ECONOMIC GROWTH CENTER
YALE UNIVERSITY
P.O. Box 208629
New Haven, CT 06520-8269
http://www.econ.yale.edu/~egcenter/

CENTER DISCUSSION PAPER NO. 969

Adaptive Experimental Design Using the Propensity Score

Jinyong Hahn
UCLA

Keisuke Hirano
University of Arizona

Dean Karlan
Yale University
M.I.T. Jameel Poverty Action Lab

January 2009
Revised July 22, 2009

Notes: Center Discussion Papers are preliminary materials circulated to stimulate discussions and critical comments.

Kyle Hood and Mario Samano provided research assistance. Hahn was supported in part by the National Science Foundation under grant SES-0819638. Hirano was supported in part by a grant from the Eller College of Management at the University of Arizona. Karlan was supported in part by the National Science Foundation under grant SES-0547898. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect those of the National Science Foundation or the Eller College of Management.

Adaptive Experimental Design using the Propensity Score

Jinyong Hahn
UCLA

Keisuke Hirano
University of Arizona

Dean Karlan
Yale University and M.I.T. Jameel Poverty Action Lab

15 April 2008

Abstract

Many social experiments are run in multiple waves, or replicate earlier social experiments. In principle, the sampling design can be modified in later stages or replications to allow for more efficient estimation of causal effects. We consider the design of a two-stage experiment for estimating an average treatment effect, when covariate information is available for experimental subjects. We use data from the first stage to choose a conditional treatment assignment rule for units in the second stage of the experiment. This amounts to choosing the propensity score, the conditional probability of treatment given covariates. We propose to select the propensity score to minimize the asymptotic variance bound for estimating the average treatment effect. Our procedure can be implemented simply using standard statistical software and has attractive large-sample properties.

JEL codes: C1, C9, C13, C14, C93

Keywords: experimental design, propensity score, efficiency bound

1 Introduction

Social experiments have become increasingly important for the evaluation of social policies and the testing of economic theories. Random assignment of individuals to different treatments makes it possible to conduct valid counterfactual comparisons without strong auxiliary assumptions. On the other hand, social experiments can be costly, especially when they involve policy-relevant treatments and a large number of individuals. Thus, it is important to design experiments carefully to maximize the information gained from them. In this paper, we consider social experiments run in multiple stages, and examine the possibility of using initial results from the first stage of an experiment to modify the design of the second stage, in order to estimate the average treatment effect more precisely. Replications of earlier social experiments can also be viewed as multiple stage experiments, and researchers may find it useful to use earlier published results to improve the design of new experiments. We suppose that in the second stage, assignment to different treatments can be randomized *conditional* on some observed characteristics of the individual. We show that data from the first wave can reveal potential efficiency gains from altering conditional treatment assignment probabilities, and suggest a procedure for using the first-stage data to construct second-stage assignment probabilities. In general, the treatment effect can be estimated with a lower variance than under pure random sampling using our sequential procedure.

This technique can be applied to two types of studies. First, many social experiments have a pilot phase or some more general multi-stage or group-sequential structure. For instance, Simester, Sun, and Tsitsiklis (2006) conduct repeated experiments with the same retailers to study price sensitivities. Karlan and Zinman (2008) conduct repeated experiments with a microfinance lender in South Africa to study interest rate sensitivities. Second, for many research questions we have seen a plethora of related social experiments, such as get-out-the-vote experiments in political science (see Green and Gerber, 2004), charitable fundraising experiments in public finance, and conditional cash transfer evaluations in development economics. To illustrate our procedure, we use data from three studies to optimize a hypothetical future wave of a similar social experiment: the first and second from two charitable fundraising experiments, and the third from a conditional cash transfer evaluation (Gertler, Martinez and Rubio-Codina, 2006). Our approach is appropriate when later stages or replications are applied to the same population and same treatments as in the initial stage; if the later replications do not satisfy this requirement, but involve similar populations or have similar treatments, then our results could still be useful to suggest alternative designs which maintain the key benefits of randomization but can improve precision.

Randomizing treatment conditional on covariates amounts to choosing the *propensity score*—the conditional treatment probability. Rosenbaum and Rubin (1983) proposed to use the propensity score to estimate treatment effects in observational studies of treatments under the assumption of unconfoundedness. Propensity score methods can also be used in pure randomized experiments

to improve precision (for example, see Flores-Lagunes, Gonzalez, and Neumann, 2006). When treatment is random conditional on covariates, the semiparametric variance bound for estimating the average treatment effect depends on the propensity score and the conditional variance of outcomes given treatment and covariates. We propose to use data from the first stage to estimate the conditional variance. Then we *choose* the propensity score in the second stage in order to minimize the asymptotic variance for estimating the average treatment effect. Finally, after data from both stages has been collected, we pool the data and construct an overall estimate of the average treatment effect. If both stages have a large number of observations, the estimation error in the first-stage preliminary estimates does not affect the asymptotic distribution of the final, pooled estimate of the treatment effect. Our procedure is "adaptive" in the sense that the design uses an intermediate estimate of the conditional variance structure, and does as well asymptotically as an infeasible procedure that uses knowledge of the conditional variances.

There is an extensive literature on sequential experimentation and experimental design, but much of this work focuses on stopping rules for sequential sampling of individuals, or on "play-the-winner" rules which increase the probability of treatments which appear to be better based on past data. Bayesian methods have also been developed for sequential experimental design; for a recent review of Bayesian experimental design see Chaloner and Verdinelli (1995). Unlike some recent work taking a simulation-based Bayesian approach, our approach is very simple and does not require extensive computations. (R code to implement our procedures is available at `http://www.u.arizona.edu/~hirano/R/hhk_scripts.R`, and Stata code is available at `http://research.yale.edu/karlan/downloads/hhk.zip`.) However, our analysis is based on asymptotic approximations where the sample size in each stage of the experiment is taken as large. Thus, our formal results would apply best to large-scale social experiments, rather than the small experiments sometimes conducted in laboratory settings.

Our approach is also closely related to the Neyman allocation formula (Neyman, 1934) for optimal stratified sampling. Manski and McFadden (1981) also discuss the possibility of using pilot or previous studies to help choose a stratification design. Some authors, such as Sukhatme (1935), have considered the problem of estimating the optimal strata sizes using preliminary samples, but in a finite-population setting where it is difficult to obtain sharp results on optimal procedures. A review of this literature is given in Solomon and Zacks (1970). Our asymptotic analysis lead to a simple adaptive rule which has attractive large-sample properties.

2 Adaptive Design Algorithm and Asymptotic Theory

2.1 Two-Stage Design Problem

We consider a two-stage social experiment comparing two treatments. In each stage we draw a random sample from the population. We assume that the population of interest remains the same across the two stages of experimentation. (We discuss extensions to replications and other settings where the population of interest changes across stages in Section 5.) For each individual we observe some background variables X, and assign the individual to one of two treatments. We will use "treatment" and "control" and "1", "0" to denote the two treatments. Let n_1 denote the number of observations in the first stage, and let n_2 denote the number of observations in the second stage, and let $n = n_1 + n_2$.

In order to develop the formal results below, we assume that the covariate X_i has finite support. If X_i is continuously distributed, we can always discretize it, and if there are multiple covariates we can form cells and define X_i accordingly. Further, since we will be making treatment assignment probabilities depend on X_i, it may be preferable to work with discretized covariates for operational purposes. All of our results to follow will still hold under discretization, although discretizing too coarsely may sacrifice some precision in estimating treatment effects. More precisely, a finer discretization may lead to a lower variance bound for estimating average treatment effects (similar in some respects to bandwidth choice problems in kernel estimation). However, it can also lead to more estimation error for a key set of conditional variance terms, leading to poor asymptotic approximations which cannot be captured by our first-order analysis. We take the level of discretization as fixed, leaving for future work the difficult question of choosing the level of discretization.

In the first stage, individuals are assigned to treatment 1 with probability π_1, which does not depend on their observed covariates. Before the second stage, the outcomes from the first stage are realized and observed by the experimental designer. In the second stage, the designer can make treatment assignment probabilities depend on the individual's covariate X. Let $\widehat{\pi}_2(x)$ denote the probability that a second-stage individual with $X_i = x$ receives treatment 1. We use the "hat" to indicate the these probabilities can depend on all the data from the first stage. The goal is to estimate the population average treatment effect with low mean-squared error.

Formally, for individuals $i = 1, 2, \ldots, (n_1+n_2)$, let (X_i, Y_{0i}, Y_{1i}) be IID from a joint distribution. We interpret X_i as the (always observed) vector of covariates, and Y_{ti} as the potential outcome under treatment $t = 0, 1$. We are interested in estimation of the average treatment effect

$$\beta := E[Y_{1i} - Y_{0i}].$$

Individuals $i = 1, \ldots, n_1$, drawn in the first stage, are assigned treatment D_i equal to 1 with

probability π_1, and 0 with probability $1 - \pi_1$. The experimental planner then observes (X_i, D_i, Y_i), where

$$Y_i := D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

Similarly, for $i = n_1 + 1, \ldots, n_1 + n_2$, we assign individuals to treatments according to $P(D_i = 1 | X_i = x) = \widehat{\pi}_2(x)$, and we observe (X_i, D_i, Y_i).

We can also consider a constrained version of the experimental design problem, where the overall probability of treatment is required to equal a fixed value p. In this case the assignment rule $\widehat{\pi}_2(\cdot)$ must satisfy

$$p = \frac{n_1}{n}\pi_1 + \frac{n_2}{n} E[\widehat{\pi}_2(X_i)],$$

where $n = n_1 + n_2$ and the expectation is with respect to the marginal distribution of X. In the sequel, we will consider both the unconstrained and constrained design problems. It would also be straightforward to extend the analysis to cases where there is an upper or lower bound on the the overall treatment probability, or other constraints.

2.2 One-Stage Problem and Optimal Propensity Score

Before giving our proposal for an adaptive experimental design rule, it is useful to consider the simpler problem of estimating the average treatment effect under a fixed treatment assignment rule.

Suppose that $(X_i, Y_{0i}, Y_{1i}, D_i)$ are IID from a population for $i = 1, \ldots, n$, and that the treatment assignment rule depends only on X_i:

$$D_i \perp (Y_{0i}, Y_{1i}) | X_i.$$

Let

$$p(x) := Pr(D_i = 1 | X_i = x).$$

The function $p(x)$ is often called the propensity score (Rosenbaum and Rubin, 1984). We also require that for all possible values of X, $0 < p(X) < 1$. In a randomized experiment, this overlap condition can be guaranteed by design.

As before, the average treatment effect $\beta = E[Y_{1i} - Y_{0i}]$ is the object of interest. Typically, there will exist estimators $\widehat{\beta}$ such that $\widehat{\beta} \xrightarrow{p} \beta$ and

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, V).$$

We wish to find an estimator with minimal asymptotic variance V. The following result, due to Hahn (1998), provides a lower bound for the variance of regular estimators. (See Chamberlain

(1986), for a discussion of regularity and semiparametric variance bounds.)

Proposition 1 *(Hahn, 1998) Let*

$$\begin{aligned}\beta(x) &:= E[Y_{1i} - Y_{0i}|X_i = x],\\ \sigma_0^2(x) &:= V[Y_{0i}|X_i = x],\\ \sigma_1^2(x) &:= V[Y_{1i}|X_i = x]\end{aligned}$$

Then any regular estimator $\widehat{\beta}$ for β has asymptotic variance

$$V \geq E\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)} + (\beta(X_i) - \beta)^2\right].$$

Estimators that achieve this bound have been constructed by Hahn (1998), Hirano, Imbens, and Ridder (2003) (hereafter HIR), and others. Consider the following two-step estimator proposed by HIR. Let $\widehat{p}(x)$ be a nonparametric regression estimate of $p(x) = E[D_i|X_i = x]$. The HIR estimator is

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}(X_i)}\right).$$

In our setting, with discrete covariate, we can simply set $\widehat{p}(x)$ equal to the empirical probability of treatment for observations with $X_i = x$. This is numerically equivalent to estimating the propensity score with a saturated logit or probit model, and is also numerically equivalent to regression-based estimators such as Hahn's estimator.

Now suppose that the researcher can choose the propensity score $p(x)$. The researcher would like to solve

$$\min_{p(\cdot)} E\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)} + (\beta(X_i) - \beta)^2\right] \tag{1}$$

If there is a constraint on the overall treatment probability, this minimization is subject to the constraint

$$E[p(X_i)] = p$$

In the constrained case, an interior solution $p(\cdot)$ will satisfy

$$-\frac{\sigma_1^2(x)}{p(x)^2} + \frac{\sigma_0^2(x)}{(1-p(x))^2} = \lambda \tag{2}$$

for all x in the support of X, where λ denotes the Lagrange multiplier.

In both the constrained and unconstrained problems, the solution depends on the conditional variances $\sigma_0(x)$ and $\sigma_1(x)$. Intuitively, if the data exhibit large differences in conditional variances by x, then allowing for different treatment probabilities for different x may permit more precise

estimation of the treatment effect. In essence, heteroskedasticity drives the possibility for improved precision.

2.3 Two-Stage Adaptive Design and Estimator

The optimization problem (1) implicitly assumes that the conditional variance functions $\sigma_1^2(X_i)$ and $\sigma_0^2(X_i)$ are known to the researcher, and therefore is not feasible in a one-stage setting. However, if the experiment is run in two stages, one can use the first stage to estimate the unknown variance functions. We propose to use the first stage results to estimate $\sigma_1^2(X_i)$ and $\sigma_0^2(X_i)$, and then use these estimates to modify the treatment assignment probabilities in the second stage. We show that if the sample sizes in both stages are large, the overall design is "adaptive" — we achieve the same overall efficiency as the infeasible version that uses knowledge of the conditional variances. Our overall design and estimation procedure is implemented in the following steps:

1. In Stage 1, we assign individuals $i = 1, \ldots, n_1$ to treatment 1 with probability π_1, irrespective of their covariate values. We collect data (D_i, X_i, Y_i) for these individuals.

2. Using data from Stage 1, we estimate the conditional variances $\sigma_0^2(x)$ and $\sigma_1^2(x)$ by their empirical analogs: $\widehat{\sigma}_0^2(x)$ is the sample variance of Y for first-stage observations with $D = 0$ and $X = x$, and $\widehat{\sigma}_1(x)$ is the sample variance of Y for first-stage observations with $D = 1$ and $X = x$. We then choose $\widehat{\pi}_2(x)$ to minimize the empirical version of the variance bound:

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left[\frac{\widehat{\sigma}_1^2(X_i)}{\pi(X_i)} + \frac{\widehat{\sigma}_0^2(X_i)}{1-\pi(X_i)} + (\beta(X_i) - \beta)^2\right]$$

   where

$$\pi(x) = \kappa\pi_1 + (1-\kappa)\widehat{\pi}_2(x),$$

$$\kappa = \lim \frac{n_1}{n_1 + n_2}.$$

   In practice we can take $\kappa = n_1/(n_1 + n_2)$. As before, if there is a constraint that the overall treatment probability is equal to p, then the minimization is subject to:

$$E[\pi(X_i)] = p.$$

   Here, all of the expectations are with respect to the marginal distribution of X_i. Note that the solution does not depend on $(\beta(X_i) - \beta)^2$, so we can drop this term from the objective function when solving the minimization problem. More detail about the computation of the $\widehat{\pi}_2(x)$ is given in the Appendix.

3. We assign individuals $i = n_1 + 1, \ldots, n_1 + n_2$ to treatment 1 with probabilities $\widehat{\pi}_2(X_i)$. We collect data (D_i, X_i, Y_i) from the second stage individuals, and estimate the average treatment effect β using the Hahn/HIR estimator

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}(X_i)}\right).$$

Note that this estimator involves estimating a propensity score. Although the propensity score is known (because it is controlled by the researcher), the estimator does not use the true propensity score. (The efficiency gain from using an estimate of the propensity score rather than the true propensity score is discussed in HIR.)

In the second step of our procedure, it is possible to have a corner solution, because the first stage randomization restricts the set of possible propensity scores achievable over the two stages. In particular, for any x, the overall conditional probability $\pi(x)$ cannot be less than $\kappa\pi_1$, and cannot be greater than $\kappa\pi_1 + (1-\kappa) = 1 - \kappa(1-\pi_1)$. However, our results to follow do not require an interior solution.

2.4 Asymptotic Theory

Our asymptotic theory is based on the regularity conditions stated below as Assumption 1:

Assumption 1 *(i) $n_1 \to \infty$ and $n_2 \to \infty$ such that $n_1/(n_1+n_2) \to \kappa$; (ii) X_i has a multinomial distribution with finite support; (iii) $\pi_2^*(\cdot)$ depend smoothly on the vectors $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$, where $\pi_2^*(x) :=$ plim $\widehat{\pi}_2(x)$; (iv) the estimators $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n}$-consistent for the true variances $\sigma_0^2(x)$ and $\sigma_1^2(x)$.*

The most notable aspect of Assumption 1 is the double asymptotics, in which n_1 and n_2 go to infinity at the same rate. This assumption can be relaxed somewhat, but we maintain it to keep the analysis relatively simple. The assumption that $\pi_2^*(\cdot)$ depends smoothly on the vectors $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$ is innocuous when the X_i has a multinomial distribution with finite support. The assumption that $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n_1}$-consistent is also harmless under the multinomial assumption. Because $n_1 = O(n)$, it follows that $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n}$-consistent.

Since the estimators $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n}$-consistent for the true variances, it follows that the second stage assignment probabilities $\widehat{\pi}_2(x)$ as defined in Step 2 of the algorithm are $\sqrt{n}$-consistent for $\pi_2^*(x)$. We also use $\pi^*(x)$ to denote the target overall propensity scores, defined as

$$\pi^*(x) := \kappa\pi_1 + (1-\kappa)\pi_2^*(x).$$

Because the assignment probabilities in the second stage depend on the realization of the first-stage data, we do not have classic IID sampling. To develop the formal results, we must take into account the dependence of the second-stage DGP on the first stage data. We do this by viewing the treatment indicators as being generated by IID uniform random variables. In the first stage,

$$D_i = 1(U_i \leq \pi_1),$$

where U_i are IID Uniform[0,1] random variables, independent of the X and Y variables. For individuals $i = n_1 + 1, \ldots, n_1 + n_2$, drawn in the second stage, treatment is determined according to an assignment rule as $\widehat{\pi}_2(X_i)$, where the "hat" indicates that the rule can depend on first-stage data. Treatment can then be defined as

$$D_i = 1(U_i \leq \widehat{\pi}_2(X_i)),$$

and we observe (X_i, D_i, Y_i) where Y_i is defined as before.

Defining the treatment indicator in this way allows us to view the underlying data as IID. Because the second-step estimator $(\widehat{\beta})$ becomes a nondifferentiable function of the first-step estimator $(\widehat{\pi}_2(x))$, we use empirical process arguments instead of standard results on two-step GMM estimators (e.g. Newey and McFadden (1994)) to derive properties of $\widehat{\beta}$ in the Appendix.

The following result shows that the two-stage design procedure, combined with the Hahn/HIR estimator, is adaptive: the estimator has asymptotic variance equal to the variance that would obtain had we used $\pi^*(x)$ to assign individuals to treatment.

Theorem 1 *Let (i)* $\pi_2^*(x) := plim\ \widehat{\pi}_2(x)$*; and (ii)* $\pi^*(x) := \kappa\pi_1 + (1-\kappa)\pi_2^*(x)$*. Assume that* $\widehat{\pi}_2(x) = \pi_2^*(x) + o_p\left(\frac{1}{\sqrt{n}}\right)$*. Further assume that* $0 < \pi^*(x) < 1$*. We then have*

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, E\left[\frac{1}{\pi^*(X_i)}\sigma_1^2(X_i) + \frac{1}{1-\pi^*(X_i)}\sigma_0^2(X_i) + (\beta(X_i) - \beta)^2\right]\right)$$

Proof: See Appendix B □

Our result requires that the conditional variances (and hence the target treatment probabilities) be estimated at a $\sqrt{n}$ rate. The result can be extended to the case where the covariate is continuous, provided that the conditional variances are parametrized so that they can be estimated at parametric rates.

Theorem 1 can be used to calculate the standard error of the estimator $\widehat{\beta}$ as follows:

- Keep $\widehat{\sigma}_0^2(x)$, $\widehat{\sigma}_1^2(x)$, and $\pi(x)$ from Step 2 in the experimental design. (The estimators of $\sigma_0^2(x)$ and $\sigma_1^2(x)$ can be recalculated using the entire sample for more accuracy. Under our

asymptotic framework, it does not make any difference whether $\sigma_0^2(x)$ and $\sigma_1^2(x)$ are estimated using only first subsample or the entire sample.)

- Keep $\widehat{p}(x)$ in Step 3.
- Calculate the standard error as

$$SE = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\widehat{\sigma}_1^2(X_i)}{\pi(X_i)} + \frac{\widehat{\sigma}_0^2(X_i)}{1-\pi(X_i)} + \left(\widehat{\beta}(X_i) - \widehat{\beta} \right)^2 \right)}$$

where

$$\widehat{\beta}(x) = \frac{\frac{\sum_{i=1}^{n} D_i Y_i 1(X_i=x)}{\sum_{i=1}^{n} 1(X_i=x)}}{\widehat{p}(x)} - \frac{\frac{\sum_{i=1}^{n} (1-D_i) Y_i 1(X_i=x)}{\sum_{i=1}^{n} 1(X_i=x)}}{1-\widehat{p}(x)}$$

The interval $\widehat{\beta} \pm 1.96SE$ then has the usual 95% (asymptotic) coverage probability.

3 Examples

In this section we give three simple numerical examples of our adaptive design algorithm, using data from recently conducted social experiments. Two of these applications were single-stage experiments. For the purpose of illustration we suppose that the researcher has the ability to carry out a second round of the same experiment. We use our adaptive algorithm, along with the data from the "first" round, to determine how the second stage should be carried out. In the first example, a charitable fundraising experiment (Karlan and List, 2007), we find significant efficiency gains from employing our adaptive treatment assignment rule. In the second example, also a charitable fundraising experiment, we also find potential efficiency gains, and have used our procedure to guide the design of the second wave of the experiment, currently underway. In the third example we use results reported in a World Bank working paper on an evaluation of a conditional cash transfer program in Mexico (Progresa) to estimate the potential efficiency gains if the study were to be replicated elsewhere. We find only a small efficiency gain, but include this as an example of how to use third-party published results to improve power in new studies.

3.1 Direct Mail Fundraising Experiment

For the first example we use data from a direct mail fundraising experiment reported in Karlan and List (2007). In this experiment a charitable organization mailed 50,083 direct mail solicitations to prior donors to their organization. Of the 50,083, two-thirds (33,396) received a matching grant offer, and one-third (16,687) received the same solicitation but without mention of a matching grant. (The two-thirds treatment assignment rate was imposed by the charitable organization.)

The matching grant test included several sub-features (i.e., the ratio of the match, the ceiling of the match, and the example amount provided), but for the sake of simplicity we will only consider the main treatment of receiving the matching grant offer. We now ask the question: in a second wave of an experiment with this organization, how should we allocate treatments, conditional on covariates?

We first consider a simple analysis with a single binary covariate, an indicator equal to one if the individual lived in a state that George W. Bush won in the 2004 presidential election ("red state"), and zero if not ("blue state"). (In the original data, the estimated treatment effect was found to be positive for residents of red states, but not significantly different from zero for residents of blue states.) We define $X_i = 1$ if the individual lives in a red state, and 0 otherwise. The outcome of interest is the individual's donation amount in dollars after receiving the direct mail solicitation. We set $\kappa = 1/2$, so that the second round will be the same size as the first round. Excluding individuals who do not live in a state (e.g., they live in a U.S. territory), we have a sample size of 50,048.

The results from this exercise are given in Table 1. Notice that for $X = 1$ (donors living in red

Table 1: Karlan-List Experiment

	#0	$\hat{\mu}_0$	$\hat{\sigma}_0^2$	#1	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	π^*
Blue State ($X = 0$)	10029	0.90	73.44	19777	0.89	67.74	0.49
Red State ($X = 1$)	6648	0.69	57.01	13594	1.06	97.67	0.57

Note: #0 is the number of observations with $D = 0$, and #1 is the number of observations with $D = 1$.

states), the variance under treatment one is considerably larger than the variance under treatment zero. This suggests that red state donors should be treated more, because it is more difficult to learn the expected outcome under treatment for this subpopulation.

We applied our algorithm without the constraint that the overall treatment probability be 2/3. The last column of Table 1 gives the overall treatment assignment probabilities calculated by our procedure. As we expected, the optimal rule gives a higher treatment probability to red state donors. For both types of donors the overall optimal treatment probabilities are below 2/3, so the second stage probabilities will be lower than the overall probabilities. Using our adaptive rule would lead to a normalized asymptotic variance of 291, compared with 320 from 2/3 random sampling in the second stage. This is a 9.1% gain in efficiency. Given that the original sample size was 50048, this implies that we could have achieved the same efficiency with adaptive sampling using a sample size of 45494, or 4554 fewer observations. We also considered adaptive treatment assignment under

the constraint that the overall treatment probability be 2/3. In this case we found that using our adaptive rule would lead to a normalized asymptotic variance of approximately 319, only a 0.04% gain in efficiency.

Next, we expand the analysis to two covariates. We use highest previous amount donated (HPA), which also featured prominently in the original analysis. We discretize HPA into four categories based on its quartiles: (1) $HPA \leq \$30$; (2) $30 < HPA \leq \$45$; (3) $45 < HPA \leq \$60$; and (4) $HPA > \$60$. Interacting the red state indicator with the HPA categories leads to eight cells. The results are shown in Table 2. Using the two covariates we can reduce the variance by

Table 2: Karlan-List Experiment: Two Covariates

State	HPA	#0	$\hat{\mu}_0$	$\hat{\sigma}_0^2$	#1	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	π^*_{unc}	π^*_{con}
Blue	1	3127	0.35	9.40	6221	0.54	14.01	0.55	0.74
Blue	2	2021	0.56	17.97	3867	0.59	40.85	0.60	0.69
Blue	3	2461	0.39	18.26	4954	0.37	16.29	0.49	0.65
Blue	4	2420	2.40	255.69	4735	2.16	212.07	0.48	0.49
Red	1	2058	0.24	5.58	4209	0.45	11.53	0.59	0.79
Red	2	1358	0.29	8.91	2687	0.56	15.16	0.57	0.74
Red	3	1661	0.28	19.39	3523	0.44	20.32	0.51	0.65
Red	4	1571	2.05	203.41	3175	2.99	362.76	0.57	0.58

Note: π^*_{unc} are unconstrained optimal treatment probabilities; π^*_{con} are optimal treatment probabilities under the constraint that the overall treatment probability equals 2/3.

about 9.3% in the unconstrained case relative to 2/3 randomization. This is only modestly better than using the red state covariate alone. However, under the constraint that the overall treatment probability is 2/3, we get a 7.5% reduction in variance. So under the constraint, using the two covariates allows for some efficiency gains that are not possible with only the red state covariate.

3.2 Freedom from Hunger Experiment

For the second example we use data from an ongoing multi-wave experiment. Freedom from Hunger, in collaboration with Dean Karlan and Michael Kremer, is testing whether fundraising letters are more effective when the letters contain anecdotal discussions of the organization's impact, or when they contain reference to research and randomized trials to measure impact. The first wave included a "research" insert, a "story" insert, and a control (no insert). We found noticeable heterogeneity in response to the treatment for those who had given more than $100 in the past, compared to those who had given less than $100. Table 3 shows the results from applying our procedure to the comparison between the research insert and the control.

For a second wave, to test "research" versus the control, the optimal proportion assigned to

treatment falls to 46.6% for the small prior donors and rises to 62.9% for the large prior donors, which represents a significant departure from the first-stage 50% assignment rule. In the actual experiment the organization decided to drop the control group entirely, and test the research versus the story treatments in the second wave. The results are given in Table 4. For this comparison, the tool still improves power, but not by as much as it would have for a research versus control test. The optimal assignment rule is 45.6% to treatment for the large prior donors, and 49.1% for the small prior donors.

Table 3: Freedom From Hunger Experiment, Research Insert vs. Control

	#0	$\hat{\mu}_0$	$\hat{\sigma}_0^2$	#1	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	π^*
Small Prior Donors (prior donation < \$100)	5044	1.09	42.64	5001	0.85	32.47	0.466
Large Prior Donors	640	5.19	1172.91	637	9.66	3369.29	0.629

We define the control as $D = 0$ and the research insert as $D = 1$.

Table 4: Freedom From Hunger Experiment, Research Insert vs. Story

	#0	$\hat{\mu}_0$	$\hat{\sigma}_0^2$	#1	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	π^*
Small Prior Donors (prior donation < \$100)	5000	0.84	34.97	5001	0.85	32.47	0.491
Large Prior Donors	641	7.64	4800.38	637	9.66	3369.29	0.456

We define the story insert as $D = 0$ and the research insert as $D = 1$.

3.3 Progresa Experiment

For the third example we use data reported in Gertler, Martinez, and Rubio-Codina (2006) on the Progresa/Oportunidades experiment in Mexico. The Progresa program randomly allocated cash and nutritional supplements to families, conditional on children attending school and visiting health clinics. The Progresa experiment was conducted only once, but similar experiments have since been conducted or begun in Colombia, Ecuador, Honduras, and Nicaragua.

We focus on one of the outcome measures: number of draft animals owned by the family. The covariate we examine takes on four values, indicating the size of the family's agricultural holdings before the program. In one of their tables, Gertler, Martinez, and Rubio-Codina report sample sizes, means and standard deviations, and treatment effects broken down by this covariate, so we are able

to calculate optimal treatment assignment probabilities directly from their tables, without requiring access to the raw data. Table 5 gives some summary statistics from the Progresa experiment, along with the variance-minimizing treatment probabilities calculated using our method. We find that our treatment probabilities differ somewhat from the ones used in the original experiment, but the reduction in variance is quite small, suggesting that the original design was not far from optimal.

Table 5: Progresa Experiment, Number of Draft Animals

	$\#0$	$\hat{\mu}_0$	$\hat{\sigma}_0^2$	$\#1$	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	p_{orig}	π^*
NoAgAssets ($X=0$)	137	0.41	0.34	110	0.34	0.07	0.45	0.31
Landless ($X=1$)	1451	0.49	0.79	714	0.44	0.37	0.33	0.41
SmallerFarm ($X=2$)	2847	0.68	1.3	1359	0.58	0.63	0.32	0.41
BiggerFarm ($X=3$)	1187	0.83	1.2	728	0.87	1.83	0.38	0.55

Note: p_{orig} denotes original treatment probabilities, and π^* gives the overall treatment probabilities selected by adaptive rule.

4 Monte Carlo Study

We conducted a set of Monte Carlo simulations to evaluate our proposed design procedure and estimator. The Monte Carlo was designed to reflect some of the salient features of the data in the Freedom from Hunger Experiment. Recall that in the real experiment, the covariate was already discretized depending on the amount of previous donations. In order to examine the effects of the degree of discretization, we assumed that the underlying covariate is in fact continuous, but the researcher imposes an arbitrary degree of discretization.

We considered two sets of DGPs. The first DGP was designed to reflect the experiment underlying Table 3. We assumed that $X_i \sim N(63.50012, 206155.4)$, and that

$$Y_{0i}|X_i \sim N(1.292709 + 0.004528X_i, \exp(0.885859 + 0.002186X_i))$$
$$Y_{1i}|X_i \sim N(1.687224807 + 0.002514941X_i, \exp(1.3139975 + 0.0007578X_i))$$

Although the normality, linearity, and log-linearity were imposed for convenience, the parameter values were derived from regression analysis of the raw data from the Freedom from Hunger Experiment. We do not have access to the exact donation amounts used to construct the binary X_i in

the original analysis. However, we have a proxy variable that nearly perfectly predicts the binary "large-donations" variable, so we used this proxy in our regressions. We set both the first and second sample sizes at 1000 in the simulation.

In order to examine the effect of discretization, we assumed that researchers discretized the X_i into several equal sized cells, and considered different degrees of discretization. For example, when the number of cells are 4, the covariates are simply indicators $1\left(X_i \leq t_1\right)$, $1\left(t_1 < X_i \leq t_2\right)$, $1\left(t_2 < X_i \leq t_3\right)$, and $1\left(t_3 < X_i\right)$ such that $\Pr\left(X_i \leq t_1\right) = \Pr\left(t_1 < X_i \leq t_2\right) = \Pr\left(t_2 < X_i \leq t_3\right) = \Pr\left(t_3 < X_i\right) = \frac{1}{4}$.

Finite sample properties of our estimator based on (estimated) optimal propensity score for the second sample are summarized in Table 6.

Table 6: Monte Carlo Experiment 1

Number of Cells	2	3	4	5	6
Bias	-0.0021	-0.0015	-0.0020	-0.0013	-0.0019
Actual Coverage Prob (of 95% CI)	0.9566	0.9532	0.9480	0.9410	0.9432
RMSE	0.1016	0.1002	0.1009	0.1006	0.0996

The second DGP was designed to reflect the experiment underlying Table 4. We assumed that

$$\begin{aligned} X_i &\sim N\left(63.50012, 206155.4\right) \\ Y_{0i}|X_i &\sim N\left(1.5775317194 + 0.0005248564X_i, \exp\left(1.088926 + 0.000152X_i\right)\right) \\ Y_{1i}|X_i &\sim N\left(1.687224807 + 0.002514941X_i, \exp\left(1.3139975 + 0.0007578X_i\right)\right) \end{aligned}$$

Again, the parameters of this DGP are based on regression estimates using the raw data. Finite sample properties of our estimator based on (estimated) optimal propensity score for the second sample are summarized in Table 7.

Table 7: Monte Carlo Experiment 2

Number of Cells	2	3	4	5	6
Bias	-0.0011	-0.0004	-0.0004	-0.0004	-0.0006
Actual Coverage Prob (of 95% CI)	0.9572	0.9538	0.9536	0.9486	0.9548
RMSE	0.0850	0.0856	0.0861	0.0866	0.0846

In both cases we see that the ATE estimator is nearly bias free. This is not surprising, because the only source of bias in our procedure is the possibility of some cells being empty. We find that the confidence intervals have coverage very close to their nomimal levels, suggesting that the

asymptotic approximations are quite accurate for this choice of DGP and sample size. Finally, the finite sample properties do not seem to be too sensitive to the degree of discretization, with very small efficiency gains from finer discretization. It is not clear if such robustness can be expected with other DGPs, though.

5 Extension to Non-Stable Populations and Replications

Our analysis up to now assumes that the population of interest, the treatments, and the effects of the treatments are stable across periods, so that it is meaningful to combine the data from both stages. In some cases the second stage might be substantially different from the first stage, for example if the treatments under consideration are modified in later time periods, or in a replication of a social experiment in a different geographic location. Then the idea of using earlier experiments to inform experimental design can still be useful, but will require additional modeling assumptions to link the data across time periods.

Suppose for example that Stage 2 is a replication of the experiment in Stage 1 in a different location. First, we would need some assumption to link the conditional variances across replications. A simple approach is to assume that the conditional variances are homogeneous across the two stages. Note that if this auxiliary assumption is incorrect, we would only sacrifice some precision; our estimator will remain consistent and our results for inference will continue to hold.

Second, the estimand of interest may be the treatment effect in the Stage 2 population. Under the assumption of time homogeneity of conditional variances, this can be handled by changing the minimization problem in Step 2 of our design procedure. The problem is now to choose $\pi_2(x)$to minimize

$$E\left[\frac{\widehat{\sigma}_1^2(X_i)}{\pi_2(X_i)} + \frac{\widehat{\sigma}_0^2(X_i)}{(1-\pi_2(X_i))} + (\beta(X_i)-\beta)^2\right].$$

As before, the term $(\beta(X_i)-\beta)^2$ can be dropped from the minimand, and the optimization can proceed along similar lines to our original procedure. Finally, the estimate of the treatment effect should only use the Stage 2 observations.

6 Conclusion

We have considered the optimal design of a two-stage experiment for estimating an average treatment effect. We propose to choose the propensity score in the second stage based on the data from the first stage, in order to minimize an estimated version of the asymptotic variance bound. We argue, using a double asymptotic approximation, that our proposal leads to an adaptive estimation procedure for the average treatment effect. Using this double asymptotics leads to a very simple,

intuitive procedure that has good theoretical properties. Extending our approach to more than two time periods is straightforward.

Manski (2001) argues that one should estimate and report subgroup treatment effects from experimental studies, because this can help policymakers target policies to those groups that would benefit from them. Our analysis suggests that it can be useful to report conditional variances and cell sizes as well, because these can be used inform future experiments on similar populations.

A Computational Details

Unconstrained Problem: In the unconstrained problem we can solve for the optimum for each value of x separately, so we do not need to estimate the marginal distribution of X_i. At an interior solution the optimal overall probabilities $\pi(x)$ satisfy, for each x:

$$\frac{\partial}{\partial \pi(x)}\left[\frac{\widehat{\sigma}_1^2(x)}{\pi(x)} + \frac{\widehat{\sigma}_0^2(x)}{1-\pi(x)}\right] = 0.$$

The solution is

$$\pi(x) = \frac{\widehat{\sigma}_1(x)}{\widehat{\sigma}_1(x) + \widehat{\sigma}_0(x)}.$$

We can then solve for the second-stage probabilities $\widehat{\pi}_2(x)$ using

$$\widehat{\pi}_2(x) = \frac{1}{1-\kappa}\left[\pi(x) - \kappa\pi_1\right],$$

and verify that the solution satisfies $\widehat{\pi}_2(x) \in [0,1]$.

Constrained Problem: For notational convenience, let the support of X_i be $\{x_1, x_2, \ldots, x_K\}$. Let $\widehat{f}(x)$ denote the sample frequency of $X_i = x$ in the first stage:

$$\widehat{f}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} 1(X_i = x).$$

We want to choose the $\pi(x_k)$ to minimize

$$\sum_{j=1}^{K} \widehat{f}(x_j)\left[\frac{\widehat{\sigma}_1^2(x_j)}{\pi(x_j)} + \frac{\widehat{\sigma}_0^2(x_j)}{1-\pi(x_j)}\right], \tag{3}$$

subject to the constraint $\sum_j \widehat{f}(x_j)\pi(x_j) = p$.

We can solve out the constraint for one of the probabilities, e.g.

$$\pi(x_K) = \frac{1}{\widehat{f}(x_K)}\left[p - \sum_{j=1}^{K-1} \widehat{f}(x_j)\pi(x_j)\right],$$

and substitute this in to the minimand (3). Then we can numerically minimize the resulting expression subject to the constraints $0 \leq \pi(x_j) \leq 1$ for all $j = 1, \ldots, K$.

B Proof of Theorem 1

Let $r_1(x) := E[Y_{1i}|X_i = x]$ and $r_0(x) := E[Y_{0i}|X_i = x]$. We can write

$$\begin{aligned}
\widehat{\beta} - \beta = & \frac{1}{n}\sum_{i=1}^{n}\left(r_1\left(X_i\right) - r_0\left(X_i\right) - \beta\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i\left(Y_i - r_1\left(X_i\right)\right)}{\widehat{\pi}\left(X_i\right)} - \frac{\left(1-D_i\right)\left(Y_i - r_0\left(X_i\right)\right)}{1-\widehat{\pi}\left(X_i\right)}\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i r_1\left(X_i\right)}{\widehat{\pi}\left(X_i\right)} - \frac{\left(1-D_i\right) r_0\left(X_i\right)}{1-\widehat{\pi}\left(X_i\right)} - \left(r_1\left(X_i\right) - r_0\left(X_i\right)\right)\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\widehat{p}\left(X_i\right)} - \frac{D_i Y_i}{\widehat{\pi}\left(X_i\right)}\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1-D_i\right) Y_i}{1-\widehat{p}\left(X_i\right)} - \frac{\left(1-D_i\right) Y_i}{1-\widehat{\pi}\left(X_i\right)}\right)
\end{aligned} \tag{4}$$

Note that

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i r_1\left(X_i\right)}{\widehat{\pi}\left(X_i\right)} - r_1\left(X_i\right)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\widehat{p}\left(X_i\right)} - \frac{D_i Y_i}{\widehat{\pi}\left(X_i\right)}\right) \\
& = \sum_{x}\left(r_1\left(x\right) - \widehat{r}_1\left(x\right)\right)\left(\frac{\widehat{p}\left(x\right) - \widehat{\pi}\left(x\right)}{\widehat{\pi}\left(x\right)}\right)\left(\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i = x\right)\right)
\end{aligned} \tag{5}$$

and

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1-D_i\right) r_0\left(X_i\right)}{1-\widehat{\pi}\left(X_i\right)} - r_0\left(X_i\right)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1-D_i\right) Y_i}{1-\widehat{p}\left(X_i\right)} - \frac{\left(1-D_i\right) Y_i}{1-\widehat{\pi}\left(X_i\right)}\right) \\
& = -\sum_{x}\left(r_0\left(x\right) - \widehat{r}_0\left(x\right)\right)\left(\frac{\widehat{p}\left(x\right) - \widehat{\pi}\left(x\right)}{1-\widehat{\pi}\left(x\right)}\right)\left(\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i = x\right)\right)
\end{aligned} \tag{6}$$

Furthermore, Lemmas 1 and 2 in Appendix C show that

$$\widehat{p}\left(x\right) - \widehat{\pi}\left(x\right) = O_p\left(n^{-1/2}\right),\ r_1\left(x\right) - \widehat{r}_1\left(x\right) = O_p\left(n^{-1/2}\right),\ r_0\left(x\right) - \widehat{r}_0\left(x\right) = O_p\left(n^{-1/2}\right),$$

which implies that (5) and (6) are $o_p\left(n^{-1/2}\right)$. We therefore obtain the following approximation for (4):

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta\right) = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(r_1\left(X_i\right) - r_0\left(X_i\right) - \beta\right) \\
& + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i\left(Y_i - r_1\left(X_i\right)\right)}{\widehat{\pi}\left(X_i\right)} - \frac{\left(1-D_i\right)\left(Y_i - r_0\left(X_i\right)\right)}{1-\widehat{\pi}\left(X_i\right)}\right) + o_p\left(1\right)
\end{aligned} \tag{7}$$

By Lemma 4 in Appendix C, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i\left(Y_i-r_1\left(X_i\right)\right)}{\widehat{\pi}\left(X_i\right)}-\frac{\left(1-D_i\right)\left(Y_i-r_0\left(X_i\right)\right)}{1-\widehat{\pi}\left(X_i\right)}\right)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i^*\left(Y_{1i}-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)+o_p\left(1\right)$$

where $D_i^*:=1\left(U_i\leq\pi_1\right)$ for the first sample, and $D_i^*:=1\left(U_i\leq\pi_2^*\left(X_i\right)\right)$ for the second sample.

Therefore, we can write

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\beta\left(X_i\right)-\beta\right)$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i^*\left(Y_{1i}-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)+o_p\left(1\right)$$

or

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\frac{\sqrt{n_1}}{\sqrt{n}}\times\left(I\right)+\frac{\sqrt{n_2}}{\sqrt{n}}\times\left(II\right)+o_p\left(1\right)$$

where

$$\left(I\right):=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left(\beta\left(X_i\right)-\beta+\frac{D_i^*\left(Y_i-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)$$
$$\left(II\right):=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\left(\beta\left(X_i\right)-\beta+\frac{D_i^*\left(Y_i-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)$$

By the Central Limit Theorem (CLT), we obtain that

$$\left(I\right)\xrightarrow{d}N\left(0,E\left[\frac{\pi_1}{\pi^*\left(X_i\right)^2}\sigma_1^2\left(X_i\right)+\frac{1-\pi_1}{\left(1-\pi^*\left(X_i\right)\right)^2}\sigma_0^2\left(X_i\right)+\left(\beta\left(X_i\right)-\beta\right)^2\right]\right)$$
$$\left(II\right)\xrightarrow{d}N\left(0,E\left[\frac{\pi_2^*\left(X_i\right)}{\pi^*\left(X_i\right)^2}\sigma_1^2\left(X_i\right)+\frac{1-\pi_2^*\left(X_i\right)}{\left(1-\pi^*\left(X_i\right)\right)^2}\sigma_0^2\left(X_i\right)\right]\right)$$

Noting that $\left(I\right)$ and $\left(II\right)$ are independent of each other, and that $\kappa\pi_1+\left(1-\kappa\right)\pi_2^*\left(X_i\right)=\pi^*\left(X_i\right)$ by definition, we obtain that $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ converges weakly to a normal distribution with mean

zero and variance equal to

$$E\left[\frac{\kappa\pi_1+(1-\kappa)\,\pi_2^*(X_i)}{\pi^*(X_i)^2}\sigma_1^2(X_i)+\frac{1-(\kappa\pi_1+(1-\kappa)\,\pi_2^*(X_i))}{(1-\pi^*(X_i))^2}\sigma_0^2(X_i)+(\beta(X_i)-\beta)^2\right]$$
$$=E\left[\frac{1}{\pi^*(X_i)}\sigma_1^2(X_i)+\frac{1}{1-\pi^*(X_i)}\sigma_0^2(X_i)+(\beta(X_i)-\beta)^2\right]$$

which proves the theorem.

C Auxiliary Results

Lemma 1 $\widehat{p}(x)-\widehat{\pi}(x)=O_p\left(n^{-1/2}\right)$

Proof: We will write

$$\begin{aligned}\widehat{p}(x)&=\frac{\sum_{i=1}^n D_i 1(X_i=x)}{\sum_{i=1}^n 1(X_i=x)}\\&=\frac{\sum_{i=1}^{n_1} D_i 1(X_i=x)+\sum_{i=n_1+1}^n D_i 1(X_i=x)}{\sum_{i=1}^n 1(X_i=x)}\\&=\frac{\sum_{i=1}^{n_1} 1(U_i\le\pi_1)1(X_i=x)+\sum_{i=n_1+1}^n 1(U_i\le\widehat{\pi}_2(x))1(X_i=x)}{\sum_{i=1}^n 1(X_i=x)}\\&=\frac{n_1}{n}\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}1(U_i\le\pi_1)1(X_i=x)}{\frac{1}{n}\sum_{i=1}^n 1(X_i=x)}+\frac{n_2}{n}\frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n}1(U_i\le\widehat{\pi}_2(x))1(X_i=x)}{\frac{1}{n}\sum_{i=1}^n 1(X_i=x)}\end{aligned}\tag{8}$$

By the law of large numbers and central limit theorem, we would have

$$\begin{aligned}\frac{1}{n_1}\sum_{i=1}^{n_1}1(U_i\le\pi_1)1(X_i=x)&=E[1(U_i\le\pi_1)1(X_i=x)]+O_p\left(\frac{1}{\sqrt{n_1}}\right)\\&=\pi_1\Pr(X_i=x)+O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}\tag{9}$$

In order to deal with the second component on the far RHS of (8), we define the empirical process

$$\xi_2(\cdot,\pi_2):=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\left(1(U_i\le\pi_2(x))1(X_i=x)-E[1(U_i\le\pi_2(x))1(X_i=x)]\right)$$

The set of functions $\{1(U_i\le\pi_2(x))1(X_i=x)\}$indexed by $\pi_2(x)$ is Euclidean, and satisfies stochas-

tic equicontinuity. We therefore have $\xi_2(\cdot, \widehat{\pi}_2) = \xi_2(\cdot, \pi_2^*) + o_p(1)$, or

$$\frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) 1(X_i = x) = \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} 1(U_i \leq \pi_2^*(x)) 1(X_i = x) + G_2 \sqrt{n_2} (\widehat{\pi}_2(x) - \pi_2^*(x)) + o_P(1) \tag{10}$$

where

$$G_2 := \frac{\partial}{\partial \pi_2} E[1(U_i \leq \pi_2(x)) 1(X_i = x)] \Big|_{\pi_2(x) = \pi_2^*(x)}$$

Because $E[1(U_i \leq \pi_2(x)) 1(X_i = x)] = \pi_2(x) \Pr(X_i = x)$, we have $G_2 = \Pr(X_i = x)$, and hence,

$$G_2 \sqrt{n_2} (\widehat{\pi}_2(x) - \pi_2^*(x)) = O_p(1) \tag{11}$$

as long as $\widehat{\pi}_2(x)$ is chosen to be a $\sqrt{n}$-consistent estimator of $\pi_2^*(x)$. We also have

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n} 1(U_i \leq \pi_2^*(x)) 1(X_i = x) = \pi_2^*(x) E[1(X_i = x)] + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{12}$$

by the law of large numbers and CLT. Combining (10), (11), and (12), we obtain

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) 1(X_i = x) = \pi_2^*(x) \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{13}$$

Now note that, by the law of large numbers and CLT, we have

$$\frac{1}{n} \sum_{i=1}^{n} 1(X_i = x) = \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{14}$$

Combining (8), (9), (13), and (14), we obtain

$$\begin{aligned} \widehat{p}(x) &= \frac{n_1}{n} \frac{\pi_1 \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)} + \frac{n_2}{n} \frac{\pi_2^*(x) \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)} \\ &= \kappa \pi_1 + (1 - \kappa) \pi_2^*(x) + O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

Therefore, as long as $\widehat{\pi}_2(x)$ is chosen to be a $\sqrt{n}$-consistent estimator of $\pi_2^*(x)$, we will have $\widehat{p}(x) = \widehat{\pi}_2(x) + O_p(1/\sqrt{n})$. □

Lemma 2 $r_1(x) - \widehat{r}_1(x) = O_p\left(n^{-1/2}\right)$, $r_0(x) - \widehat{r}_0(x) = O_p\left(n^{-1/2}\right)$

Proof: We only prove that $r_1(x) - \widehat{r}_1(x) = O_p\left(n^{-1/2}\right)$. The proof of the other equality is similar, and omitted. Our proof is based on the equality

$$\begin{aligned}
\widehat{r}_1(x) &= \frac{\sum_{i=1}^{n} D_i Y_i 1(X_i = x)}{\sum_{i=1}^{n} D_i 1(X_i = x)} \\
&= \frac{\sum_{i=1}^{n_1} D_i Y_i 1(X_i = x) + \sum_{i=n_1+1}^{n} D_i Y_i 1(X_i = x)}{\sum_{i=1}^{n_1} D_i 1(X_i = x) + \sum_{i=n_1+1}^{n} D_i 1(X_i = x)} \\
&= \frac{\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x)}{\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) 1(X_i = x) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) 1(X_i = x)}
\end{aligned}$$

We take care of the numerator first. We note that

$$\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) = E\left[1(U_i \leq \pi_1) Y_i 1(X_i = x)\right] + O_p\left(\frac{1}{\sqrt{n}}\right)$$

by the law of large numbers and central limit theorem. Because

$$\begin{aligned}
E\left[1(U_i \leq \pi_1) Y_i 1(X_i = x)\right] &= \pi_1 E\left[Y_{1i} | X_i = x\right] \Pr(X_i = x) \\
&= \pi_1 r_1(x) \Pr(X_i = x),
\end{aligned}$$

we obtain

$$\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) = \pi_1 r_1(x) \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{15}$$

In order to deal with $\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x)$, we note that the set of functions $\{1(U_i \leq \pi_2(x)) Y_i 1(X_i = x)\}$ indexed by $\pi_2(x)$ is Euclidean, and satisfies stochastic equicontinuity. We therefore have

$$\begin{aligned}
\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x) = \frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n} 1(U_i \leq \pi_2^*(x)) Y_i 1(X_i = x) \\
+ G_3\sqrt{n_2}\left(\widehat{\pi}_2(x) - \pi_2^*(x)\right) + o_P(1)
\end{aligned}$$

where

$$G_3 := \frac{\partial}{\partial \pi_2} E\left[1(U_i \leq \pi_2(x)) Y_i 1(X_i = x)\right]\Big|_{\pi_2(x) = \pi_2^*(x)}$$

Because

$$\begin{aligned}
E\left[1(U_i \leq \pi_2(x)) Y_i 1(X_i = x)\right] &= \pi_2(x) E\left[Y_{1i} | X_i = x\right] \Pr(X_i = x) \\
&= \pi_2(x) r_1(x) \Pr(X_i = x),
\end{aligned}$$

we have $G_3 = r_1(x)\Pr(X_i = x)$, and hence, $G_3\sqrt{n_2}(\widehat{\pi}_2(x) - \pi_2^*(x)) = O_p(1)$ as long as $\widehat{\pi}_2(x)$ is chosen to be a $\sqrt{n}$-consistent estimator of $\pi_2^*(x)$. We also have

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \pi_2^*(x)) Y_i 1(X_i = x) = \pi_2^*(x) r_1(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)$$

by the law of large numbers and the central limit theorem. We may therefore conclude that

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x) = \pi_2^*(x) r_1(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{16}$$

Combining (15) and (16), we obtain

$$\begin{aligned}
&\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x) \\
&= \kappa\pi r_1(x)\Pr(X_i = x) + (1-\kappa)\pi_2^*(x) r_1(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right) \\
&= \pi^*(x) r_1(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

We can take care of the denominator in a similar manner, and obtain

$$\begin{aligned}
&\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) 1(X_i = x) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x)) 1(X_i = x) \\
&= \pi^*(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

and hence, we conclude that

$$\widehat{r}_1(x) = \frac{\pi^*(x) r_1(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\pi^*(x)\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)} = r_1(x) + O_p\left(\frac{1}{\sqrt{n}}\right)$$

□

Lemma 3

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{D_i(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x) = \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{1(U_i \leq \pi_1)(Y_{1i} - r_1(x))}{\pi^*(x)} 1(X_i = x) + o_p(1)$$

$$\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{1\left(U_i\leq\pi_2^*\left(x\right)\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{\left(1-D_i\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{1\left(U_i>\pi_1\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

$$\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{\left(1-D_i\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{1\left(U_i>\pi_2^*\left(x\right)\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

Proof: We only prove the first two claims. The proof of the last two claims is identical, and omitted.

We first note that

$$\begin{aligned}
&\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\sum_{x}\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\widehat{\pi}\left(x\right)}1\left(X_i=x\right)\\
&=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\sum_{x}\frac{1\left(U_i\leq\pi_1\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)+O_p\left(\frac{1}{\sqrt{n}}\right)}1\left(X_i=x\right)\\
&=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\sum_{x}\frac{1\left(U_i\leq\pi_1\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)}1\left(X_i=x\right)+O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

as long as $\widehat{\pi}_2\left(x\right)$ is chosen to be a $\sqrt{n}$-consistent estimator of $\pi_2^*\left(x\right)$, and the latter is an interior point of $(0,1)$, which proves the first claim.

In order to prove the second claim, we define the empirical process

$$\nu_2\left(\cdot,\pi_2\right):=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\left(\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\pi\left(x\right)}1\left(X_i=x\right)-E\left[\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\pi\left(x\right)}1\left(X_i=x\right)\right]\right)$$

where $\pi\left(x\right)=\kappa\pi_1+\left(1-\kappa\right)\pi_2\left(x\right)$. Recall that $D_i=1\left(U_i\leq\pi_1\right)$ for the first sample, and $D_i=1\left(U_i\leq\widehat{\pi}_2\left(X_i\right)\right)$ for the second sample. Because the sets of functions

$$\left\{\frac{1\left(U_i\leq\pi_2\left(x\right)\right)D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\kappa\pi_1+\left(1-\kappa\right)\pi_2\left(x\right)}1\left(X_i=x\right)\right\}$$

indexed by $\pi_2\left(x\right)$ is Euclidean, we can use stochastic equicontinuity, and conclude that $\nu_2\left(\cdot,\widehat{\pi}_2\right)=$

$\nu_2(\cdot, \pi_2^*) + o_p(1)$, or

$$\frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \frac{1(U_i \leq \widehat{\pi}_2(x))(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x)$$
$$= \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \frac{1(U_i \leq \pi_2^*(x))(Y_{1i} - r_1(x))}{\pi^*(x)} 1(X_i = x) + F_2 \sqrt{n_2}(\widehat{\pi}_2 - \pi_2^*) + o_p(1)$$

where

$$F_2 = \frac{\partial}{\partial \pi_2} E\left[\frac{1(U_i \leq \pi_2(x))(Y_{1i} - r_1(x))}{\kappa\pi_1 + (1-\kappa)\pi_2(x)} 1(X_i = x)\right]\Big|_{\pi_2 = \pi_2^*}$$

Because U_i is independent of (X_i, Y_{1i}, Y_{0i}), we have

$$E\left[\frac{1(U_i \leq \pi_2(x))(Y_{1i} - r_1(x))}{\kappa\pi_1 + (1-\kappa)\pi_2(x)} 1(X_i = x)\right] = 0$$

regardless of the value of $\pi(x)$. This implies that the derivative F_2 is identically zero, from which the validity of the second claim follows. □

Lemma 4

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{D_i(Y_i - r_1(X_i))}{\widehat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)}\right)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{D_i^*(Y_{1i} - r_1(X_i))}{\pi^*(X_i)} - \frac{(1-D_i^*)(Y_{0i} - r_0(X_i))}{1 - \pi^*(X_i)}\right) + o_p(1)$$

Proof: Write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i(Y_i - r_1(X_i))}{\widehat{\pi}(X_i)} = \sum_x \left(\frac{\sqrt{n_1}}{\sqrt{n}} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n} \left(\frac{D_i(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x)\right)\right)$$
$$+ \sum_x \left(\frac{\sqrt{n_2}}{\sqrt{n}} \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \left(\frac{D_i(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x)\right)\right)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)} = \sum_x \left(\frac{\sqrt{n_1}}{\sqrt{n}} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n} \left(\frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)} 1(X_i = x)\right)\right)$$
$$+ \sum_x \left(\frac{\sqrt{n_2}}{\sqrt{n}} \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \left(\frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)} 1(X_i = x)\right)\right)$$

The conclusion then follows by using Lemma 3. □

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B., 1996, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91(434).

Chaloner, K., and Verdinelli, I., 1995, "Bayesian Experimental Design: A Review," *Statistical Science* 10(3), 273-304.

Chamberlain, G. 1986, "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 32, 189-218.

Flores-Lagunes, A., Gonzalez, A., and Neumann, T., 2006, "Learning But Not Earning? The Impact of Job Corps Training for Hispanics," working paper, University of Arizona.

Gertler, P., Martinez, S., and Rubio-Codina, M., 2006, "Investing Cash Transfers to Raise Long-Term Living Standards," World Bank Policy Research Paper 3994.

Green, D., and Gerber, A., 2004, "Get Out the Vote! How to Increase Voter Turnout," Washington, DC: Brookings Institution Press.

Hahn, J., 1998, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66(2), 315-331.

Hirano, K., Imbens, G. W., and Ridder, G., 2003, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71(4), 1161-1189.

Imbens, G. W., and Angrist, J. D., 1994, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62(2): 467-475.

Karlan, D., and List, J., 2007, "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review* 97(5), 1774-1793.

Karlan, D., and Zinman, J., 2008, "Credit Elasticities in Less-Developed Economies: Implications for Microfinance," *American Economic Review* 98(3): 1040-68.

Manski, C., 2001, "Designing Programs for Heterogeneous Populations: The Value of Covariate Information," *The American Economic Review, Papers and Proceedings* 91(2), 103-106.

Manski, C. F., and McFadden, D. L., 1981, "Alternative Estimators and Sampling Designs for Discrete Choice Analysis," Ch. 1 in *Structural Analysis of Discrete Data and Econometric Applications*, ed. C. F. Manski and D. L. McFadden, Cambridge: MIT Press.

Newey, W. K., and McFadden, D. L., 1994, "Large Sample Estimation and Hypothesis Testing," Ch. 36 in *Handbook of Econometrics, Volume IV*, ed. R. F. Engle and D. L. McFadden, Amsterdam: Elsevier.

Neyman, J., 1934, "On the Two Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Series A, 97, 558-625.

Rosenbaum, P. R., and Rubin, D. B., 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70(1): 41-55.

Simester, D. I., Sun, P., and Tsitsiklis, J. N., 2006, "Dynamic Catalog Mailing Policies," *Management Science* 52(5), 683-696.

Solomon, H., and Zacks, S., 1970, "Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas," *Journal of the American Statistical Association*, 65(330), 653-677.

Sukhatme, P. V., 1935, "Contributions to the Theory of the Representative Method," *Journal of the Royal Statistical Society*, Supplement 2, 253-268.