

# Thresholding Nonprobability Units in Combined Data for Efficient Domain Estimation\* July 2024

**Abstract:** Quasi-randomization approaches estimate latent participation probabilities for units from a nonprobability / convenience sample. Estimation of participation probabilities for convenience units allows their combination with units from the randomized survey sample to form a survey weighted domain estimate. One leverages convenience units for domain estimation under the expectation that estimation precision and bias will improve relative to solely using the survey sample; however, convenience sample units that are very different in their covariate support from the survey sample units may inflate estimation bias or variance. This paper develops a method to threshold or exclude convenience units to minimize the variance of the resulting survey weighted domain estimator. We compare our thresholding method with other thresholding constructions in a simulation study for two classes of datasets based on degree of overlap between survey and convenience samples on covariate support. We reveal that excluding convenience units that each express a low probability of appearing in *both* reference and convenience samples reduces estimation error.

**Keywords and phrases:** Survey sampling, Nonprobability sampling, Data combining, Quasi randomization, Thresholding units, Bayesian hierarchical modeling.

## Contents

1	Introduction . . . . .	2
2	Optimal Variance Thresholding . . . . .	4
2.1	Thresholding based solely on convenience sample probabilities . .	4
2.2	Thresholding using both reference and convenience sample probabilities . . . . .	7
2.3	Thresholding statistic motivated by variance structure of model score function . . . . .	9
3	Simulation study . . . . .	9
3.1	Simulation design . . . . .	9
3.2	Thresholding of convenience units . . . . .	10
3.3	Results . . . . .	12
4	Discussion . . . . .	16
A	Direct derivation of variance minimizing threshold for one-arm sample	18
	References . . . . .	19

---

\*U.S. Bureau of Labor Statistics, 4600 Silver Hill Road , Suitland, MD 20746 USA

## 1. Introduction

Declining response rates for randomized survey instruments administered by government statistical agencies (Williams and Brick, 2017) have encouraged the development of quasi-randomization processes such as those of Elliott (2009); Elliott and Valliant (2017); Wang et al. (2021); Savitsky et al. (2023) to allow inclusion of responses derived from a nonrandom convenience sample that includes responses for covariates that overlap those measured by the randomized survey or reference sample. Directly combining responses for units participating in the convenience sample with those selected into the randomized or reference sample may be expected to induce bias for inference about an underlying latent population, however, because the convenience sample is not generally representative of that population (Bethlehem, 2010; Meng, 2018; VanderWeele and Shpitser, 2011).

Quasi-randomization methods propose model formulations to estimate the convenience sample unit marginal participation probabilities as if the convenience sample is realized from a *latent* or unknown selection process. Quasi-randomization uses the reference sample and associated known inclusion probabilities to provide information about the underlying sampling frame that is, in turn, used to estimate convenience sample inclusion probabilities. The goal in using a statistical model to estimate the convenience sample inclusion probabilities is to allow inclusion of the convenience sample units in a combined (reference and convenience sample) data estimator for a domain mean (e.g., employment for computer services in New York city) with minimal bias.

Beresovsky et al. (2024) provides a comprehensive overview of quasi-randomization methods and compares the variance performances of a collection of methods for domain estimation that are mostly differentiated by assumptions about the degree of overlap in memberships in the convenience and reference samples, on the one hand, and the form of approximating inference on the non-sampled portion of the population, on the other hand. Elliott (2009) and Elliott and Valliant (2017) assume that the reference sample size is sufficiently small that there is a negligible overlap in unit inclusions with the convenience sample. This negligible overlap assumption is increasingly untenable under ever larger convenience samples. Later methods dispense with this assumption; in particular, Savitsky et al. (2023) and Wang et al. (2021) make no assumption about the degree of overlaps in units to allow more robust inference. Similarly, recent methods differ on how to estimate likelihoods specified for the population on realized (convenience and reference) samples. Wu (2022); Wang et al. (2021) use a pseudo likelihood approach by approximating unknown population units with the weighted reference sample units. The use of reference sample-weighted units may inflate estimation variance for small-sized reference samples. Savitsky et al. (2023) directly specify a likelihood for the realized samples that avoids using reference sample weights.

To motivate the focus of our paper, we highlight a key covariate balance requirement of these methods to produce combined reference and convenience sample domain estimators with reduced bias (as compared to domain estimators

obtained from solely using the reference sample).

Quasi-randomization methods require availability of the covariates used to determine the sampling design (governing the reference sample) for convenience sample units. This requirement is generally readily satisfied for sampling designs parameterized by demographic variables; for example, in the case of surveys conducted from business establishments by the U.S. Bureau of Labor Statistics these covariates might include a discretized employment size class, industry classification and metropolitan statistical area designation.

Valliant (2020) further notes that the target population units are assumed to have positive probabilities to be included into both samples conditional on the shared set of covariates among both reference and convenience samples. They refer to this condition of positive participation probabilities for all units in both samples as a requirement for “common support”. Satisfying common support requires that the support of covariate values expressed by units in the population is also expressed by units included in *both* the reference and convenience samples. This paper addresses estimation bias that arises when common support is satisfied but where a subset of population units selected into the reference sample with relatively moderate-to-large inclusion probabilities may express vanishingly low convenience sample participation probabilities. Heuristically, there are often subsets of the population purposefully emphasized in the reference sample that are poorly represented in the convenience sample.

Since a convenience sample derives from an opt-in or self-initiated participation process there will typically be some units in the realized convenience sample that are very different from those represented in the randomized reference sample. To be precise, there may be some units in the convenience samples whose covariate values don’t well overlap those for the reference sample. Gelman and Hill (2007) discuss degrees of “partial overlap” in the space of covariate values that may occur between treatment and control sample arms in the causal inference experimental set-up and the increase in bias and variance in the resulting propensity scores. The low overlap of covariate values for those convenience units with the reference sample provides less information to estimate associated participation probabilities for them, which produces estimates with large errors. Including these low overlap convenience units along with reference units to formulate a domain estimator would be expected to inflate bias and variance rather than reduce it. The error inflating effect of these low overlap convenience units on the domain estimator would partially offset the variance reduction benefit of incorporating high overlap convenience units along with the reference units discussed in Savitsky et al. (2023).

This paper introduces an approach to identify and exclude a subset of convenience sample units whose covariate values poorly overlap the reference sample in order to further reduce the error in domain estimators that incorporate convenience units (and their estimated participation probabilities). Our approach for excluding or thresholding units uses estimated reference and convenience sample inclusion and participation probabilities for the *convenience* units as a uni-dimensional summary of the overlap of multivariate covariate values. In the sequel we develop a set of alternative statistics used for thresholding where each

statistic represents distinct functional combinations of the estimated reference and convenience sample inclusion and participation probabilities for the convenience units. We note that [Savitsky et al. \(2023\)](#) specify a Bayesian modeling approach that provides estimates both convenience *and* reference sample participation and inclusion probabilities for the convenience units. The most simple example of using these estimated probabilities to threshold units would be to exclude convenience units with low reference sample inclusion probabilities below some threshold quantile. The logic for such a thresholding statistic is that convenience units with low values for estimated reference sample inclusion probabilities may be expected to express a low degree of overlap in covariate values with the reference sample.

We introduce a thresholding statistic for excluding convenience sample units that arises by minimizing of the variance of a domain mean estimator that is a function of the estimated reference and convenience sample inclusion and participation probabilities for the convenience sample units in Section 2. We begin by deriving the variance optimal thresholding statistic under the simpler set-up that composes the domain mean estimator using solely estimated convenience sample inclusion probabilities for convenience units (and excludes estimated reference sample inclusion probabilities for the convenience units). We then derive our main result under a set-up that constructs a threshold statistic composed of both estimated reference and convenience sample marginal probabilities for the convenience units. Section 2.3 introduces an additional thresholding statistic motivated by [Beresovsky et al. \(2024\)](#). We compare the reductions in bias and means squared error offered by the alternative thresholding statistics with a Monte Carlo simulation study in Section 3 and conclude with a discussion in Section 4.

## 2. Optimal Variance Thresholding

### 2.1. Thresholding based solely on convenience sample probabilities

We begin this section using only convenience sample participation probabilities (obtained from co-modeling with the reference sample) for convenience units to construct our estimator to introduce our notation under a simpler thresholding construction. This set-up contrasts with use of *both* estimated convenience and reference participation and inclusion probabilities for the convenience units to compose our domain mean estimator. We label the set-up that utilizes solely convenience sample participation probabilities (for convenience sample units) to define our thresholding statistic and set as “one-arm”. By contrast, our main result will use the more general set-up that defines the thresholding statistic from both estimated convenience and reference sample probabilities, which we label as “two-arm”.

Our main result defines a set subset of  $x \in \mathbb{X}$  where units in the convenience sample whose threshold statistic percentile (as a function of  $x$ ) is less than a some small value ( $\alpha$ ) will be excluded from the subset. Only convenience sample

units that are members of the subset will be used to render our weighted domain mean estimator,  $\hat{\mu}$ .

Let  $\delta_c \in \{0, 1\}$  index unit participation in the convenience sample where  $\delta_c = 1$  denotes participation in the sample and  $\delta_c = 0$  denotes a non-participating unit from the population frame,  $U$ , where  $|U| = N$ . Define marginal participation probability  $\pi_c(x) = \Pr[\delta_c = 1 \mid X = x]$  where  $X \in \mathbb{X}$  is a random variable. This construction for  $\pi_c(x)$  defines a marginal participation probability (rather than a propensity score). We proceed to extend and adapt a result of [Crump et al. \(2009\)](#) from the literature on causal inference that defines a threshold statistic and acceptance set for units constructed from a subset of  $x \in \mathbb{X}$  where the value of the threshold statistic is exceeded. The acceptance set formed by excluding units whose value lies below some percentile of the threshold statistic constructed by [Crump et al. \(2009\)](#) is guaranteed to produce a minimizing variance for the domain mean estimator after excluding those  $x$  not in the acceptance set. We repurpose and extend their result from treatment and control arms under their causal inference set-up to reference and convenience sampling arms under our survey sampling set-up. We begin our extension of their result with a simpler result that defines an acceptance set and formulation for a thresholding statistic for units in a convenience sample that produces a minimum variance for the domain mean estimator constructed solely from convenience sample participation probabilities.

Our population quantity of inferential interest is  $\mu = \mathbb{E}(Y)$  where  $Y$  denotes a univariate response variable of interest. Define our domain mean estimator as,

$$\hat{\mu} = \mu + \frac{1}{N} \sum_{i=1}^N \frac{z_i \delta_i}{\hat{\pi}_c(x_i)}, \quad (1)$$

where we are assuming  $N$  is known and  $z = y - \mu$ . Treating  $N$  as known may be relaxed, in practice. Let

$$\phi(Y, \delta, X, \mu, e) = \frac{z\delta}{\pi_c(X)}. \quad (2)$$

$$\hat{\mu} = \mu + \frac{1}{N} \sum_{i=1}^N \phi(y_i, \delta_i, x_i, \mu, e_i) \quad (3)$$

Then  $\phi(Y, \delta, X, \mu, e)$  has 0 expectation and variance ([Hirano et al., 2003](#), p. 1182),

$$\mathbb{E} [\phi(Y, \delta, X, \mu, e)^2] = \frac{1}{N} \mathbb{E} \left[ \frac{\sigma_1^2(X)}{\pi_c(X)} \right], \quad (4)$$

where  $\sigma_1^2 = \mathbb{V}(Y \mid \delta = 1, X = x)$ . The expectation on the LHS of Equation 4 is taken with respect to the joint distribution for  $X$  and the taking of a sample from the underlying frame on which  $\mathbb{X}$  is defined. The expectation on the RHS is taken with respect to the distribution for  $X$ .

Equation 4 may be used in combination with Corollary 1 of [Crump et al. \(2009\)](#) to produce the following result for the optimal threshold level,  $\alpha$ .

**Theorem 2.1** (One-arm extension of [Crump et al. \(2009\)](#)). Assume  $\pi_c(x) > 0 \forall x \in \mathcal{X}$ . Then set  $\mathbb{A} = \{x \in \mathbb{X} : \pi_c(x) > \alpha\}$  denotes the variance optimal subset of  $\mathbb{X}$  after thresholding units where  $\mathbb{A}$  is defined based on thresholding conditional inclusion probability,  $\pi_c(X)$ . The minimum variance quantile  $\alpha$  is constructed by,

$$\frac{1}{\alpha} = 2\mathbb{E} \left[ \frac{1}{\pi_c(X)} \middle| \frac{1}{\pi_c(X)} < \frac{1}{\alpha} \right]. \quad (5)$$

For computation of  $\alpha$  we approximate the expectation with sums over units  $i \in S_c$ , where  $S_c$  denotes the observed convenience sample,

$$\frac{1}{\alpha} = 2 \frac{\sum_{i \in S_c} \mathbf{1}(\hat{\pi}_c(x_i) > \alpha) \frac{1}{\hat{\pi}_c(x_i)}}{\sum_{i \in S_c} \mathbf{1}(\hat{\pi}_c(x_i) > \alpha)}. \quad (6)$$

*Proof.* Plugging in  $\pi_c(X)$  for  $e(X)$  into Theorem 1 of [Crump et al. \(2009\)](#) and using the result of Equation 4 for the case of where we utilize solely the convenience sample participation probabilities (for the convenience units) produces the result.  $\square$

*Remark 1.* The result of Theorem 2.1 utilizes a one-arm set-up that composes the mean estimator from solely the convenience sample. A companion, separate reference sample is required in order to estimate the convenience sample inclusion probabilities,  $\hat{\pi}_c(x_i)$ ,  $i \in (1, \dots, N)$ . In the sequel, we will further extend Theorem 2.1 by additionally estimating the reference sample inclusion probabilities for the same convenience units,  $\hat{\pi}_r(x_i)$ ,  $i \in (1, \dots, N)$  also using the reference sample inclusion probabilities estimated on the convenience units. See [Savitsky et al. \(2023\)](#) for more details on estimating  $(\hat{\pi}_c(x_i), \pi_r(x_i))$  (where subscript “r” denotes reference sample) for convenience sample units. They specify a model for the observed membership indicator in the pooled sample,  $\mathbf{1}_{z_i}$ , which is set to 1 if unit  $i$  is included in the convenience sample and 0 if the unit belongs to the reference sample. Units in the convenience and reference samples are “stacked”, which allows for a unit included in the convenience sample to also be included in the reference sample without the requirement to *know* the identity of that unit. They utilize a Bayesian hierarchical modeling approach that specifies a Bernoulli likelihood for indicator  $\mathbf{1}_{z_i}$  for all units in the pooled sample. A likelihood term is also included for  $\pi_r(X_i)$  only for units in the observed reference sample (where  $\pi_r(X_i)$  is known) to borrow further modeling strength. This modeling set-up of [Savitsky et al. \(2023\)](#) may also be performed in the frequentist paradigm. The main advantage of the Bayesian approach is that it treats values  $\pi_r(X_i)$  for the *convenience* sample as *unknown* and allows their estimation in the model. By contrast, in the frequentist set-up (see [Beresovsky et al. \(2024\)](#))  $\pi_r(X_i)$  are assumed known for *all* convenience and reference sample units.

*Remark 2.* In this one-arm case where the domain estimator is constructed solely from the estimated convenience sample inclusion probabilities, the resulting thresholding is performed on the convenience sample inclusion probabilities,  $\pi_c(x_i)$ ,  $i \in S_c \subset U$  (where  $S_c$  denotes units in frame  $U$  that participate in the

convenience sample), without accounting for the estimation quality of  $\pi_c(X)$ . So, this is a traditional regularization approach used to stabilize the variance of a survey domain estimator by excluding units with extreme weight values. This approach trades some small increase in bias for a large decrease in variance.

*Remark 3.* We include an alternative, direct derivation for the result of Theorem 2.1 in an Appendix A assuming Equation 4 is everywhere differentiable (on  $x \in \mathbb{X}$ ). We also include an illustration to show that the result of the Theorem does, indeed, produce a minimum variance estimator for  $\hat{\mu}$ .

Equation 4 can now be generalized in the manner of Section 3.1 of Crump et al. (2009) to develop an alternative to their Theorem 1 and Corollary 1 under a composite estimator that includes both reference and convenience sample inclusion and participation probabilities.

## 2.2. Thresholding using both reference and convenience sample probabilities

Let  $\delta_c$  and  $\delta_r$  denote random inclusion indicators (governed by a survey design distribution) for convenience and reference samples, respectively, and let  $\pi_c(x) = \Pr[\delta_c = 1 \mid X = x]$  and similarly for  $\pi_r$ . Define our estimator as,

$$\hat{\mu} = \mu + \frac{1}{N} \sum_{i=1}^N \frac{z_i \delta_{ci}}{\hat{\pi}_c(x_i)} + \frac{z_i \delta_{ri}}{\pi_r(x_i)}, \quad (7)$$

Although the above estimator is defined disjointly on the reference sample using  $\pi_r(X)$  and the convenience sample using  $\hat{\pi}_c(X)$ , the resulting optimal variance thresholding rule of Equation 11 applies to *only* units in the convenience sample. So, as mentioned in Remark 4, below, we may use estimated  $\hat{\pi}_c(x_i)$  and  $\hat{\pi}_r(x_i)$  for *each* unit  $i \in S_c$  to apply the thresholding rule of Equation 11. To demonstrate that this trick works, we may generate an estimator identical to Equation 7 that includes both convenience and reference sample probabilities defined *solely* for convenience units. Use  $\{\pi_c(x_i)\}_{i \in S_c}$  to generate a pseudo population of size  $N$  (from units  $i \in S_c$ , allowing for replicates). Next take a random / probability sample from this pseudo population using  $\{\pi_r(x_i)\}$  of the same size as the reference sample. Now form the same estimator as Equation 7, but the universe of units is actually confined to  $i \in S_c$ .

Let

$$\phi(Y, \delta_c, \delta_r, X, \mu, e_c, e_r) = \frac{z \delta_c}{\pi_c(X)} + \frac{z \delta_r}{\pi_r(X)} \quad (8)$$

$$\hat{\mu} = \mu + \frac{1}{N} \sum_{i=1}^N \phi(y_i, \delta_{ci}, \delta_{ri}, x_i, \mu, \pi_c(x_i), \pi_r(x_i)). \quad (9)$$

Then, from Hirano et al. (2003) the variance of our estimator is

$$\mathbb{E} [\phi(Y, \delta, X, \mu, e)^2] = \frac{1}{N} \mathbb{E} \left[ \frac{\sigma_c^2(X)}{\pi_c(X)} + \frac{\sigma_r^2(X)}{\pi_r(X)} \right], \quad (10)$$

where  $\sigma_c^2 = \mathbb{V}(Y \mid \delta_c = 1, X = x)$  and similarly for  $\sigma_r^2$ . The expectation on the LHS of Equation 4 is taken with respect to the joint distribution for  $X$  and the taking of a sample from the underlying frame on which  $\mathbb{X}$  is defined. The expectation on the RHS is taken with respect to the distribution for  $X$ . We have used the assumption of independence between the sampling arms with respect to the design distribution.

We may now use Equation 10 to extend and generalize Corollary 1 of Crump et al. (2009) in the case where  $\sigma_c^2 = \sigma_r^2 = \sigma^2$ .

**Theorem 2.2** (Two-arm extension of Crump et al. (2009)).

Assume  $(\pi_c(x) > 0, \pi_r(x) > 0), \forall x \in \mathbb{X}$ .

Then  $\mathbb{A} = \left\{x \in \mathbb{X} : \sqrt{\pi_r(X)\pi_c(X)/(\pi_r(X) + \pi_c(X))} > \alpha\right\}$  defines the optimal subset of  $\mathbb{X}$  where threshold  $\alpha$  is obtained as a solution to,

$$\frac{1}{\alpha^2} = 2\mathbb{E} \left[ \frac{1}{\pi_c(X)} + \frac{1}{\pi_r(X)} \middle| \frac{1}{\pi_c(X)} + \frac{1}{\pi_r(X)} \leq \frac{1}{\alpha^2} \right]. \quad (11)$$

*Proof.* Plugging in  $\pi_c(x)$  for  $e(X)$  and  $\pi_r(X)$  for  $1 - e(X)$  into Theorem 1 of Crump et al. (2009) and using the result of Equation 10 for the case of where we utilize both the convenience sample and reference sample participation and inclusion probabilities (for the convenience units) produces the result.  $\square$

*Remark 4.* Defining variance optimal subset,  $\mathbb{A}$ , by thresholding  $\sqrt{\pi_r(x_i)\pi_c(x_i)/(\pi_r(x_i) + \pi_c(x_i))} > \alpha$  is a harmonic mean that tends to exclude units  $i$  where  $\pi_r(x_i)$  is a very different value from  $\pi_c(x_i)$ . We may even better understand the behavior of this thresholding statistic by noting the result from Beresovsky et al. (2024) that  $\Pr[i \in S_c, i \in S_r \mid i \in S] = \pi_{ri}\pi_{ci}/(\pi_{ri} + \pi_{ci})$ , where  $S = S_c \otimes S_r$  denotes the pooled convenience and reference sample. This result reveals that convenience units with low probabilities of being in *both* the convenience and reference samples tend to be excluded. This thresholding behavior matches intuition because units with low probabilities to appear in both samples will tend to have low overlaps in their covariate supports. We further note that our derivation of this variance minimizing threshold statistic was done without explicit reference to this joint probability, which makes the concordance of the two expressions (for the thresholding statistic, on the one hand, and the joint probability of inclusion in both samples, on the other hand) to be quite fortuitous. We label this thresholding statistic as “balanced” because it favors inclusion of records for estimating the domain mean that have relatively high probabilities of participating in *both* samples.

*Remark 5.* This thresholding method can be used in practice solely directed to units  $i \in S_c$  because we have both estimated  $(\hat{\pi}_c(x_i), \hat{\pi}_r(x_i))$  available.

*Remark 6.* Theorem 2.2 assumes both  $(\pi_r(x), \pi_c(x))$  are *known* for the convenience units when, in fact, they are estimated. We explore the sensitivity to the performance of the variance minimizing thresholding statistic (for the domain mean) of this theorem to estimation uncertainty for  $(\hat{\pi}_r(x), \hat{\pi}_c(x))$  in the simulation study to follow.



### 2.3. Thresholding statistic motivated by variance structure of model score function

Our derivation of the thresholding statistic of Section 2.2 treats  $\pi_c(\mathbf{x})$  as known. By contrast, Beresovsky et al. (2024) suppose a generalized linear model,  $\text{logit}(\pi_{ci}(\boldsymbol{\beta})) = \boldsymbol{\beta}^T \mathbf{x}_i$ , with a linear form under a logit link function for logistic regression. They derive the variance of the domain mean,  $\hat{\mu}$ , that includes an additive term for variance of the score function,  $S(\boldsymbol{\beta})$ , which has two parts:

$$\begin{aligned} \text{Var}[S(\boldsymbol{\beta})] &= \text{Var}[S_c(\boldsymbol{\beta})] + \text{Var}[S_r(\boldsymbol{\beta})] =: \mathbf{A} + \mathbf{D} \\ \mathbf{D} &= \text{Var}_d \left[ \sum_{S_r} \frac{g_i}{1 + g_i} (1 - \pi_{ci}) \mathbf{x}_i \right], \end{aligned}$$

where  $g_i = \pi_c(\mathbf{x}_i)/\pi_r(\mathbf{x}_i)$  and  $\text{Var}_d$  denote the design variance. Motivated by the dependence of  $\mathbf{D}$  on  $g_i$ , we propose to use this statistic as another thresholding option.

We propose the following acceptance set that uses  $g$ :

$$\mathbb{A} = \{x \in \mathbb{X} : \pi_r(x)/\pi_c(x) > \alpha\}.$$

*Remark 7.* The use of  $\pi_r(x)/\pi_c(x)$  as a thresholding statistic may be intuitively motivated by noting that it will tend to threshold or exclude units  $i \in S_c$  where  $\pi_r(\mathbf{x}_i)$  is relatively small for each unit and  $\pi_c(\mathbf{x}_i)$  is relatively large, which may occur if the value for  $\mathbf{x}_i$  for some  $i \in S_c$  is not well covered by or represented in the reference sample,  $S_r$ .

## 3. Simulation study

### 3.1. Simulation design

We conduct a Monte Carlo simulation study that generates a finite population on each iteration to include covariates  $\mathbf{x}$  that govern both the convenience and reference sample designs. The sample designs are size-based as a linear function of  $\mathbf{x}$  where we vary the coefficients of the linear function to draw two categories of reference and convenience samples: 1. Where the covariate spaces of resulting reference and convenience samples express a *high* degree of overlap; 2. Where the two samples express a *low* degree of overlap. We also generate a response variable of interest,  $y$ , for the finite population. A domain mean,  $\mu$ , is constructed for the population and *estimated* by a combined weighted estimator over the reference and convenience samples. Finally, we compare the 3 thresholding methods we developed in Section 2 in terms of their bias, error and coverage performances. We expect that conducting thresholding of sampled convenience units using one or more of our thresholding statistics will reduce estimation error.

We utilize the simulation data generation process of Savitsky et al. (2023). We briefly summarize the procedure and refer the reader for a more detailed exposition. We generate  $M = 30$  distinct populations, each of size  $N = 4000$ .

Design covariates,  $X$ , of dimension  $K = 5$  are generated (all binary, with one continuous). Outcome variable,  $y_i$ , is generated as  $\log(y_i) \sim \mathcal{N}(\mathbf{x}_i\beta, 2)$  for  $i = 1, \dots, N$ .

A randomized reference sample of size  $n_r = 400$  is taken from the finite population under a proportion-to-size (PPS) design with size variable,  $s_{r_i} = \log(\exp(\mathbf{x}_i \times \beta) + 1)$ .

For the convenience sample, we set  $n_c \approx 800$ , which is a relatively larger sampling fraction that we choose to explore the full range of  $\pi_c \in [0, 1]$  that we would expect to see for business establishment data in the U.S. Bureau of Labor Statistics. We use a size-based Poisson sample with  $\pi_{c_i} = \text{logit}^{-1}(\mathbf{x}_i \times \beta_c + \text{offset})$ . We control ‘high’ and ‘low’ overlap by varying  $\beta_c$  compared to the reference sample.

Figure 1 presents a violin (rotated and reflected density) plot for the percentage overlap of *units* in both the convenience and reference samples over the Monte Carlo iterations. The left-hand plot represents the high overlap samples and the right-hand plot represents the low-overlap samples. We see that the number of shared units in both samples is notably higher for the high overlap samples than for the low overlap samples. We expect fewer units to be thresholded for a high overlap sample since their covariate supports express relatively more overlap such that units in the convenience sample are more similar to those in the reference sample. Since our modeling obtains information about the population from the reference sample (and reference sample inclusion probabilities) we are able to better estimate participation probabilities for convenience units that are similar in covariate values to the reference units.

### 3.2. Thresholding of convenience units

In this paper we employ the Bayesian model formulation of [Savitsky et al. \(2023\)](#) that estimates both  $(\pi_r(\mathbf{x}_i), \pi_c(\mathbf{x}_i))$ ,  $i \in S_c$ . In the sequel we use  $(\pi_{ri} = \pi(\mathbf{x}_i), \pi_{ci} = \pi_c(\mathbf{x}_i))$  for ease-of-reading and to emphasize the dependence on  $i \in S_c$ .

Within each Monte Carlo iteration,  $m \in 1, \dots, M$ , we conduct thresholding of convenience units and computation of the domain mean for each posterior/MCMC sample in the following procedure:

1. For *each* posterior/MCMC draw  $s \in 1, \dots, S$ , compute the thresholding statistic (e.g., balanced thresholding statistic) for each unit  $i \in S_c$  as a function of  $(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$ . Denote the focus thresholding statistic as,  $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$ , that allows us to provide a general exposition of how we conduct thresholding of convenience units; for example, we may set  $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi}) = \sqrt{\hat{\pi}_{rsi}\hat{\pi}_{csi}/(\hat{\pi}_{rsi} + \hat{\pi}_{csi})}$  for  $i \in S_c$ .
2. For MCMC iteration  $s$ : evaluate the distribution of the thresholding statistic  $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$  over the convenience units,  $i \in S_c$ , and compute threshold quantile,  $\alpha_s$  associated with target percentile,  $\gamma$ , below which convenience units are excluded / thresholded.

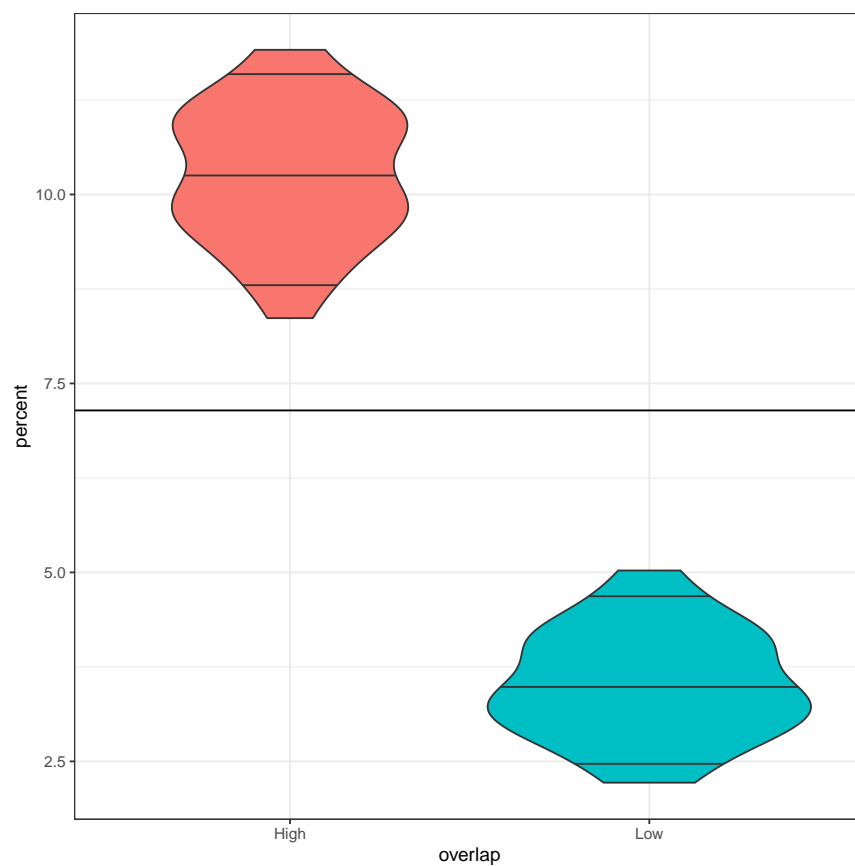


Fig 1: Distribution over  $M = 30$  Monte Carlo iterations of the percentage of units overlapping between realized reference and convenience samples (taken on each Monte Carlo iteration).

3. *Retain/accept* those *convenience* units where  $\mathbb{A}_s = \{i \in S_c : T(\hat{\pi}_{rsi}, \hat{\pi}_{csi}) > \alpha_s\}$ .
4. Use the retained units in draw  $s$  to construct the domain mean,  $\mu_s = (\sum_{i \in \mathbb{A}_s} y_i / \hat{\pi}_{csi} + \sum_{i \in S_r} y_i / \hat{\pi}_{rsi}) / (\sum_{i \in \mathbb{A}_s} 1 / \hat{\pi}_{csi} + \sum_{i \in S_r} 1 / \hat{\pi}_{rsi})$ .
5. One now has the induced posterior distribution over the  $S$  MCMC samples for  $\mu$  from which one may estimate the mean (e.g.,  $\mu = 1/S \sum_{s=1}^S \mu_s$ ).

*Remark 8.* The above procedure is a form of “soft” thresholding because a unit  $i \in S_c$  may be excluded on posterior sampling draw  $s$  in forming domain mean estimator  $\mu_s$ , but then *included* in posterior draw  $s'$  to construct  $\mu_{s'}$ . So each  $\mu_s$  may be constructed from a differing set of convenience units. This occurs because  $(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$  are parameters estimated from our model, so the distribution of  $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$  over convenience units  $i \in S_c$  will vary from over the posterior draws,  $s \in 1, \dots, S$ .

We formulate a variation to this procedure that produces a “hard” threshold to compare the performance to our main soft thresholding procedure. For the hard thresholding alternative, we construct the acceptance sets  $\mathbb{A}_s$ ,  $s = 1, \dots, S$  as described in the first 3 steps of the above procedure. We then count the percentage of over the  $S$  posterior draws that unit  $i$  is in each acceptance set  $\mathbb{A}_s$ . If the percentage is less than 50% we *exclude* or threshold unit  $i$ . In other words, we form a single acceptance set over the  $S$  MCMC draws with  $\mathbb{A} = \{i \in S_c : i \in \mathbb{A}_s \text{ for a total of } S^i = \sum_s (1 : i \in \mathbb{A}_s) > 0.5S\}$ . So, our first additional steps formulates  $\mathbb{A}$ , the set of non-thresholded convenience units. We then use this *same* set of units to compute  $\mu_s$  for each MCMC draw. So, either unit  $i$  is included to construct all the  $\mu_s$  or it is excluded. We use the label “two-step” for this hard thresholding alternative since we first threshold the units over all MCMC draws and then compute the domain mean estimator.

*Remark 9.* Although our thresholding procedure is constructed under the Bayesian model formulation of [Savitsky et al. \(2023\)](#) for developing a thresholded posterior distribution for domain mean,  $\mu$ , steps 1 – 4 of our thresholding procedure may be applied under the frequentist generalized linear formulation of [Beresovsky et al. \(2024\)](#) to obtain a thresholded estimator of  $\mu$  with no loss of generality or applicability. Instead of thresholding each MCMC draw,  $s$ , one would threshold the statistic formed from the maximum likelihood estimators of the convenience sample participation probabilities under frequentist model estimation.

### 3.3. Results

Figure 2 presents plot panels for bias, root mean squared error (RMSE), median absolute deviation (MAD) and coverage results over the  $M$  Monte Carlo iterations. The left side of each horizontal bar in the plot panels represents a result for “L” or the low overlap sample, while the right side of each horizontal bar represents a result for “H” or the high overlap sample. The top most row of bars in blue presents results using the unknown *true* values for both the reference sample inclusion probabilities for the reference sample units and the convenience

sample inclusion probabilities the convenience units as if they were known. The next row of bars down from the top in red presents the result from the model of Savitsky et al. (2023) that smooths or co-models the inclusion probabilities for the reference sample units. No thresholding is conducted for the results in these first two rows. The next two rows of bars present results for our variance optimal balanced threshold statistic: the orange bar uses our main soft thresholding procedure, while the yellow bar uses the alternative hard thresholding procedure that we label as “two-step”. The next row of light green bars presents results for thresholding  $\pi_{ri}$  while the last row of green bars presents results for the ratio  $(\pi_{ri}/\pi_{ci})$  thresholding statistic. We remind the reader that the statistics and thresholding are performed over  $i \in S_c$  (the convenience sample) and that our Bayesian model estimates both  $(\pi_{ri}, \pi_{ci})$  for each unit in the convenience sample. The vertical black dashed line in each plot panel represents the result using *only* the reference sample (and excluding the convenience sample). We use the  $\gamma = 5\%$  of the distribution over the convenience for each threshold statistic to compute the thresholding quantile,  $\alpha$ .

One notes that the estimation errors (RMSE, MAD) are little different both with and without thresholding and among the thresholding statistics for the high (H) overlap samples, which is expected because there is less need for thresholding due to the high degree of overlap in covariate spaces between the reference and convenience samples such that most convenience sample inclusion probabilities are well-estimated. By contrast, we observe that the estimation errors for the balanced statistic perform best among the different thresholding statistics and even better than the case where use the true convenience sample participation probabilities (blue bars) as is they were known. The slight increase in bias relative to the blue bar is more than offset by a decrease in variance, producing lower estimation error. There is little difference between the soft and hard thresholding alternatives under the balanced statistic, though the soft thresholding produces a slightly higher amount of bias but also a slightly lower amount of estimation error as compared to hard thresholding. Perhaps we are not surprised that the balanced threshold statistic performed best because it was derived as a minimum variance estimator for the domain mean, though it is surprising that this thresholding option performed better for low overlap (L) samples than did the domain mean estimator constructed from the true (rather than estimated) convenience sample inclusion probabilities (as if they were known).

Lastly, while the balanced threshold statistic produces only a slight improvement in error for high overlap (H) samples, the notion of whether a convenience sample is high or low overlap is relative such that the practitioner may not know whether their realized reference and convenience samples represent H or L. Nevertheless, since thresholding with the balanced statistic never produces worse errors than not thresholding and sometimes much better there is little risk to use thresholding.

We chose a reasonably small (5%) percentile for thresholding, so we next experiment with 10% and 1% under our best performing balanced thresholding statistic (under soft thresholding). Figure 3 presents the results. While the estimation errors are similar for the 3 different percentiles for low overlap (L)

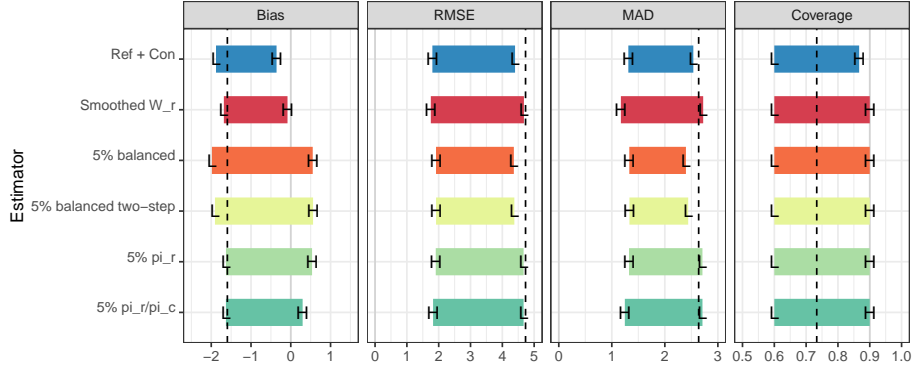


Fig 2: Performance of the weighted mean estimator between high (H) and low (L) overlapping samples using variations of the two-arm method across Monte Carlo Simulations for (top to bottom): True weights for both samples (Blue), Smoothed weights for reference sample (Red), minimum variance or balanced  $\sqrt{\pi_r(x)\pi_c(x)/(\pi_r(x) + \pi_c(x))}$  (Orange), balanced based on posterior mean (Yellow),  $\pi_r$  only (Light Green),  $\pi_r/\pi_c$  (Dark Green). Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only.

samples, we nevertheless note that the error performance is notably better for 1% balanced thresholding under high overlap (H) samples than the other two higher thresholding percentiles, and even performs slightly better than the blue bar that uses true convenience sample participation probabilities. Thresholding fewer units for high overlap samples intuitively makes sense since convenience units are relatively more similar to reference units. The low overlap sample MAD is, however, worst for the 1% threshold and best for the 10% threshold, which also accords with intuition since the convenience units in low overlap samples are less similar (in their covariate values) to reference sample units. Yet, the worsening of estimation error in the low overlap is a much smaller magnitude than the improvement in estimation error for high overlap. Our results suggest that the practitioner may generally favor a relatively lower value for the thresholding percentile.

While thresholding does notably reduce estimation errors (RMSE/MAD) on low overlap samples, as expected, uncertainty quantification is little improved (and continues to express undercoverage) even after thresholding due to the limited estimation improvement offered for a low overlap convenience sample. The fidelity of uncertainty quantification is driven by the underlying degree of overlap in the covariate supports of the reference and convenience sampling arms and is not much affected by thresholding relatively few convenience units. As a result of the low quality of uncertainty quantification under the low overlap samples, the coverage performances for all methods express little differentiation. By contrast,

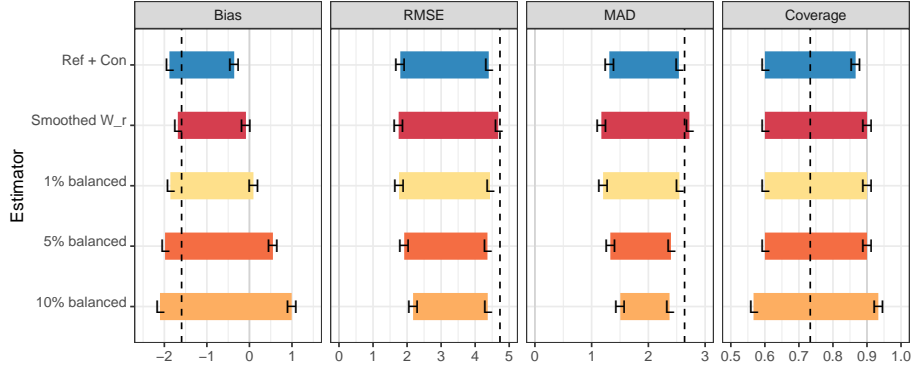


Fig 3: Comparison of the variance for the balanced threshold,  $\sqrt{\pi_r(x)\pi_c(x)/(\pi_r(x) + \pi_c(x))}$  between high (H) and low (L) overlapping samples for (top to bottom): True weights for both samples (Blue), Smoothed weights for reference sample (Red), 1% (Yellow) vs. 5% (Orange) and 10% (Light Orange). Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only.

for high overlap the coverage results are more robust and nominal coverage is achieved when thresholding relatively fewer units, as expected. Thresholding is most important for low overlap samples to prevent non-representative outliers from inducing large errors (due to biased estimation of their convenience sample inclusion probabilities). Our results show that thresholding for low overlap samples provides a notable improvement in error control over repeated sampling.

We recall that the balanced threshold statistic was derived to produce a minimum variance domain mean estimator. Yet, the result in Section 2 assumes that the reference sample inclusion and convenience sample participation probabilities for convenience sample units,  $(\pi_{ri}, \pi_{ci})$ ,  $i \in S_c$ , are *known* when, in fact, they are estimated. We seek to assess the sensitivity of the thresholding statistics to uncertainty in estimation of these inclusion and participation probabilities for convenience units.

Each curve in a each plot panel of Figure 4 presents a sequence of 90% credibility intervals of percentiles for the fit statistic estimated on each MCMC iteration. More specifically, if we fix an MCMC iteration, we next compute the estimated thresholding statistic from the probabilities for each unit and compute its percentile of the distribution of the statistic over the convenience sample units. We repeat this process for each MCMC draw, which gives us a range of percentiles of the thresholding statistic for each convenience sample unit. Each horizontal line in the curve represents the 90% credibility interval of the percentiles for a convenience sample unit. These lines are ordered along the horizontal axis by the posterior mean of estimated thresholding statistic

for each unit. The longer the horizontal lines, the greater the estimation uncertainty for the thresholding statistic. The blue-colored horizontal lines represent those units who have switched from being on one side of threshold to the other (meaning, they were sometimes included and sometimes excluded) more than 10% of the MCMC samples. The horizontal dashed lines in each panel represent 1%, 5%, 10% thresholds (from bottom-to-top).

The left-hand curve in each plot panel represents estimations under low overlap samples and the right-hand represents high overlap samples. The left plot panel represents the balanced thresholding statistic, while the right panel represents the ratio thresholding statistic.

Focusing on the left-hand panel for the balanced thresholding statistic, we see that the relatively wider horizontal lines for the low overlap sample express more estimation uncertainty than do those for the high overlap sample. That accords with our expectation because the reference sample provides less information about convenience units whose covariate values are different from those of the reference sample. Yet, we see relatively few units (colored in blue) that switch between being excluded/thresholded and included for estimating the domain mean. So, the uncertainty does not impact the thresholding set and that explains why the balanced thresholding statistic turned out to be variance optimal as compared to the other thresholding statistics despite the uncertainties in estimating inclusion and participation probabilities. By contrast, we observe a relatively higher number of units that switch between inclusion and exclusion under the ratio thresholding statistic in the right-hand plot panel. So, the performance of this thresholding statistic is less robust under uncertainty about the probabilities than is the balanced thresholding statistic.

#### 4. Discussion

The quasi-randomization method of [Savitsky et al. \(2023\)](#) that treats the non-randomized convenience sample as if it arose from a latent survey design process with an unknown sampling distribution provides a start-of-art method for producing survey-weighted domain estimates. Yet, the estimation quality of inclusion and participation probabilities for convenience units depends on the degree of overlap in the design covariate spaces between the randomized reference and convenience samples. It is typically the case that the estimated convenience sample participation probabilities for some convenience units whose design covariate values are very different from the reference sample are not well-estimated. Incorporating these units can partially defeat the purpose of leveraging the convenience sample by actually increasing bias and variance as compared to excluding them.

We devised a soft thresholding procedure for excluding convenience sample units that are very different from reference sample units and achieved a notable reduction in estimation error for low overlap (in their design covariate spaces) samples. We began by developing a new formulation for a balanced threshold statistic that minimized the resulting variance of the domain estimator. Our



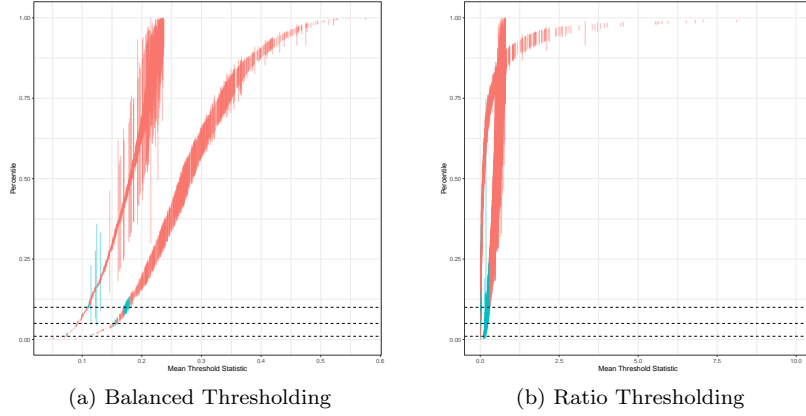


Fig 4: Vertical lines are percentiles of Threshold statistic distribution over 700 MCMC draws of  $\pi_{ci}, \pi_{ri}$  for convenience units. Left is L and Right is H. Blue denotes unit that jumped threshold  $> 10\%$  of draws

balanced thresholding statistic proposes to exclude some convenience sample units and is constructed from inclusion and participation probabilities for convenience units that effectively serve as one-dimensional summaries of the design covariates. It was particularly interesting to discover that the balanced threshold statistic derived from a theoretical exposition turns out to be a function of the joint probability that a unit is in *both* the reference and convenience samples. This formulation makes intuitive sense because our procedure proposes to exclude those convenience units that express low probabilities of being in both samples.

We motivated an additional thresholding statistic that we labeled “ratio” as the ratio of reference and convenience sample inclusion probabilities based on the variance formulation of the domain mean estimator derived in [Beresovsky et al. \(2024\)](#).

We designed a soft thresholding procedure that constructed an acceptance set for convenience units to be included in domain mean estimator on each MCMC iteration such that a unit might be included in some iterations but not others.

Our result revealed that the balanced threshold statistic produced the greatest reduction in the variance of the domain estimator, particularly for relatively lower overlap samples. We also showed that this reduction is relatively insensitive to the percentile cutoff for the estimated distribution of the balanced threshold statistic over the convenience sample units. Finally, we showed that this variance reduction result is robust against estimation uncertainty because the units that are thresholded are minimally impacted under our soft thresholding procedure.

## Appendix A: Direct derivation of variance minimizing threshold for one-arm sample

Hajek mean estimator from convenience sample  $S_c$ :

$$\hat{y} = \frac{\sum_{S_c} \frac{y(x)}{\hat{e}(x)}}{\sum_{S_c} \frac{1}{\hat{e}(x)}}$$

where  $\hat{e}(x)$  is estimated propensity score.

Model-based variance of this estimator

$$\text{var}(\hat{y}) = \frac{\sum_{S_c} \frac{\sigma_y^2(x)}{\hat{e}^2(x)}}{\left[\sum_{S_c} \frac{1}{\hat{e}(x)}\right]^2}$$

Assume that all variance  $\sigma_y^2(x) = \sigma_y^2$  are equal. Order convenience sample units by response propensity  $\hat{e}(x)$ . Units can be listed by  $\hat{e}(x)$  with density  $w(\hat{e}(x)) = \hat{e}(x)$ . Variance estimated from full convenience sample  $S_c$  without cut-off may be expressed as integral over the distribution of response propensity  $\hat{e}(x)$

$$\text{var}(\hat{y}) = \frac{\int_0^1 \frac{\sigma_y^2(x)}{\hat{e}^2(x)} w(\hat{e}(x)) d(\hat{e}(x))}{\left[\int_0^1 \frac{1}{\hat{e}(x)} w(\hat{e}(x)) d(\hat{e}(x))\right]^2} = \frac{\sigma_y^2 \int_0^1 \frac{1}{\hat{e}(x)} d(\hat{e}(x))}{\left[\int_0^1 d(\hat{e}(x))\right]^2}$$

If sample units are trimmed by response propensity at level  $\varepsilon$ , then variance depending on  $\varepsilon$  is

$$\text{var}(\hat{y}, \varepsilon) = \frac{\sigma_y^2 \int_\varepsilon^1 \frac{1}{\hat{e}(x)} d(\hat{e}(x))}{\left[\int_\varepsilon^1 d(\hat{e}(x))\right]^2} = \frac{\sigma_y^2 F(\varepsilon)}{G^2(\varepsilon)},$$

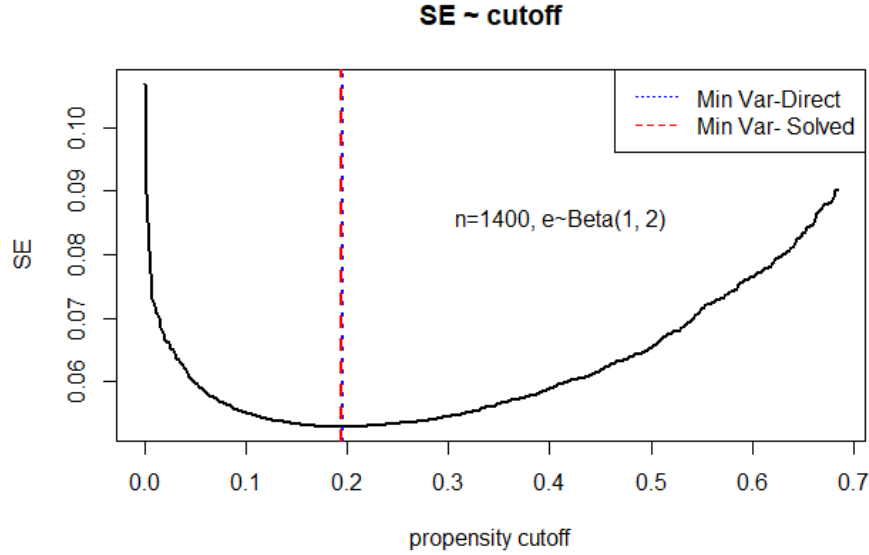
where  $F(\hat{e}(x))$  is a primitive of  $f(\hat{e}(x)) = 1/\hat{e}(x)$  and  $G(\hat{e}(x))$  is a primitive of 1.

Minimize the trimmed variance by  $\varepsilon$

$$\frac{d \text{var}(\hat{y}, \varepsilon)}{d\varepsilon} = \frac{\sigma_y^2 F'(\varepsilon) G^2(\varepsilon) - 2G'(\varepsilon) G(\varepsilon) \sigma_y^2 F(\varepsilon)}{G^4(\varepsilon)} = 0$$

Here we have:

$$\begin{aligned} F'(\varepsilon) &= \frac{d}{d\varepsilon} (F(1) - F(\varepsilon)) = 0 - \frac{1}{\varepsilon} \times 1 \\ G'(\varepsilon) &= \frac{d}{d\varepsilon} (G(1) - G(\varepsilon)) = G'(1) - G'(\varepsilon) = -1. \end{aligned}$$



Optimal propensity cut-off point  $\varepsilon$  can be estimated from the numerator null condition

$$\frac{1}{\varepsilon_c} G(\varepsilon_c) - 2F(\varepsilon_c) = 0$$

$$\frac{1}{\varepsilon_c} = \frac{2F(\varepsilon_c)}{G(\varepsilon_c)} = \frac{2 \sum_{S_c} \frac{1}{\hat{e}(x)} \mathbb{1}[\hat{e}(x) > \varepsilon_c]}{\sum_{S_c} 1 \mathbb{1}[\hat{e}(x) > \varepsilon_c]}$$

Results of simulations:

- Sample size  $n = 1,400$
- Propensity score  $\hat{e} \sim \text{Beta}(1, 2)$

## References

- Beresovsky, V., J. Gershunskaya, and T. D. Savitsky (2024). Review of quasi-randomization approaches for estimation from non-probability samples.
- Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review* 78(2), 161 — 188.
- Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009, 01). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1), 187–199.
- Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2, 813–845.

- Elliott, M. R. and R. Valliant (2017). Inference for Nonprobability Samples. *Statistical Science* 32(2), 249 – 264.
- Gelman, A. and J. Hill (2007). *Data analysis using regression and multi-level/hierarchical models*, Volume Analytical methods for social research. New York: Cambridge University Press.
- Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161–1189.
- Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics* 12(2), 685 – 726.
- Savitsky, T. D., M. R. Williams, J. Gershunskaya, and V. Beresovsky (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition* 24(5), 1–34.
- Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* 8(2), 231–263.
- VanderWeele, T. J. and I. Shpitser (2011). A new criterion for confounder selection. *Biometrics* 67(4), 1406 — 1413.
- Wang, L., R. Valliant, and Y. Li (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.* 40(4), 5237–5250.
- Williams, D. and J. M. Brick (2017, 07). Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort. *Journal of Survey Statistics and Methodology* 6(2), 186–211.
- Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology* 48(2), 283–311.