Design-Based Uncertainty for Quasi-Experiments*

Ashesh Rambachan† Jonathan Roth‡

April 13, 2022

Abstract

Social scientists are often interested in estimating causal effects in settings where all units in the population are observed (e.g. all 50 US states). Design-based approaches to uncertainty, which view the realization of treatment assignments as the source of randomness, may be more appealing than standard sampling-based approaches in such contexts. This paper develops a design-based theory of uncertainty that is suitable for analyzing difference-in-differences (DiD) and instrumental variables (IV) estimators, which are typically employed in settings where completely random assignment of treatment is implausible. We consider a model where treatment status is stochastic but the probability of receiving treatment can vary arbitrarily across units. As a building block, we first show that the simple difference-in-means (SDIM) estimator is unbiased for a design-based analog of the average treatment effect on treated (ATT) under a design-based analog to mean-independence of treatment and untreated potential outcomes. We further show that the usual standard errors for the SDIM are valid but potentially conservative, even under constant treatment effects. Our results imply that the DiD estimator is unbiased and clustered standard errors are valid (but potentially conservative) under a design-based analog to parallel trends. Likewise, the IV estimator is consistent, and its standard errors are asymptotically valid, for a re-weighted local average treatment effect (LATE) under orthogonality conditions that are weaker than complete random assignment of the instrument.

*We thank Isaiah Andrews, Iavor Bojinov, Kevin Chen, Peng Ding, Pedro Sant'Anna, Yotam Shem-Tov, Neil Shephard, Tymon Słoczyński, and Chris Walker for helpful comments and suggestions. Rambachan gratefully acknowledges support from the NSF Graduate Research Fellowship under Grant DGE1745303.

†Harvard University, Department of Economics. Email: asheshr@g.harvard.edu

‡Brown University. Email: jonathanroth@brown.edu

1 Introduction

Standard econometric analyses of causal effects typically view the data obtained by the econometrician as a random sample from a larger superpopulation. This sampling-based view may be unnatural in economic contexts where the entire population of interest is observed, such as when the researcher has state-level data on all 50 US states or administrative data for an entire state or country (Manski and Pepper, 2018). In these settings, it may be more attractive to view uncertainty as purely design-based, i.e. arising due to the stochastic nature of the treatment assignment for a finite population.

A celebrated literature in statistics, dating to Neyman (1923) and Fisher (1935), has analyzed randomized experiments from such a design-based perspective. An important classical result is that in a completely randomized experiment, the "usual" (heteroskedasticity-robust) standard errors remain valid for the average treatment effect (ATE), but are conservative if there are heterogeneous treatment effects. The design-based view of uncertainty has received substantial recent interest in both statistics (e.g. Imbens and Rubin (2015), Aronow and Middleton (2015), Savje and Delevoye (2020)) and econometrics (e.g. Abadie, Athey, Imbens and Wooldridge (2017), Abadie, Athey, Imbens and Wooldridge (2020), Bojinov, Rambachan and Shephard (2021), Roth and Sant'Anna (2021), Xu (2021)). Existing results in this literature have primarily focused on the case where treatment is randomly assigned with known probabilities, as is the case when a researcher runs a controlled experiment.

However, the assumption of random assignment of treatment with known probabilities will often be implausible in settings frequently studied by economists and social scientists. In such cases, researchers often wish to impose weaker assumptions on treatment assignment, which (from the sampling-based view) identify an average treatment effect only for a subset of the population. For example, the parallel trends assumption allows for identification of an average treatment effect on the treated (ATT), while a standard set of assumptions (Angrist and Imbens, 1994; Angrist, Imbens and Rubin, 1996) for instrumental variables analyses ensure identification of a local average treatment effect (LATE).

This paper addresses the following questions: Are there design-based analogs to the identifying assumptions used in "quasi-experimental" research designs like DiD and IV? And if so, are the usual standard errors from these designs also valid from the design-based view? In short, our answer to both of these questions is "yes." Formalizing this answer, however, requires appropriate generalizations of the identifying assumptions and target parameters to the design-based context, as well as new theoretical results to establish the validity of standard inference methods.

We begin by introducing a model of uncertainty that is suitable for analyzing quasi-experiment strategies from a finite-population perspective. Our model is design-based, in the sense that the source of randomness in the data is the stochastic assignment of treatment. However, we allow for the possibility that each unit i has idiosyncratic probability p_i of receiving a binary treatment, where p_i may not be known to the researcher. In this sense, our model allows for the possibility that the "quasi-experimental" research design may not, in fact, mimic completely random assignment.

As a building block for studying other estimators, we begin with an analysis of the simple difference-in-means estimator (SDIM) in Section 3. We first establish a finite-population analog to the omitted variable bias formula, which decomposes the expectation of the SDIM into two terms: (i) a design-based analog to the average treatment effect on the treated (ATT), and (ii) a bias term equal to the finite-population covariance between the unit-specific treatment probabilities and their untreated potential outcomes. We also derive intuitive formulas for the variance of the SDIM statistic and establish a central limit theorem under "large finite populations asymptotics"' as in Li and Ding (2017); Abadie et al. (2017, 2020). Our results imply that the usual heteroskedasticity-robust standard errors are consistent for an upper bound on the variance of the SDIM estimator. An interesting feature of our setting – which is not present in completely randomized experiments – is that the standard variance estimator may be conservative even under constant treatment effects if treatment probabilities differ across units. Thus, standard confidence intervals deliver asymptotically conservative inference for the finite-population ATT when the unit-specific treatment probabilities are orthogonal to the potential outcomes.

In Section 4, we show that our results for the SDIM have immediate implications for the difference-in-differences (DiD) estimator. We show that the DiD estimator is unbiased for the finite-population ATT under a finite-population analog to the well-known "parallel trends" assumption in the sampling-based literature (e.g., see Chapter 5 of Angrist and Pischke (2009)). Our results thus help bridge the gap between the sampling-based literature on DiD and recent work by Athey and Imbens (2018) and Roth and Sant'Anna (2021), who study DiD from a design-based perspective but assume completely random treatment timing. Importantly, our results also imply that the widely used cluster-robust standard errors (Bertrand, Duflo and Mullainathan, 2004) are asymptotically valid from the design-based view, although interestingly, may be conservative even under homogeneous treatment effects.

Finally, in Section 5, we study the properties of the two-stage least squares estimator (2SLS) with a binary instrument Z_i and binary treatment D_i. The stochastic nature of the data now arises due to the assignment of the instrument Z_i, holding fixed the potential out-

comes $Y(d)$ and the potential treatments $D(z)$. We provide an intuitive expression for the IV estimand allowing for an arbitrary relationship between the probability that $Z_i = 1$ and the potential outcomes. In the case where the instrument is completely randomly assigned, our expression reduces to a local average treatment effect (LATE), as in Angrist and Imbens (1994) and Angrist et al. (1996) from the sampling perspective, and Kang, Peck and Keele (2018) from the design-based view. Our results imply, however, that the IV estimand has an interpretation as an instrument-propensity reweighted LATE under weaker orthogonality conditions that do not impose that the instrument be completely randomly assigned. Likewise, we show that standard inference methods yield asymptotically conservative inference for this estimand under "strong instrument" asymptotics.

2 A Finite Population Model For Quasi-Experiments

Consider a finite population of N units. Let D_i denote a binary indicator for whether unit i adopts a treatment of interest. Units are associated with potential outcomes $Y_i(1), Y_i(0)$, under treatment and control respectively, and the observed outcome equals $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. Following the literature on completely random experiments (e.g. Imbens and Rubin, 2015), the potential outcomes and number of treated units ($N_1 = \sum_i D_i$) are treated as fixed (or conditioned on), and the stochastic nature of the data arises only due to the stochastic assignment of treatment.

Treatment Assignment. We assume that treatment assignment follows the following data-generating process.

Assumption 1 (Assignment of treatment)**.** *The distribution of the vector of treatment assignments $D = (D_1, ..., D_N)'$ is given by*

$$\mathbb{P}\left(D = d \,\middle|\, \sum_i D_i = N_1\right) = C \prod_i p_i^{d_i}(1 - p_i)^{1-d_i} \tag{1}$$

for all $d \in \{0, 1\}^N$ such that $\sum_i d_i = N_1$, and zero otherwise.

If each unit i is independently assigned to treatment with idiosyncratic probability p_i, then equation (1) corresponds with the distribution of D conditional on the number of treated units N_1. Importantly, we do not assume that the probabilities p_i are known to the researcher. We refer to an assignment mechanism satisfying Assumption 1 as a *rejective assignment mechanism*, since it parallels what Hajek (1964) refers to as rejective sampling, in which units are sampled from a finite population only if $D_i = 1$ and D has the distribution given

in (1). (Rejective sampling is also sometimes referred to as conditional Poisson sampling.) Assumption 1 nests as a special case a completely randomized experiment, where all units have the same probability of treatment ($p_i \equiv \bar{p}$).

However, in social science applications where a design-based perspective may be conceptually appealing, it may often be implausible that all units have the same probability of adopting treatment. Assumption 1 addresses this issue by allowing for an arbitrary relationship between the idiosyncratic probabilities p_i and the potential outcomes. For example, the D_i could be generated by a Heckman (1976)-style selection model in which

$$D_i = 1\left[g(X_i, Y_i(1), Y_i(0)) + \epsilon_i \geqslant 0\right],$$

where X_i are fixed individual characteristics and $g(\cdot)$ is a possibly unknown link function. The random variable ϵ_i is a stochastic idiosyncratic error that could correspond with preference shocks or expectational errors. We would then have that $p_i = P(\epsilon_i \geqslant -g(X_i, Y_i(1), Y_i(0)))$. For example, D_i could be whether a state adopts a policy such as an increase in the minimum wage, Y_i could be state-level employment, X_i could be the general partisan-leanings of the state, and ϵ_i could be idiosyncratic political factors or the legislature's mis-perceptions about the state of the economy. We can then analyze the distribution of D_i over possible realizations of the idiosyncratic factors ϵ_i, holding constant the potential outcomes, the fixed characteristics, and the total number of states that adopt the policy change.

Notation. Define $\pi_i := \mathbb{P}\left(D_i = 1 | \sum_i d_i = N_1\right)$ to be the probability that unit i receives treatment conditional on N_1. All probability statements will be with respect to the distribution of D *conditional* on N_1 and the potential outcomes (i.e. under the DGP described in Assumption 1). For ease of exposition, we make the conditioning implicit in our notation. For example, we write $\mathbb{E}_R\left[\cdot\right] = \mathbb{E}\left[\cdot \mid \sum_i D_i = N_1\right]$ for the expectation with respect to the randomization distribution for the treatment assignment D, conditional on the number of treated units. (The subscript 'R' makes clear that the expectation is over the randomization distribution.) Analogously, we write $\mathbb{V}_R[\cdot]$ and $\mathbb{C}\text{ov}_R\left[\cdot, \cdot\right]$ for the variance and covariance respectively.

For non-stochastic weights w_i and a non-stochastic attribute X_i (such as a potential outcome), we define

$$\mathbb{E}_w\left[X_i\right] := \frac{1}{\sum_i w_i} \sum_i w_i X_i \text{ and } \mathbb{V}_w\left[X_i\right] := \frac{1}{\sum_i w_i} \sum_i w_i \left(X_i - \mathbb{E}_w\left[X_i\right]\right)^2$$

to be the finite-population weighted expectation and variance respectively. Finally, we denote

by $N_0 = N - N_1 = \sum_i (1 - D_i)$ the number of untreated units.

3 Simple Difference-in-Means

As a building block for studying other estimators, we begin by analyzing the properties of the simple difference in means (SDIM) estimator,

$$\hat{\tau} := \frac{1}{N_1} \sum_i D_i Y_i - \frac{1}{N_0} \sum_i (1 - D_i) Y_i. \tag{2}$$

which has been studied in detail in the context of completely randomized experiments, beginning with Neyman (1923).

3.1 Bias

We first turn our attention to the expectation of $\hat{\tau}$ under the treatment assignment mechanism (1). Observe that

$$\begin{aligned} \mathbb{E}_R[\hat{\tau}] &= \frac{1}{N_1} \sum_i \pi_i \underbrace{(Y_i(0) + \tau_i)}_{=Y_i(1)} - \frac{1}{N_0} \sum_i (1 - \pi_i) Y_i(0) \\ &= \underbrace{\frac{1}{N_1} \sum_i \pi_i \tau_i}_{=:\tau_{ATT}} + \frac{N}{N_0} \frac{N}{N_1} \underbrace{\left(\frac{1}{N} \sum_i \left(\pi_i - \frac{N_1}{N} \right) Y_i(0) \right)}_{=\mathbb{C}\text{ov}_1[\pi_i, Y_i(0)]}, \end{aligned} \tag{3}$$

where $\tau_i = Y_i(1) - Y_i(0)$ is unit i's causal effect. The first term in the previous display is a weighted average of the unit-specific causal effects, with weights equal to the unit-specific treatment probabilities π_i. We interpret this object as a finite-population analog to the average treatment effect on the treated since

$$\tau_{ATT} = \frac{1}{N_1} \sum_i \pi_i \tau_i = \mathbb{E}_R \left[\underbrace{\frac{1}{N_1} \sum_i D_i \tau_i}_{=SATT} \right] \tag{4}$$

is the expected value of what Imbens (2004) and Sekhon and Shem-Tov (2020) refer to as the sample average treatment effect on the treated (SATT) — i.e., the average treatment effects for the treated units in the sample — where the expectation is taken over the stochastic realization of which units are treated. The second term in (3) is the SDIM's bias for τ_{ATT} and

equals a constant times the finite-population covariance between the treatment probabilities π_i and the untreated potential outcomes $Y_i(0)$. It is straightforward to see that the bias is zero if all units are treated with the same probability (i.e. $\pi_i = N_1/N$ for all i), in which case τ_{ATT} reduces to the average treatment effect. However, equation (3) also implies that the SDIM will be unbiased for the finite-population ATT if treatment probabilities differ across units, so long as in the the finite population the π_i are uncorrelated with $Y_i(0)$.

Equation (3) may also be interpreted as a finite population version of the omitted variables bias formula for regression analyses. Defining the errors $\varepsilon_i^Y = Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)]$ and $\varepsilon_i^\tau = \tau_i - \tau_{ATT}$, we may rewrite the observed outcome for unit i as

$$Y_i = \beta_0 + D_i\tau_{ATT} + u_i, \tag{5}$$

where $\beta_0 = \mathbb{E}_{1-\pi}[Y_i(0)]$ and $u_i = \varepsilon_i^Y + D_i\varepsilon_i^\tau$. One can show that the expression derived above for $\mathbb{E}_R[\hat{\tau} - \tau_{ATT}]$ is equivalent to $\mathbb{E}_R\left[\frac{\mathbb{C}\text{ov}_1[D_i,u_i]}{\mathbb{V}\text{ar}_1[D_i]}\right]$, which in light of equation (5) coincides with the omitted variable bias formula for the coefficient on D_i in an OLS regression of Y_i on D_i and a constant.

Remark 1 (Sensitivity analysis)**.** The characterization of the bias of the SDIM estimator in (3) may be useful for conducting sensitivity analyses. For example, researchers could report how large $\mathbb{C}\text{ov}_1[\pi_i, Y_i(0)]$ would need to be to produce a bias of a magnitude large enough to change a particular conclusion (e.g. the ATT is positive). Such a sensitivity analysis is related to, but different from existing design-based sensitivity analyses. For example, Rosenbaum (1987, 2002, 2005) places bounds on the relative odds ratio of treatment between two units (i.e., $\frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)}$ for $i \neq j$) and examines the extent to which the relative odds ratio must vary across units such that we may no longer reject a particular sharp (Fisher) null of interest. Relatedly, Aronow and Lee (2013) and Miratrix, Wager and Zubizarreta (2018) consider sensitivity analysis for the finite-population mean under unequal-probability sampling where the sampling probabilities (analogous to p_i) are restricted to an interval $[\underline{p}, \overline{p}]$. In contrast, (3) suggests a simple approach for examining how the bias of the SDIM estimator for a particular weighted average treatment effect varies with the finite population covariance between treatment probabilities and untreated potential outcomes.

3.2 Variance of the SDIM

We next turn our attention to the variance of $\hat{\tau}$. To do so, it will be useful to connect the problem of estimating treatment effects to that of sampling from a finite population with unequal probabilities, which was previously studied by Hajek (1964) (among others).

Specifically, note that $\hat{\tau}$ may be re-written as

$$\hat{\tau} = \sum_i \frac{D_i}{\pi_i}\tilde{Y}_i - \frac{1}{N_0}\sum_i Y_i(0), \tag{6}$$

where $\tilde{Y}_i := \pi_i\left(\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right)$.[1] The second term on the right-hand side of the previous display is non-stochastic. The first term, on the other hand, can be viewed as a Horvitz-Thompson estimator for the population total $\sum_{i=1}^N \tilde{Y}_i$ under what Hajek (1964) refers to as rejective sampling. We can therefore make use of results from Hajek (1964) on the distribution of the Horvitz-Thompson estimator.

As described in Hajek (1964), the exact variance of $\hat{\tau}$ depends on the second-order treatment probabilities, $\pi_{ij} = \mathbb{P}_R\left(D_i = 1, D_j = 1\right)$, which in general are complicated functions of the π_i. Fortunately, simple approximations to the variance are available which become accurate when $\sum_i \mathbb{V}_R\left[D_i\right] = \sum_i \pi_i(1-\pi_i)$ is large — that is, when the sum of the variances of the individual treatment indicators is large. We note that under an overlap condition of the form $\pi_i \in [\eta, 1-\eta]$ for some $\eta > 0$, we would have that $\sum_i \mathbb{V}_R\left[D_i\right] \geqslant \eta^2 N$, although overlap of this form is not needed for our results.

Proposition 3.1. *Under Assumption 2,*

$$\mathbb{V}_R\left[\hat{\tau}\right]\left[1 + o(1)\right] = \frac{\frac{1}{N}\sum_{k=1}^N \pi_k(1-\pi_k)}{\frac{N_0}{N}\frac{N_1}{N}}\left[\frac{1}{N_1}\mathbb{V}ar_{\tilde{\pi}}\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{\tilde{\pi}}\left[Y_i(0)\right] - \frac{1}{N}\mathbb{V}ar_{\tilde{\pi}}\left[\tau_i\right]\right], \tag{7}$$

where $o(1) \to 0$ as $\sum_i \pi_i(1-\pi_i) \to \infty$ and the weights are given by $\tilde{\pi}_i = \pi_i(1-\pi_i)$.

Proof. Since $\hat{\tau}$ can be represented as a Horvitz-Thompson estimator under rejective sampling, Theorem 6.1 in Hajek (1964) implies that

$$\mathbb{V}_R\left[\hat{\tau}\right]\left[1 + o(1)\right] = \left[\sum_{k=1}^N \pi_k(1-\pi_k)\right]\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\tilde{Y}_i/\pi_i\right] = \mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right]. \tag{8}$$

Standard decomposition arguments for completely randomized experiments (e.g. Imbens and Rubin (2015)), modified to replace unweighted variances with weighted variances, yield that

$$\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right] = \frac{N}{N_1 N_0}\left(\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[Y_i(0)\right] - \frac{1}{N}\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\tau_i\right]\right),$$

which together with the previous display yields the desired result. □

[1] The theory that follows can accommodate the case where $\pi_i = 0$ for some i, if $\frac{D_i}{\pi_i}$ is defined to be 0 whenever $\pi_i = 0$.

Proposition 3.1 shows that the asymptotic variance of $\hat{\tau}$ depends on the weighted variance of the treated and untreated potential outcomes and treatment effects, where unit i is weighted proportionally to the variance of their treatment status $\mathbb{V}_R[D_i] = \pi_i(1-\pi_i)$. The leading constant term is less than or equal to one by Jensen's inequality, with equality when π_i is constant across units. Thus, in the special case of a completely random experiment, the formula in Proposition 3.1 reduces to $(1+o(1))\left(\frac{1}{N_1}\mathbb{V}\text{ar}_1[Y_i(1)] + \frac{1}{N_0}\mathbb{V}\text{ar}_1[Y_i(0)] - \frac{1}{N}\mathbb{V}\text{ar}_1[\tau_i]\right)$, which mimics the familiar formula for completely randomized experiments up to a degrees-of-freedom correction.[2]

We next derive an expression for the expectation of the standard variance estimator. Let $\hat{s}^2 = \frac{1}{N_1}\hat{s}_1^2 + \frac{1}{N_0}\hat{s}_0^2$, where

$$\hat{s}_1^2 := \frac{1}{N_1}\sum_i D_i(Y_i - \bar{Y}_1)^2, \quad \hat{s}_0^2 := \frac{1}{N_0}\sum_i (1-D_i)(Y_i - \bar{Y}_0)^2,$$

and $\bar{Y}_1 := \frac{1}{N_1}\sum_i D_i Y_i$, $\bar{Y}_0 := \frac{1}{N_0}\sum_i (1-D_i)Y_i$. The estimator $\hat{s}^2$ is the classic Neyman variance estimator, and corresponds with the natural sample analog to the variance of $\hat{\tau}$ when treatment is completely randomly assigned ($\pi_i = \frac{N_1}{N}$) and there are constant treatment effects ($\tau_i = \tau$). The following result provides an expression for the expectation of the Neyman variance estimator (again up to an $o(1)$ approximation error).

Lemma 3.1.

$$\mathbb{E}_R\left[\hat{s}^2\right](1+o(1)) = \frac{1}{N_1}\mathbb{V}ar_\pi[Y_i(1)] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}[Y_i(0)] \tag{9}$$

Proof. We will show that $\mathbb{E}_R[\hat{s}_1^2](1+o(1)) = \mathbb{V}\text{ar}_\pi[Y_i(1)]$. The equality $\mathbb{E}_R[\hat{s}_0^2](1+o(1)) = \mathbb{V}\text{ar}_{1-\pi}[Y_i(0)]$ can be obtained analogously, from which the result is immediate. Observe that

$$\begin{aligned}
\mathbb{E}_R\left[\hat{s}_1^2\right] &= \mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i Y_i^2 - \bar{Y}_1^2\right] = \mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i Y_i^2 - (\bar{Y}_1 - \mathbb{E}_\pi[Y_i(1)] + \mathbb{E}_\pi[Y_i(1)])^2\right] \\
&= \mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i Y_i^2\right] - \mathbb{E}_\pi[Y_i(1)]^2 - 2\mathbb{E}_\pi[Y_i(1)]\,\mathbb{E}_R\left[\bar{Y}_1 - \mathbb{E}_\pi[Y_i(1)]\right] - \mathbb{E}_R\left[(\bar{Y}_1 - \mathbb{E}_\pi[Y_i(1)])^2\right] \\
&= \mathbb{V}\text{ar}_\pi[Y_i(1)] - \mathbb{V}_R\left[\bar{Y}_1\right],
\end{aligned}$$

where the last equality is obtained using the fact that $\mathbb{E}_R[D_i] = \pi_i$, and hence $\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i Y_i^2\right] = \mathbb{E}_\pi[Y_i(1)^2]$ and $\mathbb{E}_R\left[\bar{Y}_1 - \mathbb{E}_\pi[Y_i(1)]\right] = 0$. Applying Theorem 6.1 in Hajek (1964) as in the

[2]The $1+o(1)$ correction is needed here because $\mathbb{V}\text{ar}_1[Y_i(d)] = \frac{1}{N}\sum_i (Y_i(d) - \mathbb{E}_1[Y_i(d)])^2$, which differs from the usual finite population variance by the degrees-of-freedom correction factor $\frac{N}{N-1}$.

proof to Proposition 3.1, we see that

$$\mathbb{V}_R\left[\bar{Y}_1\right](1+o(1)) = \left[\sum_k \pi_k(1-\pi_k)\right]\mathbb{V}\text{ar}_{\tilde{\pi}}\left[Y_i(1)/N_1\right].$$

Next, observe that

$$\begin{aligned}\left[\sum_k \pi_k(1-\pi_k)\right]\mathbb{V}\text{ar}_{\tilde{\pi}}\left[Y_i(1)/N_1\right] &= \frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)(Y_i(1)-\mathbb{E}_{\tilde{\pi}}\left[Y_i(1)\right])^2 \\ &\leqslant \frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)(Y_i(1)-\mathbb{E}_{\pi}\left[Y_i(1)\right])^2 \\ &\leqslant \frac{1}{N_1^2}\sum_i \pi_i(Y_i(1)-\mathbb{E}_{\pi}\left[Y_i(1)\right])^2 = \frac{1}{N_1}\mathbb{V}\text{ar}_{\pi}\left[Y_i(1)\right] \\ &\leqslant \left[\sum_k \pi_k(1-\pi_k)\right]^{-1}\mathbb{V}\text{ar}_{\pi}\left[Y_i(1)\right] = o(1)\mathbb{V}\text{ar}_{\pi}\left[Y_i(1)\right]\end{aligned}$$

where the first inequality uses the fact that $\mathbb{E}_{\tilde{\pi}}\left[Y_i(1)\right] = \arg\min_u \sum_i \pi_i(1-\pi_i)(Y_i(1)-u)^2$, the second inequality uses the fact that $\pi_i(1-\pi_i) \leqslant \pi_i$, and the third inequality uses the fact that $N_1 = \sum_i \pi_i \geqslant \sum_i \pi_i(1-\pi_i)$. Combining the previous three displays, we see that $\mathbb{E}_R\left[\hat{s}_1^2\right] = (1+o(1))\mathbb{V}\text{ar}_{\pi}\left[Y_i(1)\right]$, as we wished to show. □

Lemma 3.1 shows that the expectation of $\hat{s}^2$ depends on the π_i-weighted variance of the $Y_i(1)$ and the $(1-\pi_i)$-weighted variance of $Y_i(0)$.

How does the expression for the expected estimated variance in (9) relate to the expression for the true variance in (7)? Our next result shows that it is an upper bound, and provides conditions under which it is sharp.

Proposition 3.2. *Let $\mathbb{V}_R^{approx}\left[\hat{\tau}\right]$ denote the expression on the right-hand side of (7). Then*

$$\mathbb{V}_R^{approx}\left[\hat{\tau}\right] \leqslant \frac{1}{N_1}\mathbb{V}ar_{\pi}\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}\left[Y_i(0)\right] = \mathbb{E}_R\left[\hat{s}^2\right](1+o(1)), \tag{10}$$

and the bound holds with equality if and only if

$$\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right] = \frac{1}{N_1}\mathbb{E}_{\pi}\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{E}_{1-\pi}\left[Y_i(0)\right] \tag{11}$$

$$\frac{\pi_i}{N_1/N}Y_i(1) - \frac{1-\pi_i}{N_0/N}Y_i(0) = \frac{\pi_i}{N_1/N}\mathbb{E}_{\pi}\left[Y_i(1)\right] - \frac{1-\pi_i}{N_0/N}\mathbb{E}_{1-\pi}\left[Y_i(0)\right] \text{ for all } i. \tag{12}$$

Proof. From (8), we see that the right-hand side of (7) is equivalent to

$$\sum_{i=1}^{N} \pi_i(1-\pi_i)\left(\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)-\left(\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)\right]\right)\right)^2.$$

Since for any X, $\mathbb{E}_{\tilde{\pi}}[X] = \arg\min_{\mu} \sum_{i=1}^{N} \pi_i(1-\pi_i)(X_i-\mu)^2$, it follows that this is bounded above by

$$\sum_{i=1}^{N} \pi_i(1-\pi_i)\left(\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)-\left(\mathbb{E}_{\pi}\left[\frac{1}{N_1}Y_i(1)\right]+\mathbb{E}_{1-\pi}\left[\frac{1}{N_0}Y_i(0)\right]\right)\right)^2, \tag{13}$$

and the bound is strict if and only if

$$\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)\right] = \frac{1}{N_1}\mathbb{E}_{\pi}[Y_i(1)]+\frac{1}{N_0}\mathbb{E}_{1-\pi}[Y_i(0)].$$

Let $\dot{Y}_i(1) = Y_i(1) - \mathbb{E}_{\pi}[Y_i(1)]$ and $\dot{Y}_i(0) = Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)]$. Then the expression in (13) can be written as

$$\begin{aligned}
&\sum_{i=1}^{N} \pi_i(1-\pi_i)\left(\frac{1}{N_1}\dot{Y}_i(1)+\frac{1}{N_0}\dot{Y}_i(0)\right)^2 \\
&= \left[\frac{1}{N_1^2}\sum_{i=1}^{N}\pi_i\dot{Y}_i(1)^2+\frac{1}{N_0^2}\sum_{i=1}^{N}(1-\pi_i)\dot{Y}_i(0)^2-\right. \\
&\left.\frac{1}{N_1^2}\sum_{i=1}^{N}\pi_i^2\dot{Y}_i(1)^2-\frac{1}{N_0^2}\sum_{i=1}^{N}(1-\pi)^2\dot{Y}_i(0)^2+\frac{2}{N_1N_0}\sum_{i=1}^{N}\pi_i(1-\pi_i)\dot{Y}_i(1)\dot{Y}_i(0)\right] \\
&= \left[\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\pi}[Y_i(1)]+\frac{1}{N_0}\mathbb{V}\mathrm{ar}_{1-\pi}[Y_i(0)]-\frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{\pi_i}{N_1/N}\dot{Y}_i(1)-\frac{1-\pi_i}{N_0/N}\dot{Y}_i(0)\right)^2\right],
\end{aligned}$$

from which the result is immediate. □

For the special case of a completely randomized experiment, Proposition 3.2 reduces to the classic result that the Neyman variance is conservative if and only if the variance of treatment effects is positive. Interestingly, however, Proposition 3.2 suggests that when treatment probabilities are unequal across units, the Neyman variance will typically be conservative even with homogeneous treatment effects, as captured in the following corollary.

Corollary 3.1. *If treatment effects are constant, $Y_i(1) = \tau + Y_i(0)$ for all i, and $\mathbb{E}_R[\hat{\tau}] = \tau$, then the bound in Proposition 3.2 is only strict if $\pi_i = \frac{N_1}{N}$ for all i such that $Y_i(0) \neq \mathbb{E}_{1-\pi}[Y_i(0)]$.*

Proof. Note that we can re-write (12) as

$$\frac{\pi_i}{N_1}(Y_i(1) - \mathbb{E}_\pi[Y_i(1)]) - \frac{1-\pi_i}{N_0}(Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)]) = 0 \text{ for all } i.$$

Under constant effects, $\tau_{ATT} = \tau$. Further, from display (3), we see that if $\mathbb{E}_R[\hat{\tau}] = \tau_{ATT}$, then $\mathbb{E}_\pi[Y_i(0)] = \mathbb{E}_{1-\pi}[Y_i(0)]$. Additionally, under the constant effects assumption, $Y_i(1) - \mathbb{E}_\pi[Y_i(1)] = Y_i(0) - \mathbb{E}_\pi[Y_i(0)]$, and hence $Y_i(1) - \mathbb{E}_\pi[Y_i(1)] = Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)]$. Substituting into (12) and re-arranging terms, we obtain that

$$\left(\frac{\pi_i}{N_1} - \frac{1-\pi_i}{N_0}\right)(Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)]) = 0 \text{ for all } i,$$

from which the result follows.

□

3.3 Central Limit Theorem and Variance Consistency

Our results so far imply that the typical variance estimator will be conservative in the sense that its expectation is larger than the true variance of $\hat{\tau}$ (again, up to an $o(1)$ approximation error). Intuitively, this will imply that standard confidence intervals based on $\hat{s}$ will be conservative for $\mathbb{E}_R[\hat{\tau}]$ if (i) $\hat{\tau}$ is approximately normally distributed, and (ii) $\hat{s}^2$ is close to its expectation. Our next results show that this will indeed be the case in large populations satisfying certain regularity conditions.

To formalize this intuition, we follow Hajek (1964) for sampling from a finite-population and Freedman (2008b,a), Lin (2013), and Li and Ding (2017) for randomized experiments, and consider a sequence of finite populations of increasing size. Specifically, we consider sequences of populations indexed by m of size N_m, with N_{1m} treated units, potential outcomes $\{Y_{im}(d) : d = 1, 2; i = 1, ..., N_m\}$, and assignment weights $p_{1m}, ..., p_{N_m}$. For brevity, we leave the subscript m implicit in our notation (as in the papers cited above); all limits are implicitly taken as $m \to \infty$. We then establish a central limit theorem (CLT) and variance consistency result under regularity conditions on the sequence of finite populations. These results provide an approximation to the properties of $\hat{\tau}$ for finite populations with a sufficiently large number of units. Indeed, as we show in Proposition 3.5 below, these asymptotic results translate to Berry-Esseen type bounds on the quality of the CLT in any finite population of fixed size.

Regularity conditions. We impose the following assumption on the sequence of populations.

Assumption 2. *The sequence of populations satisfies* $\sum_{i=1}^{N} \pi_i(1-\pi_i) \to \infty$.

Recall that $\pi_i(1-\pi_i)$ is the variance of the Bernouli random variable D_i, so Assumption 2 implies that the sum of the variances of the D_i grows large. Assumption 2 also implies that both N_1 and N_0 go to infinity, since $\sum_{i=1}^{N} \pi_i(1-\pi_i) \leqslant \min\{\sum_i \pi_i, \sum_i (1-\pi_i)\} = \min\{N_1, N_0\}$. Note that Assumption 2 is trivially satisfied under the overlap condition $\pi_i \in [\eta, 1-\eta]$, although overlap for all units is not necessary for Assumption 2 to hold, and indeed Assumption 2 allows for $\pi_i = 0$ or $\pi_i = 1$ for some units.

Our next assumption is similar to the Lindeberg condition for the standard Lindeberg central limit theorem, and imposes that the weighted finite-population variance of $\tilde{Y}_i$ is not dominated by a small number of observations.

Assumption 3. *Let* $\tilde{Y}_i = \frac{1}{N_1} Y_i(1) + \frac{1}{N_0} Y_i(0)$, *and assume* $\sigma^2_{\tilde{\pi}} = \mathbb{V}ar_{\tilde{\pi}}\left[\tilde{Y}_i\right] > 0$. *Suppose that for all* $\epsilon > 0$,

$$\frac{1}{\sigma^2_{\tilde{\pi}}} \mathbb{E}_{\tilde{\pi}}\left[\left(\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right]\right)^2 1\left[\left|\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right]\right| \geqslant \sqrt{\sum_i \pi_i(1-\pi_i)} \cdot \sigma_{\tilde{\pi}} \epsilon\right]\right] \to 0.$$

Finally, we introduce an assumption that bounds the influence that any single observation has on the π or $1-\pi$ weighted variances of the potential outcomes. This generalizes the assumptions in Theorem 1 in Li and Ding (2017), which establishes consistency of the Neyman variance under equal-probability sampling from a finite population, to the case of unequal treatment probabilities.

Assumption 4. *Define* $m_N(1) := \max_{1 \leqslant i \leqslant N} \left(Y_i(1) - \mathbb{E}_{\pi}\left[Y_i(1)\right]\right)^2$, *and analogously* $m_N(0) := \max_{1 \leqslant i \leqslant N} \left(Y_i(0) - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]\right)^2$. *Assume that,*

$$\frac{1}{N_1} \frac{m_N(1)}{\mathbb{V}ar_{\pi}\left[Y_i(1)\right]} \to 0 \text{ and } \frac{1}{N_0} \frac{m_N(0)}{\mathbb{V}ar_{1-\pi}\left[Y_i(0)\right]} \to 0.$$

Central limit theorem and variance consistency. Under the conditions introduced above, we can formally establish a CLT and variance consistency result.

Proposition 3.3. *Suppose Assumptions 2 and 3 hold. Then,*

$$\frac{\hat{\tau} - \mathbb{E}_R\left[\hat{\tau}\right]}{\sqrt{\mathbb{V}_R\left[\hat{\tau}\right]}} \xrightarrow{d} \mathcal{N}\left(0,\, 1\right).$$

Proof. Viewing $\hat{\tau}$ as a Horwitz-Thompson estimator under rejective sampling as in (6), the result follows immediately from Theorem 1 in Berger (1998).[3] □

[3] Hajek (1964) states a similar result where the Horvitz-Thompson estimator uses an approximation to

Proposition 3.4. *Under Assumptions 2 and 4,*

$$\frac{\hat{s}^2}{\left(\frac{1}{N_1}\mathbb{V}ar_{\pi}\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}\left[Y_i(0)\right]\right)} \xrightarrow{p} 1.$$

Proof. See Appendix. □

Lemmas 3.3 and 3.4 together with our results in the previous section immediately imply that confidence intervals of the form $\hat{\tau} \pm 1.96 \times \hat{s}/\sqrt{N}$ will have asymptotically correct (but potentially conservative) coverage of $\mathbb{E}_R\left[\hat{\tau}\right]$. If $\mathbb{C}\text{ov}_1\left[\pi_i, Y_i(0)\right] = 0$, then our results in Section 3.1 imply that $\mathbb{E}_R\left[\hat{\tau}\right] = \tau_{ATT}$, and thus standard CIs will be valid but potentially conservative for τ_{ATT}.

Finite-population bounds. In addition to the asymptotic results shown above, we can also obtain Berry-Esseen type bounds on the quality of the normal approximation in any given finite-population.

Proposition 3.5. *Let b_1, b_2 be positive constants, and define*

$$t = \frac{\hat{\tau} - \mathbb{E}_R\left[\hat{\tau}\right]}{\sqrt{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}}.$$

Then there exist constants k and $\bar{N}$ such that

$$\sup_y \left|P(t \leqslant y) - \Phi(y)\right| \leqslant \frac{k}{\sqrt{N}}$$

for any finite population of size $N \geqslant \bar{N}$ such that $N\mathbb{V}_R^{approx}\left[\hat{\tau}\right] = b_1$ and

$$\mathbb{E}_1\left[\left(\frac{N}{N_1}Y_i(1) + \frac{N}{N_0}Y_i(0)\right)^4\right] < b_2.$$

Proof. Again viewing $\hat{\tau}$ as a Horvitz-Thompson estimator under rejective sampling, the result follows immediately from Theorem 3 in Berger (1998). □

The result in Proposition 3.5 is attractive in the sense that it shows that the distribution of $\hat{\tau}$ will be approximately normally distributed in finite populations that are sufficiently large (relative to the fourth moment of the potential outcomes), without appealing to arguments involving a sequence of populations.

the π_i in terms of the p_i.

3.4 Multiple Outcomes

The results for scalar outcomes Y_i extend easily to the multiple outcome case with $\mathbf{Y}_i \in \mathbb{R}^K$. This is relevant when we observe multiple outcome measures in a cross-section, or we observe the same outcome measure for multiple periods (or both). We use the extension to multiple outcomes in our finite population analysis of difference-in-differences and instrumental variables settings later in the paper.

We extend our notation from the scalar case, so that $\mathbf{Y}_i \in \mathbb{R}^K$, and for a fixed vector-valued characteristic $\mathbf{X}_i$ (e.g a function of the potential outcomes), $\mathbb{E}_w[\mathbf{X}_i] := \frac{1}{\sum_i w_i}\sum_i w_i \mathbf{X}_i$ and $\mathbb{V}\text{ar}_w[\mathbf{X}_i] = \frac{1}{\sum_i w_i}\sum_i (\mathbf{X}_i - \mathbb{E}_w[\mathbf{X}_i])(\mathbf{X}_i - \mathbb{E}_w[\mathbf{X}_i])'$. In particular, define

$$
\begin{aligned}
&S_{1,w} := \mathbb{V}\text{ar}_w[\mathbf{Y}_i(1)], \quad S_{0,w} := \mathbb{V}\text{ar}_w[\mathbf{Y}_i(0)], \\
&S_{10,w} := \mathbb{E}_w[(\mathbf{Y}_i(1) - \mathbb{E}_w[\mathbf{Y}_i(1)])(\mathbf{Y}_i(0) - \mathbb{E}_w[\mathbf{Y}_i(0)])']
\end{aligned}
$$

to be the weighted finite population variances and covariance of $\mathbf{Y}_i(1)$ and $\mathbf{Y}_i(0)$. Additionally, the vector-valued ATT is defined as, $\boldsymbol{\tau}_{ATT} := \frac{1}{N_1}\sum_i \pi_i(\mathbf{Y}_i(1) - \mathbf{Y}_i(0))$, and consider the vector-valued SDIM estimator $\hat{\boldsymbol{\tau}} = \frac{1}{N_1}\sum_i D_i \mathbf{Y}_i(1) - \frac{1}{N_0}\sum_i (1-D_i)\mathbf{Y}_i(0)$. We also generalize the variance estimators introduced above,

$$
\begin{aligned}
&\hat{\mathbf{s}} := \frac{1}{N_1}\hat{\mathbf{s}}_1 + \frac{1}{N_0}\hat{\mathbf{s}}_0, \\
&\hat{\mathbf{s}}_1 := \frac{1}{N_1}\sum_i D_i(\mathbf{Y}_i - \bar{\mathbf{Y}}_1)(\mathbf{Y}_i - \bar{\mathbf{Y}}_1)', \quad \hat{\mathbf{s}}_0 := \frac{1}{N_0}\sum_i (1-D_i)(\mathbf{Y}_i - \bar{\mathbf{Y}}_0)(\mathbf{Y}_i - \bar{\mathbf{Y}}_0)',
\end{aligned}
$$

where $\bar{\mathbf{Y}}_1 := \frac{1}{N_1}\sum_i D_i \mathbf{Y}_i$ and $\bar{\mathbf{Y}}_0 := \frac{1}{N_0}\sum_i (1-D_i)\mathbf{Y}_i$.

We introduce the following assumptions on the sequence of finite populations.

Assumption 5. *Suppose that $N_1/N \to p_1 \in (0,1)$, and $S_{1,w}, S_{0,w}, S_{10,w}$ have finite limits for $w \in \{\pi, 1-\pi, \tilde{\pi}\}$.*

Assumption 6. *Assume that*

$$
\max_{1 \leqslant i \leqslant N} ||\mathbf{Y}_i(1) - \mathbb{E}_\pi[\mathbf{Y}_i(1)]||^2/N \to 0 \qquad \max_{1 \leqslant i \leqslant N} ||\mathbf{Y}_i(0) - \mathbb{E}_{1-\pi}[\mathbf{Y}_i(0)]||^2/N \to 0
$$

where $||\cdot||$ is the Euclidean norm.

Assumption 7. *Let $\tilde{\mathbf{Y}}_i = \frac{1}{N_1}\mathbf{Y}_i(1) + \frac{1}{N_0}\mathbf{Y}_i(0)$, and let λ_{min} be the minimal eigenvalue of*

$\Sigma_{\tilde{\pi}} = \mathbb{V}ar_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]$. *Assume* $\lambda_{min} > 0$ *and for all* $\epsilon > 0$,

$$\frac{1}{\lambda_{min}}\mathbb{E}_{\tilde{\pi}}\left[\left\|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right\|^2 \cdot 1\left[\left\|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right\| \geqslant \sqrt{\sum_i \pi_i(1-\pi_i)\cdot\lambda_{min}}\cdot\epsilon\right]\right] \to 0.$$

Assumption 5 requires that the fraction of treated units and the (weighted) variance and covariances of the potential outcomes have limits. Assumption 6 is a multivariate analog of Assumption 4 in that it requires that no single observation dominate the π or $(1-\pi)$-weighted variance of the potential outcomes. Assumption 7 is a multivariate generalization of the Lindeberg-type condition in Assumption 3.

Proposition 3.6 (Results for vector-valued outcomes)**.**

(1)

$$\mathbb{E}_R\left[\hat{\boldsymbol{\tau}}\right] = \boldsymbol{\tau}_{ATT} + \frac{N}{N_0}\frac{N}{N_1}\left(\frac{1}{N}\sum_i\left(\pi_i - \frac{N_1}{N}\right)\mathbf{Y}_i(0)\right).$$

(2) Under Assumptions 2, and 5,

$$\begin{aligned}\mathbb{V}_R\left[\hat{\boldsymbol{\tau}}\right] + o(N^{-1}) &= \frac{\frac{1}{N}\sum_{k=1}^N \pi_k(1-\pi_k)}{\frac{N_0}{N}\frac{N_1}{N}}\left[\frac{1}{N_1}\mathbb{V}ar_{\tilde{\pi}}\left[\mathbf{Y}_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{\tilde{\pi}}\left[\mathbf{Y}_i(0)\right] - \frac{1}{N}\mathbb{V}ar_{\tilde{\pi}}\left[\boldsymbol{\tau}_i\right]\right] \\ &\leqslant \frac{1}{N_1}\mathbb{V}ar_{\pi}\left[\mathbf{Y}_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}\left[\mathbf{Y}_i(0)\right]\end{aligned}$$

where $A \leqslant B$ *if* $B - A$ *is positive semi-definite.*

(3) Under Assumptions 2, 5, and 6,

$$\hat{\mathbf{s}}_1 - \mathbb{V}ar_{\pi}\left[\mathbf{Y}_i(1)\right] \xrightarrow{p} 0, \qquad \hat{\mathbf{s}}_0 - \mathbb{V}ar_{1-\pi}\left[\mathbf{Y}_i(0)\right] \xrightarrow{p} 0.$$

(4) Under Assumptions 2, 5, and 7,

$$\mathbb{V}_R\left[\hat{\boldsymbol{\tau}}\right]^{-\frac{1}{2}}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right) \xrightarrow{d} \mathcal{N}\left(0,\, I\right).$$

Assumption 5 implies $\Sigma_\tau = \lim_{N\to\infty} N\mathbb{V}_R\left[\hat{\boldsymbol{\tau}}\right]$ *exists, so the previous display can alternatively be written as*

$$\sqrt{N}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}\left(0,\, \Sigma_{\boldsymbol{\tau}}\right).$$

Proof. See Appendix. □

4 Difference-in-Differences

In this section, we apply our results to provide a design-based analysis of difference-in-differences estimators (e.g., Chapter 5 of Angrist and Pischke (2009)). Such a design-based analysis is useful since applied researchers commonly use difference-in-differences estimators in quasi-experimental settings to analyze the causal effects of state-level polices in which outcomes for all 50 US states are observed, or in which administrative data is available for the full population.

Suppose we observe panel data for a population of N units for periods $t = -\underline{T}, ..., \bar{T}$. Units with $D_i = 1$ receive a treatment of interest beginning at period $t = 1$.[4] The observed outcome for unit i at period t is $Y_{it} = Y_{it}(D_i)$. We assume the treatment has no effect prior to its implementation, so that $Y_{it}(1) = Y_{it}(0)$ for all $t < 1$. In this setting, it is common to estimate the ATT in period t by

$$\hat{\beta}_t = \hat{\tau}_t - \hat{\tau}_0 \quad \text{where} \quad \hat{\tau}_t = \frac{1}{N_1}\sum_i D_i Y_{it} - \frac{1}{N_0}\sum_i (1 - D_i) Y_{it}. \tag{14}$$

Indeed, the $\hat{\beta}_t$ correspond with the coefficients from the dynamic two-way fixed effects (TWFE) or "event-study" regression specification

$$Y_{it} = \alpha_i + \phi_t + \sum_{s \neq 0} D_i \times 1[s = t] \times \beta_s + \epsilon_{it}. \tag{15}$$

From equation (14), we see that $\hat{\beta}_t$ is the difference in the SDIM estimators for the outcome in period t and period 0. Letting $\mathbf{Y}_i = (Y_{i,-\underline{T}}, ..., Y_{i,\bar{T}})'$, (3) implies that under rejective assignment,

$$\mathbb{E}_R\left[\hat{\beta}_t\right] = \tau_t + \frac{N}{N_0}\frac{N}{N_1}\mathbb{C}\text{ov}_1\left[\pi_i, Y_{it}(0) - Y_{i0}(0)\right],$$

where $\tau_t = \frac{1}{N_1}\sum_i \pi_i Y_{it}(0)$ is the ATT in period t, and we use the fact that $\tau_0 = 0$ by the no-anticipation assumption. Thus, the bias in $\hat{\beta}_t$ is proportional to the finite population covariance between π_i and trends in the untreated potential outcomes, $Y_{it}(0) - Y_{i0}(0)$. It follows that $\hat{\beta}_t$ is unbiased for τ_t over the randomization distribution if $\mathbb{C}\text{ov}_1\left[\pi_i, Y_{it}(0) - Y_{i0}(0)\right] = 0$,

[4] We focus on the case with non-staggered treatment timing, since it may be difficult to interpret the estimand of standard two-way fixed effects models under treatment effect heterogeneity and staggered treatment timing (Borusyak and Jaravel, 2016; de Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2018; Athey and Imbens, 2018). The results in this section could potentially be extended to other estimators with a more sensible interpretation under staggered timing e.g. Callaway and Sant'Anna (2019); Sun and Abraham (2020).

or equivalently, if

$$\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i(Y_{it}(0)-Y_{i0}(0))\right]=\mathbb{E}_R\left[\frac{1}{N_0}\sum_i (1-D_i)(Y_{it}(0)-Y_{i0}(0))\right],$$

which mimics the familiar "parallel trends" assumption from the sampling-based model.

Further, if the sequence of populations satisfies the assumptions in part (4) of Proposition 3.6, then

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-(\boldsymbol{\tau}+\boldsymbol{\delta}))\to_d \mathcal{N}(0,\Sigma), \tag{16}$$

where $\hat{\boldsymbol{\beta}}$ is the vector that stacks $\hat{\beta}_t$, $\Sigma=\lim_{N\to\infty} N\mathbb{V}_R\left[\hat{\beta}_t\right]$, and $\boldsymbol{\tau}$, $\boldsymbol{\delta}$ are the vectors that stack τ_t and $\delta_t=\frac{N}{N_0}\frac{N}{N_1}\mathbb{C}\text{ov}_1\left[\pi_i, Y_{it}(0)-Y_{i0}(0)\right]$. Part (3) implies that the variance estimator $\hat{\mathbf{s}}$ is asymptotically conservative for $\hat{\boldsymbol{\beta}}$. It is easily verified that $\hat{\mathbf{s}}$ corresponds with the cluster-robust variance estimator for (15) that clusters at level i (up to degrees of freedom corrections). The normal limiting model in (16) has been studied by Roth (2019) and Rambachan and Roth (2021) from a sampling-based perspective in which parallel trends may fail; our results show that it also has a sensible interpretation from a design-based perspective.

5 Instrumental Variables

In this section, we apply our results to analyze the properties of two-stage least squares instrumental variables estimators. Let $Z_i\in\{0,1\}$ be an instrument. Let $D_i(z)\in\{0,1\}$ be the potential treatment status as a function of z. Let $Y_i(d)$ be the potential outcome as a function of $d\in\{0,1\}$. Our notation $Y(d)$ encodes the so-called "exclusion restriction" that Z affects Y only through D. We observe (Y_i, D_i, Z_i) where $Y_i=Y_i(D_i(Z_i))$ and $D_i=D_i(Z_i)$. We treat Z_i as stochastic and the potential outcomes for both D and Y as fixed. The number of units with $Z_i=1$ is denoted by N_1^Z and the number of units with $Z_i=0$ is denoted by N_0^Z.

Example 1. Researchers may have data on student outcomes for all students attending public and private schools in a particular geographic area (e.g., Goodman (2008) observes data on all high school graduates in Massachusetts from 2003-2005). The instrument Z_i could be an indicator for whether a student is offered a subsidy for attending private school, D_i could be an indicator for whether a student attends private school, and Y_i could be a student's test score. We might suspect that an organization assigns scholarships essentially as-if random, but it is also plausible that they may target their offers to students that are likely to accept if offered, or who have high benefits from private school, so that $\mathbb{P}(Z_i=1)$

may be related to $Y_i(d)$ and $D_i(z)$. It is therefore instructive to consider the distribution the 2SLS estimator when Z_i is not completely randomly assigned.

In canonical IV frameworks, it is traditionally assumed that the instrument Z is independent of the potential outcomes (see Angrist and Imbens (1994); Angrist et al. (1996) for a sampling-based model, and Kang et al. (2018) for a design-based model). We instead allow for the possibility that the probability that $Z_i = 1$ may differ across units, and be arbitrarily related to the potential outcomes. In particular, we suppose that

$$\mathbb{P}\left(Z = z \middle| \sum_i Z_i = N_1^Z\right) = C \prod_i p_i^{z_i}(1-p_i)^{1-z_i} \tag{17}$$

for all $Z \in \{0,1\}^N$ such that $\sum_i z_i = N_1^Z$, and zero otherwise. Thus, the assignment of the instrument Z_i mimics the rejective assignment of D_i in (1). We update the notation to use $\mathbb{E}_{R_Z}[\cdot], \mathbb{V}_{R_Z}[\cdot]$ to denote the expectations and variances with respect to the randomization distribution of Z conditional on the number of units assigned to $Z = 1$. We also maintain the typical monotonicity assumption that is commonly imposed in IV settings.

Assumption 8 (Monotonicity)**.** *$D_i(1) \geqslant D_i(0)$ for all i.*

A common method for estimating treatment effects in an instrumental variables setting is two-stage least squares (2SLS), defined as $\hat{\beta}_{2SLS} := \hat{\tau}_{RF}/\hat{\tau}_{FS}$ with

$$\hat{\tau}_{RF} := \frac{1}{N_1^Z}\sum_i Z_i Y_i - \frac{1}{N_0^Z}\sum_i (1-Z_i)Y_i$$
$$\hat{\tau}_{FS} := \frac{1}{N_1^Z}\sum_i Z_i D_i - \frac{1}{N_0^Z}\sum_i (1-Z_i)D_i.$$

$\hat{\tau}_{RF}$ is often referred to as the "reduced-form" coefficient, whereas $\hat{\tau}_{FS}$ is referred to as the "first-stage" coefficient.

Observe that $\hat{\tau}_{RF}$ is a SDIM for the effect of Z_i on Y_i, whereas $\hat{\tau}_{FS}$ can be viewed as a SDIM for the effect of Z_i on D_i. Equation (3) thus implies that

$$\mathbb{E}_{R_Z}[\hat{\tau}_{RF}] = \frac{1}{N}\sum_i \pi_i^Z \left(Y_i(D_i(1)) - Y_i(D_i(0))\right) + \frac{N}{N_1^Z}\frac{N}{N_0^Z}\mathbb{C}\text{ov}_1\left[\pi_i^Z, Y_i(D_i(0))\right],$$

where $\mathbb{C}\text{ov}_1\left[\pi_i^Z, Y_i(D_i(0))\right] = \frac{1}{N}\sum_i \left(\pi_i^Z - \frac{N_1^Z}{N}\right) Y_i(D_i(0))$ is the finite population covariance between π_i^Z and $Y_i(D_i(0))$. Let $\mathcal{C} = \{i : D_i(1) > D_i(0)\}$ denote the set of compliers. The

previous display along with Assumption 8 imply that

$$\mathbb{E}_{R_Z}\left[\hat{\tau}_{RF}\right] = \frac{1}{N}\sum_{i\in\mathcal{C}} \pi_i^Z \left(Y_i(1) - Y_i(0)\right) + \frac{N}{N_1^Z}\frac{N}{N_0^Z}\mathbb{C}\text{ov}_1\left[\pi_i^Z, Y_i(D_i(0))\right]. \tag{18}$$

By an analogous argument for $\hat{\tau}_{FS}$, we obtain that

$$\mathbb{E}_{R_Z}\left[\hat{\tau}_{FS}\right] = \frac{1}{N}\sum_{i\in\mathcal{C}} \pi_i^Z + \frac{N}{N_1^Z}\frac{N}{N_0^Z}\mathbb{C}\text{ov}_1\left[\pi_i^Z, D_i(0)\right]. \tag{19}$$

Define $\beta_{2SLS} := \frac{\mathbb{E}_{R_Z}[\hat{\tau}_{RF}]}{\mathbb{E}_{R_Z}[\hat{\tau}_{FS}]}$.

Our earlier results imply that under suitable regularity conditions $\hat{\beta}_{2SLS}$ is normally distributed around β_{2SLS} in large populations. Let $\mathbf{Y}_i = (Y_i, D_i)'$ and define the potential outcomes $\mathbf{Y}_i(z) = (Y_i(D_i(z)), D_i(z))$. If the sequence of populations satisfies the assumptions in Proposition 3.6, part 4 then

$$\sqrt{N}\begin{pmatrix} \hat{\tau}_{RF} - \mathbb{E}_{R_Z}\left[\hat{\tau}_{RF}\right] \\ \hat{\tau}_{FS} - \mathbb{E}_{R_Z}\left[\hat{\tau}_{FS}\right] \end{pmatrix} \to_d \mathcal{N}\left(0,\, \Sigma_\tau\right),$$

where $\Sigma_\tau = \lim_{N\to\infty} N\mathbb{V}_{R_Z}\left[\begin{pmatrix} \hat{\tau}_{RF} \\ \hat{\tau}_{FS} \end{pmatrix}\right]$. Assuming further that the sequence of populations satisfies $\left(\mathbb{E}_{R_Z}\left[\hat{\tau}_{RF}\right], \mathbb{E}_{R_Z}\left[\hat{\tau}_{FS}\right]\right) \to (\tau_{RF}^*, \tau_{FS}^*)$ with $\tau_{FS}^* > 0$, then the uniform delta method (e.g., Theorem 3.8 in van der Vaart (2000)) implies that[5]

$$\sqrt{N}(\hat{\beta}_{2SLS} - \beta_{2SLS}) \to_d N(0, g'\Sigma_\tau g),$$

where g is the gradient of $h(x, y) = x/y$ evaluated at $(\tau_{RF}^*, \tau_{FS}^*)$. Proposition 3.6 likewise implies that it is possible to obtain asymptotically conservative inference for β_{2SLS} using plug-in estimates of the variance.

How should we interpret the estimand β_{2SLS}? First, note that if $\pi_i^Z \equiv \frac{N_1^Z}{N}$, so that all units receive $Z = 1$ with equal probability, then equations (18) and (19) imply that $\beta_{2SLS} = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}(Y_i(1) - Y_i(0))$, which is a design-based analog to the canonical local average treatment effect (LATE) for compliers (Angrist et al., 1996; Kang et al., 2018). Interestingly, our results show that β_{2SLS} has a general causal interpretation under the weaker assumption

[5] It is well-known in sampling-based instrumental variables settings that the delta method fails under "weak-instrument asymptotics" in which $\mathbb{E}_{R_Z}\left[\hat{\tau}_{FS}\right]$ drifts towards zero (Staiger and Stock, 1997). Similar issues apply here. However, the test static used to form Anderson-Rubin confidence intervals, which are robust to weak identification, can be written as a quadratic form in a SDIM statistic (see, e.g., Li and Ding (2017)). Our results could thus also be applied to analyze the properties of Anderson-Rubin based CIs under weak identification asymptotics.

that $\mathbb{C}\text{ov}_1\left[\pi_i^Z, Y_i(D_i(0))\right] = \mathbb{C}\text{ov}_1\left[\pi_i^Z, D_i(0)\right] = 0$, so that the probability that $Z_i = 1$ may differ across units but the finite population covariance between treatment probabilities and $D_i(0)$ and $Y_i(D_i(0))$ is equal to zero. Under this assumption, we have that

$$\beta_{2SLS} = \frac{1}{\sum_{i \in \mathcal{C}} \pi_i^Z} \sum_{i \in \mathcal{C}} \pi_i^Z \left(Y_i(1) - Y_i(0)\right).$$

The parameter β_{2SLS} can then be interpreted as a π_i^Z-weighted local average treatment effect (LATE) for compliers. The weights given to each complier are proportional to the probability that $Z_i = 1$. This is intuitive, as a complier with a low probability of having $Z_i = 1$ should have little effect on the 2SLS estimator.

6 Conclusion

This paper analyzes quasi-experimental estimators from a design-based perspective where the population is treated as fixed and randomness in the data comes from the stochastic assignment of treatment. We show that the DiD estimator is unbiased for a design-based analog to the ATT under a design-based analog to the parallel trends assumption. We also show that standard inference methods are valid but potentially conservative from the design-based perspective as well. Similarly, we show that the 2SLS estimator is valid for a "re-weighted" LATE under orthogonality conditions between the instrument probabilities and potential outcomes, which are weaker than completely random assignment of the instrument.

The analysis in this paper could be extended in a variety of directions. First, the analysis might be extended to settings where the stochastic nature of the data arises both from the assignment of treatment and from sampling a subset of units from a finite population, as in Abadie et al. (2020). Like in Abadie et al. (2020), the analysis could also be extended to allow for clustered sampling or treatment assignment. Second, our results on the limiting distribution of the SDIM suggest that a variety of mis-specification robust tools and sensitivity analyses which have been developed under the assumption of asymptotic normality from a sampling-based perspective could also potentially be applied in finite population contexts as well (e.g., Armstrong and Kolesar (2018a,b); Bonhomme and Weidner (2020); Andrews, Gentzkow and Shapiro (2017, 2019)). However, the finite population setting studied here differs from the usual sampling-based approach in that the variance matrix is only conservatively estimated. It would be useful to study which guarantees of size control and/or optimality from the sampling literature are robust to this modification.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 2020, *88* (1), 265–296. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12675.

_ , _ , Guido W Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," Working Paper 24003, National Bureau of Economic Research November 2017. Series: Working Paper Series.

Andrews, Isaiah, Matthew Gentzkow, and Jesse Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 2017, *132* (4), 1553–1592.

_ , _ , and _ , "On the Informativeness of Descriptive Statistics for Structural Estimates," Technical Report 2019.

Angrist, Joshua and Guido Imbens, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Angrist, Joshua D. and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

_ , Guido W. Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Armstrong, Timothy and Michal Kolesar, "Optimal Inference in a Class of Regression Models," *Econometrica*, 2018, *86*, 655–683.

_ and _ , "Simple and Honest Confidence Intervals in Nonparametric Regression," Technical Report 2018.

Aronow, Peter M. and Donald K. K. Lee, "Interval estimation of population means under unknown but bounded probabilities of sample selection," *Biometrika*, 2013, *100* (1), 235–240.

_ and Joel A. Middleton, "A class of unbiased estimators of the average treatment effect in randomized experiments," *Journal of Causal Inference*, 2015, *1* (1), 135–154.

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *arXiv:1808.05293 [cs, econ, math, stat]*, August 2018.

Berger, Yves G., "Rate of convergence to normal distribution for the Horvitz-Thompson estimator," *Journal of Statistical Planning and Inference*, April 1998, *67* (2), 209–226.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, February 2004, *119* (1), 249–275.

Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard, "Panel Experiments and Dynamic Causal Effects: A Finite Population Perspective," *Quantitative Economics*, 2021, *12* (4), 1171–1196.

Bonhomme, Stéphane and Martin Weidner, "Minimizing Sensitivity to Model Misspecification," *arXiv:1808.05293 [econ.EM]*, 2020.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY August 2016.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," SSRN Scholarly Paper ID 3148250, Social Science Research Network, Rochester, NY March 2019.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *arXiv:1803.08807 [econ]*, March 2018. arXiv: 1803.08807.

Fisher, R. A., *The design of experiments* The design of experiments, Oxford, England: Oliver & Boyd, 1935. Pages: xi, 251.

Freedman, David A., "On Regression Adjustments in Experiments with Several Treatments," *The Annals of Applied Statistics*, 2008, *2* (1), 176–196.

_ , "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 2008, *40* (2), 180–193.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," Working Paper 25018, National Bureau of Economic Research September 2018.

Goodman, Joshua, "Who merits financial aid?: Massachusetts' Adams Scholarship," *Journal of Public Economics*, 2008, *92*, 2121–2131.

Hajek, Jaroslav, "Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population," *Annals of Mathematical Statistics*, December 1964, *35* (4), 1491–1523. Publisher: Institute of Mathematical Statistics.

Heckman, James, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," NBER Chapters, National Bureau of Economic Research, Inc 1976.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, February 2004, *86* (1), 4–29. Publisher: MIT Press.

_ **and Donald B. Rubin**, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press, 2015.

Kang, Hyunseung, Laura Peck, and Luke Keele, "Inference for instrumental variables: a randomization inference approach," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2018, *181* (4), 1231–1254. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12353.

Li, Xinran and Peng Ding, "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, October 2017, *112* (520), 1759–1769. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1295865.

Lin, Winston, "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's critique," *The Annals of Applied Statistics*, 2013, *7* (1), 295–318.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

Miratrix, Luke W., Stefan Wager, and Jose R. Zubizarreta, "Shape-constrained partial identification of a population mean under unknown probabilities of sample selection," *Biometrika*, 2018, *105* (1), 103–114.

Neyman, Jerzy, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1923, *5* (4), 465–472. Publisher: Institute of Mathematical Statistics.

Rambachan, Ashesh and Jonathan Roth, "An Honest Approach to Parallel Trends," Technical Report 2021.

Rosenbaum, Paul, "Sensitivity Analysis in Observational Studies," in B. S. Everitt and D. C. Howell, eds., *Encyclopedia of Statistics in Behavioral Science*, 2005.

Rosenbaum, Paul R., "Sensitivity analysis for certain permutation inferences in matched observational studies," Technical Report 1 1987.

_ , *Observational Studies*, Springer Science, 2002.

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *Working paper*, 2019.

_ **and Pedro H. C. Sant'Anna**, "Efficient Estimation for Staggered Rollout Designs," *arXiv:2102.01291 [econ, math, stat]*, June 2021. arXiv: 2102.01291.

Savje, Frederik and Angele Delevoye, "Consistency of the Horvitz-Thompson estimator under general sampling and experimental designs," *Journal of Statistical Planning and Inference*, 2020, *207*, 190–197.

Sekhon, Jasjeet S. and Yotam Shem-Tov, "Inference on a New Class of Sample Average Treatment Effects," *Journal of the American Statistical Association*, February 2020, pp. 1–18. Publisher: Taylor & Francis.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65* (3), 557–586. Publisher: [Wiley, Econometric Society].

Sun, Liyan and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2020. Forthcoming.

van der Vaart, A. W., *Asymptotic Statistics*, Cambridge University Press, June 2000.

Xu, Ruonan, "Potential outcomes and finite-population inference for M-estimators," *Econometrics Journal*, 2021, (Forthcoming).

Design-Based Uncertainty for Quasi-Experiments

Appendix

Ashesh Rambachan Jonathan Roth

April 13, 2022

A Additional Proofs

Proof of Proposition 3.4

Proof. It suffices to show that $\frac{\hat{s}_1^2}{\mathbb{V}\text{ar}_\pi[Y_i(1)]} \to_p 1$ and $\frac{\hat{s}_0^2}{\mathbb{V}\text{ar}_{1-\pi}[Y_i(0)]} \to_p 1$. We provide a proof for the former; the latter proof is analogous. For notational convenience, let $v_1 = \mathbb{V}\text{ar}_\pi[Y_i(1)]$. From the definition of $\hat{s}_1^2$, we can write

$$\frac{\hat{s}_1^2}{v_1} = \frac{1}{v_1}\left(\left(\frac{1}{N_1}\sum_i D_i(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^2\right) - (\bar{Y}_1 - \mathbb{E}_\pi[Y_i(1)])^2\right).$$

Now, $\frac{1}{N_1}\sum_i D_i(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^2$ can be viewed as a Horvitz-Thompson estimator of $\frac{1}{N_1}\sum_i \pi_i(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^2 = v_1$, and thus by Theorem 6.1 in Hajek (1964), its variance is equal to

$$(1 + o(1))\left(\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\right) \cdot \mathbb{V}\text{ar}_{\tilde{\pi}}\left[(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^2\right].$$

Note further that

$$\begin{aligned}\left(\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\right) \cdot \mathbb{V}\text{ar}_{\tilde{\pi}}\left[(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^2\right] &\leqslant \frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^4 \\ &\leqslant \frac{1}{N_1^2} m_N(1) \sum_i \pi_i(Y_i(1) - \mathbb{E}_\pi[(Y_i(1)])^2 \\ &= \frac{1}{N_1} m_N(1) \mathbb{V}\text{ar}_\pi[Y_i(1)].\end{aligned}$$

Applying Chebychev's inequality, we have

$$\frac{1}{N_1}\sum_i (D_i(Y_i(1) - \mathbb{E}_\pi[Y_i(1)])^2 - v_1 = O_p\left(\sqrt{\frac{1}{N_1} m_N(1) \mathbb{V}\text{ar}_\pi[Y_i(1)]}\right).$$

Next, viewing $\bar{Y}_1$ as a Horvitz-Thomson estimator, we see that its variance is bounded by $(1 + o(1))\left(\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\right) \cdot \mathbb{V}\text{ar}_{\tilde{\pi}}[Y_i(1)]$, which by similar logic to that above is bounded

above by $(1+o(1))\frac{1}{N_1}\mathbb{V}\text{ar}_\pi[Y_i(1)]$. Thus, by Chebychev's inequality,

$$\bar{Y}_1 - \mathbb{E}_\pi[Y_i(1)] = O_p\left(\sqrt{\frac{1}{N_1}\mathbb{V}\text{ar}_\pi[Y_i(1)]}\right).$$

Combining the results above, it follows that

$$\frac{\hat{s}_1^2}{v_1} = \frac{1}{v_1}\left(v_1 + O_p\left(\sqrt{\frac{m_N(1)v_1}{N_1}}\right) + O_p\left(\frac{1}{N_1}v_1\right)\right) = 1 + O_p\left(\sqrt{\frac{m_N(1)}{v_1 N_1}}\right) + O_p\left(\frac{1}{N_1}\right).$$

However, the first O_p term converges to 0 by assumption, and since Assumption 2 implies that $N_1 \to \infty$, the second O_p term converges to 0 as well.

□

Proof of Proposition 3.6

Proof. The proof of claim (1) is analogous to equation (3). We next prove claim (2). For simplicity, let $A_n = \mathbb{V}_R[\hat{\tau}]$, let B_n be the right-hand-side of the first equality in claim (2), and let C_n be the right-hand side of the inequality in claim (2). We first prove the inequality. Note that by the definition of a semi-definite matrix, it suffices to show that $l'B_n l \leqslant l'C_n l$ for all $l \in \mathbb{R}^K$. However, letting $Y_i(d) = l'\mathbf{Y}_i(d)$, the desired inequality follows from Proposition 3.2. Next, observe that $A_n - B_n = o(N^{-1})$ if and only if $D_n := NA_n - NB_n = o(1)$, which holds if and only if $l'D_n l = o(1)$ for all $l \in L := \{e_j \,|\, 1 \leqslant j \leqslant K\} \cup \{e_j - e_{j'} \,|\, 1 \leqslant j, j' \leqslant K\}$, where e_j is the jth basis vector in $\mathbb{R}^K$. To obtain the last equivalence, note that $e_j' D_n e_j = [D_n]_{jj}$ (the (j,j) element of D_n), whereas exploiting the fact that D_n is symmetric, $(e_j - e_{j'})'D_n(e_j - e_{j'}) = [D_n]_{jj} + [D_n]_{j'j'} - 2[D_n]_{jj'}$, and so convergence of $l'D_n l$ to zero for all $l \in L$ is equivalent to convergence of each of the elements of D_n. Next, note that if $Y_i(d) = l'\mathbf{Y}_i(d)$, then $\hat{\tau}$ as defined in (2) is equal to $l'\hat{\boldsymbol{\tau}}$ and $\mathbb{V}\text{ar}_{\tilde{\pi}}[Y_i(d)] = l'\mathbb{V}\text{ar}_{\tilde{\pi}}[\mathbf{Y}_i(d)]\,l$. It follows from Proposition 3.1 that

$$N \cdot l'\mathbb{V}_R[\hat{\boldsymbol{\tau}}]\,l[1+o(1)] = \frac{\frac{1}{N}\sum_{k=1}^N \pi_k(1-\pi_k)}{\frac{N_0}{N}\frac{N_1}{N}} l'\left[\frac{N}{N_1}\mathbb{V}\text{ar}_{\tilde{\pi}}[\mathbf{Y}_i(1)] + \frac{N}{N_0}\mathbb{V}\text{ar}_{\tilde{\pi}}[\mathbf{Y}_i(0)] - \mathbb{V}\text{ar}_{\tilde{\pi}}[\tau_i]\right] l, \tag{20}$$

which implies that $l'D_n l = l'(NA_n)l \cdot o(1)$. However, Assumption 5, together with the inequality in claim (2), implies that the right-hand side of the previous display is $O(1)$, and thus $l'(NA_n)l = O(1)$, from which the desired result follows.

The proof of (3) is similar to the proof of Lemma A3 in Li and Ding (2017), which gives a similar result in the case of completely randomized experiments. We provide a proof for the convergence of $\hat{\boldsymbol{s}}_1$; the convergence of $\hat{\boldsymbol{s}}_0$ is similar. As in the proof to claim (2), it suffices

to show that $l'\hat{\boldsymbol{s}}_1 l - l'\mathbb{V}\text{ar}_\pi\left[\mathbf{Y}_i(1)\right] l \to_p 0$ for all $l \in L$. Let $Y_i(d) = l'\mathbf{Y}_i(1)$. Then

$$\begin{aligned} l'\hat{\boldsymbol{s}}_1 l &= \frac{1}{N_1}\sum_i D_i(l'\mathbf{Y}_i(1) - \frac{1}{N_1}\sum_j D_j l'\mathbf{Y}_j(1))^2 \\ &= \left(\frac{1}{N_1}\sum_i D_i(l'\mathbf{Y}_i(1) - l'\mathbb{E}_\pi\left[\mathbf{Y}_i(1)\right])^2\right) + \left(\frac{1}{N_1}\sum_i D_i l'\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right]\right)^2, \quad (21) \end{aligned}$$

where the second line uses the bias variance decomposition. The first term can be viewed as a Horvitz-Thompson estimator of $\frac{1}{N_1}\sum_i \pi_i(l'\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^2 = \mathbb{V}\text{ar}_\pi\left[l'\mathbf{Y}_i(1)\right]$ under rejective sampling, and thus has variance equal to

$$(1+o(1))\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\mathbb{V}\text{ar}_{\tilde{\pi}}\left[(l'\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^2\right].$$

Further, observe that

$$\begin{aligned} &\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\mathbb{V}\text{ar}_{\tilde{\pi}}\left[(l'\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^2\right] \leqslant \\ &\frac{1}{N_1}\mathbb{E}_\pi\left[(l'\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^4\right] \leqslant \\ &\frac{1}{N_1}\max_i\left\{(l'\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^2\right\} \cdot \mathbb{V}\text{ar}_\pi\left[l'\mathbf{Y}_i(1)\right] \leqslant \\ &\left[||l||^2\frac{N}{N_1}\right]\left[\max_i ||\mathbf{Y}_i(1) - \mathbb{E}_\pi\left[\mathbf{Y}_i(1)\right]||^2/N\right] \cdot \left[l'\mathbb{V}\text{ar}_\pi\left[\mathbf{Y}_i(1)\right] l\right] = o(1) \end{aligned}$$

where the first inequality is obtained using the fact that $\mathbb{V}\text{ar}_{\tilde{\pi}}\left[X\right] \leqslant \mathbb{E}_{\tilde{\pi}}\left[X^2\right]$, expanding the definition of $\mathbb{E}_{\tilde{\pi}}\left[\cdot\right]$, and using the inequality $\pi_i(1-\pi_i) \leqslant \pi_i$, analogous to the argument in the proof to Proposition 3.4; the final inequality uses the Cauchy-Schwarz inequality and factors out l; and we obtain that the final term is $o(1)$ by noting that the first and final bracketed terms are $O(1)$ by Assumption 5 and the middle term is $o(1)$ by Assumption 6. Applying Chebychev's inequality, it follows that the first term in (21) is equal to $\mathbb{V}\text{ar}_\pi\left[l'\mathbf{Y}_i(1)\right] + o(1)$.

To complete the proof of the claim, we show that the second term in (21) is $o(1)$. Note that we can view $\frac{1}{N_1}\sum_i D_i l'\mathbf{Y}_i(1)$ as a Horvitz-Thompson estimator of $\mathbb{E}_\pi\left[l'\mathbf{Y}_i\right]$. Following similar arguments to that in the proceeding paragraph, we have that its variance is bounded above by $\frac{1}{N_1}l'\mathbb{V}\text{ar}_\pi\left[\mathbf{Y}_i(1)\right] l$, which is $o(1)$ by Assumption 5 combined with the fact that Assumption 2 implies $N_1 \to \infty$. Applying Chebychev's inequality again, we obtain that the second term in (21) is $o(1)$, as needed.

To prove claim (4), appealing to the Cramer-Wold device, it suffices to show that for any $l \in \mathbb{R}^K \backslash \{0\}$, $Y_i = l'\mathbf{Y}_i$, and $\hat{\tau}$ as defined in (2), $\mathbb{V}_R\left[\hat{\tau}\right]^{-\frac{1}{2}}(\hat{\tau} - \tau) \to_d \mathcal{N}(0,\, 1)$. This follows from Proposition 3.3, provided that we can show that Assumption 7 implies that Assumption 3 holds when $Y_i = l'\mathbf{Y}_i$ for any conformable vector l. Indeed, recall that $\sigma_{\tilde{\pi}}^2 = l'\Sigma_{\tilde{\pi}} l \geqslant \lambda_{min}||l||^2$,

and hence $\frac{1}{\lambda_{min}} \geqslant \frac{1}{||l||^2}\frac{1}{\sigma^2_{\tilde{\pi}}}$. From the Cauchy-Schwarz inequality

$$\left|\left|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right|\right|^2 \cdot ||l||^2 \geqslant (\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right])^2.$$

Together with the previous inequality, this implies that

$$\frac{1}{\lambda_{min}}\mathbb{E}_{\tilde{\pi}}\left[\left|\left|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right|\right|^2 \cdot 1\left[\left|\left|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right|\right| \geqslant \sqrt{\sum_i \pi_i(1-\pi_i)\cdot\lambda_{min}}\cdot\epsilon\right]\right] \geqslant$$
$$\frac{1}{\sigma^2_{\tilde{\pi}}}\mathbb{E}_{\tilde{\pi}}\left[(\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right])^2 \cdot 1\left[\left|(\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right])\right| \geqslant \sqrt{\sum_i \pi_i(1-\pi_i)}\cdot\sigma_{\tilde{\pi}}\epsilon\right]\right],$$

from which the result follows. □