# Debiased Calibration for Generalized Two-Phase Sampling

Caleb Leedy

April 18, 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 $\pi^*$-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 $\pi^*$-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 $\pi^*$-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 $\pi^*$-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

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

The first constraint is the existing calibration constraint and the second ensures that design consistency is achieved. Here, $g(w) = \partial G / \partial w$. The original method of Kwon et al. (2024) only considered having finite population quantities on the right hand side of 2. Let $z_i = (x_i/q_i, g(\pi_{2i|1}^{-1}))$. Letting $w_{1i} = \pi_{1i}^{-1}$, the the goal is to solve,

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

Let $\hat{w}_{2i|1}$ be the solution to Equation 4, then the estimate of $Y_N$ is $\hat{Y}_{DCE} = \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1} y_i$.

## 2.3 Theoretical Results

**Theorem 1** (Design Consistency)**.** *Under 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*

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

*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,

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

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, $\hat{w}_{2i}(\lambda) = g^{-1}(\lambda z_i)$ and $\hat{\lambda}$ is determined by the calibration condition. 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 $\phi^*$ 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 Studies

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

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} Unif(0,4) \\
X_{3i} &\overset{ind}{\sim} N(0,1) \\
X_{4i} &\overset{ind}{\sim} Unif(0.1,0.9) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\
\pi_{1i} &= \Phi_3(-x_{3i} - 2) \\
\pi_{2i|1} &= x_{4i}.
\end{aligned}
$$

where $\Phi_3$ 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 both the Phase 1 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 $E[n_2] \approx 550$. 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. $\pi^*$-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 $\mathbf{x}_i = (x_{1i}, x_{2i})^T$, Dr. Kim, should I modify the simulation so that the regression estimator also includes $g(\pi_{2i|1}^{-1})$ as a covariate?
3. Debiased Calibration with Population Constraints (DC-Pop): This is the method from Kwon et al. (2024) with the true population level constraints, and
4. Debiased Calibration with Estimated Population Constraints (DC-Est): This is the proposed method with the Phase 1 sample being used to estimate the population level constraints.

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 have the following variance estimate[2]:

| Estimator | Variance | Notes |
|---|---|---|
| $\pi^*$ | $N^{-2}\sum_{i\in A_2}\left(\pi_{2i\|1}^{-2}-\pi_{2i\|1}^{-1}\right)y_i^2$ | |
| TP-Reg | $N^{-2}\left(\sum_{i\in A_1}\left(\pi_{1i}^{-2}-\pi_{1i}^{-1}\right)\eta_i^2+\sum_{i\in A_2}\frac{1}{\pi_{1i}\pi_{2i\|1}}(\pi_{2i\|1}^{-1}-1)(y_i-\mathbf{x}_i'\hat{\beta})^2\right)$ | $\eta_i=\mathbf{x}_i\hat{\beta}+\frac{\delta_{2i}}{\pi_{2i\|1}}(y_i-\mathbf{x}_i\hat{\beta})$ |
| DC-Pop | $(Y-\mathbf{z}^T\hat{\gamma})^T\Pi(Y-\mathbf{z}^T\hat{\gamma})$ | $\Pi=\text{diag}(1-(\pi_1\pi_{2\|1})^{-1})\cdot\frac{\hat{w}^2}{N^2}$ |
| DC-Est | $(Y-\mathbf{z}^T\hat{\gamma})^T\Pi(Y-\mathbf{z}^T\hat{\gamma})+\hat{\gamma}_{[1:3]}^T V(\mathbf{x})\hat{\gamma}_{[1:3]}/N^2$ | |

Figure 1: This table gives the formulas for each variance estimator used in this simulation.

We run this 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. The results are in Figure 2.

| Est | Bias | RMSE | EmpCI | Ttest |
|---|---|---|---|---|
| $\pi^*$ | -4.968 | 4.972 | 0.000 | 728.544 |
| TP-Reg | 0.003 | 0.165 | 0.960 | 0.661 |
| DC-Pop | 0.022 | 0.038 | 0.758 | 21.402 |
| DC-Est | 0.022 | 0.166 | 0.957 | 4.246 |

Figure 2: This table shows the results of the simulation study. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: $\pi^*$, TP-Reg, DC-Pop, and DC-Est.

# 3 Topic 2: Non-nested Two-Phase Sampling

( Materials in Section 11.4 can be used here. I have copy-and-pasted the textbook materials below. Please modify them. ) I will do this shortly.

[2]These variance estimates use the fact that we have Poisson sampling for both phases in the simulation.

## 3.1 Background

Now we consider the sampling mechanism known as non-nested two phase sampling (Kim (2024)). 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 $X$,

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

where $\pi_{1i}$ is the probability of $i \in A_1$ and $\pi_{2i} = \Pr(i \in A_2)$. I don't know if it makes sense to change the Phase 2 selection probability as $\pi_{2i|1}$. Should I keep it at $\pi_{2i}$? But if I do, how is the two phase setup useful? We can combine these estimates to get $\hat{X}_c = W\hat{X}_1 + (1-W)\hat{X}_2$ for some matrix $W$ (see Merkouris (2004) for the optimal choice of $W$). We can then define a regression estimator as

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

where

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

Since the samples $A_1$ and $A_2$ are independent,

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

where $\beta_q^*$ 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 where $d_{2i} = \pi_{2i}^{-1}$,

$$\hat{w} = \arg\min_{w} Q(w) = \sum_{i \in A_2} (w_{2i} - d_{2i})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} x_i = \hat{X}_c \tag{5}$$

and $\hat{Y}_{NN,reg} = \sum_{i \in A_2} \hat{w}_{2i} y_i$ where $\hat{w}_{2i}$ is the solution to Equation 5.

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

### 3.3 Theoretical Results

### 3.4 Simulation Studies

## 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.

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.

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

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association 99*(468), 1131–1139.

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.