

## Introduction

In classical two-phase sampling, one first selects a sample  $A_1$  from a finite population  $U$  of size  $N$ , and observes  $(\mathbf{X}_i)_{i=1}^{n_1}$ . Then one selects a sample  $A_2$  from  $A_1$  and observes  $(\mathbf{X}_i, Y_i)_{i=1}^{n_2}$ . We take the framework of two-phase sampling and view data integration as specific case of observing a variable of interest  $Y$  within a single data set while we also observe common covariates  $(\mathbf{X}_i)$  between both the data set with  $Y$  as well as an outside auxiliary sample. We focus on the case:

- Where all of the surveys are probability samples,
- Where we only have summary level information instead of individual observation values,
- When we want the estimate the finite population total of  $Y$ ,

$$Y_T = \sum_{i \in U} y_i.$$

## Notation

- Let  $\pi_{1i}$  be the probability that element  $i$  is selected into the Phase 1 sample,  $A_1$ .
- Let  $\pi_{2i|1}$  be the probability that element  $i$  is selected into the Phase 1 sample,  $A_2$  conditional on the fact that  $i \in A_1$ .
- Let  $d_{1i} = 1/\pi_{1i}$  and  $d_{2i|1} = 1/\pi_{2i|1}$ .
- We use  $\delta_{1i}$  and  $\delta_{2i}$  to indicate if an observation is contained within  $A_1$  and  $A_2$  respectively.

## Goal

We want a two-phase sampling framework to

- Combine information from multiple data sources,
- In a way that is efficient, and
- Approximately design unbiased.

## Comparable Methods

- The **double expansion estimator** of Kott and Stukel (1997) is a Horvitz-Thompson like estimator that is design unbiased, but inefficient:

$$\hat{Y}_{\text{DEE}} = \sum_{i \in A_2} \frac{y_i}{\pi_{1i}\pi_{2i|1}}.$$

- The **two-phase regression estimator** is approximately design unbiased but not as efficient as our proposed method:

$$\hat{Y}_{\text{Reg}} = \sum_{i \in A_1} \frac{\mathbf{x}_i \hat{\beta}_q}{\pi_{1i}} + \sum_{i \in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}} (y_i - \mathbf{x}_i \hat{\beta}_q)$$

where  $q_i = q(\mathbf{x}_i)$  and

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

## Methodology

### Generalized Calibration

- In a seminal paper, Deville and Sarndal (1992) generalized the regression estimator to other loss functions besides squared-error loss for a sample  $A$  with auxiliary information about  $X$ . Their generalized loss function minimizes,

$$\sum_{i \in A} G(w_i, d_i) \text{ such that } \sum_{i \in A} d_i w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i.$$

for a non-negative, strictly convex function with respect to  $w$  function  $G$ , with a minimum at  $g(w_i, d_i) = \frac{\partial G}{\partial w}$  defined on an interval containing  $d_i$  with  $g(w_i, d_i)$  continuous.

### Calibration with Generalized Entropy

- Recently, Kwon, Kim, and Qiu (2024) proposed a calibration estimator that uses a generalized entropy function  $G(w)$  (Gneiting and Raftery (2007)) instead of the generalized loss function  $G(w, d)$  of Deville and Sarndal (1992). They separate the bias calibration from the minimization term and solve the following equation for estimated sample weights:

$$\hat{w}_i = \arg \min_w \sum_{i \in A} G(w_i) \text{ such that } \sum_{i \in A} w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i, \sum_{i \in A} w_i g(d_i) = \sum_{i \in U} g(d_i).$$

- Our proposed method extends their result to two-phase sampling.

## Proposal

Let  $\mathbf{z}_i = (\mathbf{x}_i, g(d_{2i|1}))^T$ , the proposed debiased calibration estimator is

$$\hat{Y}_{\text{DCE}} = \sum_{i \in A_2} d_{1i} \hat{w}_{2i|1} y_i \quad (1)$$

where  $d_{1i} = \pi_{1i}^{-1}$ ,  $d_{2i|1} = \pi_{2i|1}^{-1}$  and

$$\hat{w}_{2i|1} = \arg \min_{w_{2i|1}} \sum_{i \in A_2} d_{1i} G(w_{2i|1}) \text{ such that } \sum_{i \in A_2} d_{1i} w_{2i|1} \mathbf{z}_i = \sum_{i \in A_1} d_{1i} \mathbf{z}_i. \quad (2)$$

## Theoretical Results: Asymptotic Design Consistency

Under some regularity conditions,

$$\hat{Y}_{\text{DCE}} = \hat{Y}_{\ell} + O_p(N/n_2),$$

where

$$\hat{Y}_{\ell} = \hat{Y}_{\text{DEE}} + \left( \sum_{i \in A_1} d_{1i} \mathbf{z}_i - \sum_{i \in A_2} d_{1i} \pi_{2i|1}^{-1} \mathbf{z}_i \right) \phi^*$$

and

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

## Simulation Study

For a finite population of size  $N = 10,000$ , and  $n_1 = 1000$ ,

- $X_{1i} \stackrel{\text{ind}}{\sim} N(2, 1)$ ,  $X_{2i} \stackrel{\text{ind}}{\sim} \text{Unif}(0, 4)$ ,  $Z_i \stackrel{\text{ind}}{\sim} N(0, 1)$ ,  $\varepsilon_i \stackrel{\text{ind}}{\sim} N(0, 1)$
- $Y_i = 3X_{1i} + 2X_{2i} + 0.5Z_i + \varepsilon_i$
- $\pi_{1i} = n_1/N$ ,  $\pi_{2i|1} = \max(\min(\Phi_3(z_i - 1), 0.7), 0.02)$ .

where  $\Phi_3$  is the CDF of a t-distribution with 3 degrees of freedom. We compare the following algorithms:

- Double Expansion Estimator (DEE)
- Two-Phase Regression estimator (TP-Reg)
- Debiased Calibration with Population Constraints (DC-Pop): This solves

$$\arg \min_{w_{2i|1}} \sum_{i \in A_2} d_{1i} G(w_{2i|1}) \text{ such that } \sum_{i \in A_2} d_{1i} w_{2i|1} \mathbf{z}_i = \sum_{i \in U} \mathbf{z}_i.$$

- Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (2).

Est	Bias	RMSE	EmpCI	Ttest
DEE	-0.050	0.793	0.942	1.986
TP-Reg	0.005	0.153	0.947	1.131
DC-Pop	0.002	0.092	0.968	0.677
DC-Est	0.001	0.139	0.951	0.243

Table 1. 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: DEE, TP-Reg, DC-Pop, and DC-Est.

## Extensions

We also consider the following extensions:

- Non-nested two-phase sampling: when  $A_1$  and  $A_2$  are independent.
- Multi-source sampling: when  $Y$  is contained in a sample  $A_0$  that shares common covariates  $\mathbf{X}$ , with samples  $A_1, A_2, \dots, A_M$ .

## References

- Deville, Jean-Claude and Carl-Erik Sarndal (1992). "Calibration estimators in survey sampling". In: *Journal of the American statistical Association* 87(418), pp. 376–382.
- Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102(477), pp. 359–378.
- Kott, P Stukel and DM Stukel (1997). "Can the jackknife be used with a two-phase sample". In: *Survey Methodology* 23(2), pp. 81–89.
- Kwon, Yonghyun, Jae Kwang Kim, and Yumou Qiu (2024). *Debiased calibration estimation using generalized entropy in survey sampling*. arXiv: 2404.01076 [stat.ME].