and related methods of inference for Δ_N^{fp}, we document some necessary "large-N" properties of $\hat{\Delta}_n$ in the finite population paradigm. In the finite population framework we conceptualize our asymptotic approximations by imagining a sequence of increasingly larger populations on which we perform our experiment. As a consequence, our large-N results will require some discipline on how this sequence of ever larger populations evolve as a function of N. Under appropriate assumptions on the sequence of populations, it can be shown that (see, e.g., Lehmann and Romano, 2022, Theorem 12.2.5)

$$\frac{\hat{\Delta}_n - \Delta_N^{\text{fp}}}{\sqrt{\text{Var}[\hat{\Delta}_n]}} \xrightarrow{d} N(0,1) \ ,$$

as $n \to \infty$ and $N \to \infty$. Using this result, asymptotic inference on Δ_N^{fp} is straightforward once we have an estimator of $\text{Var}[\hat{\Delta}_n]$. Motivated by our decomposition in (2), a conservative variance estimator can be constructed by simply estimating the following upper bound on $\text{Var}[\hat{\Delta}_n]$:

$$V_n^{\text{obs}} = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} \ ,$$

[1] Here and throughout the rest of Section 3.1, consistency should be understood as saying that the *ratio* of the estimator to the variance converges to one in probability.

which implicitly sets S^2_Δ to its lowest possible value of zero. From this expression we see that a consistent estimator of V^{obs}_n can be obtained by replacing S^2_1 and S^2_0 by their natural estimators. Equivalently, viewing $\hat{\Delta}_n$ as the estimator of the coefficient on D_i obtained by the regression in (1), it can be shown that a consistent estimator of V^{obs}_n can simply be obtained from the resulting heteroskedasticity-robust variance estimator (see, e.g., Angrist and Pischke, 2009, Chapter 8). An asymptotically valid 95%-confidence interval for Δ^{fp}_N can therefore be constructed as

$$C_n = \left[\hat{\Delta}_n - 1.96 \cdot \text{SE}(\hat{\Delta}_n), \hat{\Delta}_n + 1.96 \cdot \text{SE}(\hat{\Delta}_n)\right] \ , \tag{3}$$

where $\text{SE}(\hat{\Delta}_n)$ is the robust standard error of the coefficient on D_i obtained from the regression in (1). Although C_n is a valid confidence interval, we emphasize that it is conservative in the sense that

$$P\{\Delta^{\text{fp}}_N \in C_n\} \to p \geq 0.95 \ ,$$

as $n, N \to \infty$, with equality only if $n/N \to 0$ or $S^2_\Delta = 0$.

We conclude this section by considering the following natural follow-up question: could we develop more precise methods of inference by constructing less conservative variance estimators of $\text{Var}[\hat{\Delta}_n]$? For the example we just presented, this would amount to considering tighter lower bounds for S^2_Δ which are themselves consistently estimable. For instance, it follows from the Cauchy-Schwarz inequality that

$$\sum_{1 \leq j \leq N} (y_j(1) - \bar{y}_N(1))(y_j(0) - \bar{y}_N(0)) \leq \left(\sum_{1 \leq j \leq N} (y_j(1) - \bar{y}_N(1))^2 \right)^{1/2} \left(\sum_{1 \leq j \leq N} (y_j(0) - \bar{y}_N(0))^2 \right)^{1/2} ,$$

which we can use to immediately verify the lower bound $S^2_\Delta \geq (S_1 - S_0)^2 \geq 0$. We thus obtain the following improved upper bound on $\text{Var}[\hat{\Delta}_n]$:

$$\frac{S^2_1}{n_1} + \frac{S^2_0}{n_0} - \frac{(S_1 - S_0)^2}{N} \ ,$$

which can be used to construct a less conservative estimator of $\text{Var}[\hat{\Delta}_n]$. In fact, it is possible to achieve even tighter upper bounds by further exploiting the structure of S^2_Δ: see Aronow et al. (2014) for details. We emphasize, however, that although we will argue in Section 3.2 that inferences based on the heteroskedasticity-robust standard error $\text{SE}(\hat{\Delta}_n)$ as in (3) will be valid under either the super-population or finite population paradigms, standard errors based on such improved upper bounds will generally be too small, and thus *invalid* when viewed from a super-population perspective.

3.2 Super-population analysis of $\hat{\Delta}_n$

Next, we repeat the above analysis within the super-population paradigm. Recall that in this case the sample $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq n\}$ is modeled as being i.i.d. according to some probability distribution.

Accordingly, in this case the average treatment effect is defined as

$$\Delta = E[Y_i(1) - Y_i(0)] \ ,$$

where the expectation is with respect to the distribution of the data.

We begin by documenting some finite-sample properties of the estimator $\hat{\Delta}_n$ in a completely randomized experiment under the super-population framework, analogous to the properties derived in Section 3.1. Using familiar properties of conditional expectations, we show in the appendix that

$$E[\hat{\Delta}_n] = E[Y_i(1) - Y_i(0)] = \Delta \ . \tag{4}$$

We thus obtain that the estimator $\hat{\Delta}_n$ is also an unbiased estimator of average treatment effect in the super-population paradigm. Following a similar line of reasoning using the properties of conditional variances, we show in the appendix that

$$\text{Var}[\hat{\Delta}_n] = \frac{\text{Var}[Y_i(1)]}{n_1} + \frac{\text{Var}[Y_i(0)]}{n_0} \ . \tag{5}$$

It is instructive to compare the limits of the variance expressions in (2) and (5). To ensure nondegenerate limits, it is useful to scale the variances by n. By doing so, we see that the limit of the super-population variance mirrors the limit of the finite-population variance in a regime where we sample a vanishing fraction of the total population, i.e., $n/N \to 0$.

Since the variance of the unit-level treatment effects does not appear in our expression for $\text{Var}[\hat{\Delta}_n]$ in (5), consistent variance estimation will be feasible in the super-population paradigm.

To discuss inference on Δ, we document some necessary large-sample properties of $\hat{\Delta}_n$ in the super-population framework. Under the assumption that $E[Y_i^2(d)] < \infty$, it can be shown that

$$\sqrt{n}(\hat{\Delta}_n - \Delta) \xrightarrow{d} N(0, V^{\text{cr}}) \ ,$$

as $n \to \infty$, where

$$V^{\text{cr}} = \frac{\text{Var}[Y_i(1)]}{\pi} + \frac{\text{Var}[Y_i(0)]}{1-\pi} \ , \tag{6}$$

(see, for instance, Bugni et al., 2018). Using the above result, asymptotic inference on Δ is straightforward once we have an estimator of V^{cr}. As in our discussion in Section 3.1, viewing $\hat{\Delta}_n$ as the estimator of the coefficient on D_i obtained by the regression in (1), it can be shown that a consistent estimator of V^{cr} can be obtained from the resulting heteroskedasticity-robust variance estimator. An asymptotically valid 95%-confidence interval for Δ is therefore once again given by C_n in (3). Moreover, since this variance estimator is consistent, C_n has *exact* asymptotic coverage, that is

$$P\{\Delta \in C_n\} \to 0.95 \ ,$$

as $n \to \infty$.

The above discussion illustrates an important feature of super-population and finite population analyses of randomized experiments: methods of inference developed in a super-population framework typically *immediately* deliver conservative methods of inference from a finite population perspective. While the preceding discussion was limited to completely randomized experiments, in Appendix A.1, we show that this feature holds much more generally under appropriate assumptions. With this in mind, the remainder of this article will focus on illustrating some of the major themes in the analysis of randomized experiments from the super-population perspective, but we will comment on any other important differences between these two perspectives whenever they arise.

3.3 Further Reading

Most of the material in this section is by-now standard, and several textbook treatments exist at various levels of formality: see in particular Imbens and Rubin (2015), Athey and Imbens (2017), and Lehmann and Romano (2022). Li and Ding (2017) provide formal statements and proofs of finite population central limit theorems. Harshaw et al. (2021) study a general method for constructing less conservative variance estimators in the design-based paradigm.

4 Stratified Randomized Experiments

In this section, we outline some benefits of *stratification* and its consequences on subsequent experimental analyses. In a stratified randomized experiment, individuals are first divided into groups (i.e., strata) sharing similar values of their baseline covariates and then assigned to treatment so as to achieve "balance" across the treatment and control groups: often, this amounts to simply performing complete randomization within each stratum. A primary motivation for stratification, going back to the work of Fisher (1935), is to ensure that the treatment and control groups are similar *in the sample*, in contrast to complete randomization which can only ensure that this will be true *in expectation*. As we will show, this property of stratification can lead to an increase in precision of the difference-in-means estimator $\hat{\Delta}_n$ relative to complete randomization, and as a result subsequent inferences will be unnecessarily conservative unless this is taken into consideration. In Section 4.1 we illustrate the benefits of stratification via a simple example. In Section 4.2 we discuss inference in stratified randomized experiments, including inference with "small" strata.

4.1 Some Benefits of Stratification

We begin by illustrating the benefits of stratification via a simple example. Suppose we have an experimental sample $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq n\}$, where for now we assume $X_i \in \{0, 1\}$ are binary variables, and for convenience we assume n is even. Consider the following two treatment assignment mechanisms for $D^{(n)}$:

1. Complete Randomization (CR): Treatment is completely randomized with $\pi = \frac{1}{2}$ (see Section 2 for a formal definition).

2. Stratified Block Randomization (SBR): Independently for each stratum (the sub-samples with $X_i = 0$ and $X_i = 1$), treatment is completely randomized with $\pi = \frac{1}{2}$.

Note that (SBR) as defined above uses the assignment proportion $\pi = \frac{1}{2}$ in both strata. Although maintaining a constant assignment proportion across all strata is the most common approach in practice, in principle we could also consider using different assignment proportions in each stratum, and we comment on the implications of doing so on subsequent analyses at the end of Section 4.2. Although both assignment mechanisms (CR) and (SBR) assign exactly $n/2$ units to treatment, (CR) does not enforce that the *composition* of the treated group across $X_i \in \{0, 1\}$ matches the composition in the total sample. For instance, suppose $n = 100$, where 40 units have $X_i = 0$ and 60 units have $X_i = 1$. Figure 1 depicts one possible assignment that could result from employing (CR) versus (SBR). Although (CR) guarantees that exactly 50 units are assigned to treatment, in this realization of the assignment, units with $X = 0$ are over-represented in the treatment group. In contrast, (SBR) reproduces the composition in the overall sample in both the treatment and control groups.

To formalize this intuition, we analyze two different notions of bias for the difference-in-means estimator $\hat{\Delta}_n$ under both designs. To simplify the exposition, in this section, we perform our analyses conditional on the observable characteristics $X^{(n)}$ (although we emphasize that this does not materially change the conclusions). To that end, we will temporarily switch our parameter of interest to

$$\Delta_n(X^{(n)}) = \frac{1}{n} \sum_{1 \leq i \leq n} E[Y_i(1) - Y_i(0)|X_i] \ ,$$

which is the average effect of the treatment conditional on the covariates in the sample.

For the estimator $\hat{\Delta}_n$, we define the ex-ante bias as

$$\text{Bias}^{\text{ante}}(X^{(n)}) = E[\hat{\Delta}_n|X^{(n)}] - \Delta_n(X^{(n)}) \ ,$$

which averages over all possible realizations of the treatment assignments. This ex-ante bias measures the average bias obtained by repeatedly running the experiment on the same realization of the covariates. In contrast, we define the ex-post bias as the bias *conditional on the realized treatment assignments*:

$$\text{Bias}^{\text{post}}(X^{(n)}, D^{(n)}) = E[\hat{\Delta}_n|X^{(n)}, D^{(n)}] - \Delta_n(X^{(n)}) \ .$$

First we compare the ex-ante biases generated by (CR) and (SBR). Note that the marginal treatment probability of each unit satisfies $E[D_i|X^{(n)}] = \frac{1}{2}$ under both designs, i.e., the conditional probability that any given unit is assigned to treatment is one half. Combining this fact with the (conditional) exogeneity of

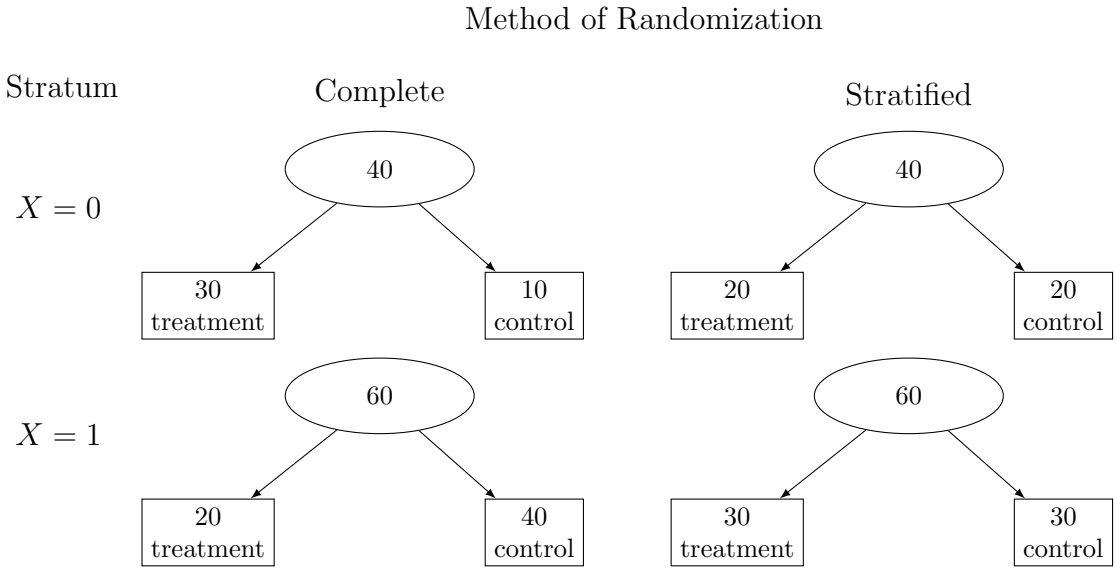

Figure 1: Assignment across treatment and control for one realization of (CR) vs (SBR)

treatment assignment, we obtain that

$$E[\hat{\Delta}_n | X^{(n)}] = \Delta_n(X^{(n)}) \ ,$$

so that $\text{Bias}^{\text{ante}}(X^{(n)}) = 0$ for both designs. To compare the ex-post biases generated by (CR) and (SBR), consider the following decomposition:

$$\begin{aligned} \text{Bias}^{\text{post}}(X^{(n)}, D^{(n)}) &= \frac{1}{n} \sum_{1 \leq i \leq n} (2D_i - 1) E[Y_i(1) + Y_i(0) | X_i] \\ &= \frac{1}{n} E[Y_i(1) + Y_i(0) | X_i = 1] \cdot \text{Imb}(1) + \frac{1}{n} E[Y_i(1) + Y_i(0) | X_i = 0] \cdot \text{Imb}(0) \ , \end{aligned}$$

where

$$\text{Imb}(x) = \#\{\text{treated units with } X_i = x\} - \#\{\text{untreated units with } X_i = x\}$$

is a measure of the *imbalance* of treatment status for each possible value of the covariate X_i. By construction, (SBR) enforces that for *any realization* of $D^{(n)}$ it is the case that $\text{Imb}(x) = 0$. In contrast, as depicted in Figure 1, this is not the case for (CR). As a consequence, it follows immediately that under (SBR), $\text{Bias}^{\text{post}}(X^{(n)}, D^{(n)}) \equiv 0$, whereas the ex-post bias under (CR) is not guaranteed to be identically zero unless X_i is an irrelevant stratification variable in the sense that $E[Y_i(1) + Y_i(0) | X_i = 1] = E[Y_i(1) + Y_i(0) | X_i = 0]$.

Next, we show that these properties of the ex-post bias have direct implications for the (ex-ante) *variance* of $\hat{\Delta}_n$ under (CR) and (SBR). By the law of total variance:

$$\text{Var}[\hat{\Delta}_n | X^{(n)}] = E[\text{Var}[\hat{\Delta}_n | X^{(n)}, D^{(n)}] | X^{(n)}] + \text{Var}[E[\hat{\Delta}_n | X^{(n)}, D^{(n)}] | X^{(n)}] \ .$$

We show in the appendix that

$$E[\text{Var}[\hat{\Delta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] = \frac{2}{n^2} \sum_{1 \leq i \leq n} (\text{Var}[Y_i(1)|X_i] + \text{Var}[Y_i(0)|X_i]) \ . \tag{7}$$

As a result, $E[\text{Var}[\hat{\Delta}_n|X^{(n)}, D^{(n)}]|X^{(n)}]$ doesn't depend on the experimental design, and thus to compare $\text{Var}[\hat{\Delta}_n|X^{(n)}]$ under (CR) and (SBR) it suffices to study

$$\text{Var}[E[\hat{\Delta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] = E[\text{Bias}^{\text{post}}(X^{(n)}, D^{(n)})^2|X^{(n)}] \ .$$

In words, comparing $\text{Var}[\hat{\Delta}_n|X^{(n)}]$ under (CR) versus (SBR) amounts to a comparison of the second moment of the ex-post bias under both designs; because the ex-post bias is always zero under (SBR), the variance of $\hat{\Delta}_n$ is always smaller under (SBR) than (CR). In the next section, we discuss the implications of this increase in precision on subsequent inferences.

4.2 Inference in Stratified Experiments

In this section, we discuss the implications of stratification for inference on Δ. For now, suppose that X_i takes a finite number of values in $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$; this could be either because X_i is naturally discrete (for instance, if X_i denotes whether or not an individual graduated college), or because the researcher has discretized some continuous variables (for example, binning students by high or low test scores). In Section 4.2.1 we consider settings where units are stratified as finely as possible based on potentially continuous covariates. To fix ideas, suppose the experiment was performed using stratified block randomization as defined in Section 4.1: independently in each sub-sample $X_i = x$, treatment is completely randomized with $\pi \in (0, 1)$. Under appropriate assumptions, it can be shown that

$$\sqrt{n}(\hat{\Delta}_n - \Delta) \xrightarrow{d} N(0, V^{\text{sbr}}) \ ,$$

where

$$V^{\text{sbr}} = \frac{\text{Var}[Y_i(1)]}{\pi} + \frac{\text{Var}[Y_i(0)]}{1-\pi} - \pi(1-\pi)\,\text{Var}\left[E\left[\frac{Y_i(1)}{\pi} + \frac{Y_i(0)}{1-\pi}\middle|X_i\right]\right] \ ; \tag{8}$$

see, for instance, Bugni et al. (2018). Comparing the variance V^{cr} obtained from complete randomization in (6) to V^{sbr}, we see that $V^{\text{sbr}} \leq V^{\text{cr}}$ with equality only if X is an irrelevant stratification variable in the sense that $E\left[\frac{Y_i(1)}{\pi} + \frac{Y_i(0)}{1-\pi}\middle|X_i\right]$ is constant; note that this is exactly the condition that guaranteed that the ex-post bias of $\hat{\Delta}_n$ under complete randomization with $\pi = 1/2$ was zero in Section 4.1.

Recall that we argued in Section 3.2 that the heteroskedasticity-robust variance estimator from the regression described in (1) is consistent for V^{cr} under complete randomization. It can be shown that the same holds under stratified block randomization, from which it follows that the robust variance estimator is generally *conservative* for V^{sbr}. In words, the robust variance estimator does not properly account for the

gain in precision obtained by having performed stratification. How then could we conduct non-conservative inference on Δ in the presence of stratification? A straightforward option is to modify the variance estimator of V^{sbr} to account for the gain in precision. When the stratification variable X takes only a finite number of values and there are at least two units in each of treatment and control, a simple solution is as follows: first, using the law of total variance, we re-write V^{sbr} as

$$V^{\text{sbr}} = E\left[\frac{\text{Var}[Y_i(1)|X_i]}{\pi}\right] + E\left[\frac{\text{Var}[Y_i(0)|X_i]}{1-\pi}\right] + \text{Var}[E[Y_i(1) - Y_i(0)|X_i]] \ .$$

Exploiting the discreteness in X, we can expand this as

$$V^{\text{sbr}} = \sum_{x \in \mathcal{X}} p(x) \left(\frac{\text{Var}[Y_i(1)|X_i = x]}{\pi} + \frac{\text{Var}[Y_i(0)|X_i = x]}{1-\pi}\right) + \sum_{x \in \mathcal{X}} p(x) \left(E[Y_i(1) - Y_i(0)|X_i = x] - \Delta\right)^2 \ ,$$

where $p(x) = P\{X_i = x\}$. From this expression it is clear how to construct a consistent estimator of V^{sbr} by simply replacing all of the unknown means and variances by their sample counterparts.

Let us briefly comment on the implications of the above discussion for design-based inference. Although we argued at the end of Section 3.2 that, in general, consistent estimators for the super-population variance immediately provide conservative estimators in the finite-population framework, it is worth mentioning that in this case the estimator is potentially "excessively" conservative: indeed, from the design-based perspective, it can be shown that a less conservative estimator is given by any consistent estimator of

$$\sum_{x \in \mathcal{X}} \frac{n(x)}{n} \left(\frac{S_1^2(x)}{\pi} + \frac{S_0^2(x)}{1-\pi}\right) \ , \tag{9}$$

where $n(x)$ denotes the number of observations in stratum x and $S_d^2(x)$ for $d \in \{0, 1\}$ are the population variances of the potential outcomes in stratum x (see, for instance, Imbens and Rubin, 2015, for details). Note that this expression mimics only the first component of V^{sbr}; this is because the second component arises from the random fluctuations in $n(x)/n$ versus $p(x)$, but these exactly coincide in a design-based framework. However, because of this, consistent estimators of (9) are *not* guaranteed to be valid if we view the sample as being drawn from a larger (finite or super-) population.

Finally, we conclude with a short discussion on settings in which the assignment proportions differ across strata. This could arise, for instance, if the cost of treatment varies across different strata. We continue to assume that treatment is assigned using stratified block randomization but we let $\pi(x)$, the fraction of units that are treated in stratum $x \in \{1, 2, \ldots, \mathcal{X}\}$, differ across strata. Our first observation is that the difference-in-means estimator is generally no longer consistent for Δ; intuitively, the experimental design induces selection bias with respect to the stratification variable. Instead, we could consider the following estimator which computes a weighted average of the stratum-specific difference-in-means estimators, with

weights determined by the strata sizes:

$$\hat{\Delta}_n^{\text{sat}} = \sum_{x \in \mathcal{X}} \frac{n(x)}{n} \hat{\Delta}_n(x) \ , \tag{10}$$

where $n(x)$ denotes the number of observations in stratum x and $\hat{\Delta}_n(x)$ denotes the difference-in-means estimator computed in stratum x. This estimator is consistent, and moreover it can be shown that its asymptotic variance is given by

$$\sum_{x \in \mathcal{X}} p(x) \left(\frac{\text{Var}[Y_i(1)|X_i = x]}{\pi(x)} + \frac{\text{Var}[Y_i(0)|X_i = x]}{1 - \pi(x)} \right) + \sum_{x \in \mathcal{X}} p(x) \left(E[Y_i(1) - Y_i(0)|X_i = x] - \Delta \right)^2 \ ,$$

for which a consistent estimator can once again be constructed by taking sample analogs; see Bugni et al. (2019) for details.

4.2.1 Variance Estimation with Small Strata

When X was discrete, a natural estimator of V^{sbr} could be constructed by computing sample analogs of the means and variances of the potential outcomes at the stratum level. If, however, X_i contains continuous components and the experiment is "finely stratified" in such a way that there is only one treated or control observation per stratum, then this logic breaks down. A leading example of when this occurs is with pairwise matching: in a matched pairs experiment, units are paired together based on their observed covariate values and then treatment is assigned such that, in each pair, one unit is selected at random to receive treatment and the other control. With only one observation per stratum with a given treatment, we cannot estimate the stratum-level variances by simply taking sample analogs as proposed in the previous section. As discussed in Klar and Donner (1997), this added challenge to variance estimation has often been perceived as a fundamental analytical limitation of matched-pair designs. In this section, we briefly illustrate how the variance of $\hat{\Delta}_n$ can still be consistently estimated by using a "collapsed-strata" estimator in the spirit of Hansen et al. (1953).

To streamline the exposition, in what follows we focus only on the setting where $\pi = \frac{1}{2}$ and units are matched into pairs. First, note that it can be shown under appropriate assumptions that, even with small strata, the limiting variance of $\hat{\Delta}_n$ is still given by (8) (see, for instance, Bai et al., 2022). The unconditional variances $\text{Var}[Y_i(1)]$ and $\text{Var}[Y_i(0)]$ in (8) are consistently estimable using their sample counterparts. We thus focus on estimating the last term on the right-hand side of (8), which in this case is $(1/2)\,\text{Var}[E[Y_i(1) + Y_i(0)|X_i]]$. To that end, note that by elementary properties of the variance and the law of iterated expectations,

$$\text{Var}[E[Y_i(1) + Y_i(0)|X_i]] = E[E[Y_i(1) + Y_i(0)|X_i]^2] - E[Y_i(1) + Y_i(0)]^2 \ .$$

The second term on the right-hand side is again unconditional and thus easy to estimate using a sample

analog. Therefore, it suffices to consistently estimate

$$E[E[Y_i(1) + Y_i(0)|X_i]^2] \ . \tag{11}$$

Intuitively, to estimate this last quantity we would want independent variation in $Y_i(1)$ and $Y_i(0)$ for each given value of X_i. Let us suppose for a moment that we did in fact have two pairs of units $\{Y_1, Y_2\}$ and $\{Y_3, Y_4\}$ sharing the same value of X. In each pair, there is exactly one treated and control observation per stratum. It follows that that $Y_1 + Y_2$ and $Y_3 + Y_4$ both share the same conditional mean $E[Y(1) + Y(0)|X]$. Moreover, across pairs, the outcomes are (conditionally) independent. As a consequence, we might conjecture that $(Y_1 + Y_2)(Y_3 + Y_4)$ could be used to construct an estimator for (11). Of course in practice, this scenario is not realistic, as units which are paired together will typically not share the same value of X. If, however, matching is performed so that we expect pairs and "adjacent" pairs to have similar characteristics, then this intuition can be formalized to construct a consistent variance estimator for (11), and thus it is possible to construct a consistent estimator of V^{sbr} even with "small" strata; for a formal exposition of this idea, see Bai et al. (2022, 2024b).

4.3 Further Reading

The exposition in Section 4.1 is most closely inspired by Bai (2022), who establishes the finite-sample optimality of certain matched-pair designs. Much of the discussion in Section 4.2 comes from Bugni et al. (2018), who study inference for covariate-adaptive randomization when the treated fraction is constant across strata; Bugni et al. (2019) extends the analysis to settings with multiple treatments and where the treatment proportions are allowed to differ across strata. Bai et al. (2022) studies inference for matched pair designs and develops the estimation strategy discussed in Section 4.2.1. Cytrynbaum (2023b) generalizes these estimation and inference procedures beyond pair-matching and jointly analyzes the stratification problem combined with the problem of selecting a representative sample based on covariates. Bai et al. (2024b) generalizes Bai et al. (2022) to settings with multiple treatments. Pashley and Miratrix (2021) provide an overview of design-based analyses of stratified experiments.

5 Regression Adjustment in Randomized Experiments

In this section, we consider the role of regression adjustment using baseline covariates in the analysis of randomized experiments. A primary motivation for regression adjustment is that it hopefully improves estimation precision in settings where the covariates are correlated with the experimental outcome. This practice has, however, often come under scrutiny. An influential paper of Freedman (2008), for instance, points out that standard linear regression adjustment only guarantees a gain in precision under strong assumptions, and concludes that in general "[...] randomization does not justify the assumptions behind

the OLS model." In Section 5.1 we review part of Freedman's critique, present a resolution popularized by Lin (2013), and briefly discuss some implications for inference. In Section 5.2 we explain how Lin's resolution is a special case of an estimation procedure based on the doubly-robust moment condition of the average treatment effect, and show how this perspective leads to more general regression adjustment strategies beyond linear adjustment.

5.1 Linear Regression Adjustment in a Completely Randomized Experiment

We revisit the setting of a completely randomized experiment with $\pi \in (0,1)$. In particular, we assume that treatment assignment is independent of the outcomes *and* the baseline covariates; we briefly comment on settings where the treatment assignment may itself depend on the baseline covariates (for instance, in a stratified randomized experiment) in Section 5.1.1 and provide further references in Section 5.3. To begin, recall from (1) that the unadjusted difference-in-means estimator can be described as the OLS estimator of the coefficient on D_i in a linear regression of Y_i on a constant and D_i. A natural starting point for regression adjustment is then to instead consider the OLS estimator of the coefficient on D_i in a linear regression of Y_i on a constant, D_i, and the covariates X_i:

$$\texttt{regress } Y_i \texttt{ on } \text{constant} + D_i + X_i \ . \tag{12}$$

Let $\tilde{\beta}_n$ be the resulting estimator of the coefficient on D_i from the above regression, and let $\tilde{\gamma}_n$ be the resulting estimator of the coefficient on X_i. We emphasize here that we do not view the regression in (12) as describing the true data generating process for the outcomes Y_i, but treat this simply as the description of an estimation procedure. Since the experimental assignment guarantees that D and X are independent, it is not surprising given standard regression logic that $\tilde{\beta}_n$ remains consistent for the average treatment effect Δ. One might expect further that $\tilde{\beta}_n$ is more efficient than the simple difference-in-means if X_i is correlated with the experimental outcomes. It can be shown, however, that $\tilde{\beta}_n$ is asymptotically normal with variance given by (see, for instance, Negi and Wooldridge, 2021; Ma et al., 2022)

$$V^{\text{pool}} = V^{\text{cr}} - \frac{1}{\pi(1-\pi)}\gamma'\Sigma_X\gamma + \frac{2(2\pi-1)}{\pi(1-\pi)}\gamma'\Sigma_X(\gamma(1)-\gamma(0)) \ ,$$

where γ denotes the probability limit of $\tilde{\gamma}_n$, Σ_X denotes the covariance matrix of X_i, and $\gamma(d)$ is the probability limit of the coefficient on X_i in a linear regression of the *potential outcome* $Y_i(d)$ on a constant and X_i. Note that the second term in this expression is always weakly negative, and is strictly negative whenever the covariates are correlated with the outcomes in the sense that $\gamma'\Sigma_X\gamma > 0$. However, the sign of the third term in the variance is ambiguous, and as a result, there is no guarantee that V^{pool} is smaller than V^{cr} in general. Note that the third term of V^{pool} is zero in the special cases where either $\pi = 1/2$ or $\gamma(1) = \gamma(0)$, so that in these cases we can in fact conclude that $V^{\text{pool}} \leq V^{\text{cr}}$. In words, regression adjustment based on the regression in (12) is not guaranteed to improve precision in general, but (weakly)

improves precision in the special cases where assignment is equal between treatment and control, or treatment effects are sufficiently homogeneous, e.g., if treatment effects are constant such that $Y_i(1) - Y_i(0) = \Delta$.

The difficulty with estimation based on (12) is that the regression pools together the observations under both treatment and control when estimating the relationship between the outcome and the covariates. To explain how this could be resolved, we first review the classical problem of estimation of a population mean using regression (see, for example, Cochran, 1977). Let us suppose for a moment that $Y_i(d)$ and $E[X_i]$ were observed, and suppose we wished to estimate $E[Y_i(d)]$ using one of the following two regressions:

$$\texttt{regress } Y_i(d) \text{ on constant} \tag{13}$$

$$\texttt{regress } Y_i(d) \text{ on constant} + (X_i - E[X_i]) \ . \tag{14}$$

Let $\hat{\alpha}_n^{(1)}$ be the resulting estimator of the coefficient on the constant from the regression in (13), and let $\hat{\alpha}_n^{(2)}$ and $\hat{\gamma}_n^{(2)}$ be the resulting estimators of the coefficient on the constant and the coefficient on $(X_i - E[X_i])$ from the regression in (14). We then obtain from elementary properties of regression that

$$\hat{\alpha}_n^{(1)} = \frac{1}{n} \sum_{1 \leq i \leq n} Y_i(d) \ ,$$

and

$$\hat{\alpha}_n^{(2)} = \frac{1}{n} \sum_{1 \leq i \leq n} Y_i(d) - \hat{\gamma}_n^{(2)} \left(\frac{1}{n} \sum_{1 \leq i \leq n} (X_i - E[X_i]) \right) \ .$$

It follows that both $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are consistent estimators of $E[Y_i(d)]$ (note here that it was important that we de-meaned the observable characteristics X_i), and that both are asymptotically normal with variances given by $\text{Var}[Y_i(d)]$ and

$$\text{Var}[Y_i(d)] - \frac{\text{Cov}[Y_i(d), X]^2}{\text{Var}[X]} \ ,$$

respectively. We thus see immediately that $\hat{\alpha}_n^{(2)}$ is a more precise estimator of $E[Y_i(d)]$ whenever the covariates X_i are correlated with the outcome $Y_i(d)$.

Lin (2013) leverages this insight for the purpose of estimating the average treatment effect in a completely randomized experiment by effectively implementing the regression (14) in each of the subsamples $D_i = d$ for $d \in \{0, 1\}$, and then taking the difference of the results. This procedure can be operationalized by estimating a linear regression model with an additional interaction term between D and X:

$$\texttt{regress } Y_i \texttt{ on constant} + D_i + X_i + D_i(X_i - \bar{X}_n) \ . \tag{15}$$

Let $\hat{\beta}_n$ be the resulting estimator of the coefficient on D_i from the above regression. It can then be shown that $\hat{\beta}_n$ is a consistent and asymptotically normal estimator of Δ and that its asymptotic variance V^{sat}

satisfies $V^{\text{sat}} - V^{cr} \leq 0$, with equality if and only if

$$\text{Cov}\left[\frac{Y_i(1)}{\pi} + \frac{Y_i(0)}{1-\pi}, X_i\right] = 0 \tag{16}$$

(see, for instance, Negi and Wooldridge, 2021; Ma et al., 2022). From this we can see that regression based on (15) always weakly improves precision regardless of the value of π, and we should expect that it strictly improves precision whenever the potential outcomes are correlated with the baseline covariates (outside of pathological cases). Further note that (16) holds if $E\left[\frac{Y_i(1)}{\pi} + \frac{Y_i(0)}{1-\pi}\middle| X_i\right]$ is constant; this is exactly the condition that guaranteed that X_i is an irrelevant stratification variable in Section 4. From this we can conclude that if X_i is an irrelevant variable for stratification, then we cannot increase estimation efficiency through adjusting for X_i either. We further explore the relationship between stratification and regression adjustment at the end of Section 5.2.

We conclude this section by noting that the above analysis suggests that, when $\pi = 1/2$, regression adjustment based on (12) may be preferred to adjustment based on (15); in this case both estimators have the same asymptotic variance and (12) involves the estimation of fewer parameters. We caution, however, that this conclusion is very specific to the case of a binary treatment, and that strategies based on interacted models tend to generalize more broadly beyond the special case we considered here.

5.1.1 Inference on Δ when using Linear Regression Adjustment: Some Caveats

We briefly discuss some complications surrounding inference on Δ based on the regressions (12) and (15) under complete randomization, and mention some related complications when generalizing beyond complete randomization. Recall that in Section 3 we explained that the robust variance estimator obtained from the regression in (1) is consistent for the asymptotic variance of the difference-in-means estimator $\hat{\Delta}_n$. A natural follow-up question is then whether or not the robust variance estimators obtained from the regressions (12) and (15) are consistent for (or at least an upper bound on) the asymptotic variances of $\tilde{\beta}_n$ and $\hat{\beta}_n$, respectively. In the case of the regression in (12), the answer is yes: the robust variance estimator is consistent. In the case of the regression in (15), however, the answer is *no* whenever we view the sample as being drawn from a larger (finite or super-) population. The issue here is similar to our discussion surrounding consistent estimators of the variance when studying stratified randomized experiments in Section 4.2. In particular, note that the regression in (15) involves a de-meaning of the baseline covariates, and when we view the sample as being drawn from a larger population, the random fluctuations in $\bar{X}_n$ versus $E[X_i]$ need to be taken into consideration. As a result, the robust variance estimator obtained from regression (15) is *not* guaranteed to be valid outside of the design-based framework (in which case $\bar{X}_n$ is non-random). With this in mind, one simple solution discussed in Negi and Wooldridge (2021) is to instead use the built-in `teffects ra` command in Stata for inference.

When moving beyond complete randomization to settings with stratified assignment, the complications

become more subtle. When stratification is based on a finite number of discrete categories, an easy solution is to perform regression adjustment in each stratum separately and then aggregate the stratum-level estimates; we provide relevant references for this and more complicated settings involving stratification in Section 5.3.

5.2 Double Robustness and General Regression Adjustments

In this section, we explain how Lin's interacted regression (15) is a special case of a more general class of estimators based on a *doubly-robust* moment function. To simplify the exposition we once again assume that treatment is completely randomized with treated fraction $\pi \in (0,1)$. Note that because the treatments are independent of the potential outcomes and covariates, the probability of assignment of any individual i satisfies $P\{D_i = 1 | X_i = x\} \equiv \pi$. Moreover, it follows that the treatments are independent of the potential outcomes *conditional* on the covariates.

Consider the parameter Δ_0 defined by the following moment equation:

$$E\left[\frac{D_i(Y_i - \tilde{\mu}_1(X_i))}{\pi} - \frac{(1-D_i)(Y_i - \tilde{\mu}_0(X_i))}{1-\pi} + \tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i) - \Delta_0\right] = 0\ , \tag{17}$$

where $\tilde{\mu}_1(x), \tilde{\mu}_0(x)$ are arbitrary researcher-defined functions of x which we call the "working models" for the conditional expectations $E[Y_i(1)|X_i = x]$ and $E[Y_i(0)|X_i = x]$. Note that

$$\begin{aligned}
E\left[\frac{D_i(Y_i - \tilde{\mu}_1(X_i))}{\pi} + \tilde{\mu}_1(X_i)\right] &= E\left[\frac{D_i Y_i(1)}{\pi} - \frac{D_i \tilde{\mu}_1(X_i)}{\pi} + \tilde{\mu}_1(X_i)\right] \\
&= E\left[E\left[\frac{D_i Y_i(1)}{\pi}\middle| X_i\right] - \frac{E[D_i|X_i]\tilde{\mu}_1(X_i)}{\pi} + \tilde{\mu}_1(X_i)\right] \\
&= E\left[\frac{\pi E[Y_i(1)|X_i]}{\pi} - \frac{\pi\tilde{\mu}_1(X_i)}{\pi} + \tilde{\mu}_1(X_i)\right] \\
&= E[Y_i(1)]\ .
\end{aligned}$$

Combining this with a similar argument for the other two symmetric terms in (17), it follows that the solution to (17) is $\Delta_0 = \Delta$ for any choice of working models $\tilde{\mu}_1(\cdot)$ and $\tilde{\mu}_0(\cdot)$. Equation (17) is the famous "doubly-robust" moment equation for the ATE due to Robins et al. (1995).[2] Since π is known in a randomized experiment by construction, estimators based on taking a sample analog of (17) will be consistent for Δ regardless of the choice of $\tilde{\mu}_1(\cdot)$ and $\tilde{\mu}_0(\cdot)$. In this way, we obtain the well-known augmented inverse-propensity weighted (AIPW) estimator of Δ:

$$\hat{\Delta}_n^{\text{AIPW}} = \frac{1}{n}\sum_{1\le i\le n}\left(\frac{D_i(Y_i - \hat{\mu}_{1,n}(X_i))}{\pi} - \frac{(1-D_i)(Y_i - \hat{\mu}_{0,n}(X_i))}{1-\pi} + \hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i)\right)\ , \tag{18}$$

where $\hat{\mu}_{1,n}(\cdot)$ and $\hat{\mu}_{0,n}(\cdot)$ are consistent estimators of the working models. $\hat{\Delta}_n^{\text{AIPW}}$ can recover a wide range of

[2]The term "doubly-robust" is due to the fact that, if conversely $\tilde{\mu}_1(x) = E[Y_i(1)|X_i = x]$ and $\tilde{\mu}_0(x) = E[Y_i(0)|X_i = x]$, equation (17) still identifies Δ even if $\pi \neq P\{D_i = 1|X_i = x\}$.

covariate-adjusted estimators of Δ by specifying different choices of $\tilde{\mu}_1(\cdot)$ and $\tilde{\mu}_0(\cdot)$ and their corresponding estimators. For instance, if we set $\hat{\mu}_{1,n}(\cdot) = \hat{\mu}_{0,n}(\cdot) = 0$ then we recover the difference-in-means estimator $\hat{\Delta}_n$. Alternatively, if we let $\hat{\gamma}_n(d)$ be the OLS estimator of the coefficient on X_i in a regression of Y_i on a constant and X_i for observations with $D_i = d$, and set

$$\hat{\mu}_{d,n}(X_i) = (X_i - \bar{X}_n)'\hat{\gamma}_n(d) \ ,$$

for $d \in \{0, 1\}$, then $\hat{\Delta}_n^{\text{AIPW}}$ recovers Lin's interacted linear regression estimator (15).

If $\hat{\mu}_{1,n}(\cdot)$ and $\hat{\mu}_{0,n}(\cdot)$ are appropriately chosen non-parametric estimators of the true conditional mean functions $E[Y_i(1)|X_i = x]$ and $E[Y_i(0)|X_i = x]$, then $\hat{\Delta}_n^{\text{AIPW}}$ becomes a *non-parametric* regression-adjusted estimator of the average treatment effect. Under appropriate assumptions, it can be shown that $\hat{\Delta}_n^{\text{AIPW}}$ is asymptotically normal with variance given by

$$V^* = E\left[\frac{\text{Var}[Y_i(1)|X_i]}{\pi}\right] + E\left[\frac{\text{Var}[Y_i(0)|X_i]}{1-\pi}\right] + \text{Var}[E[Y_i(1) - Y_i(0)|X_i]] \ ,$$

see Tu et al. (2023), Rafi (2023) for details. Importantly, this variance is the efficient variance for estimating Δ using a randomized experiment when the probability of assignment is exogenously constrained to be π (see Armstrong, 2022; Rafi, 2023; Bai et al., 2023d).

Note that V^* coincides exactly with the asymptotic variance V^{sbr} obtained when using the unadjusted estimator $\hat{\Delta}_n$ in a finely stratified randomized experiment with "small" strata (see Section 4.2.1). We have thus demonstrated two alternative methods for achieving the efficient variance V^* when estimating Δ via a randomized experiment:

1. Assign treatment using complete randomization and estimate Δ using $\hat{\Delta}_n^{\text{AIPW}}$ with suitable non-parametric estimators $\hat{\mu}_{d,n}(\cdot)$ of the conditional means $E[Y_i(d)|X_i]$.
2. Assign treatment using "fine stratification" (for instance, if $\pi = 1/2$ this could be a matched pairs design as described in Section 4.2.1) and estimate Δ using $\hat{\Delta}_n$.

This demonstrates that experiments which assign treatment using fine stratification effectively perform non-parametric regression adjustment "*by design*"; this feature of finely stratified experiments is further explored in Cytrynbaum (2023b) and Bai et al. (2023d).

5.3 Further Reading

The exposition in Section 5.1 closely follows Negi and Wooldridge (2021) and Ma et al. (2022), although some of their expressions have been modified so that they could be more easily related to Sections 3 and 4 of this paper. The use of doubly-robust estimators for regression adjustment in randomized experiments

goes back at least to the work of Tsiatis et al. (2008); Tu et al. (2023) and Rafi (2023) extend these results to general stratified experiments with finitely many strata. Bai et al. (2023b) and Cytrynbaum (2023a) consider regression adjustment in settings with "small" strata, in the sense of Section 4.2.1. Wang et al. (2023) study regression adjustment for general parameters defined by estimating equations. Recent work on design-based analyses of regression adjusted estimators can be found in Aronow and Middleton (2013), Wu and Gagnon-Bartsch (2018), Liu et al. (2021), Chang (2023), Chiang et al. (2023).

6 The Analysis of Cluster Randomized Experiments

Until now, our focus has been on experiments where treatment is assigned at the individual level and each individual's outcome depends only on their own treatment. In this section, we consider randomized experiments where treatment is assigned at an aggregated level which we call a *cluster*: for example, when evaluating an educational intervention we may observe outcomes at the student-level, but assign treatment at the school-level, so that every student in the school receives the same treatment. There are two common explanations for why a researcher would consider cluster-level assignment as opposed to individual-level assignment: first, there could be logistical constraints on the experiment that require that every individual in a cluster receives the same treatment. Second, we may be concerned that there is treatment *interference*, i.e., the treatment statuses of individuals in a cluster may affect the outcomes of others. In Section 6.1 we introduce the framework and define some relevant analogs to the average treatment effect in this setting. In Section 6.2, we discuss inference in cluster randomized experiments under complete randomization and stratified block randomization.

6.1 Defining Average Treatment Effects in Cluster Randomized Experiments

To accommodate cluster-level assignment, we first modify our notation relative to what we have considered thus far. Let $Y_{i,g}$ denote the observed outcome of the ith unit in the gth cluster, D_g denote an indicator for whether or not the gth cluster is treated, and N_g the size of the gth cluster. Further denote by $Y_{i,g}(1)$ the potential outcome of the ith unit in the gth cluster if the cluster is treated and by $Y_{i,g}(0)$ the potential outcome of the ith unit in the gth cluster if not treated. As usual, the observed outcome and potential outcomes are related to treatment assignment by the relationship

$$Y_{i,g} = Y_{i,g}(1)D_g + Y_{i,g}(0)(1 - D_g) \ .$$

In practice it is sometimes the case that the researcher does not sample all of the units in a given cluster. To allow for this possibility, define $\mathcal{M}_g$ to be the subset of $\{1, \ldots, N_g\}$ corresponding to the observations within the gth cluster that are sampled by the researcher. For example, in the event that all observations in a cluster are sampled, $\mathcal{M}_g = \{1, \ldots, N_g\}$ and $|\mathcal{M}_g| = N_g$.

With this updated notation, our sampling framework models

$$\{((Y_{i,g}(1), Y_{i,g}(0) : 1 \leq i \leq N_g), \mathcal{M}_g, Z_g, N_g) : 1 \leq g \leq G\} \ ,$$

as a collection of G independent and identically distributed draws from a distribution of clusters. Importantly, we note here that the cluster sizes N_g are modeled as random variables which are potentially related to the potential outcomes. In this framework, we introduce two natural parameters that arise as generalizations of the average treatment effect Δ which we focused on in earlier sections. These parameters differ in the way they aggregate, or average, the individual level treatment effects.

Both parameters of interest we consider can be written as weighted averages of the cluster-level average treatment effects:

$$E\left[\omega_g \left(\frac{1}{N_g} \sum_{1 \leq i \leq N_g} (Y_{i,g}(1) - Y_{i,g}(0))\right)\right] \ ,$$

for different choices of (possibly random) weights ω_g satisfying $E[\omega_g] = 1$. The first parameter of interest corresponds to the choice of $\omega_g = 1$, thus weighting the average effect of the treatment across clusters equally:

$$\Delta^{\text{eq}} := E\left[\frac{1}{N_g} \sum_{1 \leq i \leq N_g} (Y_{i,g}(1) - Y_{i,g}(0))\right] \ .$$

We refer to this quantity as the equally-weighted cluster-level average treatment effect. Since Δ^{eq} assigns an equal weight to each cluster regardless of size, it can be thought of as the average treatment effect where the clusters themselves are the units of interest. The second parameter of interest corresponds to the choice of $\omega_g = N_g/E[N_g]$, thus weighting the average effect of the treatment across clusters in proportion to their size:

$$\Delta^{\text{size}} := E\left[\frac{1}{E[N_g]} \sum_{1 \leq i \leq N_g} (Y_{i,g}(1) - Y_{i,g}(0))\right] \ .$$

We refer to this quantity as the size-weighted cluster-level average treatment effect. Since Δ^{size} assigns a weight proportional to each cluster's size, it can be thought of as the average treatment effect where individuals are the units of interest.

Note that in empirical settings with treatment effect heterogeneity (so that $Y_{i,g}(1) - Y_{i,g}(0)$ is random) and cluster-size heterogeneity, we should expect that Δ^{eq} and Δ^{size} are indeed distinct parameters with differing policy interpretations. For instance, suppose the experiment studies the effect of an educational intervention on students' reading level. If the policy-maker is interested in raising the average reading level across all students, then the magnitude of Δ^{size} is the relevant parameter. If, on the other hand, the policy-maker has concerns about raising the average reading level across all schools, then the magnitude of Δ^{eq} would also be important to consider.

6.2 Inference in Cluster Randomized Experiments

In this section, we study estimation and inference on Δ^{eq} and Δ^{size}. We begin with the setting of a completely randomized experiment; note that in the context of a cluster randomized experiment this means that a fraction $\pi \in (0,1)$ of *clusters* is assigned to treatment and the rest to control.

We begin by first studying the probability limit of the difference-in-means estimator obtained from a regression of the individual-level outcomes $Y_{i,g}$ on a constant and the cluster-level treatment D_g. With our new notation, this is given by

$$\hat{\Delta}_G := \frac{\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} Y_{i,g}D_g}{\sum_{1\le g\le G}|\mathcal{M}_g|D_g} - \frac{\sum_{1\le g\le G}\sum_{i\in\mathcal{M}_g} Y_{i,g}(1-D_g)}{\sum_{1\le g\le G}|\mathcal{M}_g|(1-D_g)} \ .$$

It can be shown under appropriate assumptions that

$$\hat{\Delta}_G \xrightarrow{P} E\left[\frac{1}{E[|\mathcal{M}_g|]}\sum_{i\in\mathcal{M}_g}(Y_{i,g}(1)-Y_{i,g}(0))\right] =: \vartheta \ ,$$

as $G\to\infty$; see Bugni et al. (2022) for details. This parameter corresponds to a *sample*-weighted cluster-level average treatment effect and in general does not coincide with either Δ^{eq} or Δ^{size}. Of course, in some special cases, ϑ does coincide with Δ^{eq} or Δ^{size}. For instance, if we sample the same number of observations in every cluster, then ϑ coincides with Δ^{eq}. If instead we sample observations proportionally to the size of each cluster, then ϑ coincides with Δ^{size}.

In order to do inference on Δ^{eq} and Δ^{size} regardless of the specific choice of sampling design, we now present estimators which are consistent for these parameters more generally. In the case of Δ^{eq}, we consider the following difference-in-means estimator computed on the cluster-average outcomes:

$$\hat{\Delta}^{\text{eq}}_G := \frac{\sum_{1\le g\le G}\bar{Y}_g D_g}{\sum_{1\le g\le G} D_g} - \frac{\sum_{1\le g\le G}\bar{Y}_g(1-D_g)}{\sum_{1\le g\le G}(1-D_g)} \ ,$$

where $\bar{Y}_g = \frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}$. Note that $\hat{\Delta}^{\text{eq}}_G$ may be obtained as the estimator of the coefficient on D_g from the following regression:

$$\texttt{regress } \bar{Y}_g \texttt{ on } \text{constant} + D_g \ ; \tag{19}$$

as such, it is exactly the difference-in-means estimator obtained from viewing the clusters as the experimental units of interest, with outcomes defined by their cluster averages. Under appropriate assumptions it can be shown that

$$\sqrt{G}(\hat{\Delta}^{\text{eq}}_G - \Delta^{\text{eq}}) \xrightarrow{d} N(0, V^{\text{eq}}) \ ,$$

as $G\to\infty$, where

$$V^{\text{eq}} = \frac{\text{Var}[\bar{Y}_g(1)]}{\pi} + \frac{\text{Var}[\bar{Y}_g(0)]}{1-\pi} \ ,$$

with $\bar{Y}_g(d) = \frac{1}{|\mathcal{M}_g|}\sum_{i\in\mathcal{M}_g} Y_{i,g}(d)$ (see Bugni et al., 2022, for details). We thus obtain that $\hat{\Delta}^{\text{eq}}_G$ is a consistent and asymptotically normal estimator of Δ^{eq} with asymptotic variance which exactly mirrors the asymptotic variance V^{cr} obtained in Section 3.2, but with the individual level outcomes $Y_i(d)$ replaced with the cluster-level average outcomes $\bar{Y}_g(d)$. It is therefore not surprising that a consistent estimator of V^{eq} can be obtained from the resulting heteroskedasticity-robust variance estimator of the regression in (19). More generally, when studying Δ^{eq} in a cluster randomized experiment, all of the tools introduced in Sections 3–5 can be applied by simply studying the experiment as if the clusters were individuals with outcomes given by the cluster averages $\bar{Y}_g$.

In the case of Δ^{size}, we consider the following cluster size weighted difference-in-means estimator:

$$\hat{\Delta}^{\text{size}}_G = \frac{\sum_{1\le g\le G}\bar{Y}_g N_g D_g}{\sum_{1\le g\le G} N_g D_g} - \frac{\sum_{1\le g\le G}\bar{Y}_g N_g (1-D_g)}{\sum_{1\le g\le G} N_g (1-D_g)} \ ,$$

Note that $\hat{\Delta}^{\text{size}}_G$ may be obtained as the estimator of the coefficient on D_g in the following *weighted* least squares regression:

$$\texttt{regress } Y_{i,g} \texttt{ on } \text{constant} + D_g \texttt{ using weights } N_g/|\mathcal{M}_g| \ . \tag{20}$$

Under appropriate assumptions it can be shown that

$$\sqrt{G}(\hat{\Delta}^{\text{size}}_G - \Delta^{\text{size}}) \xrightarrow{d} N(0, V^{\text{size}}) \ ,$$

as $G\to\infty$, where

$$V^{\text{size}} := \frac{\text{Var}[\widetilde{Y}_g(1)]}{\pi} + \frac{\text{Var}[\widetilde{Y}_g(0)]}{1-\pi} \ ,$$

with

$$\widetilde{Y}_g(d) \ := \ \frac{N_g}{E[N_g]}\left(\bar{Y}_g(d) - \frac{E[\bar{Y}_g(d)N_g]}{E[N_g]}\right) \ ,$$

see Bugni et al. (2022) for details. We thus obtain that $\hat{\Delta}^{\text{size}}_G$ is a consistent and asymptotically normal estimator of Δ^{size} with asymptotic variance which again mirrors V^{cr}, but now with the individual-level outcomes $Y_i(d)$ replaced with the *re-scaled* cluster-level outcomes $\tilde{Y}_g(d)$. It turns out that in this case, a consistent estimator of V^{size} can be obtained from the regression (20) by computing the resulting *cluster-robust* variance estimator. Equivalently, it is straightforward to construct a sample analog estimator of V^{size} where the infeasible re-scaled outcomes $\widetilde{Y}_g(d)$ are replaced by feasible analogs:

$$\hat{Y}_g := \frac{N_g}{\frac{1}{G}\sum_{1\le j\le G} N_j}\left(\bar{Y}_g - \frac{\frac{1}{G}\sum_{1\le j\le G}\bar{Y}_j I\{D_j = D_g\}N_j}{\frac{1}{G}\sum_{1\le j\le G} I\{D_j = D_g\}N_j}\right) \ .$$

We find this latter approach particularly useful when moving beyond completely randomized experiments to settings with stratification: for example, under stratified block randomization, the limiting variance of $\hat{\Delta}^{\text{size}}_G$

equals

$$\frac{\text{Var}[\widetilde{Y}_g(1)]}{\pi} + \frac{\text{Var}[\widetilde{Y}_g(0)]}{1-\pi} - \pi(1-\pi)\,\text{Var}\left[E\left[\frac{\widetilde{Y}_g(1)}{\pi} + \frac{\widetilde{Y}_g(0)}{1-\pi}\Big| X_g\right]\right] ,$$

where X_g denotes the variables used in stratification (see Bugni et al., 2022; Bai et al., 2023c, for details). This variance mirrors the variance V^{sbr} for individual-level stratified experiments defined in (8), and therefore similar inference procedures apply to cluster randomized experiments by modifying the individual-level outcomes to suitable cluster-level counterparts such as $\hat{Y}_g$.

6.3 Further Reading

The material in this section is most closely related to Bugni et al. (2022) and Bai et al. (2023c). Donner and Klar (2000) contains a textbook treatment of early work in cluster randomized experiments. Su and Ding (2021) and Wang et al. (2024) study regression adjustment for cluster randomized experiments without stratification (Bugni et al. (2022) and Bai et al. (2023c) discuss extensions of these results to settings with stratification). Other recent work on cluster randomized experiments (mostly from a design-based perspective) includes Imai et al. (2009), Schochet (2013), Middleton and Aronow (2015), Schochet et al. (2021), de Chaisemartin and Ramirez-Cuellar (2024).

7 Other Topics

7.1 Treatment Effect Heterogeneity and Quantile Treatment Effects

This article focused exclusively on estimation and inference of the unconditional average treatment effect. Recent work on the analysis of randomized experiments has studied inference for quantile treatment effects: see in particular Zhang and Zheng (2020), Jiang et al. (2021), and Jiang et al. (2023). We may also be interested in heterogeneity of the average treatment effect as a function of the observable characteristics (i.e., the conditional average treatment effect, or CATE). Although estimation and inference on the CATE has been an extremely active area of research in causal inference more broadly (see, for instance, Kennedy, 2023, for an in-depth discussion and further references), to our knowledge almost all of this work maintains the assumption that treatment assignment is independent and identically distributed across individuals; this precludes, for instance, stratified block randomization or pair matching. One exception is Zhang and Ma (2023), who present tests for treatment-covariate interactions and study their properties under general covariate-adaptive stratified randomization schemes.

7.2 Re-randomization

Re-randomization is a method of experimental assignment which, in a similar spirit to stratification, is intended to enforce "balance" in the covariate distributions between treatment and control groups. In a standard re-randomization procedure, researchers specify a balance criterion for the covariate values and then repeatedly generate assignments using a completely randomized design until an assignment is found which achieves an acceptable covariate distribution according to the balance criterion. An excellent historical summary is given in Morgan and Rubin (2012). Similarly to stratification, re-randomization leads to an increase in precision relative to complete randomization, and as a result inferences will be unnecessarily conservative unless this is taken into consideration. However, unlike with stratification, where correcting this issue usually amounts to simply modifying the standard errors, corrected inferences for re-randomization are sometimes further complicated by the fact that the limiting distribution of the difference-in-means estimator is not asymptotically normal. For an in-depth theoretical discussion of re-randomization, see Li et al. (2018), Li et al. (2020), Zhao and Ding (2021b), Lu et al. (2023), Branson et al. (2024), and Cytrynbaum (2024).

7.3 Multiple Testing

Our discussion has so far mostly focused on inference about a single parameter of interest, namely the average effect of a binary treatment on an outcome of interest. In many experiments, however, there may be many parameters of interest: there may be multiple treatments, and so it may be of interest to compare the average effect of each of these treatments with the control or with each other; there may be multiple outcomes of interest, and so it may be of interest to examine these effects for each of these outcomes of interest; finally, there may be multiple subgroups of interest (defined by observed, baseline characteristics, as in Section 7.1), and so it may be of interest to examine these effects separately for these different subgroups. In many cases, the methods described previously can be modified in a straightforward fashion for inference about any one of these parameters, but it is often of interest to examine at least some subset of them simultaneously in order to determine, e.g., which of the parameters are equal to zero or not. This naturally leads to a problem of testing multiple null hypotheses simultaneously. If one were to test each of these null hypotheses in the usual way (i.e., ensuring that the probability of Type I error is controlled adequately for each null hypothesis separately), then the probability of *some* false rejection across all of the null hypotheses may be quite high. A conventional solution to this problem is to require control of the familywise erorr rate – the probability of *any* false rejection across all of the null hypotheses under consideration. Methods for this problem for experiments in which treatment status is assigned in an i.i.d. fashion across units are developed in List et al. (2019, 2023); see also Lee and Shaikh (2014). A very complicated treatment assignment scheme that arises in the context of a well known experiment in the early childhood education literature is treated in Heckman et al. (2024). These results build upon general results in the multiple testing literature described in Romano and Wolf (2005, 2010). In some cases, especially when the number of null hypotheses under consideration is very large, it may be desirable to consider error rates that penalize false rejections less severely, such as the

k familywise error rate (defined as the probability of k or more false rejections), the (tail probability of) the false discovery proportion (defined as the fraction of total rejections that are false), or the false discovery rate (defined as the expected value of the false discovery proportion). Some relevant results for the control of such error rates are described in Romano and Wolf (2010); see also Romano and Wolf (2007) and Romano et al. (2008).

7.4 Imperfect Compliance and Attrition

Even the most well designed experiments encounter challenges that can complicate subsequent analyses. Two common issues that arise in practice are imperfect compliance and attrition.

Imperfect compliance arises when units assigned to the treatment group end up not taking up treatment, and/or units assigned to the control group manage to obtain the treatment. Of course, if researchers are simply interested in the effect of the *assignment* to treatment as opposed to the true *receipt* of the treatment (i.e., the intention-to-treat estimand), then everything we have discussed thus far applies directly. However, if the decision to comply with the treatment is not exogenous but is instead determined by the unobserved characteristics of the units, then the experiment no longer point identifies the causal effect of the true receipt of treatment on the outcome (i.e., the average treatment effect). Most recent work on the analysis of randomized experiments with imperfect compliance has adopted the framework of Imbens and Angrist (1994), where treatment assignment is used an an *instrument* for the receipt of the treatment and the primary focus is on the so-called "local" average treatment effect (i.e., the average treatment effect for those units who comply with the treatment assignment): see, in particular, Ansel et al. (2018), Bugni and Gao (2023), Ren (2023), and Bai et al. (2023a). Alternatively, a similar instrumental variables strategy could be used to *partially* identify the average treatment effect: see Manski (1990), Balke and Pearl (1997), Bhattacharya et al. (2008), Bhattacharya et al. (2012), Machado et al. (2019), Bugni et al. (2024).

Attrition arises when outcomes are not observed for some subset of the experimental units. This situation could arise, for instance, if researchers lose track of subjects in the experiment. As with imperfect compliance, if the decision for the unit to drop out of the experiment is not exogenous, but is instead determined by their unobserved characteristics, then the experiment no longer point identifies the average treatment effect. Standard resolutions to this issue include modelling the selection process (as in Heckman, 1979) or partial identification (as in Horowitz and Manski, 2000). A textbook treatment on the analysis of randomized experiments with attrition is given in Gerber and Green (2012), and DiNardo et al. (2006) review different methods for dealing with attrition within the context of the Moving to Opportunity (MTO) experiment. Ghanem et al. (2023) develop a test for attrition bias in randomized experiments. Fukumoto (2022) and Bai et al. (2024a) study the problem of attrition in the setting of matched pair experiments, and revisit common recommendations about whether or not to drop pairs with an attrited unit.

7.5 Network Experiments and Experiments with Interference

Other than in Section 6, we have so far maintained that each individual's outcome depends on only their own treatment. A rapidly growing literature considers the analysis of experiments in the presence of *interference*, i.e., that the outcome of a given individual in the experiment may be affected by the treatment statuses of others (see Halloran and Struchiner, 1995, for an early discussion). The simplest example of such a setting is what is called *partial* interference, where units are grouped into a collection of disjoint clusters, and interference is possible between individuals within the same cluster; Section 6 discussed a special case of such a setting where every unit in the cluster receives the same treatment. Hudgens and Halloran (2008), Basse and Feller (2018), Imai et al. (2021), Vazquez-Bare (2023), Leung (2023), and Liu (2023) study *two-stage experiments* in settings with partial interference, where first clusters are randomly assigned to different treatment *proportions*, and then individuals within the clusters are assigned to treatment with (marginal) probability according to their cluster's assigned proportion.

More generally, we could consider settings with complex patterns of interference, for instance, if individuals interact on a large network. Manski (2013) and Aronow and Samii (2017) develop the concept of *effective treatments* or *exposure mappings*, which summarize how a given unit's outcome is affected by the treatments of other units. For example, in a network context, the exposure mapping may dictate that only the treatments of a unit's direct links affect their outcome. The exposure mapping formulation of interference has had a major influence on the subsequent literature: examples include Leung (2020), Viviano (2020), Forastiere et al. (2021), Auerbach and Tabord-Meehan (2021), Munro et al. (2021), Li and Wager (2022), Leung (2022), Gao and Ding (2023), Park and Kang (2023), Viviano et al. (2023), Sävje (2024). Some recent papers that study general forms of interference without employing the formalism of exposure mappings are Wager and Xu (2021), Sävje et al. (2021), and Hu et al. (2022).

7.6 Randomization Inference

An extremely important topic in the analysis of randomized experiments which we did not cover in this article is the idea of *randomization inference*. Mechanically, randomization inference generates the null distribution of the test statistic by repeatedly re-assigning treatments to the experimental sample and re-computing the resulting test statistic. If the true value of the test-statistic is too "large" relative to this null distribution then the null hypothesis is rejected. The primary strength of these types of tests is that, for appropriate null hypotheses, they can be shown to be *finite sample* valid.

One such type of test, going back to the work of Hoeffding (1952), is where the data satisfies some form of *group invariance* under the null hypothesis, with respect to a group of transformations of the data. For example, consider a completely randomized experiment where we wish to test the null hypothesis

$$H_0 : Y_i(1) \stackrel{d}{=} Y_i(0) \ .$$

In this case, a valid group of transformations is given by the transformations which permute the treatment assignments of the individuals. Lehmann and Romano (2022) provide a comprehensive textbook introduction to these types of tests in very general settings (even outside the context of randomized experiments). Romano et al. (2024) in this issue provide a survey of recent advances.

A similar but distinct concept, going back to the work of Fisher (1925) and typically employed in the design-based paradigm, is where the null hypothesis is "sharp" in the sense that the test statistic under any counterfactual treatment assignment can be imputed from the data. For example, consider a completely randomized experiment conducted on a finite population of n individuals, then the canonical "sharp" null is given by

$$H_0 : Y_i(1) = Y_i(0) \quad \text{for all } 1 \leq i \leq n \ .$$

In this case, as we permute the treatment assignments of the individuals, we can perfectly impute the value of the test statistic under the null hypothesis. Imbens and Rubin (2015) provide a comprehensive textbook introduction to these "Fisher-style" tests. Recent work (particularly in settings with interference) includes Athey et al. (2018), Basse et al. (2019b), Basse et al. (2019a).

It is important to emphasize that, in both cases, finite-sample validity is only guaranteed for *very specific* choices of the null hypothesis. In particular, in either case considered above, if we instead consider the null hypothesis that the average treatment effect is equal to zero, then we would have neither the required group invariance property nor a "sharp" null which would guarantee finite-sample validity. However, a recent series of papers establishes that the randomization test can be *asymptotically* valid for these "weak" nulls, while retaining its finite-sample validity for the "sharp" null, as long as the test-statistic is constructed appropriately. See Chung and Romano (2013), Bugni et al. (2018), Bai et al. (2022), and Bai et al. (2023c), for examples in a super-population context, and Zhao and Ding (2021a), Wu and Ding (2021) for examples in a design-based context.

7.7 Policy Learning

Our discussion has focused on the problem of estimation and inference of treatment effect parameters. A recent literature, popularized in econometrics by Manski (2004), seeks to instead use the data in order to directly inform the allocation of treatments over the population as a function of the observable characteristics in order to maximize welfare. Formally, the problem is framed as a statistical decision problem in the framework of Wald (1949) (see Hirano and Porter, 2020, for a comprehensive introduction). Extensive work on this topic, which is now often referred to as "policy learning", exists at the intersection of econometrics, statistics, and computer science; important contributions to this literature which also provide comprehensive overviews are Kitagawa and Tetenov (2018) and Athey and Wager (2021). Relevant work in econometrics on this topic (primarily focused on settings where data are obtained from a randomized experiment) includes Dehejia (2005), Stoye (2009), Hirano and Porter (2009), Bhattacharya and Dupas (2012), Viviano (2019),

Ananth (2020), Azevedo et al. (2020), Kitagawa and Tetenov (2021), Mbakop and Tabord-Meehan (2021), Sun (2021), Viviano (2022), Kitagawa and Wang (2023), Higbee (2023), Kock et al. (2023).

7.8 Response-Adaptive Designs and Bandit Experiments

Until now, we have assumed that the experimental design being employed by the researcher does not use information from earlier waves of experimentation. In this section, we briefly comment on a rapidly growing literature which considers *response-adaptive* experimental designs, which are designs that can adapt throughout the experiment as a result of data that have already accrued. Response-adaptive designs have a long history in the analysis of clinical trials; Hu and Rosenberger (2006) and Rosenberger and Lachin (2015) provide textbook introductions.

Some recent work on response-adaptive designs in econometrics and statistics studies procedures to construct feasible analogs of the Neyman allocation, in order to efficiently estimate treatment effect parameters; examples include Hahn et al. (2011), Tabord-Meehan (2023), Blackwell et al. (2022), Li and Owen (2023), Wei et al. (2024). Cai and Rafi (2022) demonstrate that some of these procedures may have poor finite-sample properties when the data used to estimate the optimal treatment assignment proportions is not sufficiently large.

Much of the recent work on adaptive designs is related to bandit problems and/or best arm identification. These are essentially policy learning problems (in the sense defined in Section 7.6) where the researcher wishes to choose a policy to maximize welfare, either for the participants in the experiment itself or for the broader population of interest. Bubeck et al. (2012) and Lattimore and Szepesvári (2020) provide comprehensive textbook introductions. Some recent work on this topic includes Russo (2016), Russo and Van Roy (2016), Agrawal and Goyal (2017), Kasy and Sautmann (2021), Adusumilli (2021), Lieber (2022), Kuang and Wager (2023), Kato et al. (2024). The analysis of treatment effect parameters in these contexts can be particularly challenging since the experiment was not necessarily designed to facilitate inference. Recent work on inference in adaptive experiments includes Bibaut et al. (2021), Hadad et al. (2021), Zhang et al. (2021), Adusumilli (2023), Chen and Andrews (2023), Hirano and Porter (2023).

A Additional Details

A.1 A General Comparison of Super-population and Finite-population Variances

In this section we formalize a claim made in Section 3.1 about the general relationship between the finite population and super-population variance of $\hat{\Delta}_n$.

When presenting the regularity conditions maintained in a finite-population analysis, researchers will often use the motivation that their conditions hold with probability one when the observations are in fact i.i.d. draws from a super-population (see for instance the discussion following Theorem 5 in Li and Ding, 2017). With this in mind, finite-population results about the limiting distribution of $\hat{\Delta}_n$ can often be conceptualized as first imagining a collection of i.i.d. draws $W^{(N)} = \{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq N\}$ from some distribution P, sampling a subset of n units from $W^{(N)}$, and then deriving the limiting distribution of $\hat{\Delta}_n$ conditional on $W^{(N)}$:

$$\sqrt{n}(\hat{\Delta}_n - \Delta_N)|W^{(N)} \xrightarrow{d} N(0, \sigma^2_{1,\lambda}(W^{(\infty)})) \ , \tag{21}$$

where $W^{(\infty)} = \{W_i : i \geq 1\}$ and $n/N \to \lambda \in [0, 1]$. Formally, this conditional convergence can be defined as

$$\sup_{t \in \mathbf{R}} |P\{\sqrt{n}(\hat{\Delta}_n - \Delta_N) \leq t|W^{(N)}\} - \Phi(t/\sigma_{1,\lambda}(W^{\infty}))| \xrightarrow{P} 0 \ .$$

We now relate (21) to the unconditional limiting distribution of $\hat{\Delta}_n$ is a super-population analysis. First, we note that while in principle $\sigma_{1,\lambda}(W^{(\infty)})$ could vary with different realizations of the sequence of finite populations $W^{(\infty)}$, if $W^{(\infty)}$ in fact arise from a super-population, then $\sigma_1(W^{(\infty)})$ will often be the same for (almost all) realizations of $W^{(\infty)}$. As an example, if we consider complete randomization as defined in Section 2 and if $W^{(\infty)}$ is drawn from a super-population P, then it follows from the strong law of large numbers that for almost every sequence $W^{(\infty)}$, as $n \to \infty$,

$$n \operatorname{Var}[\hat{\Delta}_n|W^{(N)}] \to \frac{\operatorname{Var}_P[Y_i(1)]}{\pi} + \frac{\operatorname{Var}_P[Y_i(0)]}{1-\pi} - \lambda \operatorname{Var}_P[Y_i(1) - Y_i(0)] \ , \tag{22}$$

where $n/N \to \lambda \in [0, 1]$ (we illustrate a similar property for stratified block randomization at the end of this section). Note that it follows immediately by the central limit theorem that $\sqrt{n}(\Delta_N - \Delta) \xrightarrow{d} N(0, \lambda\sigma_2^2)$, where $\sigma_2^2 = \operatorname{Var}_P[Y_i(1) - Y_i(0)]$. Because Δ_N is a function of $W^{(\infty)}$, if (21) holds with $\sigma^2_{1,\lambda}(W^{(\infty)}) \equiv \sigma^2_{1,\lambda}$, then it follows from Lemma S.1.2 in Bai et al. (2022) that the distributions of $\sqrt{n}(\hat{\Delta}_n - \Delta_N)|W^{(N)}$ and $\sqrt{n}(\Delta_N - \Delta)$ are "asymptotically independent" and moreover that

$$\sqrt{n}(\hat{\Delta}_n - \Delta) \xrightarrow{d} N(0, \sigma^2_{1,\lambda} + \lambda\sigma_2^2) \ . \tag{23}$$

In other words, the super-population variance given in (23) and the finite-population variance given in

(21) differ by $\lambda \operatorname{Var}_P[Y_i(1) - Y_i(0)]$ in the limit, which is the fraction of people sampled from the finite population times the variance of the individual level treatment effects. When $n/N \to \lambda = 0$, so that we sample a negligible fraction of the finite population in the limit, the difference is zero. Indeed, we can see this for complete randomization by comparing (6) and (22). It thus follows immediately that constructing a consistent estimator of $\sigma^2_{1,\lambda} + \lambda\sigma^2_2$ delivers a conservative variance estimator of $\sigma^2_{1,\lambda}$ (in fact, $\sigma_{1,\lambda} + \lambda\sigma^2_2$ doesn't vary across λ, as can be seen for example in complete randomization.)

As another example, consider stratified block randomization with discrete $x \in \{1, 2, \ldots, \mathcal{X}\}$ as defined in Section 4. Here we suppose $n = N$ so that $\lambda = 1$. For $d \in \{0, 1\}$ and $x \in \{1, 2, \ldots, \mathcal{X}\}$, define

$$\begin{aligned} n_d(x) &= \sum_{1 \le i \le n} I\{D_i = d, X_i = x\} \\ n(x) &= n_1(x) + n_0(x) \\ \bar{Y}_n(d, x) &= \frac{1}{n_d(x)} \sum_{1 \le i \le n} Y_i(d) I\{D_i = d, X_i = x\} \\ \Delta_n(x) &= \bar{Y}_n(1, x) - \bar{Y}_n(0, x) \ . \end{aligned}$$

The finite population variance of the fully saturated estimator $\hat{\Delta}^{\text{sat}}_n$ is

$$\operatorname{Var}[\hat{\Delta}^{\text{sat}}_n | W^{(N)}] = \sum_x \frac{n(x)}{n} \left(\frac{S^2_1(x)}{n_1(x)} + \frac{S^2_0(x)}{n_0(x)} - \frac{S^2_\Delta(x)}{n(x)} \right) \ ,$$

where $S^2_d(x)$ and $S^2_\Delta(x)$ are the within-stratum counterpart of S^2_d and S^2_Δ in (2):

$$\begin{aligned} S^2_d(x) &= \frac{1}{n_d(x) - 1} \sum_{1 \le i \le n} (Y_i(d) - \bar{Y}_n(d, x))^2 I\{D_i = d, X_i = x\} \\ S^2_\Delta(x) &= \frac{1}{n(x) - 1} \sum_{1 \le i \le n} (Y_i(1) - Y_i(0) - \Delta_n(x))^2 I\{X_i = x\} \ . \end{aligned}$$

If the finite populations are realizations from a super-population P, then it can be shown that with probability one,

$$n \operatorname{Var}[\hat{\Delta}^{\text{sat}}_n | W^{(N)}] \to \sum_x p(x) \left(\frac{\operatorname{Var}[Y_i(1) | X_i = x]}{\pi(x)} + \frac{\operatorname{Var}[Y_i(0) | X_i = x]}{1 - \pi(x)} - \operatorname{Var}[Y_i(1) - Y_i(0) | X_i = x] \right) \ ,$$

where $p(x) = P\{X_i = x\}$.

On the other hand, the super-population variance satisfies

$$\begin{aligned} n \operatorname{Var}[\hat{\Delta}^{\text{sat}}_n] \to \sum_x p(x) \Bigg(& \frac{\operatorname{Var}[Y_i(1) | X_i = x]}{\pi(x)} + \frac{\operatorname{Var}[Y_i(0) | X_i = x]}{1 - \pi(x)} \\ & + (E[Y_i(1) - Y_i(0) | X_i = x] - E[Y_i(1) - Y_i(0)])^2 \Bigg) \ . \end{aligned}$$

Therefore, once again we have

$$n \operatorname{Var}[\hat{\Delta}_n^{\text{sat}}] - n \operatorname{Var}[\hat{\Delta}_n^{\text{sat}} | W^{(N)}] \to \operatorname{Var}[Y_i(1) - Y_i(0)] \ .$$

B Proofs of claims in main text

B.1 Derivations of $E[\hat{\Delta}_n]$ and $\operatorname{Var}[\hat{\Delta}_n]$ under complete randomization in the finite population framework

Define the set $\mathcal{C}_{n,N}$ to index the random subset of n observations sampled without replacement from the population of size N. Consider the following expansion:

$$\hat{\Delta}_n = A_n - B_n \ ,$$

where

$$A_n = \frac{1}{n_1} \sum_{1 \leq i \leq n} \left(Y_i(1) + Y_i(0) \frac{n_1}{n_0} \right) D_i$$

$$B_n = \frac{1}{n_0} \sum_{1 \leq i \leq n} Y_i(0) \ .$$

In this re-writing, we have partitioned $\hat{\Delta}_n$ into two components A_n and B_n. Conditional on $\mathcal{C}_{n,N}$, and given our definition of $D^{(n)}$, the first component A_n can be interpreted as the sample average when sampling n_1 units without replacement from the finite population of observations $\{(Y_i(1) + Y_i(0)\frac{n_1}{n_0}) : 1 \leq i \leq n\}$, whose population average is given by

$$\frac{1}{n} \sum_{1 \leq i \leq n} \left(Y_i(1) + Y_i(0) \frac{n_1}{n_0} \right) \ .$$

When viewed from this perspective, the treatment indicators D_i are now *sampling* indicators which determine which of the n_1 units are sampled from our population of n units. Using standard results from the literature on survey sampling (see, for instance, Theorem 2.1 in Cochran, 1977), the sample average is unbiased for the population average:

$$E[A_n | \mathcal{C}_{n,N}] = \frac{1}{n} \sum_{1 \leq i \leq n} \left(Y_i(1) + Y_i(0) \frac{n_1}{n_0} \right) \ ,$$

where we emphasize that the expectation is with respect to the sampling indicators $D^{(n)}$, which are indeed the only source of randomness once we condition on $\mathcal{C}_{n,N}$. Combining this result with our decomposition for $\hat{\Delta}_n$, we obtain that

$$\begin{aligned} E[\hat{\Delta}_n | \mathcal{C}_{n,N}] &= E[A_n | \mathcal{C}_{n,N}] - E[B_n | \mathcal{C}_{n,N}] \\ &= \frac{1}{n} \sum_{1 \leq i \leq n} \left(Y_i(1) + Y_i(0) \frac{n_1}{n_0} \right) - B_n \end{aligned}$$

$$= \frac{1}{n} \sum_{1 \leq i \leq n} (Y_i(1) - Y_i(0)) \ ,$$

where the second equality used our above derivation of $E[A_n|\mathcal{C}_{n,N}]$ and the fact that B_n is non-random conditional on $\mathcal{C}_{n,N}$, and the third equality follows from some additional algebra. Then by the law of iterated expectations:

$$E[\hat{\Delta}_n] = E[E[\hat{\Delta}_n|\mathcal{C}_{n,N}]] = E\left[\frac{1}{n} \sum_{1 \leq i \leq n} (Y_i(1) - Y_i(0))\right] = \Delta_N^{\text{fp}} \ ,$$

where the last equality again follows from Theorem 2.1 in Cochran (1977). Using the same decomposition, we obtain that

$$\begin{aligned}
\text{Var}[\hat{\Delta}_n|\mathcal{C}_{n,N}] &= \text{Var}[A_n|\mathcal{C}_{n,N}] \\
&= \text{Var}\left[\frac{1}{n_1} \sum_{1 \leq i \leq n} \left(Y_i(1) + Y_i(0)\frac{n_1}{n_0}\right) D_i\right] \\
&= \left(\frac{1}{n_1} - \frac{1}{n}\right) \frac{1}{n-1} \sum_{1 \leq i \leq n} \left[\left(Y_i(1) + Y_i(0)\frac{n_1}{n_0}\right) - \frac{1}{n} \sum_{1 \leq i \leq n} \left(Y_i(1) + Y_i(0)\frac{n_1}{n_0}\right)\right]^2 \\
&= \frac{\zeta_1^2}{n_1} + \frac{\zeta_0^2}{n_0} - \frac{1}{n(n-1)} \sum_{1 \leq i \leq n} \left[(Y_i(1) - \bar{Y}_n(1))^2 + (Y_i(0) - \bar{Y}_n(0))^2 - 2(Y_i(1) - \bar{Y}_n(1))(Y_i(0) - \bar{Y}_n(0))\right] \\
&= \frac{\zeta_1^2}{n_1} + \frac{\zeta_0^2}{n_0} - \frac{\zeta_\Delta^2}{n} \ ,
\end{aligned}$$

with

$$\begin{aligned}
\zeta_d^2 &= \frac{1}{n-1} \sum_{1 \leq i \leq n} \left(Y_i(d) - \bar{Y}_n(d)\right)^2 \\
\zeta_\Delta^2 &= \frac{1}{n-1} \sum_{1 \leq i \leq n} \left(Y_i(1) - Y_i(0) - \left(\bar{Y}_n(1) - \bar{Y}_n(0)\right)\right)^2 \ ,
\end{aligned}$$

where the first equality follows since B_n is non-random conditional on $\mathcal{C}_{n,N}$, the third equality follows from Theorem 2.2 in Cochran (1977) and the last two equalities from additional algebra. We thus obtain from the law of total variance that

$$\begin{aligned}
\text{Var}[\hat{\Delta}_n] &= E[\text{Var}[\hat{\Delta}_n|\mathcal{C}_{n,N}]] + \text{Var}[E[\hat{\Delta}_n|\mathcal{C}_{n,N}]] \\
&= E\left[\frac{\zeta_1^2}{n_1} + \frac{\zeta_0^2}{n_0} - \frac{\zeta_\Delta^2}{n}\right] + \text{Var}\left[\frac{1}{n} \sum_{1 \leq i \leq n} (Y_i(1) - Y_i(0))\right] \ .
\end{aligned}$$

Repeated applications of Theorems 2.1 and 2.2 in Cochran (1977) along with additional algebra reveals that

$$E\left[\zeta_d^2\right] = \frac{n}{n-1}\left(\left(\frac{N-1}{N}\right)S_d^2 - \left(\frac{1}{n}-\frac{1}{N}\right)S_d^2\right) = S_d^2 \ ,$$

and similarly

$$E\left[\zeta_\Delta^2\right] = S_\Delta^2 \ .$$

By another application of Theorem 2.2 in Cochran (1977),

$$\text{Var}\left[\frac{1}{n}\sum_{1\leq i\leq n}\left(Y_i(1)-Y_i(0)\right)\right] = \left(\frac{1}{n}-\frac{1}{N}\right)S_\Delta^2 \ .$$

Putting this all together, we obtain

$$\begin{aligned}\text{Var}[\hat{\Delta}_n] &= \frac{S_1^2}{n_1}+\frac{S_1^2}{n_0}+\frac{S_\Delta^2}{n}+\left(\frac{1}{n}-\frac{1}{N}\right)S_\Delta^2 \\ &= \frac{S_1^2}{n_1}+\frac{S_1^2}{n_0}-\frac{S_\Delta^2}{N} \ ,\end{aligned}$$

as desired. ■

B.2 Derivations of (4) and (5)

For the first claim,

$$\begin{aligned}E[\hat{\Delta}_n] &= E\left[E\left[\hat{\Delta}_n|D^{(n)}\right]\right] \\ &= E\left[\frac{1}{n_1}\sum_{1\leq i\leq n}E[Y_i(1)|D^{(n)}]D_i - \frac{1}{n_0}\sum_{1\leq i\leq n}E[Y_i(0)|D^{(n)}](1-D_i)\right] \\ &= E\left[E[Y_i(1)]\left(\frac{1}{n_1}\sum_{1\leq i\leq n}D_i\right) - E[Y_i(0)]\left(\frac{1}{n_0}\sum_{1\leq i\leq n}(1-D_i)\right)\right] \\ &= E[Y_i(1)-Y_i(0)] = \Delta \ ,\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second from properties of conditional expectations and the definition of the observed outcome Y_i, the third from the exogeneity of treatment assignment and the fact that the sample is i.i.d., and the final equality by the definition of n_1 and n_0.

For the second claim,

$$\begin{aligned}\text{Var}[\hat{\Delta}_n] &= E\left[\text{Var}\left[\hat{\Delta}_n|D^{(n)}\right]\right] + \text{Var}\left[E\left[\hat{\Delta}_n|D^{(n)}\right]\right] \\ &= E\left[\text{Var}\left[\hat{\Delta}_n|D^{(n)}\right]\right] + \text{Var}[\Delta]\end{aligned}$$

$$
\begin{aligned}
&= E\left[\frac{1}{n_1^2}\sum_{1\leq i\leq n}\mathrm{Var}[Y_i(1)|D^{(n)}]D_i + \frac{1}{n_0^2}\sum_{1\leq i\leq n}\mathrm{Var}[Y_i(0)|D^{(n)}](1-D_i)\right] \\
&= E\left[\mathrm{Var}[Y_i(1)]\frac{1}{n_1^2}\sum_{1\leq i\leq n}D_i + \mathrm{Var}[Y_i(0)]\frac{1}{n_0^2}\sum_{1\leq i\leq n}(1-D_i)\right] \\
&= \frac{\mathrm{Var}[Y_i(1)]}{n_1} + \frac{\mathrm{Var}[Y_i(0)]}{n_0}\ ,
\end{aligned}
$$

where the first equality follows from the law of total variance, the second from the derivation of the expectation, the third from the properties of conditional variances and the fact that D_i is binary, the fourth by the exogeneity of treatment assignment and the fact that the sample is i.i.d., and the final equality by the definition of n_1 and n_0 under complete randomization.

B.3 Derivation of (7)

We have

$$
\begin{aligned}
&E[\mathrm{Var}[\hat{\Delta}_n|X^{(n)},D^{(n)}]|X^{(n)}] \\
&= E\left[\frac{4}{n^2}\sum_{1\leq i\leq n}\left(D_i\,\mathrm{Var}[Y_i(1)|X_i] + (1-D_i)\,\mathrm{Var}[Y_i(0)|X_i]\right)\middle|X^{(n)}\right] \\
&= \frac{2}{n^2}\sum_{1\leq i\leq n}\left(\mathrm{Var}[Y_i(1)|X_i] + \mathrm{Var}[Y_i(0)|X_i]\right)\ ,
\end{aligned}
$$

where in the first equality we use the fact that the potential outcomes are independent across units conditional on $X^{(n)}$ and $D^{(n)}$. ■

References

Adusumilli, K. (2021). Risk and optimal policies in bandit experiments. *arXiv preprint arXiv:2112.06363*.

Adusumilli, K. (2023). Optimal tests following sequential experiments. *arXiv preprint arXiv:2305.00403*.

Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, **64** 1–24.

Ananth, A. (2020). Optimal treatment assignment rules on networked populations. Tech. rep., working paper.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ansel, J., Hong, H. and Jessie Li, a. (2018). Ols and 2sls in randomized and conditionally randomized experiments. *Jahrbücher für Nationalökonomie und Statistik*, **238** 243–293.

Armstrong, T. B. (2022). Asymptotic Efficiency Bounds for a Class of Experimental Designs. ArXiv:2205.02726 [stat], URL http://arxiv.org/abs/2205.02726.

Aronow, P. M., Green, D. P. and Lee, D. K. K. (2014). Sharp Bounds on the Variance in Randomized Experiments. *The Annals of Statistics*, **42** 850–871.

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

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

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

Athey, S. and Imbens, G. W. (2017). The Econometrics of Randomized Experiments. In *Handbook of Economic Field Experiments*, vol. 1. Elsevier, 73–140.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, **89** 133–161.

Atkinson, A., Donev, A. and Tobias, R. (2007). *Optimum experimental designs, with SAS*, vol. 34. OUP Oxford.

Auerbach, E. and Tabord-Meehan, M. (2021). The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*.

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J. and Weyl, E. G. (2020). A/b testing with fat tails. *Journal of Political Economy*, **128** 4614–000.

Bai, Y. (2022). Optimality of Matched-Pair Designs in Randomized Controlled Trials. *American Economic Review*, **112** 3911–3940.

Bai, Y. (2023). Why randomize? Minimax optimality under permutation invariance. *Journal of Econometrics*, **232** 565–575.

Bai, Y., Guo, H., Shaikh, A. M. and Tabord-Meehan, M. (2023a). Inference in Experiments with Matched Pairs and Imperfect Compliance. ArXiv:2307.13094 [econ, math, stat], URL http://arxiv.org/abs/2307.13094.

Bai, Y., Hsieh, M. H., Liu, J. and Tabord-Meehan, M. (2024a). Revisiting the analysis of matched-pair and stratified experiments in the presence of attrition. *Journal of Applied Econometrics*, **39** 256–268.

Bai, Y., Jiang, L., Romano, J. P., Shaikh, A. M. and Zhang, Y. (2023b). Covariate Adjustment in Experiments with Matched Pairs. ArXiv:2302.04380 [econ], URL http://arxiv.org/abs/2302.04380.

Bai, Y., Liu, J., Shaikh, A. M. and Tabord-Meehan, M. (2023c). Inference in Cluster Randomized Trials with Matched Pairs. ArXiv:2211.14903 [econ, stat], URL http://arxiv.org/abs/2211.14903.

Bai, Y., Liu, J., Shaikh, A. M. and Tabord-Meehan, M. (2023d). On the Efficiency of Finely Stratified Experiments. ArXiv:2307.15181 [econ, math, stat], URL http://arxiv.org/abs/2307.15181.

Bai, Y., Liu, J. and Tabord-Meehan, M. (2024b). Inference for Matched Tuples and Fully Blocked Factorial Designs. *Quantitative Economics*.

Bai, Y., Romano, J. P. and Shaikh, A. M. (2022). Inference in Experiments With Matched Pairs. *Journal of the American Statistical Association*, **117** 1726–1737.

Balke, A. and Pearl, J. (1997). Bounds on Treatment Effects from Studies with Imperfect Compliance. *Journal of the American Statistical Association*, **92** 1171–1176.

Basse, G., Ding, P., Feller, A. and Toulis, P. (2019a). Randomization tests for peer effects in group formation experiments. *arXiv preprint arXiv:1904.02308*.

Basse, G. and Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, **113** 41–55.

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

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, **167** 168–196.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2008). Treatment effect bounds under monotonicity assumptions: an application to swan-ganz catheterization. *American Economic Review*, **98** 351–356.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2012). Treatment effect bounds: An application to swan–ganz catheterization. *Journal of Econometrics*, **168** 223–243.

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A. and van Der Laan, M. (2021). Post-contextual-bandit inference. *Advances in neural information processing systems*, **34** 28548–28559.

Blackwell, D. and Girshick, M. A. (1954). *Theory of games and statistical decisions*. John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London.

Blackwell, M., Pashley, N. E. and Valentino, D. (2022). Batch adaptive designs to improve efficiency in social science experiments.

Branson, Z., Li, X. and Ding, P. (2024). Power and sample size calculations for rerandomization. *Biometrika*, **111** 355–363.

Bruhn, M. and McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, **1** 200–232.

Bubeck, S., Cesa-Bianchi, N. et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, **5** 1–122.

Bugni, F., Canay, I., Shaikh, A. and Tabord-Meehan, M. (2022). Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes. ArXiv:2204.08356 [econ, stat], URL http://arxiv.org/abs/2204.08356.

Bugni, F., Gao, M., Obradović, F. and Velez, A. (2024). Identification and inference on treatment effects under covariate-adaptive randomization and imperfect compliance. *Working Paper*.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2018). Inference Under Covariate-Adaptive Randomization. *Journal of the American Statistical Association*, **113** 1784–1796.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, **10** 1747–1785.

Bugni, F. A. and Gao, M. (2023). Inference under covariate-adaptive randomization with imperfect compliance. *Journal of Econometrics*, **237** 105497.

Cai, Y. and Rafi, A. (2022). On the Performance of the Neyman Allocation with Small Pilots. ArXiv:2206.04643 [econ], URL http://arxiv.org/abs/2206.04643.

Chang, H. (2023). Design-based estimation theory for complex experiments. *arXiv preprint arXiv:2311.06891*.

Chen, J. and Andrews, I. (2023). Optimal conditional inference in adaptive experiments. *arXiv preprint arXiv:2309.12162*.

Chiang, H. D., Matsushita, Y. and Otsu, T. (2023). Regression adjustment in randomized controlled trials with many covariates. *arXiv preprint arXiv:2302.00469*.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, **41** 484–507.

Cochran, W. G. (1977). *Sampling techniques*. 3rd ed. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York-London-Sydney.

Cox, D. and Reid, N. (2000). *The Theory of the Design of Experiments*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.

Cytrynbaum, M. (2023a). Covariate Adjustment in Stratified Experiments. ArXiv:2302.03687 [econ, stat], URL http://arxiv.org/abs/2302.03687.

Cytrynbaum, M. (2023b). Optimal Stratification of Survey Experiments. ArXiv:2111.08157 [econ, math, stat], URL http://arxiv.org/abs/2111.08157.

Cytrynbaum, M. (2024). Estimation and inference under stratified rerandomization. *Working Paper*.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2024). At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments? *American Economic Journal: Applied Economics*, **16** 193–212.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, **125** 141–173.

DiNardo, J., McCrary, J. and Sanbonmatsu, L. (2006). Constructive proposals for dealing with attrition: An empirical example. Tech. rep., Working paper, University of Michigan.

Donner, A. and Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Duflo, E., Glennerster, R. and Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. In *Handbook of Development Economics*, vol. 4. Elsevier, 3895–3962.

Fisher, R. A. (1925). Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **22** 700–725.

Fisher, R. A. (1935). The design of experiments. *The design of experiments*.

Forastiere, L., Airoldi, E. M. and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, **116** 901–918.

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

Fukumoto, K. (2022). Nonignorable attrition in pairwise randomized experiments. *Political Analysis*, **30** 132–141.

Gao, M. and Ding, P. (2023). Causal inference in network experiments: regression-based analysis and design-based properties. *arXiv preprint arXiv:2309.07476*.

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

Ghanem, D., Hirshleifer, S. and Ortiz-Beccera, K. (2023). Testing attrition bias in field experiments. *Journal of Human Resources*.

Glennerster, R. and Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S. and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, **118** e2014602118.

Hahn, J., Hirano, K. and Karlan, D. (2011). Adaptive Experimental Design Using the Propensity Score. *Journal of Business & Economic Statistics*, **29** 96–108.

Halloran, M. E. and Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* 142–151.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). Sample survey methods and theory. vol. i. methods and applications.

Harshaw, C., Middleton, J. A. and Sävje, F. (2021). Optimized variance estimation under interference and complex experimental designs. *arXiv preprint arXiv:2112.01709*.

Heckman, J., Pinto, R. and Shaikh, A. M. (2024). Dealing with imperfect randomization: Inference for the highscope perry preschool program. *Journal of Econometrics* 105683.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* 153–161.

Higbee, S. (2023). Policy learning with new treatments. *arXiv preprint arXiv:2210.04703*.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, **77** 1683–1701.

Hirano, K. and Porter, J. R. (2020). Asymptotic analysis of statistical decision rules in econometrics. In *Handbook of econometrics*, vol. 7. Elsevier, 283–354.

Hirano, K. and Porter, J. R. (2023). Asymptotic representations for sequential decisions, adaptive experiments, and batched bandits. *arXiv preprint arXiv:2302.03117*.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. In *The Collected Works of Wassily Hoeffding*. Springer, 247–271.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, **95** 77–84.

Hu, F. and Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons.

Hu, Y., Li, S. and Wager, S. (2022). Average direct and indirect causal effects under interference. *Biometrika*, **109** 1165–1172.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, **103** 832–842.

Imai, K., Jiang, Z. and Malani, A. (2021). Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, **116** 632–644.

Imai, K., King, G. and Nall, C. (2009). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, **24** 29–53.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62** 467–475.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jiang, L., Liu, X., Phillips, P. C. and Zhang, Y. (2021). Bootstrap Inference for Quantile Treatment Effects in Randomized Experiments with Matched Pairs. *The Review of Economics and Statistics* 1–47.

Jiang, L., Phillips, P. C., Tao, Y. and Zhang, Y. (2023). Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics*, **234** 758–776.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 85–112.

Kallus, N. (2021). On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **83** 404–409.

Karlan, D. and Appel, J. (2017). *Failing in the field: What we can learn when field research goes wrong*. Princeton University Press.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, **89** 113–132.

Kato, M., Okumura, K., Ishihara, T. and Kitagawa, T. (2024). Contextual fixed-budget best arm identification: Adaptive experimental design with policy learning. *arXiv preprint arXiv:2401.03756*.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, **17** 3008–3049.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, **21** 272–304.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, **86** 591–616.

Kitagawa, T. and Tetenov, A. (2021). Equality-minded treatment choice. *Journal of Business & Economic Statistics*, **39** 561–574.

Kitagawa, T. and Wang, G. (2023). Who should get vaccinated? individualized allocation of vaccines over sir network. *Journal of Econometrics*, **232** 109–131.

Klar, N. and Donner, A. (1997). The merits of matching in community intervention trials: a cautionary tale. *Statistics in medicine*, **16** 1753–1764.

Kock, A. B., Preinerstorfer, D. and Veliyev, B. (2023). Treatment recommendation with distributional targets. *Journal of Econometrics*, **234** 624–646.

Kuang, X. and Wager, S. (2023). Weak signal asymptotics for sequentially randomized experiments. *Management Science*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lee, S. and Shaikh, A. M. (2014). Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of Progresa on School Enrollment. *Journal of Applied Econometrics*, **29** 612–626.

Lehmann, E. and Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer Texts in Statistics, Springer International Publishing, Cham.

Leung, M. P. (2020). Treatment and spillover effects under network interference. *Review of Economics and Statistics*, **102** 368–380.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, **90** 267–293.

Leung, M. P. (2023). Design of cluster-randomized trials with cross-cluster interference. *arXiv preprint arXiv:2310.18836*.

Li, H. H. and Owen, A. B. (2023). Double machine learning and design in batch adaptive experiments. *arXiv preprint arXiv:2309.15297*.

Li, K.-C. (1983). Minimaxity for Randomized Designs: Some General Results. *Annals of Statistics*, **11** 225–239.

Li, S. and Wager, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, **50** 2334–2358.

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** 1759–1769.

Li, X., Ding, P. and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, **115** 9157–9162.

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

Lieber, J. (2022). Estimating concentration parameters for bandit algorithms.

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

List, J. A. (2023). The voltage effect. *Business Economics*, **58** 3–8.

List, J. A., Shaikh, A. M. and Vayalinkal, A. (2023). Multiple testing with covariate adjustment in experimental economics. *Journal of Applied Econometrics*, **38** 920–939.

List, J. A., Shaikh, A. M. and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, **22** 773–793.

Liu, H., Ren, J. and Yang, Y. (2021). Randomization-based joint central limit theorem and efficient covariate adjustment in stratified 2^K factorial experiments. *arXiv preprint arXiv:2103.04050*.

Liu, J. (2023). Inference for two-stage experiments under covariate-adaptive randomization. *arXiv preprint arXiv:2301.09016*.

Lu, X., Liu, T., Liu, H. and Ding, P. (2023). Design-based theory for cluster rerandomization. *Biometrika*, **110** 467–483.

Ma, W., Tu, F. and Liu, H. (2022). Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective. *Statistics in Medicine*, **41** 5645–5661.

Machado, C., Shaikh, A. M. and Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*, **212** 522–555.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *The American Economic Review*, **80** 319–323.

Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, **72** 1221–1246.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, **16** S1–S23.

Mbakop, E. and Tabord-Meehan, M. (2021). Model Selection for Treatment Choice: Penalized Welfare Maximization. *Econometrica*, **89** 825–848.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, **6** 39–75.

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

Munro, E., Wager, S. and Xu, K. (2021). Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*.

Muralidharan, K. and Niehaus, P. (2017). Experimentation at Scale. *Journal of Economic Perspectives*, **31** 103–124.

Negi, A. and Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, **40** 504–534.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, **5** 465–472.

Park, C. and Kang, H. (2023). Assumption-lean analysis of cluster randomized trials in infectious diseases for intent-to-treat effects and network effects. *Journal of the American Statistical Association*, **118** 1195–1206.

Pashley, N. E. and Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, **46** 271–296.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Classics in Applied Mathematics, Society for Industrial and Applied Mathematics.

Rafi, A. (2023). Efficient Semiparametric Estimation of Average Treatment Effects Under Covariate Adaptive Randomization. ArXiv:2305.08340 [econ] version: 1, URL http://arxiv.org/abs/2305.08340.

Reichardt, C. S. and Gollob, H. F. (1999). Justifying the use and increasing the power of at test for a randomized experiment with a convenience sample. *Psychological methods*, **4** 117.

Ren, J. (2023). Model-assisted complier average treatment effect estimates in randomized experiments with noncompliance. *Journal of Business & Economic Statistics* 1–12.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, **90** 106–121.

Romano, J. P., Ritzwoller, D. and Shaikh, A. M. (2024). Randomization inference: Theory and applications.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, **24** 404–447.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, **100** 94–108.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *Annals of Statistics*, **38** 598–633.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons.

Russo, D. (2016). Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*. PMLR, 1417–1418.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, **17** 1–30.

Savage, L. J. (1951). The Theory of Statistical Decision. *Journal of the American Statistical Association*, **46** 55–67.

Sävje, F. (2024). Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, **111** 1–15.

Sävje, F., Aronow, P. and Hudgens, M. (2021). Average treatment effects in the presence of unknown interference. *Annals of statistics*, **49** 673.

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

Schochet, P. Z., Pashley, N. E., Miratrix, L. W. and Kautz, T. (2021). Design-based ratio estimators and central limit theorems for clustered, blocked rcts. *Journal of the American Statistical Association* 1–12.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, **151** 70–81.

Su, F. and Ding, P. (2021). Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **83** 994–1015.

Sun, L. (2021). Empirical welfare maximization with constraints. *arXiv preprint arXiv:2103.15298*, **2**.

Tabord-Meehan, M. (2023). Stratification Trees for Adaptive Randomisation in Randomised Controlled Trials. *The Review of Economic Studies*, **90** 2646–2673.

Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*, **27** 4658–4677.

Tu, F., Ma, W. and Liu, H. (2023). A unified framework for covariate adjustment under stratified randomization. *arXiv preprint arXiv:2312.01266*.

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, **237** 105237.

Viviano, D. (2019). Policy targeting under network interference. *arXiv preprint arXiv:1906.10258*.

Viviano, D. (2020). Experimental design under network interference. *arXiv preprint arXiv:2003.08421*.

Viviano, D. (2022). Policy design in experiments with unknown interference. Tech. rep., working paper.

Viviano, D., Lei, L., Imbens, G., Karrer, B., Schrijvers, O. and Shi, L. (2023). Causal clustering: design of cluster experiments under network interference. *arXiv preprint arXiv:2310.14983*.

Wager, S. and Xu, K. (2021). Experimenting in equilibrium. *Management Science*, **67** 6694–6715.

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics* 165–205.

Wang, B., Park, C., Small, D. S. and Li, F. (2024). Model-robust and efficient covariate adjustment for cluster-randomized experiments. *Journal of the American Statistical Association* 1–13.

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M. and Rosenblum, M. (2023). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, **118** 1152–1163.

Wei, W., Ma, X. and Wang, J. (2024). Fair adaptive experiments. *Advances in Neural Information Processing Systems*, **36**.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*. John Wiley & Sons.

Wu, E. and Gagnon-Bartsch, J. A. (2018). The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation review*, **42** 458–488.

Wu, J. and Ding, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, **116** 1898–1913.

Zhang, K., Janson, L. and Murphy, S. (2021). Statistical inference with m-estimators on adaptively collected data. *Advances in neural information processing systems*, **34** 7460–7471.

Zhang, L. and Ma, W. (2023). Interaction tests with covariate-adaptive randomization. *arXiv preprint arXiv:2311.17445*.

Zhang, Y. and Zheng, X. (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics*, **11** 957–982.

Zhao, A. and Ding, P. (2021a). Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*, **225** 278–294.

Zhao, A. and Ding, P. (2021b). No star is good news: A unified look at rerandomization based on p-values from covariate balance tests. *arXiv preprint arXiv:2112.10545*.