Debiased Calibration for Generalized Two-Phase Sampling

Caleb Leedy

April 29, 2024

1 Introduction

Combining information from several sources is an important practical problem. (CITEME) We want to incorporate information from external data sources to reduce the bias in our estimates or improve the estimator's efficiency. For many problems, the additional information consists of summary statistics with standard errors. The goal of this project is to incorporate external information with existing data to create more efficient estimators using calibration weighting.

To model this scenaro, we formulate the problem as a generalized two-phase sample where the first phase sample consists of data from multiple sources. The second phase sample contains our existing data. To motivate this setup, we consider the following approach: first, we consider the classical two-phase sampling setup where the second phase sample is a subset of the first phase sample; then, we extend this setup to consider non-nested two-phase samples; and finally, we consider the more general approach of having multiple sources.

2 Topic 1: Classical Two-Phase Sampling

2.1 Background

Consider a finite population of size N containing elements (X_i, Y_i) where an initial (Phase 1) sample of size n_1 is selected and X_i is observed. Then from the Phase 1 sample of elements, a (Phase 2) sample of size $n_2 < n_1$ is selected and Y_i is observed. This is two-phase sampling

(See Fuller (2009), Kim (2024) for general references.) The goal of two-phase sampling is to construct an estimator of $\bar{Y}_N$ that uses both the observed information in the Phase 2 sample and also the extra auxiliary information from X in the Phase 1 sample. The challenge is doing this efficiently.

An easy-to-implement unbiased estimator in the spirit of a Horvitz-Thompson (HT) estimator (Horvitz and Thompson (1952), Narain (1951)) is the π^*-estimator. Let $\pi_i^{(2)}$ be the response probability of element i being observed in the Phase 2 sample. Then, allowing the elements in the Phase 1 sample to be represented by A_1 and the elements in the Phase 2 sample to be denoted as A_2, if we define $\pi_{2i|1} = \sum_{A_2:i\in A_2} \Pr(A_2 \mid A_1)$ and $\pi_{1i} = \sum_{A_1:i\in A_1} \Pr(A_1)$ then,

$$\pi_i^{(2)}(A_1) = \pi_{2i|1}\pi_{1i}.$$

This means that we can define the π^*-estimator as the following design unbiased estimator:

$$\hat{Y}_{\pi^*} = \sum_{i\in A_2} \frac{y_i}{\pi_{2i|1}\pi_{1i}}.$$

While unbiased (see Kim (2024)), the π^*-estimator does not account for the additional information contained in the auxiliary Phase 1 variable X. The two-phase regression estimator $\hat{Y}_{reg,tp}$ does incorporate information for X by using the estimate $\hat{X}_1$ from the Phase 1 sample. This is how we can leverage the external information $\hat{X}_1$ to improve the initial π^*-estimator in the second phase sample. The two-phase regression estimator has the form,

$$\hat{Y}_{reg,tp} = \sum_{i\in A_1} \frac{1}{\pi_{1i}} x_i\hat{\beta}_q + \sum_{i\in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}}(y_i - x_i\hat{\beta}_q)$$

where for $q_i = q(x_i)$ and is a function of x_i,

$$\hat{\beta}_q = \left(\sum_{i\in A_2} \frac{x_i x_i'}{\pi_{1i}q_i}\right)^{-1} \sum_{i\in A_2} \frac{x_i y_i}{\pi_{1i}q_i}.$$

The regression estimator is the minimum variance design consistent linear estimator which is easily shown to be the case because $\hat{Y}_{reg,tp} = \sum_{i \in A_2} \hat{w}_{2i} y_i / \pi_{1i}$ where

$$\hat{w}_{2i} = \arg\min_{w} \sum_{i \in A_2} (w_{2i} - \pi_{2i|1}^{-1})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} x_i / \pi_{1i} = \sum_{i \in A_1} x_i / \pi_{1i}.$$

This means that $\hat{Y}_{reg,tp}$ is also a calibration estimator. The idea that regression estimation is a form of calibration was noted by Deville and Sarndal (1992) and extended by them to consider loss functions other than just squared loss. Their generalized loss function minimizes $\sum_i G(w_i, d_i) q_i$ for weights w_i and design-weights d_i where $G(\cdot)$ is a non-negative, strictly convex function with respect to w, defined on an interval containing d_i, with $g(w_i, d_i) = \partial G / \partial w$ continuous.[1] This generalization includes empirical likelihood estimation, and maximum entropy estimation among others. The variance estimation is based on a linearization that shows that minimizing the generalized loss function subject to the calibration constraints is asymptotically equivalent to a regression estimator.

Furthermore, the regression estimator has a nice feature that its two terms can be thought about as minimizing the variance and bias correction,

$$\hat{Y}_{reg,tp} = \underbrace{\sum_{i \in A_1} \frac{x_i \hat{\beta}_q}{\pi_{1i}}}_{\text{Minimizing the variance}} + \underbrace{\sum_{i \in A_2} \frac{1}{\pi_{1i} \pi_{2i|1}} (y_i - x_i \hat{\beta}_q)}_{\text{Bias correction}}.$$

The Deville and Sarndal (1992) method incorporates the design weights into the loss function, which is the part minimizing the variance. We would rather separate have bias calibration separate from the minimizing the variance so that we can control each in isolation. In Kwon et al. (2024), the authors show that for a generalized entropy function $G(w)$, including a term of $g(\pi_{2i|1}^{-1})$ into the calibration for $g = \partial G / \partial w$ not only creates a design consistent estimator, but it also has better efficiency than the generalized regression estimators of Deville and Sarndal (1992).

The method of Kwon et al. (2024) requires known finite population calibration levels. It

[1] The Deville and Sarndal (1992) paper considers regression estimators for a single phase setup, which we apply to our two-phase example.

does not handle the two-phase setup where we need to estimate the finite population total of x from the Phase 1 sample. In the rest of the section, we extend this method to two phase sampling so that we have a valid estimator when including estimated Phase 1 weights with appropriate variance estimation.

2.2 Methodology

We follow the approach of Kwon et al. (2024) for the debiased calibration method. We consider maximizing the generalized entropy Gneiting and Raftery (2007),

$$H(w) = -\sum_{i \in A_2} \frac{1}{\pi_{1i}} G(w_{2i}) q_i \tag{1}$$

where $G : \mathcal{V} \to \mathbb{R}$ is strictly convex, differentiable function subject to the constraints:

$$\sum_{i \in A_2} \frac{x_i w_{2i} q_i}{\pi_{1i}} = \sum_{i \in A_1} \frac{x_i q_i}{\pi_{1i}} \tag{2}$$

and

$$\sum_{i \in A_2} \frac{g(\pi_{2i|1}^{-1}) w_{2i} q_i}{\pi_{1i}} = \sum_{i \in A_1} \frac{g(\pi_{2i|1}^{-1}) q_i}{\pi_{1i}}, \tag{3}$$

where $g(w) = \partial G / \partial w$.

The first constraint is the existing calibration constraint and the second ensures that design consistency is achieved. The original method of Kwon et al. (2024) only considered having finite population quantities on the right hand side of (2).

Writing $w_{1i} = \pi_{1i}^{-1}$, the the goal is to solve

$$\arg\min_{w_{2|1}} \sum_{i \in A_2} \frac{1}{\pi_{1i}} G(w_{2i}) q_i \text{ such that } \sum_{i \in A_2} w_{1i} w_{2i|1} z_i q_i = \sum_{i \in A_1} w_{1i} z_i q_i \tag{4}$$

where $z_i = (x_i/q_i, g(\pi_{2i|1}^{-1}))$. Let $\hat{w}_{2i|1}$ be the solution to Equation (4). The resulting estimator of Y_N is $\hat{Y}_{DCE} = \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1} y_i$. To solve this problem we can use the method of Lagrange multipliers. We need to minimize the Lagrangian function

$$L(w_{2i|1}, \lambda) = \sum_{i \in A_2} w_{1i} G(w_{2i|1}) q_i + \lambda \left(\sum_{i \in A_1} w_{1i} z_i q_i - \sum_{i \in A_2} w_{1i} w_{2i|1} z_i q_i \right). \quad (5)$$

Differntiating with respect to $w_{2i|1}$ and setting this expression equal to zero, yields the fact that $\hat{w}_{2i|1}$ satisfies

$$\hat{w}_{2i|1}(\hat{\lambda}) = g^{-1}(\hat{\lambda}^T z_i)$$

where $\hat{\lambda}$ is the solution to

$$\left(\sum_{i \in A_1} w_{1i} z_i q_i - \sum_{i \in A_2} w_{1i} w_{2i|1}(\hat{\lambda}) z_i q_i \right) = 0. \quad (6)$$

2.3 Theoretical Results

Theorem 1 (Design Consistency)**.** *Let λ^* be the probability limit of $\hat{\lambda}$. Under some regularity conditions,*

$$\hat{Y}_{DCE} = \hat{Y}_\ell(\lambda^*, \phi^*) + O_p(N/n_2)$$

where

$$\hat{Y}_\ell(\hat{\lambda}, \phi^*) = \hat{Y}_{DCE}(\hat{\lambda}) + \left(\sum_{i \in A_1} w_{1i} z_i q_i - \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1}(\hat{\lambda}) z_i q_i \right) \phi^*$$

and

$$\phi^* = \left[\sum_{i \in U} \frac{\pi_{2i|1} z_i z_i^T q_i}{g'(d_{2i|1})} \right]^{-1} \sum_{i \in U} \frac{\pi_{2i|1} z_i y_i}{g'(d_{2i|1})}.$$

$$\hat{Y}_\ell(\lambda^*, \phi^*) \;=\; \hat{Y}_{\pi^*} + \left(\sum_{i \in A_1} w_{1i} x_i - \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} x_i \right)^T \phi_1^* + \left(\sum_{i \in A_1} w_{1i} g_i - \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} g_i \right)^T \phi_2^*$$

and

$$\begin{pmatrix}\phi_1^* \\ \phi_2^*\end{pmatrix} = \left[\sum_{i\in U}\frac{\pi_{2i|1}}{g'(d_{2i|1})q_i}\begin{pmatrix}\mathbf{x}_i\mathbf{x}_i^T & \mathbf{x}_i g_i \\ g_i\mathbf{x}_i^T & g_i^2\end{pmatrix}\right]^{-1}\sum_{i\in U}\frac{\pi_{2i|1}}{g'(d_{2i|1})q_i}\begin{pmatrix}\mathbf{x}_i \\ g_i\end{pmatrix} y_i$$

with $g_i = g(\pi_{2i|1}^{-1})q_i$.

Proof. In this proof, we derive the solution to Equation 4 and show that it is asymptotically equivalent to a regression estimator. Using the method of Lagrange multipliers, to solve Equation (4) we need to minimize the Lagrangian in Equation (5). The first order conditions show that

$$\frac{\partial\mathcal{L}}{\partial w_{2i|1}} : g(w_{2i|1})w_{1i}q_i - \lambda w_{1i}z_i q_i = 0.$$

Hence, $\hat{w}_{2i}(\lambda) = g^{-1}(\lambda^T z_i)$ and $\hat{\lambda}$ is determined by Equation (6). When the sample size gets large, we have $\hat{w}_{2i|1}(\hat{\lambda}) \to d_{2i|1}$ which means that $\hat{\lambda} \to \lambda^*$ where $\lambda^* = (\mathbf{0}, 1)$. Then using the linearization technique of Randles (1982), we can construct a regression estimator,

$$\hat{Y}_\ell(\hat{\lambda}, \phi) = \hat{Y}_{DCE}(\hat{\lambda}) + \left(\sum_{i\in A_1} w_{1i}z_i q_i - \sum_{i\in A_2} w_{1i}\hat{w}_{2i|1}(\hat{\lambda})z_i q_i\right)\phi.$$

Notice that $\hat{Y}_\ell(\hat{\lambda}, \phi) = \hat{Y}_{DCE}(\hat{\lambda})$ for all $\phi \in \mathbb{R}$. We choose ϕ^* such that

$$E\left[\frac{\partial}{\partial\lambda}\hat{Y}_\ell(\lambda^*, \phi^*)\right] = 0.$$

Using the fact that $g^{-1}(\lambda^* z_i) = g^{-1}(g(d_{2i|1})) = d_{2i|1}$ and $(g^{-1})'(x) = 1/g'(g^{-1}(x))$, we have

$$\begin{aligned}\phi^* &= E\left[\sum_{i\in A_2}\frac{w_{1i}z_i z_i^T q_i}{g'(d_{2i|1})}\right]^{-1} E\left[\sum_{i\in A_2}\frac{w_{1i}z_i y_i}{g'(d_{2i|1})}\right] \\ &= \left[\sum_{i\in U}\frac{\pi_{2i|1}z_i z_i^T q_i}{g'(d_{2i|1})}\right]^{-1}\left[\sum_{i\in U}\frac{\pi_{2i|1}z_i y_i}{g'(d_{2i|1})}\right]\end{aligned}$$

Thus, the linearization estimator is

$$\hat{Y}_\ell(\lambda^*, \phi^*) = \sum_{i \in A_1} w_{1i} q_i z_i \phi^* + \sum_{i \in A_2} w_{1i} d_{2i|1} (y_i - q_i z_i \phi^*).$$

By construction using a Taylor expansion yields,

$$\begin{aligned}\hat{Y}_{DCE}(\hat{\lambda}) &= \hat{Y}_\ell(\lambda^*, \phi^*) + E\left[\frac{\partial}{\partial \lambda} \hat{Y}_\ell(\lambda^*, \phi^*)\right](\hat{\lambda} - \lambda^*) + \frac{1}{2} E\left[\frac{\partial}{\partial \lambda^2} \hat{Y}_{DCE}(\lambda^*)\right](\hat{\lambda} - \lambda^*)^2 \\ &= \hat{Y}_\ell(\lambda^*, \phi^*) + O(N) O_p(n_2^{-1}).\end{aligned}$$

The final equality comes from the fact that $E\left[\frac{\partial}{\partial \lambda} \hat{Y}_\ell(\lambda^*, \phi^*)\right] = 0$, $\frac{\partial}{\partial \lambda^2} \hat{w}_{2i|1}(\lambda^*)$ is bounded and $|\hat{\lambda} - \lambda^*| = O_p(n_2^{-1/2})$, which proves our result. □

2.4 Simulation Study 1

We run a simulation testing the proposed method. In this approach we have the following simulation setup:

$$\begin{aligned} X_{1i} &\overset{ind}{\sim} N(2, 1) \\ X_{2i} &\overset{ind}{\sim} Unif(0, 4) \\ X_{3i} &\overset{ind}{\sim} N(0, 1) \\ \varepsilon_i &\overset{ind}{\sim} N(0, 1) \\ Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\ \pi_{1i} &= n_1 / N \\ \pi_{2i|1} &= \min(\Phi_3(-x_{1i} + 1), 0.9). \end{aligned}$$

where Φ_3 is the CDF of a t-distribution with 3 degrees of freedom. This is a two-phase extension of the setup in Kwon et al. (2024). We consider a finite population of size $N = 10,000$ with the Phase 1 sampling being a simple random sample (SRS) and the Phase 2 sampling occuring under Poisson sampling. This yields a Phase 1 sample size of $n_1 = 1000$ and a Phase 2 sample size of $E[n_2] \approx 267$. In the Phase 1 sample, we observe (X_1, X_2) while in the Phase 2 sample we observe (X_1, X_2, Y). This simulation does not deal with

model misspecification, and we compare the proposed method for the parameter $\bar{Y}_N$ with four approaches:

1. π^*-estimator: $\hat{Y}_{\pi^*} = N^{-1}\sum_{i\in A_2}\frac{y_i}{\pi_{1i}\pi_{2i|1}}$,
2. Two Phase Regression estimator (TP-Reg): $\hat{Y}_{reg} = \sum_{i\in A_1}\frac{\mathbf{x}_i'\hat{\beta}}{\pi_{1i}} + \sum_{i\in A_2}\frac{1}{\pi_{1i}\pi_{2i|1}}(y_i - \mathbf{x}_i'\hat{\beta})$ where $\hat{\beta} = \left(\sum_{i\in A_2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A_2}\mathbf{x}_i y_i$ and $\mathbf{x}_i = (x_{1i}, x_{2i})^T$,
3. Debiased Calibration with Population Constraints (DC-Pop): This solves

$$\arg\min_{w_{2|1}} \sum_{i\in A_2}\frac{1}{\pi_{1i}}G(w_{2i}) \text{ such that } \sum_{i\in A_2} w_{1i}w_{2i|1}z_i = \sum_{i\in U} z_i. \tag{7}$$

4. Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (4) with $q_i = 1$.

In addition to estimating the mean parameter $\bar{Y}_N$, we also construct variance estimates $\hat{V}(\hat{Y})$ for each estimate $\hat{Y}$. For each approach we give the variance estimate in Table 1.

Estimator	Estimated Variance	Notes
π^*	$\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{s}_Y^2 + \sum_{i\in A_2}\frac{w_{1i}^2}{N^2}\frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2}y_i^2$	$\hat{s}_Y^2 = \frac{1}{n_2-1}\sum_{i\in A_2}(y_i-\bar{y}_i)^2$
TP-Reg	$\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{s}_\eta^2 + \sum_{i\in A_2}\frac{w_{1i}}{N^2}\frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2}\varepsilon_i^2$	$\eta_i = x_i'\beta + \delta_{2i}w_{2i\|1}\varepsilon_i,\ \varepsilon_i = (y_i - x_i'\beta)$
DC-Pop	$\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{s}_\varepsilon^2 + \sum_{i\in A_2}\frac{w_{1i}}{N^2}\frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2}\varepsilon_i^2$	$\eta_i = z_i'\phi + \delta_{2i}w_{2i\|1}\varepsilon_i,\ \varepsilon_i = (y_i - z_i'\phi)$
DC-Est	$\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{s}_\eta^2 + \sum_{i\in A_2}\frac{w_{1i}}{N^2}\frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2}\varepsilon_i^2$	$\eta_i = z_i'\phi + \delta_{2i}w_{2i\|1}\varepsilon_i,\ \varepsilon_i = (y_i - z_i'\phi)$

Table 1: This table gives the formulas for each variance estimator used in Simulation 1. For the derivation of the estimated variance of the DC-Pop estimatator see Appendix A.

We run the simulation 1000 times for each of these methods and compute the Bias ($E[\hat{Y}]-\bar{Y}_N$), the RMSE ($\sqrt{\text{Var}(\hat{Y}-\bar{Y}_N)}$), a 95% empirical confidence interval ($\sum_{b=1}^{1000}|\hat{Y}^{(b)} - \bar{Y}_N| \leq \Phi(0.975)\sqrt{\hat{V}(\hat{Y}^{(b)})^{(b)}}$), and a T-test that assesses the unbiasedness of each estimator where $\hat{Y}^{(b)}$ is the result from the bth simulation replicate. The results are in Table 2.

Dr. Kim is this confidence interval good enough? It seems slightly high for me.

I also give the mean variance estimates, the Monte Carlo mean, as well as a test for bias.

Est	Bias	RMSE	EmpCI	Ttest
π^*	-0.064	1.196	0.918	1.701
TP-Reg	-0.005	0.145	0.985	1.065
DC-Pop	-0.003	0.096	0.969	1.002
DC-Est	-0.005	0.148	0.976	1.076

Table 2: This table shows the results of Simulation Study 1. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: π^*, TP-Reg, DC-Pop, and DC-Est.

Est	MCVar	EstVar	VarVar	Ttest
π^*	1.4278	1.3457	0.5088	3.6412
TP-Reg	0.0209	0.0389	0.0002	45.7082
DC-Pop	0.0093	0.0119	0.0000	58.2888
DC-Est	0.0219	0.0343	0.0001	47.8689

Table 3: This table shows the variance estimates from Simulation Study 1. It displays the Monte Carlo variance of $\hat{Y}^{(b)}$, the average estimated variance, the variance of the variance estimator, and a t-statistic assessing the unbiasedness of each variance estimator.

3 Topic 2: Non-nested Two-Phase Sampling

3.1 Background

Now we consider the sampling mechanism known as non-nested two phase sampling (Hidiroglou (2001)). In the last section, we considered two phase sampling in which the Phase 2 sample was a subset of the Phase 1 sample. With non-nested two phase sampling the Phase 2 sample is independent of the Phase 1 sample. It is a separate independent sample of the same population. Like traditional two phase sampling, we consider the Phase 1 sample, A_1, to consist of observations of $(X_i)_{i=1}^{n_1}$ and the Phase 2 sample, A_2, to consist of observations of $(X_i, Y_i)_{i=1}^{n_2}$.

Whereas the classical two phase estimator uses a single Horvitz-Thompson estimator of the Phase 1 sample to construct estimates for calibration totals, in the non-nested two phase sample we have two independent Horvitz-Thompson estimators of the total of X,

$$\hat{X}_1 = \sum_{i \in A_1}^{n_1} d_{1i} x_i \text{ and } \hat{X}_2 = \sum_{i \in A_2}^{n_2} d_{2i} x_i$$

where $d_{1i} = \pi_{1i}^{-1}$, $d_{2i} = \pi_{2i}^{-1}$, π_{1i} is the probability of $i \in A_1$ and $\pi_{2i} = \Pr(i \in A_2)$. We can combine these estimates using the effective sample size (Kish (1965)) to get $\hat{X}_c = (n_{1,eff}\hat{X}_1 + n_{2,eff}\hat{X}_2)/(n_{1,eff} + n_{2,eff})$ where $n_{1,eff}$ and $n_{2,eff}$ are the effective sample size for A_1 and A_2 respectively. Then we can define a regression estimator as

$$\hat{Y}_{NN,reg} = \hat{Y}_2 + (\hat{X}_c - \hat{X}_2)^T \hat{\beta}_q = \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)^T W^T \hat{\beta}_q$$

where, $W = n_{1,eff}/(n_{1,eff} + n_{2,eff})$, and

$$\hat{\beta}_q = \left(\sum_{i \in A_2} \frac{x_i x_i^T}{q_i} \right)^{-1} \sum_{i \in A_2} \frac{x_i y_i}{q_i} \text{ and } \hat{Y}_2 = \sum_{i \in A_2} d_{2i} y_i.$$

Since the samples A_1 and A_2 are independent,

$$V(\hat{Y}_{NN,reg}) = V\left(\sum_{i \in A_2} \frac{1}{\pi_{2i}} (y_i - x_i^T W \beta_q^*) \right) + (\beta_q^*)^T W V(\hat{X}_1) W \beta_q^*$$

where β_q^* is the probability limit of $\hat{\beta}_q$. Like the two phase sample this regression estimator can be viewed as the solution to the following calibration equation

$$\hat{w}_2 = \arg\min_{w_2} Q(w_2) = \sum_{i \in A_2} (w_{2i} - d_{2i})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} x_i = \hat{X}_c \tag{8}$$

and $\hat{Y}_{NN,reg} = \sum_{i \in A_2} \hat{w}_{2i} y_i$ where $\hat{w}_{2i}$ is the solution to Equation 8.

We extend the debiased calibration estimator of Kwon et al. (2024) to the non-nested two phase sampling case where we use a combined estimate $\hat{X}_c$ as the calibration totals instead of using the true totals from the finite population.

3.2 Methodology

The methodology for the non-nested two phase sample is very similar to the setup described as part of Topic 1. Given a strictly convex differentiable function, $G : \mathcal{V} \to \mathbb{R}$, the goal is to

solve

$$\hat{w}_2 = \arg\min_{w} \sum_{i \in A_2} G(w_{2i}) q_i \text{ such that } \sum_{i \in A_2} w_{2i} x_i = \hat{X}_c \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) q_i = \sum_{i \in U} g(d_{2i}) q_i \tag{9}$$

for $g(x) = G'(x)$ and a known choice of $q_i \in \mathbb{R}$.

The difference between solving Equation 9 and Equation 4 is that the estimator $\hat{X}_c$ is estimated from the combined sample $A_c = A_1 \cup A_2$. Before using $\hat{X}_c$ in the debiased calibration estimator, we need to estimate it from the non-nested samples.

We can get multiple estimates of $\hat{X}$,

$$\hat{X}_1 = \sum_{i \in A_1} d_{1i} x_i \text{ and } \hat{X}_2 = \sum_{i \in A_2} d_{2i} x_i.$$

Let V_1 and V_2 be known variance matrices for $\hat{X}_1$ and $\hat{X}_2$, then the optimal combined estimate is

$$\hat{X}_c = \frac{V_1^{-1} \hat{X}_1 + V_2^{-1} \hat{X}_2}{V_1^{-1} + V_2^{-1}}.$$

We can constuct a non-nested two phase estimator $\hat{Y}_{NNE}$ for Y_N where $\hat{Y}_{NNE} = \sum_{i \in A_2} \hat{w}_{2i} y_i$ and $\hat{w}_{2i}$ solves Equation 9. Like the classical two phase approach, to solve this setup we minimize the Lagrangian,

$$L(w_{2i}, \lambda) = \sum_{i \in A_2} G(w_{2i}) q_i + \lambda \left(\hat{T} - \sum_{i \in A_2} w_{2i} z_i q_i \right). \tag{10}$$

with

$$\hat{T} = \begin{bmatrix} \hat{X}_c \\ \sum_{i \in U} g(d_{2i}) q_i \end{bmatrix}.$$

I am a little worried about the second term of $\hat{T}$. It is somewhat weird to assume that $T_g = \sum_{i \in U} g(d_{2i}) q_i$ is known. One simple way is to use a HT estimator of T_g from sample A_2. That is, use $\hat{T}_{g,\text{HT}} = \sum_{i \in A_2} d_{2i} g(d_{2i}) q_i$ or even use a GREG estimator of T_g.

Differntiating with respect to w_{2i} and setting this expression equal to zero, yields the fact that $\hat{w}_{2i}$ satisfies

$$\hat{w}_{2i}(\hat{\lambda}) = g^{-1}(\hat{\lambda}^T z_i)$$

where $\hat{\lambda}$ is the solution to

$$\left(\hat{T} - \sum_{i \in A_2} w_{2i}(\hat{\lambda}) z_i q_i\right) = 0. \tag{11}$$

3.3 Theoretical Results

Theorem 2 (Design Consistency)**.** *Allowing λ^* to be the probability limit of $\hat{\lambda}$, under some regularity conditions, $\hat{Y}_{NNE} = \hat{Y}_{\ell,NNE}(\lambda^*, \phi^*) + O_p(Nn_2^{-1})$ where*

$$\hat{Y}_{\ell,NNE}(\lambda^*, \phi^*) = \sum_{i \in A_2} \hat{w}_{2i}(\lambda^*) + \left(\hat{T} - \sum_{i \in A_2} \hat{w}_{2i}(\lambda^*) z_i q_i\right) \phi^*$$

and

$$\phi^* = \left(\sum_{i \in U} \frac{\pi_{2i} q_i}{g'(d_{2i})} \begin{bmatrix} x_i^2/q_i & x_i g(d_{2i})/q_i \\ x_i g(d_{2i})/q_i & g(d_{2i})^2 \end{bmatrix}\right)^{-1} \sum_{i \in U} \frac{\pi_{2i} y_i}{g'(d_{2i})} \begin{bmatrix} x_i/q_i \\ g(d_i) \end{bmatrix}.$$

Proof. The proof of this result is very similar to the proof the Theorem 1. The biggest difference is that the total for X is estimated from both samples using $\hat{X}_c$ instead of $\hat{X}_{HT}$ from the Phase 1 sample.

Since $\hat{Y}_{NNE} = \sum_{i \in A_2} \hat{w}_{2i}(\hat{\lambda}) y_i$ where $\hat{\lambda}$ solves

$$\sum_{i \in A_2} \hat{w}_{2i}(\lambda) q_i \underbrace{\begin{bmatrix} x_i/q_i \\ g(d_i) \end{bmatrix}}_{z_i} = T \tag{12}$$

we have

$$\hat{Y}_{\ell,NNE}(\hat{\lambda}, \phi) = \sum_{i \in A_2} \hat{w}_{2i}(\hat{\lambda}) + \left(T - \sum_{i \in A_2} \hat{w}_{2i}(\hat{\lambda}) z_i q_i\right) \phi.$$

If we choose ϕ^* such that $E\left[\frac{\partial}{\partial\lambda}\hat{Y}_{\ell,NNE}(\lambda^*,\phi^*)\right]=0$, then

$$\phi^*=\begin{bmatrix}\phi_1^*\\ \phi_2^*\end{bmatrix}=\left(\sum_{i\in U}\frac{\pi_{2i}q_i}{g'(d_{2i})}\begin{bmatrix}x_i^2/q_i & x_ig(d_{2i})/q_i\\ x_ig(d_{2i})/q_i & g(d_{2i})^2\end{bmatrix}\right)^{-1}\sum_{i\in U}\frac{\pi_{2i}y_i}{g'(d_{2i})}\begin{bmatrix}x_i/q_i\\ g(d_i)\end{bmatrix}.$$

Hence, by a Taylor expansion around $\hat{\lambda}$,

$$\hat{Y}_{NNE}(\hat{\lambda})=\hat{Y}_{\ell,NNE}(\lambda^*,\phi^*)+O_p(Nn_2^{-1}).$$

□

Theorem 3 (Variance Estimation)**.** *The variance of $\hat{Y}_{NNE}$ is*

$$\begin{aligned}Var(\hat{Y}_{NNE}(\hat{\lambda})=(\phi_1^*)^T\,Var(\hat{X}_c)\phi_1^*&+\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i-z_i\phi^*q_i)(y_j-z_j\phi^*q_j)\\&+(1-W)\phi_1^*\sum_{i\in U}\sum_{j\in U}\Delta_{2ij}d_{2i}x_id_{2j}(y_j-z_j\phi_1^*q_j)\end{aligned}$$

We can estimate the variance using

$$\begin{aligned}\hat{V}_{NNE}=(\hat{\phi}_1)^T\,Var(\hat{X}_c)\hat{\phi}_1&+\sum_{i\in A_2}\sum_{j\in A_2}\frac{\Delta_{2ij}}{\pi_{2ij}\pi_{2i}\pi_{2j}}(y_i-z_i\hat{\phi}q_i)(y_j-z_j\hat{\phi}q_j)\\&+(1-W)\hat{\phi}_1\sum_{i\in A_2}\sum_{j\in A_2}\frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_i}{\pi_{2i}}\frac{(y_j-z_j\hat{\phi}_1q_j)}{\pi_{2j}}\end{aligned}$$

where

$$\hat{\phi}=\begin{bmatrix}\hat{\phi}_1\\ \hat{\phi}_2\end{bmatrix}=\left(\sum_{i\in A_2}\frac{q_i}{g'(d_{2i})}\begin{bmatrix}x_i^2/q_i & x_ig(d_{2i})/q_i\\ x_ig(d_{2i})/q_i & g(d_{2i})^2\end{bmatrix}\right)^{-1}\sum_{i\in A_2}\frac{y_i}{g'(d_{2i})}\begin{bmatrix}x_i/q_i\\ g(d_i)\end{bmatrix}.$$

Proof. From Theorem 2, we know that $\hat{Y}_{NNE}(\hat{\lambda})=\hat{Y}_{\ell,NNE}(\lambda^*,\phi^*)+O_p(Nn_2^{-1})$. Hence, the variance of $\hat{Y}_{NNE}(\hat{\lambda})$ is

$$
\begin{aligned}
\mathrm{Var}(\hat{Y}_{NNE}(\hat{\lambda})) &= \mathrm{Var}(\hat{Y}_{\ell,NNE}(\lambda^*,\phi^*) + O_p(Nn_2^{-1})) \\
&= \mathrm{Var}\left(\sum_{i\in A_2}\hat{w}_{2i}(\lambda^*)y_i + \left(T - \sum_{i\in A_2}\hat{w}_{2i}(\lambda^*)z_iq_i\right)\phi^*\right) \\
&= (\phi_1^*)^T\mathrm{Var}(\hat{X}_c)\phi_1^* + \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i - z_i\phi^*q_i)(y_j - z_j\phi^*q_j) \\
&\quad + 2\mathrm{Cov}\left(\hat{X}_c\phi_1^*, \sum_{i\in A_2}\frac{(y_i - z_i\phi^*q_i)}{\pi_{2i}}\right) \\
&= (\phi_1^*)^T\mathrm{Var}(\hat{X}_c)\phi_1^* + \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i - z_i\phi^*q_i)(y_j - z_j\phi^*q_j) \\
&\quad + (1-W)\phi_1^*\sum_{i\in U}\sum_{j\in U}\Delta_{2ij}\frac{x_i}{\pi_{2i}}\frac{(y_j - z_j\phi_1^*q_j)}{\pi_{2j}}
\end{aligned}
$$

where the last equality comes from the fact that $\hat{X}_c = W\hat{X}_1 + (1-W)\hat{X}_2$. To have an unbiased estimator of the variance we can use:

$$
\begin{aligned}
\hat{V}_{NNE} &= (\hat{\phi}_1)^T\mathrm{Var}(\hat{X}_c)\hat{\phi}_1 + \sum_{i\in A_2}\sum_{j\in A_2}\frac{\Delta_{2ij}}{\pi_{2ij}\pi_{2i}\pi_{2j}}(y_i - z_i\hat{\phi}q_i)(y_j - z_j\hat{\phi}q_j) \\
&\quad + (1-W)\hat{\phi}_1\sum_{i\in A_2}\sum_{j\in A_2}\frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_i}{\pi_{2i}}\frac{(y_j - z_j\hat{\phi}_1q_j)}{\pi_{2j}}
\end{aligned}
$$

where

$$
\hat{\phi} = \begin{bmatrix}\hat{\phi}_1 \\ \hat{\phi}_2\end{bmatrix} = \left(\sum_{i\in A_2}\frac{q_i}{g'(d_{2i})}\begin{bmatrix}x_i^2/q_i & x_ig(d_{2i})/q_i \\ x_ig(d_{2i})/q_i & g(d_{2i})^2\end{bmatrix}\right)^{-1}\sum_{i\in A_2}\frac{y_i}{g'(d_{2i})}\begin{bmatrix}x_i/q_i \\ g(d_i)\end{bmatrix}.
$$

□

3.4 Simulation Study 2

We run a simulation testing the proposed method. This is very similar to Simulation 1. We have the following simulation setup:

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} Unif(0,4) \\
X_{3i} &\overset{ind}{\sim} N(0,1) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\
\pi_{1i} &= n_1/N \\
\pi_{2i} &= \min(\Phi_3(-x_{1i}+1), 0.9).
\end{aligned}
$$

where Φ_3 is the CDF of a t-distribution with 3 degrees of freedom. We consider a finite population of size $N = 10,000$ with the Phase 1 sampling being a simple random sample (SRS) of size $n_1 = 1000$. The Phase 2 sample is a Poisson sample with an expected sample size of

and Phase 2 sampling occuring under Poisson (Bernoulli) sampling. This yields a Phase 1 sample size of $E[n_1] \approx 1100$ and a Phase 2 sample size of $E[n_2] \approx 300$. In the Phase 1 sample, we observe (X_1, X_2) while in the Phase 2 sample we observe (X_1, X_2, Y). This simulation does not deal with model misspecification, and we compare the proposed method for the parameter $\bar{Y}_N$ with four approaches:

1. HT-estimator: $\hat{Y}_{HT} = N^{-1} \sum_{i \in A_2} \frac{y_i}{\pi_{2i}}$,

2. Regression estimator (Reg): Let $\hat{Y}_{NN,reg} = \hat{Y}_{HT} + (\hat{X}_c - \hat{X}_{2,HT})\hat{\beta}_2$ where $\hat{Y}_{HT} = \sum_{i \in A_2} d_{2i} y_i$, $\hat{X}_c = W\hat{X}_{1,HT} + (1-W)\hat{X}_{2,HT}$, $W = n_{1,eff}/(n_{1,eff} + n_{2,eff})$, $\hat{X}_{1,HT} = \sum_{i \in A_1} d_{1i} x_i$, $\hat{X}_{2,HT} = \sum_{i \in A_2} d_{2i} x_i$ and

$$
\hat{\beta}_2 = \left(\sum_{i \in A_2} x_i x_i^T \right)^{-1} \sum_{i \in A_2} x_i y_i.
$$

   Then $\hat{\bar{Y}}_{NN,reg} = \hat{Y}_{NN,reg}/N$.

3. Debiased Calibration with Population Constraints (DC-Pop): This solves

$$\hat{w}_2 = \arg\min_{w} \sum_{i \in A_2} G(w_{2i}) q_i \text{ with } \sum_{i \in A_2} w_{2i} x_i = \sum_{i \in U} x_i \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) q_i = \sum_{i \in U} g(d_{2i}) q_i$$

4. Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (9) with $q_i = 1$.

In addition to estimating the mean parameter $\bar{Y}_N$, we also construct variance estimates $\hat{V}(\hat{Y})$ for each estimate $\hat{Y}$.

We run the simulation 1000 times for each of these methods and compute the Bias ($E[\hat{Y}] - \bar{Y}_N$), the RMSE ($\sqrt{\text{Var}(\hat{Y} - \bar{Y}_N)}$), a 95% empirical confidence interval ($\sum_{b=1}^{1000} |\hat{Y}^{(b)} - \bar{Y}_N| \leq \Phi(0.975)\sqrt{\hat{V}(\hat{Y}^{(b)})^{(b)}}$), and a T-test that assesses the unbiasedness of each estimator where $\hat{Y}^{(b)}$ is the result from the bth simulation replicate. The results are in Table 4.

Est	Bias	RMSE	EmpCI	Ttest
HT	-0.007	0.372	0.949	0.630
Reg	-0.004	0.111	0.957	1.233
DC-Pop	0.000	0.028	0.939	0.031
DC-Est	-0.004	0.111	0.955	1.252

Table 4: This table shows the results of Simulation Study 2. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: HT, Reg, DC-Pop, and DC-Est.

4 Topic 3: Multi-source Two-Phase Sampling

References

Deville, J.-C. and C.-E. Sarndal (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association 87*(418), 376–382.

Fuller, W. A. (2009). *Sampling statistics*. John Wiley & Sons.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Hidiroglou, M. (2001). Double sampling. *Survey methodology 27*(2), 143–154.

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

Kim, J. K. (2024). *Statistics in Survey Sampling*. arXiv.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, Inc.

Kwon, Y., J. K. Kim, and Y. Qiu (2024). Debiased calibration estimation using generalized entropy in survey sampling.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics 3*(2), 169–175.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 462–474.

A Derivation of the Estimated Variance of DC-Pop in Simulation 1

We know that $\hat{Y}_{DC-Pop} = \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1}(\hat{\lambda}) y_i$ where $\hat{w}_{2i|1}$ solves

$$\arg\min_{w_{2|1}} \sum_{i \in A_2} \frac{1}{\pi_{1i}} G(w_{2i}) \text{ such that } \sum_{i \in A_2} w_{1i} w_{2i|1} z_i = \sum_{i \in U} z_i. \tag{13}$$

This means that using the technique of Randles (1982), the linearization of $\hat{Y}_{DC-Pop}$ is

$$\hat{Y}_{DC-Pop,\ell} = \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1}(\hat{\lambda}) y_i + \left(\sum_{i \in U} z_i - \sum_{i \in A_2} w_{1i} w_{2i|1}(\hat{\lambda}) z_i \right) \phi.$$

Since ϕ^* satisfies $E\left[\frac{\partial}{\partial \lambda} \hat{Y}_{DC-Pop,\ell}(\lambda^*, \phi^*)\right] = 0$, ϕ^* is the same as it is for $\hat{Y}_{DC-Est,\ell}$ with $q_i = 1$,

$$\begin{aligned} \phi^* &= E\left[\sum_{i \in A_2} \frac{w_{1i} z_i z_i^T}{g'(d_{2i|1})}\right]^{-1} E\left[\sum_{i \in A_2} \frac{w_{1i} z_i y_i}{g'(d_{2i|1})}\right] \\ &= \left[\sum_{i \in U} \frac{\pi_{2i|1} z_i z_i^T}{g'(d_{2i|1})}\right]^{-1} \left[\sum_{i \in U} \frac{\pi_{2i|1} z_i y_i}{g'(d_{2i|1})}\right]. \end{aligned}$$

Hence,

$$\hat{Y}_{DC-Pop,\ell} = \sum_{i \in U} z_i \phi^* + \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1}(\lambda^*)(y_i - z_i \phi^*).$$

Unlike the two-phase regression estimator and $\hat{Y}_{DC-Est}$ this linearization only has uncertainty in the second term. The first term is a population estimate. Let $\varepsilon_i = (y_i - z_i \phi^*)$. This means that

$$
\begin{aligned}
V(\hat{Y}_{DC-Pop,\ell}) &= V(E[\hat{Y}_{DC-Pop,\ell} \mid A_1]) + E[V(\hat{Y}_{DC-Pop,\ell} \mid A_1)] \\
&= V\left(\sum_{i\in A_1} w_{1i}\pi_{2i|1}\hat{w}_{2i|1}(\lambda^*)\varepsilon_i\right) + E\left[\sum_{i\in A_1}\pi_{2i|1}(1-\pi_{2i|1})w_{1i}^2\hat{w}_{2i|1}(\lambda^*)^2\varepsilon_i^2\right] \\
&= \sum_{i\in U}\pi_{1i}(1-\pi_{1i})w_{1i}^2\pi_{2i|1}^2\hat{w}_{2i|1}(\lambda^*)^2\varepsilon_i^2 + \sum_{i\in U}\pi_{1i}\pi_{2i|1}(1-\pi_{2i|1})w_{1i}^2\hat{w}_{2i|1}(\lambda^*)^2\varepsilon_i^2 \\
&= \sum_{i\in U} w_{1i}w_{2i|1}(\lambda^*)^2\pi_{2i|1}((1-\pi_{1i})\pi_{2i|1} + (1-\pi_{2i|1}))\varepsilon_i^2.
\end{aligned}
$$

Since $w_{2i|1}(\lambda^*)\pi_{2i|1} = 1$, there are two potential estimators of the variance,

$$
\begin{aligned}
\hat{V}_1 &= \sum_{i\in A_2} w_{1i}^2 w_{2i|1}(\hat{\lambda})^2((1-\pi_{1i})\pi_{2i|1} + (1-\pi_{2i|1}))\varepsilon_i^2 \\
\hat{V}_2 &= \sum_{i\in A_2} w_{1i}^2 d_{2i|1}^2((1-\pi_{1i})\pi_{2i|1} + (1-\pi_{2i|1}))\varepsilon_i^2.
\end{aligned}
$$

In the current simulation, we use $\hat{V}_1$, but I have tried $\hat{V}_2$ and the results are similar.