

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### Correcting Selection Bias in a Non-probability Two-phase Payment Survey

by Heng Chen and John Tsang

Release date: September 8, 2025



# Correcting Selection Bias in a Non-probability Two-phase Payment Survey

Heng Chen and John Tsang<sup>1</sup>

## Abstract

This paper employs the pseudo maximum likelihood (PML) estimator to the non-probability two-phase sampling when relevant auxiliary information is available from both probability survey sample and non-probability survey sample. To accommodate various weight adjustments and estimates variance beyond totals and means such as medians and quantiles, a simplified pseudo-population bootstrap procedure is proposed to approximately estimate the second-phase variance. Specifically, the simplification ignores the second phase sampling variability (i.e., treated as fixed, while in fact it is random), if the first-phase sampling fraction of the non-probability sample is negligible. Using the Bank of Canada 2020 Cash Alternative Survey Wave 2, the performance of this proposed method was compared to alternative methods, which either do not explicitly model the selection probability (i.e., raking) or ignore the valuable information from Phase 1 (i.e., Phase-2-Only). The results show that the PML-based approach performs better than raking and Phase-2-Only estimates in terms of reducing the selection bias for both phases' payment-related variables, especially for the low-response youth group. Estimated variances of the PML-based estimates are stable.

Key Words: Data integration; Non-probability survey sample; Resampling methods; Two-phase sampling; Variance estimation.

## 1. Introduction

In non-probability sampling, a two-phase design allows us to collect auxiliary information that is important to characterize the selection mechanism and the study variable when such information from the sampling frame is limited. Empirical applications and payment surveys under this setup include Henry et al. (2022) and Welte and Wu (2023). Section 2 outlines how to produce weighting schemes for the first and the second phases of a two-phase non-probability survey sample. In Section 3, we propose a pseudo-population bootstrap (PPB) procedure for variance estimation with weighing schemes from Section 2. Section 4 applies methods described in Sections 2 and 3. Specifically, we use the 2020 Cash Alternative Survey Wave 2 (CASW2) from the Bank of Canada to showcase these methods. We also attempt to address the selection bias from the low response rate in the 18 – 34 age group in the CASW2. Compared with the current weighting scheme of the CASW2, our proposal performs better in reducing selection bias. Estimates from our proposed method are stable. Section 5 concludes and discusses future research.

## 2. Weighting Non-probability Two-phase Survey Samples by Data Integration

The Phase-1 sample  $S_{NP,1}$  comes from a finite population  $U = \{1, 2, \dots, N\}$  of size  $N$  ( $S_{NP,1} \subset U$ ), while the Phase-2 sample  $S_{NP,2}$  comes from the first phase ( $S_{NP,2} \subset S_{NP,1}$ ). The dataset of  $S_{NP,1}$  contains auxiliary variables  $\mathbf{x}_k$ , study variables  $\mathbf{y}_{1k}$  and indicator  $i_{2k}$  of whether unit  $k \in S_{NP,1}$  participated in Phase 2:  $\{(\mathbf{x}_k, \mathbf{y}_{1k}, i_{2k}) : k \in S_{NP,1}\}$ . After data collection of Phase 1,  $\mathbf{y}_{1k}$  is available as auxiliary information for Phase 2. Therefore, auxiliary variables available for Phase 2 are  $\mathbf{z}_k = (\mathbf{x}_k, \mathbf{y}_{1k})$ ,  $k \in S_{NP,1}$ . Together with study variables  $\mathbf{y}_{2k}$  from each unit  $k \in S_{NP,2}$ , the dataset of  $S_{NP,2}$  is  $\{(\mathbf{z}_k, \mathbf{y}_{2k}) : k \in S_{NP,2}\}$ . We also have a reference probability sample  $S_p$  at the individual level whose dataset is  $\{(\mathbf{x}_k, d_k) : k \in S_p\}$ , where  $d_k$  is the (design) weight of unit  $k \in S_p$  and  $\mathbf{x}_k$  denotes the set of variables (excluding  $\mathbf{y}_{1k}$ ) that  $S_p$  and  $S_{NP,1}$  have in common.

---

<sup>1</sup>Heng Chen, Currency, Bank of Canada, 234 Wellington St. W., Ottawa, Ontario, Canada, K1A 0G9 (hchen@bank-banque-canada.ca); John Tsang, Department of Mathematics and Statistics, University of Ottawa, 150 Louis-Pasteur Pvt, Ottawa, Ontario, Canada, K1N 6N5 (John.Tsang@uottawa.ca).

In this setup, we use the pseudo-maximum-likelihood (PML) method to estimate the probabilities for each sampled unit  $k$  to be selected into Phase 1 ( $\pi_{1k}$ ) and to be selected into Phase 2 provided that  $k$  has been selected into Phase 1 ( $\pi_{2|1k}$ ). To facilitate the estimation, we use logistic regression models for all probabilities. Then, we use inverse probability weighting (IPW) to obtain weighting systems to estimate finite population quantities, such as means and medians, for Phases 1 and 2.

## 2.1 Weighting Scheme for Phase 1

In Phase 1, we use the PML method to estimate  $\pi_{1k}$  by integrating  $S_{NP,1}$  and  $S_p$ , and then obtain a weighting scheme from IPW (Chen et al., 2020). This scheme is valid under three main assumptions. (1) The selection into Phase 1 and the study variable  $y_1$  are independent conditional on auxiliary variables. (2) Every unit in the population has a chance to be selected. (3) The selection into Phase 1 among sampled units are independent, conditional on their auxiliary variables.  $S_{NP,1}$  and  $S_p$  are also independent. Under these assumptions, if we use a logistic regression to model the selection probabilities, the PML method leads to the following system of estimating equations.

$$\sum_{k \in S_{NP,1}} \mathbf{x}_k^T \boldsymbol{\alpha} = \sum_{k \in S_p} d_k \log[1 + \exp(\mathbf{x}_k^T \boldsymbol{\alpha})]$$

This system of equations matches the first moments of auxiliary variables  $\mathbf{x}_k$  from  $S_{NP,1}$  and  $S_p$ . The solution  $\hat{\boldsymbol{\alpha}}$  to this system leads to the Phase-1 PML weighting scheme:  $\{w_{1k}^{\text{PML}} = \pi_1(\hat{\boldsymbol{\alpha}}; \mathbf{x}_k) = \hat{\pi}_{1k} : k \in S_{NP,1}\}$ .

## 2.2 Weighting Scheme for Phase 2

Each assumption for Phase 2 is analogous to those for Phase 1. Then, the PML for Phase 2 conditional on Phase 1 reduces to the following system of estimating equations and balances the first moments of variables  $\mathbf{z}_k$  from Phases 1 and 2.

$$\sum_{k \in S_{NP,2}} \mathbf{z}_k \boldsymbol{\beta} = \sum_{k \in S_{NP,1}} \log[1 + \exp(\mathbf{z}_k^T \boldsymbol{\beta})],$$

The two-phase PML weighting scheme is  $\{w_{2k}^{\text{PML}} = [\pi_1(\hat{\boldsymbol{\alpha}}; \mathbf{x}_k) \pi_{2|1}(\hat{\boldsymbol{\beta}}; \mathbf{z}_k, \mathbf{i}_1)]^{-1} = (\hat{\pi}_{1k} \hat{\pi}_{2|1k})^{-1} : k \in S_{NP,2}\}$ , where  $\hat{\boldsymbol{\beta}}$  is the solution to this system,  $\mathbf{i}_1$  is a vector of Phase-1 selection indicators and  $\hat{\pi}_{1k}$  comes from Section 2.1.

## 3. Variance Estimation by Pseudo-Population Bootstrap

Generally, a pseudo-population bootstrap (PPB) procedure first uses the weight of each sampled unit to estimate the population. The estimated population is called the pseudo-population. Then, we draw bootstrap samples from the pseudo-population and calculate bootstrap statistics from such samples. Mashreghi et al. (2016) provide an excellent overview of variance estimation under the PPB approach. Our suggested procedure conforms to the same idea. The following describes step-by-step how to use the PPB to approximately estimate variances of estimated quantities  $\hat{\theta}_1$  for Phase 1 and  $\hat{\theta}_2$  for Phase 2.

**Step 1:** Repeat unit  $k \in S_p$   $[d_k]$  times to create the pseudo-population  $U_p^*$ .

**Step 2:** Repeat unit  $k \in S_{NP,1}$   $[w_{1k}^{\text{PML}}] = [\hat{\pi}_{1k}^{-1}]$  times to create a pseudo-population  $U_{NP}^*$  including  $\mathbf{z}_k$  and  $i_{2k}$ .

**Step 3:** Repeat Steps (3a) and (3b)  $B$  times to create  $B^2$  pairs of bootstrap samples. Each pair has a bootstrap sample from each pseudo-population. For  $b = 1, 2, 3, \dots, B$ ,

- **Step 3a:** Draw bootstrap samples  $S_p^{*(b)}$  from  $U_p^*$  according to Poisson sampling with inclusion probabilities  $d_k^{-1}$ .
- **Step 3b:** Draw bootstrap samples  $S_{NP,1}^{*(b)}$  from  $U_{NP}^*$  according to Poisson sampling with inclusion probabilities  $\hat{\pi}_{1k}$ . We do not update the value of  $i_{2k}$  of each selected unit.

**Step 4:**

- For Phase 1, use Phase-1 PML (Section 2.1) to produce a weighting scheme from every pair of bootstrap samples. Then, estimate  $\theta_1$  from all  $B^2$  weighting schemes, and then use the usual variance formula for variance estimation.
- For Phase 2, use two-phase PML (Section 2.2) to produce a weighting scheme from every pair of bootstrap samples. Then, estimate  $\theta_2$  from all  $B^2$  weighting schemes, and then apply the usual variance formula for variance estimation.

The above procedure can serve as a one-stop shop to estimate variance of finite population quantities such as totals, means, medians and other quantiles, while incorporating final weight adjustments such as calibration. Estimated variance from the above procedure is an approximation due to three simplifying assumptions below:

1. In Step 1, we have assumed Poisson sampling for  $S_p$ . If available, we should use bootstrap replicate weights of  $S_p$  instead.
2. As Beaumont et al. (2015) have shown, if we assume that the first-phase sampling fraction ( $|S_{NP,1}|/N$ ) is negligible and that Phase 2 is fixed (while Phase 2 is indeed random), the contribution from the omitted randomization to the total variance is negligible. Therefore, in Step 3, we extend this idea from two-phase probability samples to two-phase non-probability samples. In the context of a pseudo-population bootstrap, variances for Phase 2 can be approximately estimated even if the bootstrap samples from the non-probability samples are based solely on Phase-1 randomization.
3. In Steps 1 and 2, we ignore the variability from the fractional part of weights  $d_k$  and  $w_{1k}^{PML}$  while creating pseudo-populations  $U_p^*$  and  $U_{NP}^*$ . Because, in the second point above, we have assumed that the Phase-1 sampling fraction is negligible, the omitted variability from this simplification is very small.

## 4. Application: The 2020 Cash Alternative Survey Wave 2

The Bank of Canada conducted the 2020 Cash Alternative Survey Wave 2 (CASW2) to collect information about cash holdings and Canadians' daily use of cash. This survey is a two-phase non-probability survey consisting of two phases. The first phase is a survey questionnaire (SQ). The second phase is a three-day Diary Survey Instrument (DSI). The DSI is a strict sub-sample of the SQ. In the current weighting system for the CASW2, each of the SQ and the DSI has its own weighting scheme produced from raking. These two weighting schemes regard each phase as separate survey samples. Readers are referred to Chen et al. (2021) for further details about the CASW2.

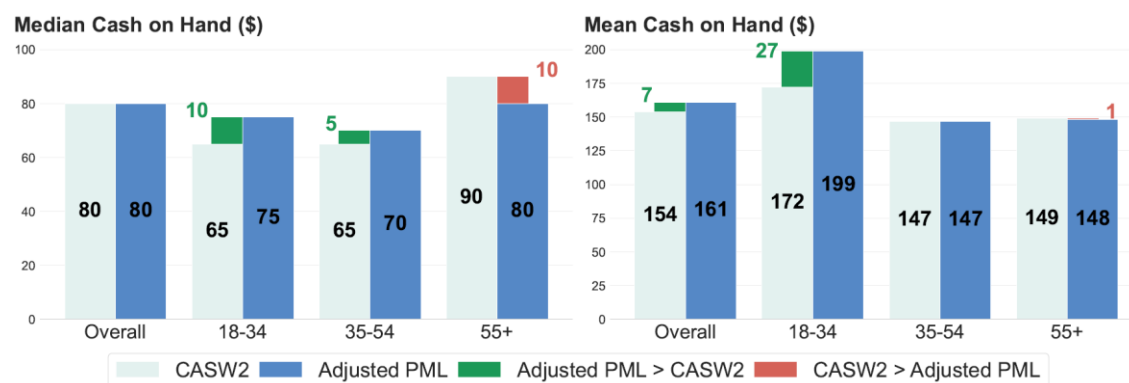
Like many other surveys, the CASW2 also has a low response rate in the younger age group (18 – 34). The participation rate from the SQ into the DSI of the same age group is also low. We use weighting schemes from Section 2 to reduce the consequent selection bias. To estimate Phase-1 selection probabilities with the PML, we choose the Canadian Perspectives Survey Series 5 (CPSS 5) administered by Statistics Canada as the reference probability sample. The CPSS 5 was chosen because the CASW2 and the CPSS 5 are similar in terms of the target population, survey mode, and data collection period. In this manuscript, we focus on estimating the median and the mean cash on hand from the SQ (Phase 1) and the DSI (Phase 2).

## 4.1 Phase 1: The Survey Questionnaire

In application, after estimating weights with Phase-1 PML (Section 2.1), we use raking to calibrate these weights to known population totals from the Census (adjusted PML) for further bias reduction and efficiency improvement. We have validated these adjusted PML weights and the current CASW2 weighting scheme by comparing estimates of eight variables about online shopping behaviour during COVID-19 with those from the CPSS 5 at both the overall and the age-group levels.

Using these estimates from the CPSS 5 as benchmarks, weights from the PML outperform the current CASW2 weighting scheme in terms of average and maximum relative deviation among all age groups, especially the low-response 18 – 34 age group. This larger bias from the current CASW2 weighting scheme has led to a 14-per-cent underestimation in cash on hand for the 18 – 34 age group (Figure 4.1-1). Variation estimated with the PPB (Section 3) shows that overall and age-group level estimates from PML followed by raking are stable (Table 4.1-1).

**Figure 4.1-1**  
**Estimated Cash Holdings by Age Group: PML Followed by Raking vs. Raking for Phase 1 of the CASW2**



Note: Adjusted PML denotes Phase-1 PML followed by raking.

**Table 4.1-1**  
**Coefficients of Variation for Cash Holdings Estimated by PML followed by Raking, the Survey Questionnaire, at the Overall and the Age-group Level, per cent**

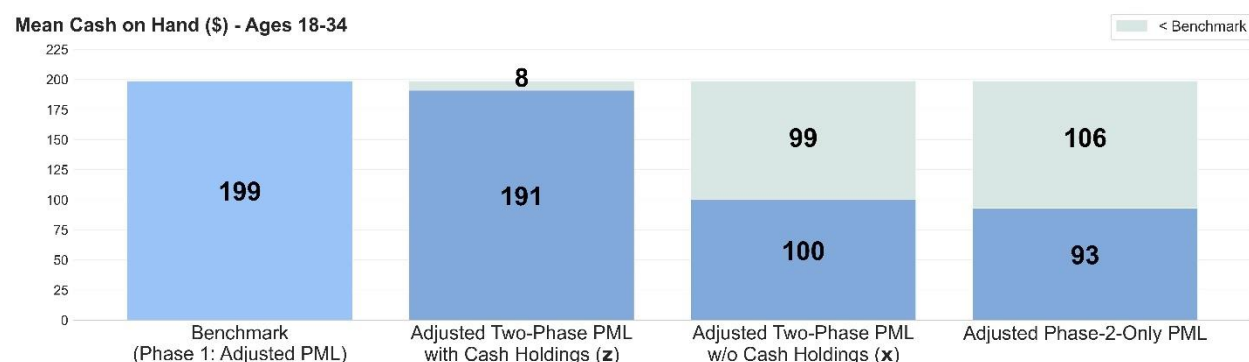
	Mean	Median
<b>Overall</b>	3.92	3.66
<b>18 – 34</b>	8.90	9.03
<b>35 – 54</b>	6.25	7.11
<b>55+</b>	5.14	5.53

## 4.2 Phase 2: The Diary Survey Instrument

In Phase 2, we consider four candidate weighting schemes. Each of them is followed by raking. The first two weighting schemes treat the second phase as standalone from the first phase: (1) the current CASW2 weighting scheme and (2) the weighting scheme from directly applying PML to Phase 2 (Phase-2-Only PML). The other two schemes are based on two-phase PML (Section 2.2). The difference between these remaining schemes is whether they include study variables from Phase 1 to model  $\pi_{2|1k}$  (two-phase PML with or without cash holdings).

To compare these four weighting schemes, we use cash on hand estimated by adjusted PML from Phase 1 as the benchmark. The comparison shows that the two-phase PML with cash holdings outperforms the other three candidates by significant margins (Figure 4.2-1). Therefore, the two-phase PML with study variables from Phase 1 can improve the effort to reduce bias for the second phase. Moreover, the overall and the age-group level estimates from the two-phase PML with cash holdings are stable, as shown by the variances estimated from the PPB (Table 4.2-1).

**Figure 4.2-1**  
**Estimated Cash Holdings by Age Group: Two-Phase PMLs with Different Sets of Variables in the Second-Phase Selection Probability Model and Phase-2-Only PML**



Note: Adjusted weighting schemes denote weighting schemes adjusted by raking.

**Table 4.2-1**  
**Coefficients of Variation for Cash Holdings Estimated by Two-Phase PML (with Cash Holdings) followed by Raking, the Diary Survey Instrument, at the Overall and the Age-group Level, per cent**

	Mean	Median
<b>Overall</b>	8.31	6.81
<b>18 – 34</b>	23.11	14.48
<b>35 – 54</b>	6.72	9.73
<b>55+</b>	5.84	7.33

## 5. Discussion and Next Steps

Based on the Pseudo Maximum Likelihood (PML) approach, we create both Phase-1 and Phase-2 weighting schemes for non-probability two-phase samples. We use the Bank of Canada 2020 Cash Alternative Survey Wave 2 (CASW2) to showcase bias reduction from our proposal and compare our proposal with other alternatives. Three main empirical findings are that: (i) the PML-based weights are better than raked weights (from the CASW2) to reduce the selection bias, especially when the selection probabilities are heterogeneous across different demographics groups; (ii) two-phase PML, which separately models Phase-1 and Phase-2 selection probabilities (conditional on Phase 1), provides less biased estimates than the alternative Phase-2-Only weighting scheme of ignoring the first-phase information; and (iii) it is desirable to include relevant Phase-1 variables into the Phase-2 selection model as shown in Section 4.2. Estimates from Sections 4.1 and 4.2 are stable.

In the future, we plan to collect more para-data, such as response time and duration, during Phase-1 data collection to better characterize respondents' willingness to participate in Phase 2 (Liu and Netzer, 2023). For future research, we are investigating how to apply data integration to other Bank of Canada surveys to overcome unknown or complicated selection probabilities, such as indirect sampling without using the generalized weight-sharing method.

## 6. Acknowledgements

This work is financially supported by the Bank of Canada and Mitacs through the Mitacs Accelerate program. The views expressed are those of the authors and do not necessarily represent the official views of the Bank of Canada. Estimates in this manuscript are preliminary. All remaining errors are solely the responsibility of the authors.

## References

- Beaumont, J. F., Béliveau, A., and Haziza, D. (2015), “Clarifying some aspects of variance estimation in two-phase sampling”, *Journal of Survey Statistics and Methodology*, 3(4), pp. 524-542.
- Chen, H., Engert, W., Felt, M. H., Huynh, K., Nicholls, G., O’Habib, D., and Zhu, J. (2021), “Cash and COVID-19: The impact of the second wave in Canada”, (No. 2021-12), Bank of Canada.
- Chen, Y., Li, P., and Wu, C. (2020), “Doubly robust inference with nonprobability survey samples”, *Journal of the American Statistical Association*, 115(532), pp. 2011-2021.
- Henry, C., Shimoda, M., and Zhu, J. (2022), *2021 Methods-of-Payment Survey Report* (No. 2022-23), Bank of Canada Staff Discussion Paper.
- Liu, S., and Netzer, N. (2023), “Happy times: Measuring happiness using response times”, *American Economic Review*, 113(12), pp. 3289-3322.
- Mashreghi, Z., Haziza, D., and Léger, C. (2016), “A survey of bootstrap methods in finite population sampling”, *Statistics Surveys*, 10, pp.1-52.
- Welte, A., and Wu, J. (2023). *The 2021–22 Merchant Acceptance Survey Pilot Study* (No. 2023-1), Bank of Canada.