Vol. 0 (2023) 1–2

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

Terrance D. Savitsky[1], and Matthew R. Williams[2]. and Vladislav Beresovsky[3], and Julie Gershunskaya[4]

[1]*Office of Survey Methods Research, U.S. Bureau of Labor Statistics, e-mail:* Beresovsky.Vladislav@bls.gov

[2]*Office of Survey Methods Research, U.S. Bureau of Labor Statistics, e-mail:* Savitsky.Terrance@bls.gov

[3]*RTI International, e-mail:* mrwilliams@rti.org

[4]*OEUS Statistical Methods Division, U.S. Bureau of Labor Statistics, e-mail:* Gershunskaya.Julie@bls.gov

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, Inclusion probabilities, Thresholding units, Bayesian hierarchical modeling.

Contents

*U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E, Washington, D.C. 20212 USA

1. Introduction

Declining response rates for randomized survey instruments administered by government statistical agencies have encouraged the development of quasi-randomization processes such as those of Wu (2022); Wang et al. (2021); Savitsky et al. (2023) that leverage 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 precisely because the convenience sample is not generally representative of that population.

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 to produce a domain estimator (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.

Yet, because the 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. 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 (α) 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 partcipation 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) 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 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,

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

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

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

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

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

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

where $\sigma_1^2 = \mathbb{V}\left(Y \mid \delta = 1, X = x\right)$. The expectation on the LHS of Equation 4is 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, α.

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* α *is constructed by,*

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

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

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

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. □

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, $\hat{\pi}_c(x_i),\ i \in (1, \ldots, N)$. In the sequel, we will further extend Theorem 2.1 by additionally estimating the reference sample inclusion probabilities for the same convenience units, $\hat{\pi}_r(x_i),\ i \in (1, \ldots, N)$ 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.

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 **??** 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 δ_c and δ_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 π_r. Define our estimator as,

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

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

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

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

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

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

where $\sigma_c^2 = \mathbb{V}\left(Y \mid \delta_c = 1, X = x\right)$ and similarly for σ_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* α *is obtained as a solution to,*

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

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. □

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 $S = S_c \bigotimes S_r$ 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.

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 Beresovsky et al. (2024)

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 linear model, $\text{logit}(\pi_{ci}(\boldsymbol{\beta})) = \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}_i$. 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:

$$\text{Var}[S(\boldsymbol{\beta})] = \text{Var}[S_c(\boldsymbol{\beta})] + \text{Var}[S_r(\boldsymbol{\beta})] =: \mathbf{A} + \mathbf{D}$$

$$\boldsymbol{D} = \text{Var}_d\left[\sum\nolimits_{S_r} \frac{g_i}{1 + g_i} (1 - \pi_{ci})\, \mathbf{x}_i \right],$$

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

In particular, in this paper we employ the Bayesian model formulation of Savitsky et al. (2023) that estimates both $(\pi_r(\mathbf{x}_i), \pi_c(\mathbf{x}_i))\,,\ i \in S_c$. So, 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

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, μ, 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 $i = 1, \ldots, N$.

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 β_c compared to the reference sample.

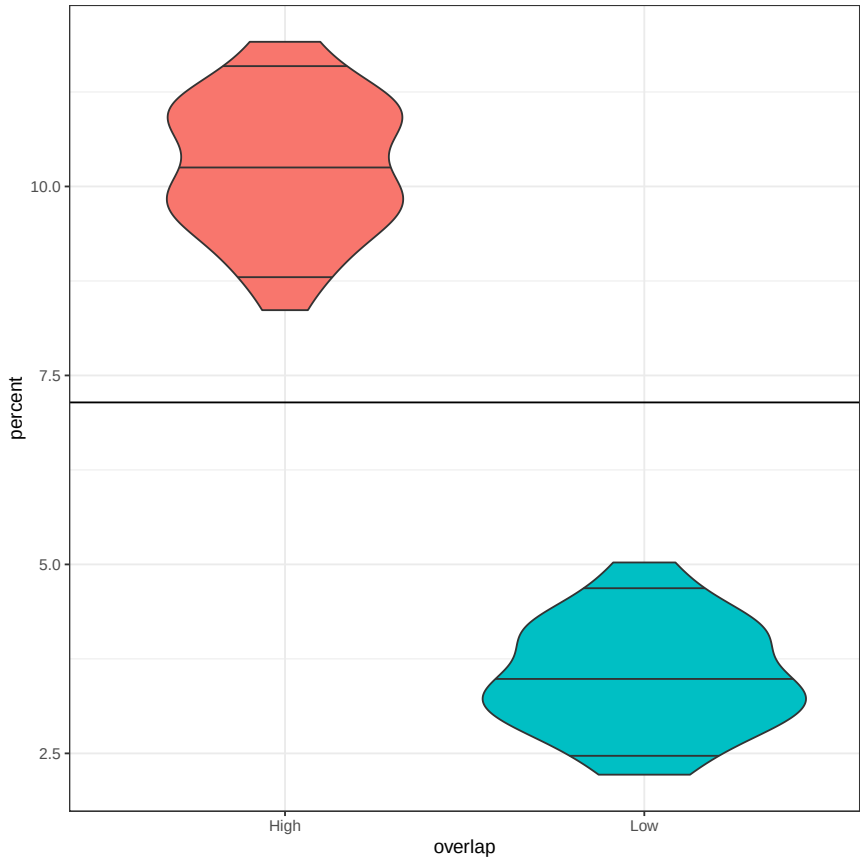


Fig 1: Distribution over $M = 30$ Monte Carlo iterations of the percentage of units overlapping between realized reference and convenience samples (taken on each Monte Carlo iteration).

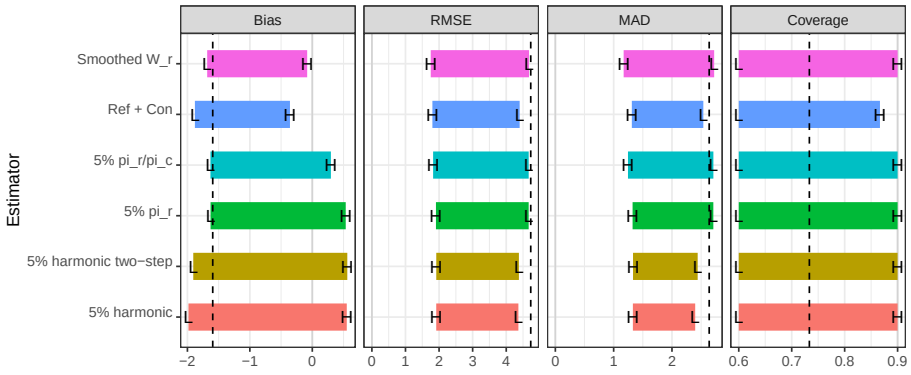


Fig 2: Performance of the weighted mean estimator between high (H) and low (L) overlapping samples using variations of the two-arm method across Monte Carlo Simulations for (top to bottom): True weights for both samples (Blue), Smoothed weights for reference sample (Pink), π_r/π_c (Turquoise), π_r only (Green), minimum variance or harmonic $\sqrt{\pi_r(x)\pi_c(x)/(\pi_r(x)+\pi_c(x))}$ (Red), harmonic based on posterior mean (Gold). 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.

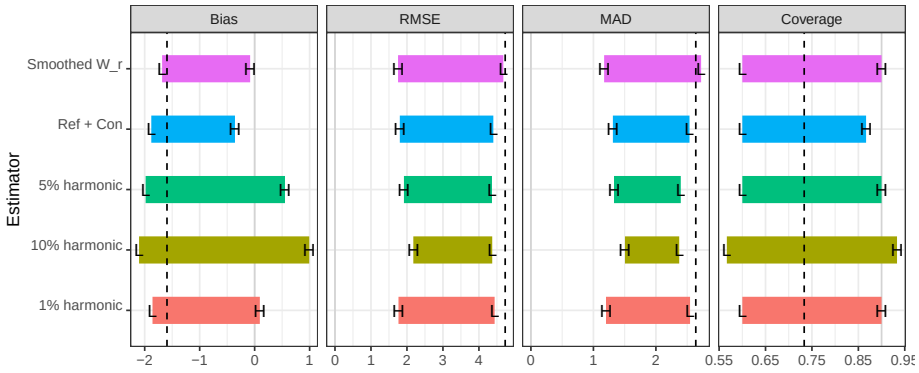


Fig 3: Comparison of the variance for the harmonic threshold, $\sqrt{\pi_r(x)\pi_c(x)/(\pi_r(x)+\pi_c(x))}$ between high (H) and low (L) overlapping samples for (top to bottom): True weights for both samples (Blue), Smoothed weights for reference sample (Pink), 5% (Green) vs. 10% (Gold) and 1% (Red). 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.

4. Discussion

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

Hajek mean estimator from convenience sample S_c:

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

where $\hat{e}(x)$ is estimated propensity score.
Model-based variance of this estimator

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

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

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

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

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

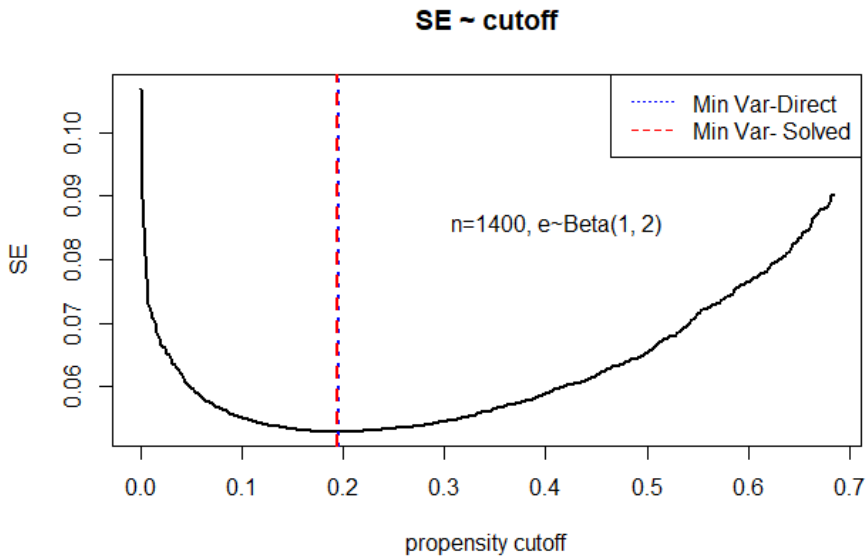
where $F(\hat{e}(x))$ is a primitive of $f(\hat{e}(x)) = 1/\hat{e}(x)$ and $G(\hat{e}(x))$ is a primitive of 1.

Minimize the trimmed variance by ε

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

Here we have:

$$F'\left(\varepsilon\right) = \frac{d}{d\varepsilon}\left(F(1) - F(\varepsilon)\right) = 0 - \frac{1}{\varepsilon} \times 1$$
$$G'\left(\varepsilon\right) = \frac{d}{d\varepsilon}\left(G(1) - G(\varepsilon)\right) = G'(1) - G'(\varepsilon) = -1.$$



Optimal propensity cut-off point ε can be estimated from the numerator null condition

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

$$\frac{1}{\varepsilon_c} = \frac{2F\left(\varepsilon_c\right)}{G\left(\varepsilon_c\right)} = \frac{2\sum_{S_c} \frac{1}{\widehat{e}(x)} \left|\widehat{e}\left(x\right) > \varepsilon_c\right.}{\sum_{S_c} 1 \left|\hat{e}\left(x\right) > \varepsilon_c\right.}$$

Results of simulations:

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

References

Beresovsky, V., J. Gershunskaya, and T. D. Savitsky (2024). Review of quasi-randomization approaches for estimation from non-probability samples.

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.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

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.

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.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology 48*(2), 283–311.