Estimating Average Causal Effects Under General Interference

Peter M. Aronow and Cyrus Samii*

March 5, 2012

Abstract

This paper presents randomization-based methods for estimating average causal effects under arbitrary interference of known form. Conservative estimators of the randomization variance of the average treatment effects estimators are presented, as is justification for confidence intervals based on a normal approximation. Examples relevant to research in environmental protection, networks experiments, "viral marketing," two-stage disease prophylaxis trials, and stepped-wedge designs are presented.

1 Introduction

Experimental and observational studies often involve treatments with effects that "interfere" (Cox, 1958) across units through spillover or other forms of dependency. Such interference is typically considered a nuisance, and researchers often strive to design studies that isolate units as much as possible from the effects of the interference. However, such designs are not always possible. Furthermore, researchers have become increasingly interested in estimation of the spillover effects themselves, as these effects may be of substantive importance. Treatments may be applied to individuals in an existing network, and we may wish to study how effects transmit to peers in the network. For example, an urban renewal program applied to one town may divert capital from other towns, in which case the overall effect of the program may be ambiguous. In these cases, we need a method to estimate effects of both direct and *indirect* exposure to a treatment. Another form of interference

*Peter M. Aronow is PhD student, Department of Political Science, Yale University, 77 Prospect St., New Haven, CT 06520 (Email: peter.aronow@yale.edu). Cyrus Samii is Assistant Professor, Department of Politics, New York University, 19 West 4th St., New York, NY 10012 (Email: cds2083@nyu.edu). The authors are grateful to Bernd Beber, Winston Lin, Don Green, Allison Sovey, Teppei Yamamoto, and participants at the Yale Field Experiments Seminar, the Workshop on Information in Networks 2011, the NYU Development Economics Workshop, and the NYU-CESS Experimental Political Science Conference 2012 for helpful feedback. The authors express special gratitude to Joel Middleton for particularly valuable conversations.

occurs when there are multiple varieties of a treatment but these varieties are not explicitly randomized. For example, there might exist two teams administering a treatment and the teams could each have some probability of being the one to administer treatment to a given unit. We may then want to know whether one team is more effective than another. Finally, it may be that treatment effects are time-varying and units have some chance of receiving treatment at any one of a set of points in time.

In this paper, we develop general, randomization-based methods for estimating average causal effects under these and other forms of interference. In so doing, we develop a general framework for estimation and inference that naturally draws from the design of the experiment. Interference represents a departure from the traditional randomized experiment wherein units are randomly assigned directly to treatment or control and the potential outcomes that would be observed for a unit in either the treatment or control condition are fixed and do not depend on the overall set of treatment assignments. The latter condition is what Rubin (1990) refers to as the "stable unit treatment value assumption" (SUTVA). In the examples above, the traditional randomized experiment model is clearly inadequate, as SUTVA would be violated. A more sophisticated model of treatment exposure must be specified.

The estimators that we propose in this paper are simple and jointly leverage known features of the design of a randomized experiment with an arbitrary model of indirect exposure under which SUTVA is assured. They provide unbiased point estimates of the average causal effects induced by treatment exposure. We also provide estimators for the randomization variance of the estimated average causal effects. These variance estimators are assured of being either conservative or unbiased. We then provide justification for the use of large-sample confidence intervals based on a normal approximation.

Our contribution is novel and important in three ways. First, once a treatment "exposure model" has been fixed, the methods we present are based entirely on known randomization distributions that arise from the experiment's design. Essentially, we use the randomization plan defined by the experiment's design to determine the probability of exposure for each unit. We can use these probabilities to obtain estimates of average potential outcomes under the different exposure conditions. No ancillary modeling assumptions are needed, although in some cases they may be introduced to gain efficiency. Second, we move beyond approaches that presume both hierarchical treatment assignment and interference only within groups at the lowest level of some hierarchy. As we discuss in the next section, nearly all treatments of interference in the applied statistics literature presume this setting. However, interference can come in many less structured forms. The methods that we propose are general to a wide range of interference settings of which hierarchical settings are but one example. Third, we provide a conservative basis for inference on estimated causal quantities. Our variance estimators are based directly on the randomization distribution and they account for "non-measurability" problems. (Details on this issue are given below.) Unlike measures of uncertainty that are based on either inverted exact tests or approximate methods, our variance estimator is guaranteed to have a weakly positive bias.

2 Related literature

Current results on estimation under interference focus on estimating group-level causal effects under "partial" interference. Rosenbaum (2007) discusses methods for testing sharp null hypotheses in the presence of interference in trials where random assignment occurs within groups and interference does not cross group boundaries. Sobel (2006) analyzes the potential for bias in such settings when non-interference is mistakenly assumed, and then defines a number of direct and indirect effects that, in principle, may be identifiable. Hudgens and Halloran (2008) extend this work to define direct, indirect, total, and overall causal effects defined at the group level in two-stage randomized trials in which some groups are randomly assigned to host treatments, and then treatments are assigned at random within the selected groups. Interference is presumed to operate only within groups. These authors provide randomization-based estimators for these group level effects. Tchetgen-Tchetgen and VanderWeele (2010) extend Hudgens and Halloran's results, providing conservative variance estimators, a framework for finite sample inference, and extensions to observational studies.

Hierarchical treatment assignment, interference limited to within groups, and the establishment of group-level effects as the estimands greatly simplify the estimation problem. With this analytical framework, inference can proceed assuming independence across groups, in which case we effectively obtain a completely randomized experiment on groups. In some settings, however, such rigid hierarchical structuring may not be valid, as in the case of experiments carried out over networks of actors that share links as a result of a complex, endogenous process. A key contribution of this paper is to generalize estimation and inference theory to settings that exhibit arbitrary forms of interference and treatment assignment dependencies. In addition, our inferential targets are unit-level causal effects. Unit-level causal effects are often the estimand of primary interest, as is the case, for example, when exploring unit-level characteristics associated with the magnitude of treatment effects.

3 Treatment assignment and exposure models

The first step is to distinguish between (1) treatment assignments over the set of experimental units and (2) each unit's treatment exposure under a given assignment. Treatment assignments can be manipulated arbitrarily with the experimental design. However, treatment exposures may be constrained on the basis of the varying interference or heterogeneity potential of different experimental units. For example, interference or spillover effects may spread over a spatial gradient. If so, different treatment assignments may result in different patterns of interference depending on where treatments are applied on the spatial plane.

Formally, suppose we have a finite population U of units indexed by $i = 1, ..., N$ on which a randomized experiment is performed. Define a treatment assignment vector, $\mathbf{z} = (z_1, ..., z_N)'$, where $z_i \in \{1, .., M\}$ specifies which of M possible treatment values that unit i receives. An experimental design contains a plan for randomly selecting a particular value

of $\mathbf{z}$ from the M^N different possibilities with predetermined probability $p_{\mathbf{z}}$. Restricting our attention only to treatment assignments that can be generated by a given experimental design, define $\Omega = \{\mathbf{z} : p_{\mathbf{z}} > 0\}$, so that $\mathbf{Z} = (Z_1, ..., Z_N)'$ is a random vector with support Ω and $\Pr(\mathbf{Z} = \mathbf{z}) = p_{\mathbf{z}}$.

We define a unit-specific onto function that maps an assignment vector and unit specific traits to an exposure value: $f : \Omega \times \Theta \to \Delta$, where $\theta_i \in \Theta$ quantifies relevant traits of unit i. The codomain Δ contains all of the sorts of treatment exposures (e.g., "direct" or "indirect" exposure) that might be induced in the experiment. The contents of Δ depend on the nature of interference or treatment heterogeneity. These exposures may be represented as vectors, discrete classes, or real numbers.

The estimation of causal effects under interference or treatment heterogeneity amounts to using information about treatment *assignments*, which come from the experiment's design, to estimate the *effects of treatment exposures*, which result from the interaction of the design (captured by $\mathbf{Z}$) and other underlying features of the population (captured by f and the θ_is). The key step in estimating causal quantities under general interference or treatment heterogeneity is in specifying f and Δ. These two elements constitute an *exposure model*. In a traditional randomized experiment, the exposure model simplifies to one in which we can set $\Delta = \{1, ..., M\}$ and $f(\mathbf{z}, \theta_i) = f(\mathbf{z}) = z_i$ for all i. This model has been the workhorse for most causal effects estimation under the Neyman (1923)-Rubin paradigm. An exposure model that allowed for arbitrary interference or treatment heterogeneity would be one for which $|\Delta| = |\Omega| \times N$, in which case each unit has a unique type of exposure under each treatment assignment, and $f(\mathbf{z}, \theta_i)$ would be unique for each $\mathbf{z}$. We can call this the "arbitrary exposure" model. If such an exposure model were valid, then it is clear that there would be no meaningful way to use the results of the experiment. We must use substantive judgment to fix a model somewhere between the traditional randomized experiment and arbitrary exposure models in order to carry out analyses under interference or treatment heterogeneity.

Define $D_i \equiv f(\mathbf{Z}, \theta_i)$, a random variable with support $\Delta_i \subseteq \Delta$ and for which $\Pr(D_i = d) \equiv \pi_i(d)$. Note that because $|\Delta| \leq |\Omega| \times N$, Δ is a finite set of $K \leq |\Omega| \times N$ values, such that $\Delta = \{d_1, ..., d_K\}$. Then for each unit, i, we have a vector of probabilities, $(\pi_i(d_1), ..., \pi_i(d_K))' = \boldsymbol{\pi}_i$. Invoking Imbens (2000)'s *generalized propensity score*, we call $\boldsymbol{\pi}_i$ the *generalized probability of treatment* (GPT) for i. A unit i's GPT tells us the probability of i being subject to each of the possible exposures in $\{d_1, ..., d_K\}$. Because f is onto,

$$\pi_i(d_k) = \sum_{\mathbf{z} \in \Omega} \mathbf{I}(f(\mathbf{z}, \theta_i) = d_k) \Pr(\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{z} \in \Omega} p_{\mathbf{z}} \mathbf{I}(f(\mathbf{z}, \theta_i) = d_k).$$

Thus the GPT for unit i is known exactly, with each component probability, $\pi_i(d_k)$, equal to the expected proportion of treatment assignments that induce exposure d_k for unit i.

Below, we will refer to joint exposure probabilities when discussing variance estimators. That is, we define $\pi_{ij}(d_k)$ as the probability of the joint event that both units i and j are subject to exposure d_k, and we define $\pi_{ij}(d_k, d_l)$ as the probability of the joint event that units i and j are subject to exposures d_k and d_l, respectively. To compute both individual and joint exposure probabilities from the experiment's design, first define the $N \times |\Omega|$

matrix,

$$\mathbf{I}_k = [\mathbf{I}(f(\mathbf{z},\theta_i) = d_k)]_{\substack{\mathbf{z}\in\Omega \\ i=1,\ldots,N}} =$$

$$\begin{bmatrix} \mathbf{I}(f(\mathbf{z}_1,\theta_1) = d_k) & \mathbf{I}(f(\mathbf{z}_2,\theta_1) = d_k) & \ldots & \mathbf{I}(f(\mathbf{z}_N,\theta_1) = d_k) \\ \mathbf{I}(f(\mathbf{z}_1,\theta_2) = d_k) & \mathbf{I}(f(\mathbf{z}_2,\theta_2) = d_k) & \ldots & \mathbf{I}(f(\mathbf{z}_N,\theta_2) = d_k) \\ \vdots & \vdots & \ddots & \\ \mathbf{I}(f(\mathbf{z}_1,\theta_N) = d_k) & \mathbf{I}(f(\mathbf{z}_2,\theta_N) = d_k) & & \mathbf{I}(f(\mathbf{z}_N,\theta_N) = d_k) \end{bmatrix},$$

which is a matrix of indicators for whether units are in exposure condition k over possible assignment vectors. Define the $|\Omega| \times |\Omega|$ diagonal matrix $\mathbf{P} = \text{diag}(p_{\mathbf{z}_1}, p_{\mathbf{z}_2}, ..., p_{\mathbf{z}_{|\Omega|}})$. Then,

$$\mathbf{I}_k\mathbf{P}\mathbf{I}_k' = \begin{bmatrix} \pi_1(d_k) & \pi_{12}(d_k) & \ldots & \pi_{N1}(d_k) \\ \pi_{12}(d_k) & \pi_2(d_k) & \ldots & \pi_{N2}(d_k) \\ \vdots & \vdots & \ddots & \\ \pi_{N1}(d_k) & \pi_{N2}(d_k) & & \pi_N(d_k) \end{bmatrix},$$

is an $N \times N$ symmetric matrix with individual exposure probabilities, the $\pi_i(d_k)$'s, on the diagonal and joint exposure probabilities, the $\pi_{ij}(d_k)$'s, on the off-diagonals. The non-symmetric $N \times N$ matrix,

$$\mathbf{I}_k\mathbf{P}\mathbf{I}_l' = \begin{bmatrix} 0 & \pi_{12}(d_k,d_l) & \ldots & \pi_{1N}(d_k,d_l) \\ \pi_{21}(d_k,d_l) & 0 & \ldots & \pi_{2N}(d_k,d_l) \\ \vdots & \vdots & \ddots & \\ \pi_{N1}(d_k,d_l) & \pi_{N2}(d_k,d_l) & & 0 \end{bmatrix},$$

yields all joint probabilities across exposure conditions k and l. The zeroes on the diagonal are due to the fact that a unit cannot be subject to multiple exposure conditions at once.[1]

We have indicated that $\Delta_i \subseteq \Delta$. That is to say that unit i may have zero probability of being subjected to some subset of the K exposures. For example, under the arbitrary exposure model, Δ_i would contain only the $|\Omega|$ exposure types that apply to unit i, whereas Δ would contain the $|\Omega| \times N$ exposure types that apply to all N units. Typically we will work with an exposure model such that $K \ll |\Omega| \times N$. In fact, we may limit the number of exposure types to only a handful for which all units have some non-zero chance of being

[1] In practice, $|\Omega|$ may be so large that it is impractical to construct Ω to compute the π_is and the joint probability matrices exactly. One may nonetheless approximate the π_is and joint probabilities with arbitrary precision through replication (Fattorini, 2006). That is, produce R random replicate $\mathbf{z}$s based on the randomization plan. From these R replicates, we can construct an $N \times R$ indicator matrix, $\widehat{\mathbf{I}}_k$, for each of the $k = 1, ..., K$ exposure conditions. Then an estimator for $\mathbf{I}_k\mathbf{P}\mathbf{I}_k'$ is $\widehat{\mathbf{I}}_k\widehat{\mathbf{I}}_k'/R$, and similarly for $\mathbf{I}_k\mathbf{P}\mathbf{I}_l'$. The replication procedure would be equivalent to drawing a random sample without replacement from Ω with probabilities of selection equal to those which are defined in the randomization plan. As such, the resulting exposure probability and joint probability estimates would be unbiased. Chen et al. (2010) apply a similar approach.

subjected.[2] (Recall that for the traditional randomized experiment, $K = M$.) But it may still be the case that some units have zero probability of being subjected to some of the exposure values. In the discussion below, we offer refinements to account for this possibility.

4 Average potential outcomes and causal effects

Current work on causal inference under interference expends considerable effort on defining various forms of "direct," "indirect," and "total" effects. The ability to estimate any such effects depends on whether one is able to recover reliable estimates of average potential outcomes under the various treatment exposure conditions. We focus on the problem of estimating average potential outcomes under each of the exposure conditions. With that, the analyst is then free to compute arbitrary causal quantities of interest. Estimators for the randomization variance of average effects estimates are developed in the section that follows.

For the moment, we restrict attention to the case where all units have non-zero probability of being subject to each of the K exposures. That is, $0 < \pi_i(d_k) < 1$ for all i and k. (We discuss the alternative case below.) Then, each unit i has K potential outcomes associated with each of the K exposures, which we denote by $\{Y_i(d_1), ..., Y_i(d_K)\}$. We seek estimates for all k of $\mu(d_k) = \frac{1}{N}\sum_{i=1}^{N} Y_i(d_k) = \frac{1}{N}Y^T(d_k)$, where $Y^T(d_k)$ is the total of the potential outcomes under d_k.[3] The number of units in the population, N, is fixed, but we cannot estimate $Y^T(d_k)$ directly, as we only observe $Y_i(d_k)$ for those with $D_i = d_k$. However, by design, the collection of units for which we observe $Y_i(d_k)$ is a unequal-probability without-replacement sample from $\{Y_1(d_k), ..., Y_N(d_k)\}$, with the sampling probabilities known exactly. By Horvitz and Thompson (1952), an unbiased estimator for $Y^T(d_k)$ is given by the inverse probability weighted estimator,

$$\widehat{Y^T_{HT}}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{Y_i(d_k)}{\pi_i(d_k)}. \tag{1}$$

With potential outcomes and the randomization plan fixed, the exact variance for $\widehat{Y^T_{HT}}(d_k)$

[2] Nonetheless, sometimes the number of relevant exposures, K, is truly very large. Then, it may make sense to specify a function that maps D_i to a random variable $\widetilde{D}_i$ that is governed by a "continuous" GPT function, $\Pr(\widetilde{D}_i \leq d) = \int_{-\infty}^{d} \widetilde{\pi}(x, \theta_i; \psi)dx$, where ψ parameterizes a probability density function, $\widetilde{\pi}(.)$. Computing continuous GPT values cannot rely on the design alone, but rather requires the input of the analyst in determining the appropriate density function. We do not consider continuous exposure models in this paper, but note that their development could follow many of the same principles as the discrete exposure case that we do consider.

[3] This construction owes greatly to Joel Middleton.

is,

$$\begin{aligned}
\text{Var}[\widehat{Y^T_{HT}}(d_k)] &= \sum_{i=1}^{N}\sum_{j=1}^{N} \text{Cov}\left[\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_k)\right] \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_k)}{\pi_j(d_k)} \\
&= \sum_{i=1}^{N} \text{Var}\left[\mathbf{I}(D_i = d_k)\right] \left[\frac{Y_i(d_k)}{\pi_i(d_k)}\right]^2 \\
&\quad + \sum_{i=1}^{N}\sum_{j\neq i} \text{Cov}\left[\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_k)\right] \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_k)}{\pi_j(d_k)} \\
&= \sum_{i=1}^{N} \pi_i(d_k)[1-\pi_i(d_k)] \left[\frac{Y_i(d_k)}{\pi_i(d_k)}\right]^2 \\
&\quad + \sum_{i=1}^{N}\sum_{j\neq i} [\pi_{ij}(d_k) - \pi_i(d_k)\pi_j(d_k)] \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_k)}{\pi_j(d_k)}. \qquad (2)
\end{aligned}$$

The estimator for the mean of all N potential outcomes under exposure d_k is thus,

$$\widehat{\mu_{HT}}(d_k) = (1/N)\widehat{Y^T_{HT}}(d_k), \qquad (3)$$

with exact variance,

$$\text{Var}\left(\widehat{\mu_{HT}}(d_k)\right) = (1/N^2)\text{Var}[\widehat{Y^T_{HT}}(d_k)]. \qquad (4)$$

Then,

$$\widehat{\tau_{HT}}(d_k, d_l) = \hat{\mu}_{HT}(d_k) - \hat{\mu}_{HT}(d_l) = \frac{1}{N}\left[\widehat{Y^T_{HT}}(d_k) - \widehat{Y^T_{HT}}(d_l)\right]$$

estimates $\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N}[Y_i(d_k) - Y_i(d_l)]$, the average causal effect of having all units subject to exposure k versus having all units subject to exposure l. The exact variance of the difference in estimated means is given by,

$$\text{Var}\left(\widehat{\tau_{HT}}(d_k, d_l)\right) = \frac{1}{N^2}\left\{\text{Var}[\widehat{Y^T_{HT}}(d_k)] + \text{Var}[\widehat{Y^T_{HT}}(d_l)] - 2\text{Cov}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)]\right\}, \qquad (5)$$

where (Wood, 2008),

$$\begin{aligned}
\text{Cov}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)] &= \sum_{i=1}^{N}\sum_{j=1}^{N} \text{Cov}\left[\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_l)\right] \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_l)}{\pi_j(d_l)} \\
&= \sum_{i=1}^{N}\sum_{j\neq i} \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_l)}{\pi_j(d_l)} \left[\pi_{ij}(d_k, d_l) - \pi_i(d_k)\pi_j(d_l)\right] \\
&\quad - \sum_{i=1}^{N} Y_i(d_k)Y_i(d_l), \qquad (6)
\end{aligned}$$

with $\pi_{ij}(d_k, d_l) = \Pr[\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_l)]$, and the last line follows from the fact that $\pi_{ii}(d_k, d_l) = 0$.

Expressions (2) and (6) allow us to see the conditions under which exact variances are identified. So long as all joint exposure probabilities are non-zero (that is, $\pi_{ij}(d_k) > 0$ for all i, j), unbiased estimators for $\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]$ are identified for the population U. Because we only observe one potential outcome for each unit, the last term in (6) is always unidentified, and thus $\widehat{\text{Cov}}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)]$ is always unidentified. This is a familiar problem in estimating the randomization variance for the average treatment effect—e.g., Freedman et al. (1998, A32-A34). If $\pi_{ij}(d_k) = 0$ for some i, j, $\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]$ is unidentified. Similarly, if $\pi_{ij}(d_k, d_l) = 0$ for some i, j, then additional components of $\widehat{\text{Cov}}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)]$ are unidentified.[4] Nonetheless, we can always identify estimators for $\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]$ and $\widehat{\text{Cov}}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)]$ that are guaranteed to have weakly positive bias. Thus, we can always identify a conservative approximation to the exact variances. We take this and related issues up in the next section.

5 Variance estimators

We derive conservative estimators for $\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]$ and $\widehat{\text{Var}}(\widehat{\tau_{HT}}(d_k, d_l))$. Although not necessarily unbiased, the estimators we present here are guaranteed to have a weakly positive bias relative to the randomization distributions of the estimated averages or effects.[5]

Then, given $\pi_{ij}(d_k) > 0$ for all i, j, the unbiased Horvitz-Thompson estimator for $\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]$ is,

$$\begin{aligned}
\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)] &= \sum_{i \in U} \sum_{j \in U} \mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_k) \frac{\text{Cov}\left[\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_k)\right]}{\pi_{ij}(d_k)} \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_k)}{\pi_j(d_k)} \\
&= \sum_{i \in U} \mathbf{I}(D_i = d_k)[1 - \pi_i(d_k)] \left[\frac{Y_i(d_k)}{\pi_i(d_k)}\right]^2 \\
&\quad + \sum_{i \in U} \sum_{j \in U \setminus i} \mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_k) \frac{\pi_{ij}(d_k) - \pi_i(d_k)\pi_j(d_k)}{\pi_{ij}(d_k)} \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_k)}{\pi_j(d_k)}.
\end{aligned} \tag{7}$$

Then an unbiased estimator for the variance of $\widehat{\mu}_{HT}(d_k)$ is given by,

$$\widehat{\text{Var}}[\widehat{\mu}_{HT}(d_k)] = (1/N^2)\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)].$$

In the case where $\pi_{ij}(d_k) = 0$ for some i, j, there exist no unbiased estimators for $\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]$. As demonstrated in Aronow and Samii (2011, Proposition 1), the bias of

[4] In this case, expression (6) can be further refined to account for the zero pairwise probabilities. We examine this scenario in the next section.

[5] If any units i, j appear as co-members of a cluster (i.e., $\Pr(D_i = D_j) = 1$), then there are advantages to totaling these units' outcome values, $Y_i(d_k)$ and $Y_j(d_k)$, into one larger unit before variance estimation. We discuss this refinement in Appendix A below.

$\widehat{\text{Var}}[\widehat{\mu}_{HT}(d_k)]$, is given by,

$$A = \sum_{i \in U} \sum_{j \in \{U \setminus i : \pi_{ij}(d_k) = 0\}} Y_i(d_k) Y_j(d_k).$$

Thus, $\widehat{\text{Var}}[\widehat{\mu}_{HT}(d_k)]$ is guaranteed to have weakly positive bias when $Y_i(d_k)Y_j(d_k) \geq 0$ for all i, j with $\pi_{ij}(d_k) = 0$, and the bias will be small when the terms in the sum tend to offset each other.[6] Another option is to use the following correction term (derived via Young's inequality):

$$\widehat{A_2}(d_k) = \sum_{i \in U} \sum_{j \in \{U \setminus i : \pi_{ij}(d_k) = 0\}} \left[\frac{\mathbf{I}(D_i = d_k) Y_i(d_k)^2}{2\pi_i(d_k)} + \frac{\mathbf{I}(D_j = d_k) Y_j(d_k)^2}{2\pi_j(d_k)} \right],$$

noting that $\widehat{A_2}(d_k) = 0$ if $\pi_{ij}(d_k) > 0$ for all i, j. By Aronow and Samii (2011, Corollary 2),

$$\text{E}\left[\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)] + \widehat{A_2}(d_k) \right] \geq \text{Var}[\widehat{Y^T_{HT}}(d_k)],$$

in which case,

$$\widehat{\text{Var}}_A[\widehat{\mu}_{HT}(d_k)] = (1/N^2) \left[\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)] + \widehat{A_2}(d_k) \right],$$

provides an either unbiased or weakly conservative estimator for the variance of the estimated average of potential outcomes under exposure d_k.

As discussed above, $\text{Cov}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)]$ is unidentified, which is to say that there exist no unbiased or consistent estimators for this quantity. However, we can compute an approximation that is guaranteed to have expectation less than or equal to the true covariance, providing a conservative (i.e., weakly positively biased) estimator for $\text{Var}(\widehat{\tau_{HT}}(d_k, d_l))$. For the case where $\pi_{ij}(d_k, d_l) > 0$ for all i, j such that $i \neq j$, we propose the following Horvitz-Thompson-type estimator for the covariance,

$$\begin{aligned} \widehat{\text{Cov}}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)] = &\sum_{i \in U} \sum_{j \in U \setminus i} \frac{\mathbf{I}(D_i = d_k)\mathbf{I}(D_j = d_l)}{\pi_{ij}(d_k, d_l)} \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_l)}{\pi_j(d_l)} \\ &- \sum_{i \in U} \left[\frac{\mathbf{I}(D_i = d_k) Y_i(d_k)^2}{2\pi_i(d_k)} + \frac{\mathbf{I}(D_i = d_l) Y_i(d_l)^2}{2\pi_i(d_l)} \right]. \end{aligned} \tag{8}$$

The term on the second line in expression (8) has expected value less than or equal to the quantity in the last line of expression (6), again via Young's inequality. This estimator is exactly unbiased if, for all $i \in U$, $Y_i(d_l) = Y_i(d_k)$, implying no effect associated with condition l relative to condition k. For the case where $\pi_{ij}(d_k, d_l) = 0$ for some i, j and k, l, we can refine the expression for the covariance given in (6) to be,

$$\begin{aligned} \text{Cov}[\widehat{Y^T_{HT}}(d_k), \widehat{Y^T_{HT}}(d_l)] = &\sum_{i \in U} \sum_{j \in \{U \setminus i : \pi_{ij}(d_k, d_l) > 0\}} \frac{Y_i(d_k)}{\pi_i(d_k)} \frac{Y_j(d_l)}{\pi_j(d_l)} [\pi_{ij}(d_k, d_l) - \pi_i(d_k)\pi_j(d_l)] \\ &- \sum_{i \in U} \sum_{j \in \{U : \pi_{ij}(d_k, d_l) = 0\}} Y_i(d_k) Y_j(d_l), \end{aligned} \tag{9}$$

[6] This notation requires that we maintain the assumption that $\frac{0}{0} = 0$. Aronow and Samii (2011) provide simulations showing a variety of cases when A is very large and other cases when it is close to zero.

where the term on the last line subsumes the term on the last line in expression (6). This leads us to propose a more general estimator for the covariance,

$$\widehat{\text{Cov}}_A[\widehat{Y^T_{HT}}(d_k),\widehat{Y^T_{HT}}(d_l)] = \sum_{i\in U}\sum_{j\in\{U\setminus i:\pi_{ij}(d_k,d_l)>0\}} \frac{\mathbf{I}(D_i=d_k)\mathbf{I}(D_j=d_l)}{\pi_{ij}(d_k,d_l)}\frac{Y_i(d_k)}{\pi_i(d_k)}\frac{Y_j(d_l)}{\pi_j(d_l)} - \sum_{i\in U}\sum_{j\in\{U:\pi_{ij}(d_k,d_l)=0\}}\left[\frac{\mathbf{I}(D_i=d_k)Y_i(d_k)^2}{2\pi_i(d_k)}+\frac{\mathbf{I}(D_j=d_l)Y_j(d_l)^2}{2\pi_j(d_l)}\right]. \tag{10}$$

Again, the term in the last line in (10) has expected value no greater than the term in the last line of (9) by Young's inequality.

Combining expressions, a conservative variance estimator for $\text{Var}(\widehat{\tau_{HT}}(d_k,d_l))$ is given by,

$$\widehat{\text{Var}}[\widehat{\tau_{HT}}(d_k,d_l)] = \frac{1}{N^2}\left\{\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_k)]+\widehat{A_2}(d_k)+\widehat{\text{Var}}[\widehat{Y^T_{HT}}(d_l)]+\widehat{A_2}(d_l) - 2\widehat{\text{Cov}}_A[\widehat{Y^T_{HT}}(d_k),\widehat{Y^T_{HT}}(d_l)]\right\}. \tag{11}$$

6 Asymptotics and intervals

Consistency and confidence intervals are always defined relative to some notion of asymptotic growth. We adopt a conceptualization of asymptotic growth, based on Brewer (1979), that is analogous to obtaining estimates by aggregating results from repeated experimentation (or sampling, as the case may be) on a fixed finite population. Suppose that the original population, U, is replicated $B-1$ times, yielding B copies of U, indexed by $b = 1,...,B$. Each of the B subpopulations hosts its own realization of $\mathbf{Z}^{(b)}$ to which the exposure model is applied, generating B separate realizations of $\mathbf{I}_k^{(b)}$, independent across the B subpopulations. The B subpopulations are collected into a population of BN units, and estimates are produced using values from this population. We let B tend to infinity. The conceptualization is simple to visualize, embeds straightforward moment assumptions (e.g., potential outcomes are bounded given finite U), and allows us to treat the $\mathbf{I}_k^{(b)}$ as iid over $b = 1,...,B$.[7] Consistency of $\widehat{\mu_{HT}}(d_k)$ and $\widehat{\tau_{HT}}(d_k,d_l)$ follow straightforwardly from the weak law of large numbers. Asymptotic normality for $\widehat{\mu_{HT}}(d_k)$ and $\widehat{\tau_{HT}}(d_k,d_l)$ follows from the classical central limit theorem (Lehmann, 1999, p. 73).

[7] We have studied results based on weaker assumptions on the asymptotic growth process, namely, (i) bounded potential outcomes, (ii) bounded exposure probabilities, and (iii) constraints on the level of correlation between exposure indicators as the finite population $N \to \infty$. The constraints on the level of correlation between exposure indicators assumes an $N \times N$ dependency matrix for each exposure condition, where each slot represents non-zero correlation in exposure conditions. We assume that the sum on any row or column of this matrix is bounded above by a fixed m. Under these assumptions, consistency of the mean and variance estimators above can be established, and an analogy to central limit theorems for bounded m-dependent series would seem to justify asymptotic normality (Hoeffding and Robbins, 1948; Mittelhammer, 1996, pp. 281-282), although demonstrating this definitively is beyond the scope of this paper.

To establish normal approximation confidence intervals for the estimated causal effect, $\widehat{\tau_{HT}}(d_k, d_l)$, consider the limiting behavior of $\widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]$. Define

$$\widehat{\mathrm{Var}}[\widehat{Y^T_{HT}}(d_k) - \widehat{Y^T_{HT}}(d_l)] \equiv N^2 \widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)].$$

Based on the properties of the conservative variance and covariance estimators given above, define on the population U,

$$\bar{V}_U \equiv N^2 \mathrm{E}\left\{\widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]\right\} = \mathrm{E}\left\{\widehat{\mathrm{Var}}[\widehat{Y^T_{HT}}(d_k) - \widehat{Y^T_{HT}}(d_l)]\right\} \geq \mathrm{Var}\left[\widehat{Y^T_{HT}}(d_k) - \widehat{Y^T_{HT}}(d_l)\right].$$

Now, under the asymptotic growth process that we have established,

$$\begin{aligned} NB\widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)] &= \frac{1}{NB}\sum_{b=1}^{B} \widehat{\mathrm{Var}}[\widehat{Y^T_{HT}}^{(b)}(d_k) - \widehat{Y^T_{HT}}^{(b)}(d_l)] \\ &= \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{\mathrm{Var}}[\widehat{Y^T_{HT}}^{(b)}(d_k) - \widehat{Y^T_{HT}}^{(b)}(d_l)]}{N} \xrightarrow{p} \frac{\bar{V}_U}{N}, \end{aligned}$$

where the simple summation in the first line is due to the independence of the $\mathbf{I}_k^{(b)}$ and the last line follows from the weak law of large numbers. This establishes that the variance estimator for the average causal effect converges to a quantity that is at least as large as the true variance. Finally, define,

$$t = \frac{\widehat{\tau_{HT}}(d_k, d_l) - \tau_{HT}(d_k, d_l)}{\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]}} = \frac{\widehat{\tau_{HT}}(d_k, d_l) - \tau_{HT}(d_k, d_l)}{\sqrt{\mathrm{Var}[\widehat{\tau_{HT}}(d_k, d_l)]}} \left(\frac{\mathrm{Var}[\widehat{\tau_{HT}}(d_k, d_l)]}{\widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]}\right)^{1/2}.$$

Under the given conditions, $\left(\widehat{\tau_{HT}}(d_k, d_l) - \tau_{HT}(d_k, d_l)\right) / \sqrt{\mathrm{Var}[\widehat{\tau_{HT}}(d_k, d_l)]}$ is asymptotically $\mathrm{N}(0,1)$ by the classical central limit theorem, while $(\mathrm{Var}[\widehat{\tau_{HT}}(d_k, d_l)] / \widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)])^{1/2}$ converges to a quantity less than one. Therefore, t is asymptotically normal, and intervals constructed as

$$\widehat{\tau_{HT}}(d_k, d_l) \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}[\widehat{\tau_{HT}}(d_k, d_l)]}$$

will tend to cover $\tau_{HT}(d_k, d_l)$ at least $100(1-\alpha)\%$ of the time for large B.

7 Zero exposure probabilities

So far we have restricted attention to the case where all units have a non-zero probability of being subject to any of K exposure conditions. To the extent possible, this is something that one should try to ensure when designing an experiment. But in some cases, there may be no way to avoid having some units with zero probability of being subject to some of the

exposure conditions. For example, in a spatial experiment, when an exposure is defined as the number of directly treated units within some radius, then a unit with no neighbors within this radius has zero probability of indirect exposure, even though this may not be the case universally.

Without making any assumptions about the functional form of exposure effects, the problem of zero exposure probabilities is side-stepped entirely only if we redefine the population on which we are estimating effects. When $\pi_i(d_k) = 0$, then the potential outcome $Y_i(d_k)$ can never be observed in the experiment, and so any quantity involving $Y_i(d_k)$ is unidentified without making auxiliary assumptions. Thus, causal effects associated with exposure d_k are unidentified for the full population U. Estimation in the manner presented in the preceding section must be restricted to the subpopulation of units for which exposures probabilities are non-zero. To compute an average causal effect of exposure k relative to exposure l, we would proceed by defining a new subpopulation, $U^{kl} = \{i \in U : \pi_i(d_k) > 0 \text{ and } \pi_i(d_l) > 0\}$, and estimation would proceed as above. All cardinalities from the full population, e.g, $N = |U|$, would be replaced by cardinalities from the subpopulation, $N = |U^{kl}|$. The joint exposure probability matrices and outcome data would be reconfigured to exclude rows and columns associated with units for which $\pi_i(d_k) = 0$ or $\pi_i(d_l) = 0$. The estimated effects could be considered as "local" average effects for the subpopulation, U^{kl}.

8 Refinements

The mean and difference-in-means estimators presented thus far are unbiased by sample theoretic arguments, and we have derived conservative variance estimators. However, we may wish to improve efficiency by incorporating auxiliary covariate information. In addition, by analogy to results from the unequal probability sampling literature, other forms of inverse probability weighting may significantly reduce mean square error with little cost in terms of bias (Särndal et al., 1992, pp. 181-184). Finally, by imposing working assumptions about the nature of treatment effects (e.g., constant effects), variance calculations are simplified, leading to estimators that may provide reasonable coverage while being more stable and potentially less biased than those based on conservative approximations. We discuss such refinements in the subsections that follow.

8.1 Covariance adjustment

Auxiliary covariate information may help to improve efficiency. A first method of covariance adjustment is based on the so-called "difference estimator" (Raj, 1965; Särndal et al., 1992, Ch. 6). Covariance adjustment of this variety can reduce the randomization variance of the estimated exposure means and average causal effects without compromising unbiasedness. In addition, the difference estimator addresses the problem of location non-invariance that afflicts Horvitz-Thompson-type estimators (Fuller, 2009, 9-10). The

estimator requires prior knowledge about how outcomes relate to covariates, perhaps obtained from analysis of auxiliary datasets.

Assume an auxiliary covariate vector $\mathbf{X_i}$ is observed for each i. We have some predefined function $g(\mathbf{X_i}, \xi_\mathbf{i}(d_k)) \to \mathbb{R}$, where $\xi_\mathbf{i}$ is a parameter vector. Ideally $g(.)$ is calibrated on auxiliary data to produce values that approximate $Y_i(d_k)$. We assume $\text{Cov}\left[g(\mathbf{X_i}, \xi_\mathbf{i}(d_k)), Z_i\right] = 0$ as a sufficient condition for unbiasedness.[8] Define

$$\widehat{Y_G^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{Y_i(d_k)}{\pi_i(d_k)} - \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{g(\mathbf{X_i}, \xi_\mathbf{i}(d_k))}{\pi_i(d_k)} + \sum_{i=1}^{N} g(\mathbf{X_i}, \xi_\mathbf{i}(d_k)), \quad (12)$$

which is unbiased for $Y^T(d_k)$ by,

$$\text{E}\left[-\sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{g(\mathbf{X_i}, \xi_\mathbf{i}(d_k))}{\pi_i(d_k)} + \sum_{i=1}^{N} g(\mathbf{X_i}, \xi_\mathbf{i}(d_k))\right] = 0.$$

Define $G_i(d_k) = Y_i(d_k) - g(\mathbf{X_i}, \xi_\mathbf{i}(d_k))$. Then, by substitution,

$$\widehat{Y_G^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{G_i(d_k)}{\pi_i(d_k)} + \sum_{i=1}^{N} g(\mathbf{X_i}, \xi_\mathbf{i}(d_k)). \quad (13)$$

Estimation proceeds as above using $\widehat{Y_G^T}(d_k)$ in place of $\widehat{Y^T}(d_k)$ to estimate $Y^T(d_k)$. In Appendix B, we demonstrate that $\widehat{Y_G^T}(d_k)$ is location invariant. Variance estimation proceeds as in section 5, using $G_i(d_k)$ in place of $Y_i(d_k)$ so long as $g(\mathbf{X_i}, \xi_\mathbf{i}(d_k))$ is fixed.

An approximation to the difference estimator is given by regression adjustment using the sample at hand. Regression can be thought of as a way to automate selection of the parameters in the difference estimator. In doing so, unbiasedness is compromised although the regression estimator is typically consistent (Särndal et al., 1992, pp. 225-239). As a general operating principle, we may use weighted nonlinear least squares to estimate a sensible parameter vector. For some common experimental designs, the least squares criterion will be optimal (Lin, 2011), and weighting by $1/\pi_i(d_k)$ ensures that the regression proceeds on a sample representative of the population of potential outcomes. With additional details on $\mathbf{I}_k$ and $g(.)$, it is possible to estimate optimal parameter vectors (Särndal et al., 1992, 219-244), though such values will typically be close to those produced by the weighted nonlinear least squares estimator (barring unusual and extreme forms of clustering).

Define an estimated parameter vector associated with exposure condition d_k,

$$\widehat{\xi}(d_k) = \arg\min_{\xi(d_k)} \sum_{i:D_i=d_k} \frac{1}{\pi_i(d_k)}\left[Y_i(d_k) - g(\mathbf{X_i}, \xi(d_k))\right]^2,$$

[8] This allows for the possibility that $\xi_\mathbf{i}(d_k)$ is a random variable. The condition $\xi_\mathbf{i}(d_k) \perp\!\!\!\perp D_i$ is also a sufficient condition for unbiasedness. Aronow and Middleton (2011) provide greater discussion of conditions for unbiased effect estimation and conservative variance estimation.

where $g(.)$ is the specification for the regression of $Y_i(d_k)$ on $\mathbf{I}(D_i = d_k)$ and $\mathbf{X_i}$. Then the regression estimator for the total is given by,

$$\widehat{Y_R^T}(d_k) = \sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{Y_i(d_k) - g\left(d_k, \mathbf{X_i}, \widehat{\xi}(d_k)\right)}{\pi_i(d_k)} + \sum_{i=1}^{N} g\left(d_k, \mathbf{X_i}, \widehat{\xi}(d_k)\right), \qquad (14)$$

Estimation proceeds as above using $\widehat{Y_R^T}(d_k)$ in place of $\widehat{Y_{HT}^T}(d_k)$ to estimate $Y^T(d_k)$. Under weak regularity conditions on $g(.)$, a variance estimator may be produced using a Taylor linearized form of $\widehat{Y_R^T}(d_k)$ (Särndal et al., 1992, 236-237).

8.2 Hajek ratio estimation

The Hajek (1971) ratio estimator is a refinement of the standard Horvitz-Thompson estimator that often facilitates efficiency gains at the cost of some finite sample bias and complications in variance estimation. Let us first consider the problem that the Hajek estimator is designed to resolve. The high variance of $\widehat{\mu_{HT}}(d_k)$ is often driven by the fact that some randomizations may yield an unusually large or small number of units or, depending on the nature of $\mathbf{I}_k$, an unusually large or small number of units with high values of the weights $1/\pi_i(d_k)$. The Hajek refinement allows the denominator of the estimator to vary according to the sum of the weights $1/\pi_i(d_k)$, thus shrinking the magnitude of the estimator when its value is large, and raising the magnitude of the estimator when its value is small.

$$\widehat{\mu_H}(d_k) = \frac{\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{Y_i(d_k)}{\pi_i(d_k)}}{\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{1}{\pi_i(d_k)}}. \qquad (15)$$

Note that $\mathrm{E}[\sum_{i=1}^{N} \mathbf{I}(D_i = d_k) \frac{1}{\pi_i(d_k)}] = N$, so that the Hajek estimator is the ratio of two unbiased estimators. It is well known that the ratio of two unbiased estimators is not an unbiased estimator of the ratio. However, the bias will tend to be small relative to the estimator's sampling variability, and we may place bounds on its magnitude.

By Hartley and Ross (1954) and Särndal et al. (1992, 176),

$$\left|\mathrm{E}[\widehat{\mu_H}(d_k)] - \mu(d_k)\right| \leq \sqrt{\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{I}(D_i = d_k)\frac{1}{\pi_i(d_k)}\right) \mathrm{Var}\left(\widehat{\mu_H}(d_k)\right)}$$

Under the asymptotic growth process hypothesized in section 6, both variances will converge to zero, and thus the bias (and, in fact, bias ratio) will converge to zero. Variance estimation again proceeds via Taylor linearization.

8.3 Constant effects variance estimation

Although the variance estimators proposed in section 5 are guaranteed to be conservative, the estimators may be imprecise or highly biased when A is large. An alternative approach

is to maintain a hypothesis of constant effects in order to derive a variance estimator. Such an estimator may be considerably more precise (as it pools variance across exposure conditions) and less biased, although the bias may be of unknown sign. For construction of confidence intervals, particularly in small experiments, such efficiency gains may be desirable.

To construct an approximate variance estimator, we may impute the full set of potential outcomes using the observed outcomes and estimated treatment effects. The imputed potential outcome under exposure d_k for unit i, $\widehat{Y}_i(d_k) = Y_i(d_l) + \widehat{\tau}_{HT}(d_k, d_l), \forall i, k, l$. We then substitute $\widehat{Y}_i(d_k)$ for $Y_i(d_k)$ in equation 5, and derive the estimator's variance using the imputed potential outcomes. Then the constant effects variance estimator, $\widehat{\text{Var}_{CE}}(\widehat{\tau_{HT}}(d_k, d_l)) =$

$$\frac{1}{N^2}\left\{\sum_{i=1}^{N}\pi_i(d_k)[1-\pi_i(d_k)]\left[\frac{\widehat{Y}_i(d_k)}{\pi_i(d_k)}\right]^2 + \sum_{i=1}^{N}\sum_{j\neq i}[\pi_{ij}(d_k) - \pi_i(d_k)\pi_j(d_k)]\frac{\widehat{Y}_i(d_k)}{\pi_i(d_k)}\frac{\widehat{Y}_j(d_k)}{\pi_j(d_k)}\right.$$
$$+\sum_{i=1}^{N}\pi_i(d_l)[1-\pi_i(d_l)]\left[\frac{\widehat{Y}_i(d_l)}{\pi_i(d_l)}\right]^2 + \sum_{i=1}^{N}\sum_{j\neq i}[\pi_{ij}(d_l) - \pi_i(d_l)\pi_j(d_l)]\frac{\widehat{Y}_i(d_l)}{\pi_i(d_l)}\frac{\widehat{Y}_j(d_l)}{\pi_j(d_l)}$$
$$\left.+\sum_{i=1}^{N}\sum_{j\neq i}\frac{\widehat{Y}_i(d_k)}{\pi_i(d_k)}\frac{\widehat{Y}_j(d_l)}{\pi_j(d_l)}\left[\pi_{ij}(d_k,d_l) - \pi_i(d_k)\pi_j(d_l)\right] - \sum_{i=1}^{N}\widehat{Y}_i(d_k)\widehat{Y}_i(d_l)\right\}.$$

Results for this estimator (including equivalencies to alternative, common estimators) are established in Samii and Aronow (2012) in the case of complete random assignment of a binary treatment variable. We have preliminary results for an arbitrary design suggesting that, under modest regularity conditions, $\widehat{\text{Var}_{CE}}$ is consistent if effects are indeed constant, though its properties are not as well established in the case of varying treatment effects.

9 Illustrations

Externalities, spill-over effects, carry-over effects, and treatment heterogeneity are ubiquitous in social scientific settings. Available methods either suppress consideration of that fact, accepting some bias in the estimation of unit-level causal effects for the sake of simplicity, or this issue is handled by defining causal effects at the group level, where wholly untreated groups provide a baseline condition (a "uniformity trial", cf. Rosenbaum (2007)). In this paper, we have shown how careful selection of an exposure model, combined with known information about direct treatment assignment, allows for consistent estimation of unit level causal effects associated with various forms of "indirect" exposure. These methods provide a principled basis for estimation any time one "indirectly" randomizes the assignment of units to exposure conditions, a situation that arises in a broad range of scenarios of substantive interest. We provide illustrations in the subsections below.

9.1 Spatial spillover in an environmental protection experiment

An interesting set of applications comes from when the effects of experimental treatments have the potential to transmit over space or through networks, and treatments are allocated to point locations in the space or network. For example, consider an environmental protection experiment in which forest monitoring stations are positioned at fixed points around the perimeter of a protected forest.[9] The goal is to determine an optimal allocation of monitoring stations so as to reduce risks (such as illegal cutting) sufficiently while not committing excessive resources. In this case, the units of analysis are segments of the forest, and exposure might be defined in terms of whether the segment centroid very close, moderately close, or far from the nearest monitoring station. In most cases, there will be irregularities in the places where stations could be established as well as irregularities in the spacing and orientation of the segments. For this reason, some segments may be in close proximity to multiple potential stationing points, whereas other may be in close proximity to only a few. Suppose the research design randomly selects M out of S potential locations to receive a monitoring station. Then, a forest segment's probability of being very close, moderately close, or far from a monitoring station will be determined by the combination of this random assignment ($\mathbf{Z}$) and the segment's location relative to the different potential stationing sites (θ_i). Using the methods above, one could generate the set of all stationing possibilities, record the exposure profile of the segments for each of the stationing possibilities, and then empirically determine the GPTs for each segment.

9.2 Network persuasion experiments

Another example of this sort comes from experiments with "viral" marketing and network persuasion campaigns, where the goal is to assess the effectiveness of different types of messages in inducing a cascade of persuasion to get people to support some position, purchase a product, or adopt a technology (Chen et al., 2010; Aral and Walker, 2011; Paluck, 2011). In this case, we might have baseline information on the network of communication and friendship ties between individuals. An exposure model might define exposures such as direct exposure to the message, indirect exposure by having a linked peer receive the message, or no direct or indirect exposure of this kind. Then, the first order and second order effects of the campaign can be measured in terms of contrasts between the last condition and each of the first and second conditions, respectively. In this case, if direct exposure ($\mathbf{Z}$) is assigned via a uniform random assignment mechanism, then a unit's probability of indirect exposure will depend on the number of ties that the unit has to others in the network (θ_i), and this in turn will affect that unit's probability of no exposure. The results of an experiment such as this could help to determine how much direct exposure is needed to induce the desired overall level of persuasion.

[9] The example is based on an actual evaluation in which the authors have been involved.

9.3 Multi-stage trials with grouped interference

As discussed above, Hudgens and Halloran (2008) and Tchetgen-Tchetgen and VanderWeele (2010) develop methods for estimating causal effects under "partial" interference in two-stage randomized trials of disease prophylaxis treatments. Interference due to disease transmission across units is presumed to be limited to taking place within, and not between, groups, and direct treatment is withheld altogether for some groups. Two-stage trials of this sort combine issues associated with interference and cluster-level randomization. The methods developed here straightforwardly accommodate both of these issues. For example, consider a two-stage trial, where groups are first assigned to be either treatment sites or control sites with probabilities $(\lambda, 1-\lambda)$, respectively, and then units within treatment sites are assigned to receive treatment with probability γ. Then, a reasonable exposure model might assume three exposure conditions: (i) control site, (ii) treated site, but no treatment, (iii) treated site with treatment. Identification of causal effects under this model is straightforward, as the design yields homogenous GPTs equal to $(1-\lambda, \lambda(1-\gamma), \lambda\gamma)$, respectively, for all units. Inference for effect estimates is complicated by the clustered nature of the treatment assignment: for units i, j in the same group, $\pi_{ij}(d_k, d_l) = 0$, for $d_k \neq d_l$. This adds an additional complication to estimating (6). The issue was noted by Hudgens and Halloran (2008). The theoretical development above provides a way to gain additional traction on the problem. The solution we propose above is to use the conservative approximation given in (10) or to use the constant effects variance estimator.

9.4 Dynamic experiments

Dynamic experiments have time-varying treatment assignments. Exposure in this context could be defined in terms of a unit's treatment history. A prominent example of a dynamic experiment is the "stepped-wedge" design, in which there are a fixed number of periods, and in each period, some proportion of non-treated subjects are permanently assigned to treatment for all future periods. Outcomes are observed for all subjects in each period, so that the unit of inference is the subject-period. For analyzing per-period effects, one would want to account for possible interference due to effects from a subject's assignment in previous periods carrying over into the current period. For example, suppose a stepped-wedge experiment with three periods. An exposure model in this case might define, $\Delta_1 = \{(1,1,1),(0,1,1),(0,0,1),(0,0,0)\}$, to define treatment initiated in periods 1, 2, 3, or never, respectively. Then, simple random assignment to each of these three exposure conditions provides for very straightforward identification and inference. Suppose, however, that there is good reason to believe that carry-over effects are likely to last only one period. The analyst then may use an alternative exposure model, $\Delta_2 = \{(1,1),(0,1),(0,0)\}$, to indicate two consecutive periods of exposure, only one period of exposure, and no exposure, respectively. If the experiment randomly assigned the histories enumerated in Δ_1, then the probabilities of assignment to the conditions in Δ_2 would vary over the Δ_1 conditions. Brown and Lilford (2006) review applications of stepped-wedge designs in medical research, and Gerber et al. (2011) is an application from political science.

10 Conclusion

We have proposed methods for estimating average causal effects under general, but known, forms of interference. Our approach combines a known randomization process with the analyst's model of treatment exposure, thus permitting inference under clear and defensible assumptions. Importantly, the union of the design of the experiment and the exposure model may imply unequal probabilities of exposure and forms of dependence between units that may not be obvious ex ante. In constructing parsimonious estimators that naturally account for such complications, we have demonstrated how obtaining reliable causal inference may depend critically on the *joint* role of randomization and modeling assumptions, however minimal.

We develop estimators based on results from the literature on unequal probability sampling rooted in the foundational insights of Horvitz and Thompson (1952). The estimators are clearly derived from the known sampling distribution of the "direct" treatment, $\mathbf{Z}$, and provide a basis for unbiased effect estimation and weakly conservative variance estimation. Normal approximation intervals straightforwardly characterize the distribution of the effect estimates for a setting analogous to the aggregation of results from repeated experimentation on a fixed population. Nonetheless, it is well known that Horvitz-Thompson-type estimators may be volatile in cases where selection probabilities vary greatly or exhibit strong inverse correlation with outcome values (Basu, 1971). Thus, we provide refinements that allow for variance control via covariance adjustment and Hajek estimation. In addition, we provide a method of variance estimation based on hypotheses of the nature of causal effects which may be preferred when design-based estimators are unstable.

Some readers may raise objections or have concerns about how the methods proposed here rely on an exposure model. Does this not introduce arbitrariness to the analysis? The question is misguided: there is no escaping specification of exposure models for causal analysis. Consider the classical approach to inference under the Neyman-Rubin model. Here, analysts typically assume a very specific exposure model—namely, one that assumes no interference relative to unit-level treatment assignments. This typical Neyman-Rubin model is nested within more general exposure models that allow for some forms of interference, which are in turn nested within other exposure models that place fewer restrictions on the form of the interference. Our approach permits estimation and testing under an arbitrarily general exposure model. Generally speaking, unless $|\Delta| = |\Omega| \times N$, the analyst may always estimate average potential outcomes under a less restrictive exposure model that allows for additional forms of interference, thus allowing for the enumeration of additional potential outcomes. Then the analyst may test for significant differences between the hypothesized potential outcomes and those associated with a nested model. Rejection of the null hypothesis of no mean difference between potential outcomes provides support for the more complex exposure model.[10] While issues of model specification may be unavoid-

[10] A related issue arises when one thinks that one has the appropriate *form* of the exposure model but one is uncertain about certain structural features that precisely determine a unit's exposure probabilities. A concrete example comes from experiments on social networks. Suppose we assign a treatment to people in a network and we expect that those directly tied to the treated people will receive a meaningful dose of indirect

able, the proposed framework allows for inference under arbitrarily flexible (and testable) assumptions on the exposure model.

To summarize, our approach combines a limited set of modeling assumptions with randomization-based estimation and may be characterized as a design consistent, but also "model assisted," approach to causal effect estimation, similar in spirit to Särndal et al. (1992)'s approach to general survey sampling. It is an alternative to parametric approaches that are often employed with little substantive justification for all of the modeling assumptions. These methods greatly extend the reach of randomization-based estimation of causal effects, allowing researchers to explore causal effects on units other than those for which treatment is manipulated directly. Our illustrations suggest that number of substantively important applications is vast.

exposure. The problem may be that we do not know for sure the links between people. Such uncertainty could be formalized in terms of a probability distribution over the θ_i's, in which case one could marginalize over a range of estimates that first condition on a set of θ_i's. Such is beyond the scope of this paper, but for a related approach based on parametric data augmentation, see Chandrasekhar and Lewis (2012).

References

Aral, S. and D. Walker (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* (online advance access), 1–17.

Aronow, P. M. and J. A. Middleton (2011). Unbiased estimation of the average treatment effect in randomized experiments. Manuscript, Yale University and New York University.

Aronow, P. M. and C. Samii (2011). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. Unpublished manuscript, Yale University and New York University.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part i. In V. Godambe and D. Sprott (Eds.), *Foundations of Statistical Inference*, pp. s. Toronto: Holt, Rinehart, and Winston.

Brewer, K. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association 74*(368), 911–915.

Brown, C. A. and R. J. Lilford (2006). The stepped wedge trial design: A systematic review. *BMC Medical Research Methodology 6*(54), (online).

Chandrasekhar, A. G. and R. Lewis (2012). Econometrics of sampled networks. Manuscript, Massachussetts Institute of Technology.

Chen, J., M. Humphreys, and V. Modi (2010). Technology diffusion and social networks: Evidence from a field experiment in Uganda. Manuscript, Columbia University.

Cox, D. R. (1958). *Planning of Experiments*. Wiley.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika 93*, 269–278.

Freedman, D., R. Pisani, and R. Purves (1998). *Statistics, 3rd Ed.* New York: W.W. Norton.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken: Wiley.

Gerber, A. S., J. G. Gimpel, D. P. Green, and D. R. Shaw (2011). How large and long-lasting are the persuasive effects of television campaign ads? Results from a randomized field experiment. *American Political Science Review 105*(1), 135–150.

Hajek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one,". In V. Godambe and D. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Hartley, H. O. and A. Ross (1954, 08). Unbiased ratio estimators. *Nature 174*(4423), 270–271.

Hoeffding, W. and H. Robbins (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal 15*(3), 773–780.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*, 706–710.

Lehmann, E. L. (1999). *Elements of Large Sample Theory*. New York: Springer-Verlag.

Lin, W. (2011). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. Manuscript, Department of Statistics, University of California, Berkeley.

Middleton, J. A. and P. M. Aronow (2011). Unbiased estimation of the average treatment effect in cluster-randomized experiments. Paper Presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

Mittelhammer, R. C. (1996). *Mathematical Statistics for Economics and Business*. New York: Springer.

Neyman, J. S. (1990 [1923], November). On the application of probability theory to agricultural experiments, essay on principles, section 9 (reprint). *Statistical Science 5*(4), 465–472.

Paluck, E. L. (2011). Peer pressure against prejudice: A high school field experiment examining social network change. *Journal of Experimental Psychology 47*, 350–358.

Raj, D. (1965). On a method of using multi-auxiliary information in sample surveys. *Journal of the American Statistical Association 60*, 270–277.

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

Rubin, D. B. (1990). Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference 25*(3), 279–292.

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

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

Tchetgen-Tchetgen, E. J. and T. J. VanderWeele (2010). On causal inference in the presence of interference. *Statistical Methods in Medical Research Online early content*, 1–21.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics 24*, 53–78.

Appendix

A Equivalence of using combined clusters

We rewrite $\widehat{Y_{HT}^T}(d_k)$ to account for clustering in exposure. While the treatment effect estimators are identical, such a notational switch will allow us to simplify and reduce the bias of our eventual variance estimators.) Noting that if, for some units i, j, $\rho\left(\mathbf{I}(D_i = d_k), \mathbf{I}(D_j = d_k)\right) = \rho_{ij}(d_k, d_k) = 1$, then the total estimators may be equivalently rewritten. Define $M_m \in \mathbf{M}$ as the set of unit indices i, j that satisfy $\rho_{ij}(d_k, d_k) = 1$, where $\mathbf{I}'(D_m = d_k)$ is indexed over all $|\mathbf{M}|$ unique random variables in $\{\mathbf{I}(D_i = d_k) : i = (1, 2, ..., N)\}$. Define $\pi'_m(d_k)$ as the value of $\pi_i(d_k), \forall i \in M_m$. Joint probabilities $\pi'_{kl}(d_k, d_l)$ are defined analogously. Given these definitions, we can rewrite the HT estimator of the total of treatment potential outcomes as

$$\begin{aligned}
\widehat{Y_{HT}^T}(d_k) &= \sum_{i=1}^{N} \frac{1}{\pi_i} \mathbf{I}(D_i = d_k) Y_i(d_k) = \sum_{m=1}^{|\mathbf{M}|} \sum_{i \in M_m} \frac{1}{\pi_i(d_k)} \mathbf{I}(D_i = d_k) Y_i(d_k) \\
&= \sum_{m=1}^{|\mathbf{M}|} \sum_{i \in M_m} \frac{1}{\pi'_m(d_k)} \mathbf{I}'(D_m = d_k) Y_i(d_k) = \sum_{m=1}^{|\mathbf{M}|} \frac{1}{\pi'_m(d_k)} \mathbf{I}'(D_m = d_k) \sum_{i \in M_m} Y_i(d_k) \\
&= \sum_{m=1}^{|\mathbf{M}|} \frac{1}{\pi'_m(d_k)} \mathbf{I}'(D_m = d_k) Y'_m(d_k),
\end{aligned}$$

where $Y'_m(d_k) = \sum_{i \in M_m} Y_i(d_k)$. Since clustered units will always be observed together, they can be summed prior to estimation. The equivalency of these totaled and untotaled HT estimators serves as the basis for the estimation approach in Middleton and Aronow (2011).

B Location invariance of the difference estimator

We follow the formulations of Aronow and Middleton (2011). Invariance to location shifts implies that the estimates do not substantively change with linear transformations of the outcome. If the outcome changes such that $Y_i^*(d_k) = b_0 + b_1 Y_i(d_k)$, we will assume that the predictive function will also change by the identical transformation such that

$$g^*\left(\mathbf{X_i}, \xi_{\mathbf{i}}(d_k)\right) = b_0 + b_1 g\left(\mathbf{X_i}, \xi_{\mathbf{i}}(d_k)\right).$$

Since $g\left(\mathbf{X_i}, \xi_{\mathbf{i}}(d_k)\right)$ is a function designed to predict the value of Y_i, then the predicted outcome would change according to the scale of the outcome.

The difference estimator for covariance adjustment is then therefore invariant to location shifts, so that, when the estimator is applied to $Y_i^*(d_k) = b_0 + b_1 Y_i(d_k)$,

$$\widehat{Y_G^T}^*(d_k) = N b_0 + b_1 \widehat{Y_G^T(d_k)}.$$

By algebraic manipulations,

$$
\begin{aligned}
\widehat{Y_G^T}^*(d_k) &= \sum_{i=1}^{N}\left[\mathbf{I}(D_i=d_k)\frac{Y_i^*(d_k)}{\pi_i(d_k)}-\mathbf{I}(D_i=d_k)\frac{g^*\left(\mathbf{X_i},\xi_{\mathbf{i}}(d_k)\right)}{\pi_i(d_k)}\right]+\sum_{i=1}^{N}g^*\left(\mathbf{X_i},\xi_{\mathbf{i}}(d_k)\right)\\
&= \sum_{i=1}^{N}\mathbf{I}(D_i=d_k)\left[\frac{b_0+b_1Y_i(d_k)}{\pi_i(d_k)}-\frac{b_0+b_1g\left(\mathbf{X_i},\xi_{\mathbf{i}}(d_k)\right)}{\pi_i(d_k)}\right]+\sum_{i=1}^{N}\left[b_0+b_1g\left(\mathbf{X_i},\xi_{\mathbf{i}}(d_k)\right)\right]\\
&= \sum_{i=1}^{N}\mathbf{I}(D_i=d_k)\left[\frac{b_1Y_i(d_k)}{\pi_i(d_k)}-\frac{b_1g\left(\mathbf{X_i},\xi_{\mathbf{i}}(d_k)\right)}{\pi_i(d_k)}\right]+Nb_0+\sum_{i=1}^{N}b_1g\left(\mathbf{X_i},\xi_{\mathbf{i}}(d_k)\right)\\
&= Nb_0+b_1\widehat{Y_G^T(d_k)}.
\end{aligned}
$$