BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel—2025-17

Last updated: June 25, 2025

Correcting Selection Bias in a Non-Probability Two-Phase Payment Survey

Heng Chen
Currency Department
Bank of Canada
hchen@bankofcanada.ca

John Tsang
University of Ottawa
John.Tsang@uottawa.ca

DOI: https://doi.org/10.34989/swp-2025-17 | ISSN 1701-9397

Acknowledgements

We would like to thank Andrew Mercer, David Haziza, Kim P. Huynh, Jean-François Beaumont, Marcel Voia and Mehdi Dagdoug for their valuable comments and suggestions. We also appreciate the comments received from the 2024 CIPHER at Washington, D.C., Statistics Canada's 2024 International Methodology Symposium, and the 2024 International Total Survey Error Workshop. This work was supported by Mitacs through the Mitacs Accelerate program.

Abstract

We develop statistical inferences for a non-probability two-phase survey sample when relevant auxiliary information is available from a probability survey sample. To reduce selection bias and gain efficiency, both selection probabilities of Phase 1 and Phase 2 are estimated, and two-phase calibration is implemented. We discuss both analytical plug-in and pseudo-population bootstrap variance estimation methods that account for the effects of using estimated selection probabilities and calibrated weights. The proposed method is assessed by simulation studies and used to analyze a non-probability two phase payment survey.

Topics: Bank notes; Econometric and statistical methods
JEL codes: C, C8, C83

Résumé

Nous élaborons des inférences statistiques pour un échantillon non probabiliste à deux phases quand il est possible d'obtenir des informations auxiliaires pertinentes à partir d'un échantillon probabiliste. Afin de réduire le biais de sélection et de faire des gains d'efficience, nous estimons les probabilités de sélection des échantillons de la première et de la deuxième phases, et faisons un calage à deux phases. Nous examinons deux méthodes d'estimation de la variance qui tiennent compte des effets découlant de l'utilisation de probabilités de sélection estimées et de poids calés : la méthode de substitution analytique ainsi que la méthode d'autoamorçage à partir de pseudo-populations. Nous évaluons la méthode proposée au moyen de simulations et l'utilisons pour analyser un échantillon non probabiliste à deux phases utilisé dans une enquête sur les paiements.

Sujets : Billets de banque ; Méthodes économétriques et statistiques
Codes JEL : C, C8, C83

1 Introduction

A two-phase sampling design is useful when we lack auxiliary information from the original sampling frame to reach the target population effectively. In the first phase, we select a sample from the original frame and collect data on variables related to study variables. Then, we use the extra information collected to build a pseudo-sampling frame. According to this new frame, we collect the second-phase sample from the first-phase sample. In practice, this sampling design helps survey hard-to-reach or rare populations. For instance, Statistics Canada employs a two-phase design in the Aboriginal Peoples Survey (Cloutier and Langlet, 2014) to gather information about the Indigenous population.

A two-phase sampling design can also serve as a conceptual framework for addressing unit nonresponse. Specifically, we can view respondents who finish every survey task as a second-phase sample. Payment surveys under this setup include Henry et al. (2022) and Welte and Wu (2023). Consumers and merchants are asked to complete two survey instruments on two different dates in these applications. Those who complete both instruments constitute the second-phase sample. Since Neyman's seminal work on design-based inference, probability sampling designs have become the standard for most two-phase surveys. Therefore, a vast body of literature explores the theoretical foundations of probability-based two-phase sampling (e.g., Kim and Kim (2007), Beaumont et al. (2015), Kim et al. (2006), Binder et al. (2000), Hidiroglou and Särndal (1998)).

In recent years, however, non-probability sampling has emerged as a convenient and important tool due to its efficient recruitment process, quick responses and low maintenance expenses. It has been used to sample first-phase respondents. Unlike probability sampling, the probabilities of being selected into the first phase are unknown. Poor estimates of these unknown selection probabilities can lead to substantial selection bias. In this paper, we develop statistical inferences for non-probability two-phase survey samples by estimating

selection probabilities of Phase 1 when relevant auxiliary information is available from a probability survey sample. In addition, we integrate the estimation of Phase 2 selection probabilities and two-phase calibration into the construction of Phase 2 weights, which we employ to estimate various finite population parameters, such as totals, means, medians and other quantiles.

It is not enough just to produce (asymptotically) unbiased estimates; it is also important to provide indicators of the quality of those estimates. We discuss both analytical plug-in and pseudo-population bootstrap variance estimation methods that account for the effects of using estimated selection probabilities. In order to flexibly extend to other finite population parameters (i.e., median) and allow for calibration adjustment, we suggest using a pseudo-population bootstrap approach adapted to our non-probability two-phase setup, where we resample indicators of Phase 1 respondents from the pseudo-population but retain indicators of Phase 2 selections from the original sample. This is related to the simplified variance estimator of Beaumont et al. (2015) where Phase 2 selection indicators are treated as fixed.

The literature on estimating unknown selection probabilities from non-probability survey sampling has focused on single-phase samples (Nevo, 2003, Chen et al., 2020, Rao, 2021, Yang and Kim, 2020, Wu, 2022, Elliot, 2009, Wang et al., 2021). So far, researchers have paid less attention to the problem of statistical inferences for two-phase designs. Our paper tries to fill this gap. We use simulation studies to assess our proposed method in terms of biases and variances. We then use our method to analyze a non-probability two-phase payment survey collected by the Bank of Canada during COVID-19 (the November 2020 Cash Alternative Survey), based on auxiliary information from Statistics Canada's probability survey Canadian Perspectives Survey Series 5 (CPSS 5).

The organization of this paper is as follows. Section 2 derives two weighting schemes

for the non-probability two-phase sample: one without calibration, and the other with calibration. Section 3 presents analytical plug-in and pseudo-population bootstrap variance estimators. Section 4 studies the performance of our proposed method via simulation studies. In Section 5, we apply our approach to the non-probability two-phase diary survey. Section 6 concludes. The Appendix provides proofs of theoretical results in this paper.

2 Weighting Non-probability Two-phase Sampling

For two-phase sampling, we first select a Phase 1 sample from a finite population $U = \{1, 2, \ldots, N\}$ of size N and then select a sub-sample, called Phase 2 sample, from Phase 1. Let $\boldsymbol{I}_1 = [I_{11}, I_{12}, \ldots, I_{1N}]$ be the vector of first-phase sample selection indicators such that $I_{1k} = 1$ if k is selected in Phase 1, and $I_{1k} = 0$ otherwise, and let I_{2k} be the Phase 2 selection indicator such that $I_{2k} = 1$ if k is selected in Phase 2, and $I_{2k} = 0$ otherwise. For unit k, the inclusion probability into the Phase 1 sample is $\pi_{1k} := \Pr[I_{1k} = 1]$, and the inclusion probability in Phase 2 conditional on Phase 1 sample is $\pi_{2k}(\boldsymbol{I}_1) := \Pr[I_{2k} = 1 \mid \boldsymbol{I}_1]$.

When the first-phase sample is selected through probability sampling, the probability π_{1k} is known. If $\pi_{2k}(\boldsymbol{I}_1)$ is also known, we can use weights $d_k^* := [\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)]^{-1}$ for each Phase 2 sampled unit k to estimate finite population parameters. In the case of a population total, the weights d_k^* lead to the double expansion (DE) estimator. On the other hand, if $\pi_{2k}(\boldsymbol{I}_1)$ is unknown, we can replace $\pi_{2k}(\boldsymbol{I}_1)$ with an estimate $\widehat{\pi}_{2k}(\boldsymbol{I}_1)$ and create the alternative weight $\widehat{d}_k^* := [\pi_{1k}\widehat{\pi}_{2k}(\boldsymbol{I}_1)]^{-1}$ for each Phase 2 sampled unit k. This is common in the treatment of unit non-response; see, e.g., Kim and Kim (2007). In the case of a population total, the weights $\widehat{d}_k^*$ lead to the so-called *empirical* double expansion (EDE) estimator (Haziza and Beaumont, 2007).

However, under non-probability two-phase sampling, the probability of being selected into

the Phase 1 sample, π_{1k}, is unknown. Since π_{1k} cannot be estimated from the non-probability sample alone, information on the rest of the finite population is required. Suppose there exists an additional probability sample. Then, we can apply the data integration approach from Chen et al. (2020) and Wu (2022) to estimate π_{1k}. Based on the estimated $\widehat{\pi}_{1k}$, we propose two weighting schemes for each Phase 2 sampled unit:

- The first one is a non-probability two-phase weight **without** calibration, that is, $\widehat{w}^*_{2k} := [\widehat{\pi}_{1k}\widehat{\pi}_{2k}(\boldsymbol{I}_1)]^{-1}$.
- The second one is a non-probability two-phase weight **with** calibration, that is, $\widetilde{w}_{2k} = \widehat{w}^*_{2k} g_{1k} g_{2k}$ where g_{1k} and g_{2k} are calibration factors for Phase 1 and Phase 2, respectively.[1]

2.1 Non-probability two-phase weight without calibration

Under the data integration scenario of Chen et al. (2020), a probability sample $S_{\mathrm{P}} \subset U$ is available. For each unit $k \in S_{\mathrm{P}}$, the auxiliary variables $\boldsymbol{x}_{1k}$ and survey weights d_k are observed, but the variables of interest are missing in the probability sample. Thus, the dataset for S_{P} of size n_{P} is $\{(\boldsymbol{x}_{1k}, d_k) : k \in S_{\mathrm{P}}\}$. Now let us consider the non-probability two-phase sample: Phase 1 sample $S_{\mathrm{NP},1}$ of size n_1 is selected from a finite population, and then Phase 2 sample $S_{\mathrm{NP},2}$ of size n_2 is selected from Phase 1 ($S_{\mathrm{NP},2} \subset S_{\mathrm{NP},1}$). For Phase 1, the dataset for $S_{\mathrm{NP},1}$ is $\{(\boldsymbol{x}_{1k}, \boldsymbol{z}_{1k}, \boldsymbol{y}_{1k}, I_{2k}) : k \in S_{\mathrm{NP},1}\}$. Here, $\boldsymbol{x}_{1k}$ is a vector of auxiliary variables that S_{P} and $S_{\mathrm{NP},1}$ share together for estimating Phase 1 selection probabilities, $\boldsymbol{z}_{1k}$ is a vector of auxiliary variables used for Phase 1 calibration, and $\boldsymbol{y}_{1k}$ is a vector of study variables in Phase 1. Note that the probability of being selected into Phase 1, $\pi_{1k} := \Pr[I_{1k} = 1]$, is not observed in $S_{\mathrm{NP},1}$. Next, the dataset for $S_{\mathrm{NP},2}$ is $\{(\boldsymbol{x}_{2k}, \boldsymbol{z}_{2k}, \boldsymbol{y}_{2k}) : k \in S_{\mathrm{NP},2}\}$. Here $\boldsymbol{x}_{2k}$ is a vector of auxiliary variables for estimating Phase 2 selection probabilities, and $\boldsymbol{z}_{2k}$ is

[1]Notice that we follow Hidiroglou and Särndal (1998) and Cohen et al. (2017) to use the superscript "*" to denote overall weights, i.e., weights taking all phases into account, and use the superimposed symbol "$\sim$" to denote calibrated weights.

a vector of auxiliary variables used for Phase 2 calibration, and $\boldsymbol{y}_{2k}$ is a vector of Phase 2 study variables. Notice that both $\boldsymbol{x}_{2k}$ and $\boldsymbol{z}_{2k}$ could contain a subset of variables from $S_{\text{NP},1}$, which could be common between the two.[2]

To estimate the probabilities $\pi_{1k} = \Pr[I_{1k} = 1 \mid \boldsymbol{x}_{1k}, \boldsymbol{y}_{1k}]$ and $\pi_{2k}(\boldsymbol{I}_1) = \Pr[I_{2k} = 1 \mid \boldsymbol{I}_1, \boldsymbol{x}_{2k}, \boldsymbol{y}_{2k}]$, we make the following assumptions:

- **A1:** In each phase, selection indicators and study variables are independent, conditional on auxiliary variables. i.e., $I_{1k} \perp \boldsymbol{y}_{1k} \mid \boldsymbol{x}_{1k}$ and $I_{2k} \perp \boldsymbol{y}_{2k} \mid \boldsymbol{x}_{2k}$.
- **A2:** $\pi_{1k} > 0$ and $\pi_{2k}(\boldsymbol{I}_1) > 0$ for all k.
- **A3:** Selection indicators at each phase are independent. *i.e.*, $I_{1j} \perp I_{1k}$ and $I_{2j} \perp I_{2k}$ for all $j \neq k$.
- **A4:** The probability sample S_{P} and non-probability Phase 1 sample $S_{\text{NP},1}$ are independent.

Under Assumption A1 with both probabilities being the form of logistic regressions, we have

$$\pi_{1k} = \Pr[I_{1k} = 1 \mid \boldsymbol{x}_{1k}, \boldsymbol{y}_{1k}] = \Pr[I_{1k} = 1 \mid \boldsymbol{x}_{1k}] =: \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k}) \text{ and}$$
$$\pi_{2k}(\boldsymbol{I}_1) = \Pr[I_{2k} = 1 \mid \boldsymbol{I}_1, \boldsymbol{x}_{2k}, \boldsymbol{y}_{2k}] = \Pr[I_{2k} = 1 \mid \boldsymbol{I}_1, \boldsymbol{x}_{2k}] =: \pi_2(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k}),$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of regression coefficients of the logistic regressions. Under Assumptions A2, A3 and A4, we construct the joint pseudo-likelihood function for both non-probability samples $S_{\text{NP},1}$ and $S_{\text{NP},2}$, with the help of the probability sample S_{P} to replace

[2]Notationally, in this paper, we use the letter x to denote auxiliary information for estimating selection probabilities. The letter z represents auxiliary variables used for calibration, and the letter y is used for study variables.

unknown quantities:

$$\left\{ \sum_{k \in S_{\text{NP},1}} \log \left[\frac{\pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k})}{1 - \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k})} \right] + \sum_{k \in S_{\text{P}}} d_k \log \left[1 - \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k}) \right] \right\} +$$
$$\left\{ \sum_{k \in S_{\text{NP},1}} I_{2k} \log \left[\pi_2(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k}) \right] + \sum_{k \in S_{\text{NP},1}} (1 - I_{2k}) \log \left[1 - \pi_2(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k}) \right] \right\}. \tag{1}$$

In (1), the first line corresponds to $\pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k})$, while the second line corresponds to $\pi_2(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k})$. Note that the second term in the first line follows Chen et al. (2020) where we use the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) $\sum_{k \in S_{\text{P}}} d_k \log \left[1 - \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k}) \right]$ from the probability sample S_{P} to replace the unknown population quantity $\sum_{k \in U} \log \left[1 - \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k}) \right]$. The second line in (1) is related to Kim and Kim (2007) where they estimate the response probabilities of Phase 2 conditional on Phase 1 sample. The score equations from (1) are

$$\begin{cases} \displaystyle\sum_{k \in S_{\text{NP},1}} \boldsymbol{x}_{1k}^{\intercal} = \sum_{k \in S_{\text{P}}} d_k \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k}) \boldsymbol{x}_{1k}^{\intercal} \\ \displaystyle\sum_{k \in S_{\text{NP},2}} \boldsymbol{x}_{2k}^{\intercal} = \sum_{k \in S_{\text{NP},1}} \pi_{2k}(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k}) \boldsymbol{x}_{2k}^{\intercal}. \end{cases} \tag{2}$$

Solutions $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ to (2) lead to $\widehat{\pi}_{1k} := \pi_1(\widehat{\boldsymbol{\alpha}}; \boldsymbol{x}_{1k})$ and $\widehat{\pi}_{2k}(\boldsymbol{I}_1) := \pi_2(\widehat{\boldsymbol{\beta}}; \boldsymbol{I}_1, \boldsymbol{x}_{2k})$. Under inverse probability weighting, weights for non-probability Phase 1 sample are $\{\widehat{w}_{1k} := \widehat{\pi}_{1k}^{-1} : k \in S_{\text{NP},1}\}$. Furthermore, weights for non-probability Phase 2 sample are $\{\widehat{w}_{2k}^{*} := [\widehat{\pi}_{1k} \widehat{\pi}_{2k}(\boldsymbol{I}_1)]^{-1} = \widehat{w}_{1k} \widehat{w}_{2k} : k \in S_{\text{NP},2}\}$ where $\widehat{w}_{2k} := \widehat{\pi}_2(\boldsymbol{I}_1)^{-1}$.

2.2 Non-probability two-phase weight with calibration

This section discusses calibration in the context of the non-probability two-phase sampling design. The advantages of calibrations are two-fold (Hidiroglou and Särndal, 1998): one is to ensure external and internal consistency, and the other is to improve the efficiency when there is a strong correlation between calibration variables and variables of interest. Building

upon $\widehat{w}_{1k}$ and $\widehat{w}^*_{2k}$, we employ calibrations to both Phase 1 and Phase 2, respectively.

(a) First-phase Calibration The first calibration uses Phase 1 weights $\{\widehat{w}_{1k} := \widehat{\pi}_{1k}^{-1} : k \in S_{\mathrm{NP},1}\}$ as initial weights and adjusts it to match known population totals $\sum_{k\in U} \boldsymbol{z}_{1k}$. We denote the calibrated weight as $\{\widetilde{w}_{1k} := \widehat{w}_{1k} g_{1k} : k \in S_{\mathrm{NP},1}\}$. The calibration factor is $g_{1k} = \exp\left(\boldsymbol{z}_{1k}^{\intercal}\widehat{\boldsymbol{\lambda}}_1\right)$, where the Lagrange multiplier $\widehat{\lambda}_1$ comes from the following minimization problem:

$$\min \sum_{k\in S_{\mathrm{NP},1}} \widehat{w}_{1k} G(\widetilde{w}_{1k}/\widehat{w}_{1k}) \text{ subject to } \sum_{k\in S_{\mathrm{NP},1}} \widetilde{w}_{1k}\boldsymbol{z}_{1k} = \sum_{k\in U} \boldsymbol{z}_{1k}, \tag{3}$$

where the function $G(\widetilde{w}/\widehat{w}) = (\widetilde{w}/\widehat{w})\log(\widetilde{w}/\widehat{w}) - (\widetilde{w}/\widehat{w}) + 1$ is the Kullback-Leibler information distance.

(b) Second-phase Calibration The initial weights for the second-phase calibration are $\{\widetilde{w}_{1k}\widehat{w}_{2k} : k \in S_{\mathrm{NP},2}\}$, and we adjust them to match totals $\sum_{k\in S_{\mathrm{NP},1}} \widetilde{w}_{1k}\boldsymbol{z}_{2k}$ available up to Phase 1 when auxiliary information is available at the population level or at both the population and the first-phase levels. Thus the calibrated weights for the Phase 2 sample are denoted by $\widetilde{w}_2 := \{\widetilde{w}_{2k} = \widetilde{w}_{1k}\widehat{w}_{2k}g_{2k} : k \in S_{\mathrm{NP},2}\}$. The calibration factor is $g_{2k} = \exp\left(\boldsymbol{z}_{2k}^{\intercal}\widehat{\boldsymbol{\lambda}}_2\right)$, where the Lagrange multiplier $\widehat{\lambda}_2$ comes from the following minimization problem:

$$\min \sum_{k\in S_{\mathrm{NP},2}} [\widetilde{w}_{1k}\widehat{w}_{2k}] G(\widetilde{w}_{2k}/[\widetilde{w}_{1k}\widehat{w}_{2k}]) \text{ subject to } \sum_{k\in S_{\mathrm{NP},2}} \widetilde{w}_{2k}\boldsymbol{z}_{2k} = \sum_{k\in S_{\mathrm{NP},1}} \widetilde{w}_{1k}\boldsymbol{z}_{2k}. \tag{4}$$

2.3 Weighted Finite Population Estimates for Phase 2

From the above two weighting systems, we can estimate $t_2 := \sum_{k\in \mathrm{U}} y_{2k}$, the total of a (scalar) Phase 2 variable of interest y_2, respectively, by

$$\widehat{t}^*_2 = \sum_{k\in S_{\mathrm{NP},2}} \widehat{w}^*_{2k} y_{2k} \text{ and } \widetilde{t}_2 = \sum_{k\in S_{\mathrm{NP},2}} \widetilde{w}_{2k} y_{2k}. \tag{5}$$

The Hájek estimators (Hájek, 1971) for the population mean of y_2 are

$$\widehat{\mu}_2^* = \frac{\widehat{t}_2^*}{\sum_{k \in S_{\text{NP},2}} \widehat{w}_{2k}^*} \text{ and } \widetilde{\mu}_2 = \frac{\widetilde{t}_2}{\sum_{k \in S_{\text{NP},2}} \widetilde{w}_{2k}} \tag{6}$$

Estimated median for y_2 can be computed as

$$\widehat{m}_2^* = \left\{ y_{2k}, k \in \dot{S} : \sum_{i=1}^{k-1} \dot{w}_i^* \leq \frac{1}{2} \text{ and } \sum_{i=k+1}^{n} \dot{w}_i^* \leq \frac{1}{2} \right\} \text{ and } \widetilde{m}_2 = \left\{ y_{2k}, k \in \dot{S} : \sum_{i=1}^{k-1} \dot{\widetilde{w}}_i \leq \frac{1}{2} \text{ and } \sum_{i=k+1}^{n} \dot{\widetilde{w}}_i \leq \frac{1}{2} \right\}, \tag{7}$$

where $\dot{S}$ is $S_{\text{NP},2}$ in a non-decreasing order of y_2, $\dot{w}_i^* := \widehat{w}_{2i}^* / \sum_{j \in \dot{S}} \widehat{w}_{2j}^*$ for $i \in \dot{S}$ and $\dot{\widetilde{w}}_i := \widetilde{w}_{2i} / \sum_{j \in \dot{S}} \widetilde{w}_{2j}$ for $i \in \dot{S}$. Similarly, we can estimate other functionals of the above estimators, such as a ratio of totals.

Remark on consistency: The consistency of the weighted total and mean are derived by accounting for unknown $\widehat{\pi}_{1k}$ and $\widehat{\pi}_{2k}(\boldsymbol{I}_1)$ via the linearization. For example, Theorem 1 in Section 3 establishes the consistency of $\widehat{t}_2^*$ for the population total of Phase 2. The proof of the design consistency of the weighted median is derived in a similar fashion as Proposition 2 of Huber (2014) by using the quantile implicit function.

3 Variance Estimation

We now discuss variance estimation for the six weighted finite population estimates described in Section 2.3. This includes two weighting schemes—one with calibration and one without—combined with three parameters of interest: total, mean and median. We focus on the asymptotic variance of $\widehat{t}_2^*$, the estimated total of (non-probability) Phase 2, based on non-calibrated weights $\widehat{w}_{2k}^*$. In total, we examine three different variance estimators for $\text{Var}[\widehat{t}_2^*]$: The first is a plug-in estimator that utilizes the joint randomization of $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$. The second is another plug-in estimator, but it treats $\boldsymbol{I}_2$ as fixed despite being a random vector. Lastly,

the third estimator employs the pseudo-population bootstrap, also treating $\boldsymbol{I}_2$ as fixed.[3]

We further discuss the choice among three variance estimators and recommend the third option: a pseudo-population bootstrap (PPB) approach. In this approach, we treat $\boldsymbol{I}_2$ as fixed, even though it is random. The primary reason for favouring PPB over these plug-in methods is that this resampling-based technique is easier to extend to different finite population parameters, such as total, mean, median and some other complex statistics. Additionally, it allows for the inclusion of other weight adjustments, like calibration discussed in Section 2.2. In contrast, the plug-in options require the derivation of first-order Taylor expansions, which makes it more involved. As for why we treat $\boldsymbol{I}_2$ as fixed, the motivation comes from the work of Beaumont et al. (2015). In the probability two-phase setup, they propose a simplified variance estimator. Their approach aims to avoid the need for specialized software for two-phase sampling and the joint inclusion probabilities of the second phase. We adapt this idea to our framework of non-probability two-phase sampling, where the selection probabilities for both Phase 1 and Phase 2 (conditional) are unknown.

3.1 Asymptotic Variance of $\widehat{t}_2^*$ and Two Plug-in Estimators

We follow the asymptotic framework from Chen et al. (2020) and Kim and Kim (2007) to derive the asymptotic variance of the estimator $\widehat{t}_2^* = \sum_{k \in S_{\mathrm{NP},2}} \widehat{w}_{2k}^* y_{2k}$.

Theorem 1. Under Assumptions A1 – A4 and regularity conditions C1 – C4 in the Appendix, if π_{1k} and $\pi_{2k}(\boldsymbol{I}_1)$ assume the form of logistic regressions, we have $\widehat{t}_2^* = t_2 +$

[3] In addition to the joint randomization of $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, we also take the sampling design of the probability sample S_{P} into account.

$O_p(Nn_1^{-1/2})$, $\mathrm{Var}[\widehat{t}_2^*] = \mathrm{Var}^{\mathrm{L}}[\widehat{t}_2^*] + o(N^2 n_1^{-1})$ with

$$\begin{aligned}\mathrm{Var}^{\mathrm{L}}[\widehat{t}_2^*] =& \sum_{k\in U}(1-\pi_{1k})\pi_{1k}\left(\frac{y_{2k}}{\pi_{1k}} - \boldsymbol{b}_1^\intercal \boldsymbol{x}_{1k}\right)^2 + \boldsymbol{b}_1^\intercal \boldsymbol{D}\boldsymbol{b}_1, \\ &+ \sum_{k\in U}(1-\pi_{2k}(\boldsymbol{I}_1))\pi_{2k}(\boldsymbol{I}_1)\pi_{1k}\left(\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} - \boldsymbol{c}_2^\intercal \boldsymbol{x}_{2k}\right)^2, \end{aligned} \tag{8}$$

where $\boldsymbol{b}_1^\intercal = \left[N^{-1}\sum_{i\in U}(1-\pi_{1i})y_{2i}\boldsymbol{x}_{1i}^\intercal\right]\left[N^{-1}\sum_{i\in U}\pi_{1i}(1-\pi_{1i})\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^\intercal\right]^{-1}$ and $\boldsymbol{c}_2^\intercal = \left[N^{-1}\sum_{i\in U}(1-\pi_{2i}(\boldsymbol{I}_1))\right.$ $\left.y_{2i}\boldsymbol{x}_{2i}^\intercal\right]\left[N^{-1}\sum_{i\in U}\pi_{1i}\pi_{2i}(\boldsymbol{I}_1)(1-\pi_{2i}(\boldsymbol{I}_1))\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}^\intercal\right]^{-1}$, and $\boldsymbol{D} = \mathrm{Var}_p\left[\sum_{i\in \mathrm{S_P}} d_i\pi_{1i}\boldsymbol{x}_{1i}\right]$ where Var_p denotes the variance under the sampling design of S_{P}.

Remark on Theorem 1: The first line in (8) is identical to Equation (9) in Chen et al. (2020). The second line in (8) captures the variability arising from the non-probability two-phase sample. Based on (8), we obtain a plug-in variance estimator:

$$\begin{aligned}\widehat{\mathrm{Var}}^{L}[\widehat{t}_2^*] =& \sum_{k\in S_{\mathrm{NP},2}}\frac{(1-\widehat{\pi}_{1k})}{\widehat{\pi}_{2k}(\boldsymbol{I}_1)}\left(\frac{y_{2k}}{\widehat{\pi}_{1k}} - \widehat{\boldsymbol{b}}_1^\intercal \boldsymbol{x}_{1k}\right)^2 + \widehat{\boldsymbol{b}}_1^\intercal \widehat{\boldsymbol{D}}\widehat{\boldsymbol{b}}_1 \\ &+ \sum_{k\in S_{\mathrm{NP},2}}(1-\widehat{\pi}_{2k}(\boldsymbol{I}_1))\left(\frac{y_{2k}}{\widehat{\pi}_{1k}\widehat{\pi}_{2|1k}} - \widehat{\boldsymbol{c}}_2^\intercal \boldsymbol{x}_{2k}\right)^2, \end{aligned} \tag{9}$$

where $\widehat{\boldsymbol{b}}_1^\intercal = \left[\sum_{i\in S_{\mathrm{NP},2}}\frac{(1-\widehat{\pi}_{1i})}{\widehat{\pi}_{1i}\widehat{\pi}_{2i}(\boldsymbol{I}_1)}y_{2i}\boldsymbol{x}_{1i}^\intercal\right]\left[\sum_{i\in S_{\mathrm{P}}} d_i\widehat{\pi}_{1i}(1-\widehat{\pi}_{1i})\boldsymbol{x}_{1i}^\intercal\boldsymbol{x}_{1i}\right]^{-1}$, $\widehat{\boldsymbol{D}} = \widehat{\mathrm{Var}}_p\left[\sum_{i\in \mathrm{S_P}} d_i\widehat{\pi}_{1i}\boldsymbol{x}_{1i}\right]$ and $\widehat{\boldsymbol{c}}_2^\intercal = \left[\sum_{i\in S_{\mathrm{NP},1}}\frac{1-\widehat{\pi}_{2i}(\boldsymbol{I}_1)}{\widehat{\pi}_{1i}\widehat{\pi}_{2i}(\boldsymbol{I}_1)}y_{2i}\boldsymbol{x}_{2i}^\intercal\right]\left[\sum_{i\in S_{\mathrm{NP},2}}(1-\widehat{\pi}_{2i}(\boldsymbol{I}_1))\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}^\intercal\right]^{-1}$.

As Beaumont et al. (2015) discuss, the plug-in variance estimator suffers from two drawbacks: its computation requires specialized software designed for two-phase sampling, and the plug-in depends on the second-phase joint inclusion probabilities, which may be difficult to obtain. To overcome these drawbacks, they suggest the simplified plug-in variance estimator by treating $\boldsymbol{I}_2$ as fixed when, in fact, $\boldsymbol{I}_2$ is random and construct the usual design-biased variance estimator for the single-phase sampling design. Here, we adopt the framework of Beaumont et al. (2015) from their probability two-phase to our non-probability two-phase

survey sample, and we impose Assumption A5 where the first-phase sampling fraction is negligible.[4]

- **A5**: The Phase 1 sampling fraction n_1/N is negligible.

Proposition 2. Under Assumptions A1 – A4 and regularity conditions C1 – C4 in the Appendix, and if π_{1k} and $\pi_{2k}(\boldsymbol{I}_1)$ assume the form of logistic regressions.
(i) We have the Taylor linearization of

$$\widehat{t}_2^* = \left[\sum_{k \in S_{\mathrm{NP},1}} \frac{1}{\pi_{1k}} a_k + \boldsymbol{b}_1^\intercal \sum_{k \in S_{\mathrm{P}}} d_k \pi_{1k} \boldsymbol{x}_{1k} \right] + o_p(N n_1^{-1/2}), \tag{10}$$

where $a_k := \dfrac{y_{2k} I_{2k}}{\pi_{2k}(\boldsymbol{I}_1)} - \pi_{1k} \boldsymbol{b_1^T x_{1k}} - \pi_{1k} I_{2k} \boldsymbol{c_2^T x}_{2k} + \pi_{1k} \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{c_2^T x_{2k}}$.

(ii) With the additional assumption A5, $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$ is an alternative estimator of $\mathrm{Var}[\widehat{t}_2^*]$:

$$\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*] = \sum_{k \in S_{\mathrm{NP},1}} \frac{1 - \widehat{\pi}_{1k}}{\widehat{\pi}_{1k}^2} \widehat{a}_k^2 + \widehat{\boldsymbol{b}}_1^\intercal \widehat{\boldsymbol{D}} \widehat{\boldsymbol{b}}_{\boldsymbol{1}}, \tag{11}$$

where

$$\widehat{a}_k := \frac{y_{2k} I_{2k}}{\widehat{\pi}_{2k}(\boldsymbol{I}_1)} - \widehat{\pi}_{1k} \widehat{\boldsymbol{b}}_{\boldsymbol{1}}^\intercal \boldsymbol{x_{1k}} - \widehat{\pi}_{1k} I_{2k} \widehat{\boldsymbol{c}}_{\boldsymbol{2}}^\intercal \boldsymbol{x_{2k}} + \widehat{\pi}_{1k} \widehat{\pi}_{2k}(\boldsymbol{I}_1) \widehat{\boldsymbol{c}}_{\boldsymbol{2}}^\intercal \boldsymbol{x_{2k}}. \tag{12}$$

Remark on Part (i) of Proposition 2: Note that the first sum of our linearized version in $\widehat{t}_2^*$ is over $S_{\mathrm{NP},1}$ instead of over $S_{\mathrm{NP},2}$, which helps construct a simplified variance estimator for single-phase sampling design. Moreover, let us compare the linearization form of our Proposition 2 to Theorem 1 in Kim and Kim (2007). Rearranging the right-hand side of (10), we have

[4]Similar to Beaumont et al. (2015), we do not require the two-phase sampling design to be invariant (Beaumont and Haziza, 2016).

$$\widehat{t}_2^* = \sum_{k \in S_{\mathrm{NP},1}} \frac{1}{\pi_{1k}} \left[\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{c}_{\boldsymbol{2}}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{2k}} + \frac{I_{2k}}{\pi_{2k}(\boldsymbol{I}_1)} \left(y_{2k} - \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{c}_{\boldsymbol{2}}^{\boldsymbol{T}}\boldsymbol{x}_{2k} \right) \right]$$
$$+ \boldsymbol{b}_1^{\intercal} \left[\sum_{k \in S_{\mathrm{P}}} d_k(\pi_{1k}\boldsymbol{x}_{1k}) - \sum_{k \in S_{\mathrm{NP},1}} \frac{1}{\pi_{1k}}(\pi_{1k}\boldsymbol{x}_{1k}) \right] + o_p(Nn_1^{-1/2}). \tag{13}$$

In (13), the first line matches Theorem 1 in Kim and Kim (2007). In their study, they base the linearization on a probability two-phase sample in which the inclusion probabilities for Phase 1 are known, but those for Phase 2 are unknown. In contrast, all Phase 1 selection probabilities are unknown in our non-probability two-phase framework. Consequently, the second line introduces an additional term that results from employing the data integration approach to estimating the unknown Phase 1 selection probabilities by combining both $S_{\mathrm{NP},1}$ and S_{P}.

Remark on Part (ii) of Proposition 2: When the second phase is Poisson sampling (Assumptions A1–A4) and the first-phase sampling fraction is negligible (Assumption A5), our simplified plug-in variance estimator $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$ provides a good approximation of the total variance of $\widehat{t}_2^*$ (still design-biased, though). Because non-response can be viewed as a special second phase, our $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$ is also useful under the single-phase non-probability survey when treating unit non-response. Furthermore, we can compare (11) to Equation (7) in Beaumont et al. (2015). Under Poisson sampling at the second phase, Equation (7) in Beaumont et al. (2015) reduces to $\sum_{k \in S_1} \frac{1-\pi_{1k}}{\pi_{1k}^2} \left[\frac{y_{2k}I_{2k}}{\pi_{2k}(\boldsymbol{I}_1)} \right]$. As shown, the main differences are that: (i) our $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$ is using estimated $\widehat{\pi}_{1k}$ and $\widehat{\pi}_{2k}(\boldsymbol{I}_1)$ rather than π_{1k} and $\pi_{2k}(\boldsymbol{I}_1)$ in Beaumont et al. (2015); (ii) our $\widehat{a}_k$ in (12) has extra terms $-\widehat{\pi}_{1k}\widehat{\boldsymbol{b}}_{\boldsymbol{1}}^{\intercal}\boldsymbol{x}_{\boldsymbol{1k}} - \widehat{\pi}_{1k}I_{2k}\widehat{\boldsymbol{c}}_{\boldsymbol{2}}^{\intercal}\boldsymbol{x}_{\boldsymbol{2k}} + \widehat{\pi}_{1k}\widehat{\pi}_{2k}(\boldsymbol{I}_1)\widehat{\boldsymbol{c}}_{\boldsymbol{2}}^{\intercal}\boldsymbol{x}_{\boldsymbol{2k}}$ arising from the first-order Taylor-expansions of $\widehat{\pi}_{1k}$ and $\widehat{\pi}_{2k}(\boldsymbol{I}_1)$; (iii) in the end, our $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$ has an extra term $\widehat{\boldsymbol{b}}_1^{\intercal}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{b}}_1$ from the data-integration approach to estimate π_1 (Chen et al., 2020), which borrows the strength from the probability sample S_{P} to facilitate the estimation.

3.2 Pseudo-Population Bootstrap for Estimating $\text{Var}[\widehat{t}_2^*]$

Although the plug-in variance estimator $\widehat{\text{Var}}^{\text{Alt}}[\widehat{t}_2^*]$ simplifies variance estimation by fixing $\boldsymbol{I}_2$, its underlying linearization lacks flexibility. If we have other finite population parameters, such as the median and other functionals of totals, or if we apply additional weight adjustments, we need to derive the variance estimator again. As such, we require new computation procedures in these cases. To address this concern, we propose a pseudo-population bootstrap (PPB) approach as a third option. Our proposed method is a resampling approach that follows the spirit of $\widehat{\text{Var}}^{\text{Alt}}[\widehat{t}_2^*]$ to treat $\boldsymbol{I}_2$ as fixed. We incorporate this idea into the creation of bootstrap non-probability Phase 2 samples.[5] To illustrate the procedure for the PPB, we focus on estimating the variance of $\widehat{t}_2^*$.

The pseudo-population bootstrap (PPB) allows the first two moments of the HT-estimator to be consistently estimated for sampling designs with large entropy, such as Poisson sampling. A typical PPB approach to variance estimation involves a few key steps (Mashreghi et al., 2016). First, we use each sampled unit's (non-negative) weight to create pseudo-populations. Next, according to the original sampling design, we draw B bootstrap samples from the pseudo-population. Then, we repeat all estimation steps for each bootstrap sample, including the estimation of selection probabilities and calibration, to compute the bootstrap estimate. In the end, we use all B bootstrap estimates to compute the estimated variance. Our proposed variance estimation procedure for Phase 2 follows this general framework. The following is a high-level overview of our approach.

[5]The idea of not bootstrapping Phase 2 selection in our PPB is similar to the multi-stage resampling-based variance estimation from Bessonneau et al. (2021), where indicators of the second-stage sampling and non-response step after the first stage remain fixed in the with-replacement bootstrap. This idea is also analogous to Kim et al. (2006)'s probability two-phase replicated variance estimation, where indicators of second-phase sampling are treated as fixed in their jackknife approach. For the case of the probability two-stage sampling, Beaumont et al. (2015) show that treating $\boldsymbol{I}_2$ fixed (when, in fact, it is random) will lead to a simplified plug-in variance estimator.

High-Level Overview of the Pseudo-Population Bootstrap for Estimating Var$[\widehat{t}_2^*]$

Step 1: Create a pseudo-population U_{P} from S_{P} and another pseudo-population U_{NP} from $S_{\mathrm{NP},1}$

Step 2: Draw B bootstrap samples from U_{P} and then draw another B bootstrap samples from U_{NP}. Pair up each bootstrap sample from these two sets to form a set $\Omega = \{\Omega_1, \ldots, \Omega_C\}$ of size $C = B^2$, where the c^{th} pair in Ω is $\Omega_c = \left(S_{\mathrm{P}}^{(c)}, S_{\mathrm{NP},1}^{(c)}\right)$, and $S_{\mathrm{P}}^{(c)}$ and $S_{\mathrm{NP},1}^{(c)}$ are the bootstrap samples from U_{P} and U_{NP}, respectively. Construct $S_{\mathrm{NP},2}^{(c)}$ from units whose $I_{2k} = 1$ in $S_{\mathrm{NP},1}^{(c)}$.

Step 3: For each $c = 1, 2, \ldots, C$, calculate the bootstrap two-phase weights $\widehat{w}_2^{*(c)}$ and bootstrap estimate $\widehat{t}_2^{*(c)}$ from $\left(S_{\mathrm{P}}^{(c)}, S_{\mathrm{NP},1}^{(c)}, S_{\mathrm{NP},2}^{(c)}\right)$.

Step 4: Apply the simulation-based variance estimator based on Step 3.

The remainder of this section discusses each step in detail. Steps 1 and 2 reproduce the sampling design. Step 3 reproduces the estimation steps (estimation of selection probabilities and calibration).

Step 1: Create Pseudo-populations U_{P} and U_{NP}

The weight is the number of duplications the sampled unit recreates in the pseudo-population.

- For each unit $k \in S_{\mathrm{P}}$, we replicate $\lfloor d_k \rfloor$ times for the pseudo-population $U_{\mathrm{P}} = \{1, 2, \ldots, N_{\mathrm{P}}\}$ where $N_{\mathrm{P}} = \sum_{k \in S_{\mathrm{P}}} \lfloor d_k \rfloor$.
- For each unit $k \in S_{\mathrm{NP},1}$, we replicate $\lfloor \widehat{w}_{1k} \rfloor$ times for the pseudo-population $U_{\mathrm{NP}} = \{1, 2, \ldots, N_{\mathrm{NP}}\}$ where $N_{\mathrm{NP}} = \sum_{k \in S_{\mathrm{NP},1}} \lfloor \widehat{w}_{1k} \rfloor$.

If both sampling fractions of $S_{\mathrm{NP},1}$ and S_{P} are negligible, we can ignore fractional parts of weights d_k and $\widehat{w}_{1k}$ when creating U_{P} and U_{NP}. Otherwise, we could complete the "fixed" part of the pseudo-populations following Chen et al. (2019) and account for bootstrap randomness

induced by completing these pseudo-populations via Monte Carlo approximation.

Step 2: Draw Bootstrap Samples from U_{P} and U_{NP}

This step creates bootstrap probability samples and bootstrap non-probability Phase 1 and Phase 2 samples.

- **Creation of bootstrap probability sample**: From U_{P}, we draw B bootstrap samples according to the original sampling design of S_{P}. If the sampling design of S_{P} is unavailable, we assume Poisson sampling (Beaumont and Patak, 2012). Under Poisson sampling, each unit $k \in U_{\mathrm{P}}$ is independently selected into a bootstrap sample with inclusion probability d_k^{-1}.

- **Creation of bootstrap non-probability Phase 1 sample**: From U_{NP}, we draw B bootstrap samples by Poisson sampling with inclusion probability of each unit $k \in U_{\mathrm{NP}}$ being $\widehat{\pi}_{1k}$.

- **Creation of bootstrap non-probability Phase 2 sample**: From each $S_{\mathrm{NP},1}^{(c)}$, we construct $S_{\mathrm{NP},2}^{(c)}$ from $k \in S_{\mathrm{NP},1}^{(c)}$ whose $I_{2k} = 1$. Notice that we create a bootstrap non-probability Phase 2 sample by retaining the respondents of $S_{\mathrm{NP},1}^{(c)}$ with $I_{2k} = 1$, instead of resampling Phase 2 respondents from the bootstrap Phase 1 sample. This is analogous to treating $\boldsymbol{I}_2$ fixed in the plug-in estimator $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$. Unlike the vector of first-phase sample selection indicators $\boldsymbol{I}_1$ being bootstrapped, every second-phase sample selection indicator I_{2k} remains fixed in the PPB process.

Then, we pair up bootstrap samples from U_{P} and those from U_{NP} into the set $\Omega := \{\Omega_1, \ldots, \Omega_C\}$ of size $C = B^2$ and $\Omega_c = \left(S_{\mathrm{P}}^{(c)}, S_{\mathrm{NP},1}^{(c)}\right)$. In the end, we append $S_{\mathrm{NP},2}^{(c)}$ to Ω_c to obtain the bootstrap samples $\left(S_{\mathrm{P}}^{(c)}, S_{\mathrm{NP},1}^{(c)}, S_{\mathrm{NP},2}^{(c)}\right)$. For clarity, Steps 1 and 2 are illustrated in the following example.

Illustrative Example of Steps 1 and 2: Suppose the non-probability sample and the probability sample are

Non-probability Sample

ID	$\widehat{w}_1$	$\widehat{\pi}_1 = 1/\widehat{w}_1$	y_1	y_2	x_1	x_2	I_2
1	2.4	1/2.4	20	-	4	-	0
2	3.6	1/3.6	14	20	7	3	1

Probability Sample

ID	d	$\pi = 1/d$	x_1
1	1.7	1/1.7	5
2	2.3	1/2.3	6

To create U_{NP} from the non-probability sample, we repeat the row with ID = 1 twice ($\lfloor 2.4 \rfloor = 2$) and the row with ID = 2 three times ($\lfloor 3.6 \rfloor = 3$). As for U_{P}, we duplicate the row with ID = 1 once ($\lfloor 1.7 \rfloor = 1$) and the row with ID = 2 twice ($\lfloor 2.3 \rfloor = 2$) from the probability sample. Note that the weights d_k are treated as an intrinsic variable of S_{P}, so that both d_k and x_{1k} are kept in the pseudo-population U_{P}. The datasets for U_{NP} and U_{P} are as follows (the output of Step 1).

U_{NP}

ID	$\widehat{w}_1$	$\widehat{\pi}_1 = 1/\widehat{w}_1$	y_1	y_2	x_1	x_2	I_2
1	2.4	1/2.4	20	-	4	-	0
1	2.4	1/2.4	20	-	4	-	0
2	3.6	1/3.6	14	20	7	3	1
2	3.6	1/3.6	14	20	7	3	1
2	3.6	1/3.6	14	20	7	3	1

U_{P}

ID	d	$\pi = 1/d$	x_1
1	1.7	1/1.7	5
2	2.3	1/2.3	6
2	2.3	1/2.3	6

To create bootstrap non-probability Phase 1 and probability samples, each row of U_{NP} and U_{P} are selected according to Poisson sampling. Selections into $S^{(c)}_{\text{NP},1}$ are based on the column $\widehat{\pi}_1$ of U_{NP}, and selections into $S^{(c)}_{\text{P}}$ are based on the column π of U_{P}. Suppose the first, fourth and fifth rows of U_{NP} are selected, and this forms $S^{(c)}_{\text{NP},1}$. At the same time, suppose the first and second rows of U_{P} are selected, and this forms $S^{(c)}_{\text{P}}$. Examples of $S^{(c)}_{\text{NP},1}$ and $S^{(c)}_{\text{P}}$ are shown below.

$S_{\text{NP},1}^{(c)}$

ID	y_1	y_2	x_1	x_2	I_2
1	20	-	4	-	0
2	**14**	**20**	**7**	**3**	**1**
2	**14**	**20**	**7**	**3**	**1**

$S_{\text{P}}^{(c)}$

ID	d	$\pi = 1/d$	x_1
1	1.7	1/1.7	5
2	2.3	1/2.3	6

In the end, we create the bootstrap non-probability Phase 2 sample $S_{\text{NP},2}^{(c)}$ by selecting rows of $S_{\text{NP},1}^{(c)}$ with $I_{2k} = 1$. In the above example, blue rows correspond to $S_{\text{NP},2}^{(c)}$. And this concludes Step 2.

Step 3: Compute Bootstrap Weights and Estimates

This step is to compute bootstrap two-phase weights $\widehat{w}_2^{*(c)}$ and bootstrap estimates $\widehat{t}_2^{*(c)}$ from $\left(S_{\text{P}}^{(c)}, S_{\text{NP},1}^{(c)}, S_{\text{NP},2}^{(c)}\right)$ for $c = 1, 2, \ldots, C$.

- **Creation of bootstrap two-phase weight $\widehat{w}_2^{*(c)}$:**
  Based on (2) but applied to $\left(S_{\text{P}}^{(c)}, S_{\text{NP},1}^{(c)}, S_{\text{NP},2}^{(c)}\right)$, we have:

$$\begin{cases} \displaystyle\sum_{k \in S_{\text{NP},1}^{(c)}} \boldsymbol{x}_{1k}^{\intercal} = \sum_{k \in S_{\text{P}}^{(c)}} d_k \pi_1(\boldsymbol{\alpha}^{(c)}; \boldsymbol{x}_{1k}) \boldsymbol{x}_{1k}^{\intercal} \\ \displaystyle\sum_{k \in S_{\text{NP},2}^{(c)}} \boldsymbol{x}_{2k}^{\intercal} = \sum_{k \in S_{\text{NP},1}^{(c)}} \pi_2(\boldsymbol{\beta}^{(c)}; \boldsymbol{I}_1^{(c)}, \boldsymbol{x}_{2k}) \boldsymbol{x}_{2k}^{\intercal} \end{cases}, \tag{14}$$

  where $\boldsymbol{I}_1^{(c)}$ is an N_{NP}-vector indicating Phase 1 selection based on $S_{\text{NP},1}^{(c)}$. Vectors of estimated coefficients $\widehat{\boldsymbol{\alpha}}^{(c)}$ and $\widehat{\boldsymbol{\beta}}^{(c)}$ from (14) lead to Phase 1 weights $\widehat{w}_1^{(c)}$ and Phase 2 weights $\widehat{w}_2^{*(c)}$, respectively. In particular, we have $\widehat{w}_1^{(c)} := \{\widehat{w}_{1k}^{(c)} := \pi_1(\widehat{\boldsymbol{\alpha}}^{(c)}; \boldsymbol{x}_{1k})^{-1} : k \in S_{\text{NP},1}^{(c)}\}$, and $\widehat{w}_2^{*(c)} := \{\widehat{w}_{2k}^{*(c)} := \widehat{w}_{1k}^{(c)} \pi_2(\widehat{\boldsymbol{\beta}}; \boldsymbol{I}_1^{(c)}, \boldsymbol{x}_{2k})^{-1} : k \in S_{\text{NP},2}^{(c)}\}$.

- **Creation of bootstrap estimate $\widehat{t}_2^{*(c)}$:**

$$\widehat{t}_2^{*(c)} = \sum_{k \in S_{\text{NP},2}^{(c)}} \widehat{w}_{2k}^{*(c)} y_{2k}. \tag{15}$$

Step 4: Simulation-based Variance Estimator

Given bootstrap estimates $\widehat{t}_2^{*(1)}, \ldots, \widehat{t}_2^{*(C)}$ from Step 3, the PPB version of the variance estimator of $\widehat{t}_2^{*}$ is

$$\widehat{\text{Var}}^{\text{PPB}}[\widehat{t}_2^{*}] = \frac{1}{C-1} \sum_{c=1}^{C} \left[\widehat{t}_2^{*(c)} - \left(\frac{1}{C} \sum_{c=1}^{C} \widehat{t}_2^{*(c)} \right) \right]^2.$$

Proposition 3. Under Assumptions A1 – A5 and regularity conditions C1 – C4 in the Appendix, and if π_{1k} and $\pi_{2k}(\boldsymbol{I}_1)$ assume the form of logistic regressions, $\widehat{\text{Var}}^{\text{PPB}}[\widehat{t}_2^{*}]$ is a consistent estimator of $\text{Var}[\widehat{t}_2^{*}]$, for a large number of B bootstrap samples.

Remark on other resampling-based methods: Our paper presents a pseudo-population bootstrap (PPB)[6] method to estimate variance in a non-probability two-phase sampling design. Chen et al. (2020) use a with-replacement bootstrap in a non-probability single-phase setup and show good results in simulations. Kim et al. (2006) also suggest a jackknife method for two-phase probability sampling. Therefore, in the future, it would be interesting to compare these methods to our PPB estimator.

3.3 Pseudo-Population Bootstrap for Estimating Variances of Other Weighted Finite Population Parameters

When both models for π_1 and $\pi_2(\mathbf{I}_1)$ are valid, Section 3.2 offers a detailed PPB procedure to estimate the variance of weighted totals from non-calibrated Phase 2 weights $\widehat{w}_2^{*}$. If, for each bootstrap sample, we follow Section 2.2 to compute calibrated Phase 2 weights, PPB

[6]The PPB originates from Gross (1980) and has recently been extended to various complex sampling designs by Wang et al. (2022).

can be used to estimate the variance of weighted totals from calibrated Phase 2 weights $\widetilde{w}_2^*$. We can also use PPB for other weighted finite population parameters and complex statistics presented in Section 2.3 (i.e., Hájek mean and median). In the next Section, we conduct a series of simulation studies to support our suggested PPB approach.

4 Simulation Study

By simulation, this section showcases estimators for population totals, means and medians from Section 2, as well as variance estimation methods from Section 3. We repeatedly draw samples from a known finite population based on the mechanisms we set up. In total, we conduct $R = 12\ 000$ repetitions. For population totals, from each repetition $r = 1, 2, \ldots, R$, we estimate population totals in two ways: (1) $\widehat{t}_2^{*(r)}$ from the non-probability two-phase weighting system without calibration and (2) $\widetilde{t}_2^{(r)}$ from the calibrated version. Then, we evaluate the performance of $\widehat{t}_2^*$ and $\widetilde{t}_2$ by

- the Monte Carlo relative biases (in percent):

$$\%\mathrm{RB}(\widehat{t}_2^*) = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{t}_2^{*(r)} - t_2}{t_2} \times 100\% \text{ and } \%\mathrm{RB}(\widetilde{t}_2) = \frac{1}{R}\sum_{r=1}^{R}\frac{\widetilde{t}_2^{(r)} - t_2}{t_2} \times 100\%;$$

- the Monte Carlo relative efficiency:

$$\mathrm{RE}(\widehat{t}_2^*) = \frac{\mathrm{MSE}_{\mathrm{MC}}[\widehat{t}_2^*]}{\mathrm{MSE}_{\mathrm{MC}}[\widetilde{t}_2]} \times 100 = \frac{\frac{1}{R}\sum_{r=1}^{R}\left[\widehat{t}_2^{*(r)} - t_2\right]^2}{\frac{1}{R}\sum_{r=1}^{R}\left[\widetilde{t}_2^{(r)} - t_2\right]^2} \times 100 \text{ and } \mathrm{RE}(\widetilde{t}_2) = 100,$$

  where $\mathrm{MSE}_{\mathrm{MC}}$ denotes the Monte Carlo mean squared error (MSE).

To assess the performance of variance estimators of $\widehat{t}_2^*$, we compute the relative bias and the coverage probability. In particular,

- the Monte Carlo relative biases (in percent):

$$\%\mathrm{RB}(\widehat{\mathrm{Var}}^{\mathrm{L}}[\widehat{t_2}^*]) = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{\mathrm{Var}}^{\mathrm{L}}[\widehat{t_2}^*]^{(r)} - \mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*]}{\mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*]} \times 100\%,$$

$$\%\mathrm{RB}(\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t_2}^*]) = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t_2}^*]^{(r)} - \mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*]}{\mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*]} \times 100\%,$$

$$\%\mathrm{RB}(\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widehat{t_2}^*]) = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widehat{t_2}^*]^{(r)} - \mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*]}{\mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*]} \times 100\%,$$

  where the Monte Carlo variance of $\widehat{t_2}^*$ is

$$\mathrm{Var}_{\mathrm{MC}}[\widehat{t_2}^*] = \frac{1}{R}\sum_{r=1}^{R}\left[\widehat{t_2}^{*(r)} - \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{t_2}^{*(r)}\right)\right]^2.$$

- the Monte Carlo coverage probabilities (in percent):

$$\%\mathrm{CP}^{\mathrm{L}}(\widehat{t_2}^*) = \frac{1}{R}\sum_{r=1}^{R} I\left[t_2 \in \left(\widehat{t_2}^{*(r)} \pm 1.96\sqrt{\widehat{\mathrm{Var}}^{\mathrm{L}}[\widehat{t_2}^*]^{(r)}}\right)\right] \times 100\%,$$

$$\%\mathrm{CP}^{\mathrm{Alt}}(\widehat{t_2}^*) = \frac{1}{R}\sum_{r=1}^{R} I\left[t_2 \in \left(\widehat{t_2}^{*(r)} \pm 1.96\sqrt{\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t_2}^*]^{(r)}}\right)\right] \times 100\%,$$

$$\%\mathrm{CP}^{\mathrm{PPB}}(\widehat{t_2}^*) = \frac{1}{R}\sum_{r=1}^{R} I\left[t_2 \in \left(\widehat{t_2}^{*(r)} \pm 1.96\sqrt{\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widehat{t_2}^*]^{(r)}}\right)\right] \times 100\%.$$

Similarly, we also measure these Monte Carlo statistics for $\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widetilde{t_2}]$, estimated means and their PPB variance estimators, and estimated medians and their PPB variance estimators.

4.1 Simulation Set-up

The finite population $U = \{1, 2, \ldots, N\}$ for this simulation contains $N = 20\,000$ population units. Each population unit $k \in U$ has survey variables y_{1k} for Phase 1 and y_{2k} for Phase 2 and a set of auxiliary variables a_{1k}, a_{2k} and a_{3k}. The following regression models relate

these two types of variables.

$y_{1k} = 800 + a_{1k} + 5a_{2k} + 10a_{3k} + \epsilon_{1k}$ and $y_{2k} = 20 + a_{1k} + 0.2a_{2k} + 0.5a_{3k} + 0.0025y_{1k} + \epsilon_{2k}$ for $k \in U$,

where $a_{1k} = \gamma_{1k}(v_{1k} + v_{2k})$, $a_{2k} = \gamma_{1k}v_{3k} + \gamma_{2k}v_{1k}$ and $a_{3k} = \gamma_{2k}(v_{2k} + v_{3k})$, $\gamma_{1k} = I[u_{1k} < 0.29]u_{1k}$ and $\gamma_{2k} = I[u_{1k} \geq 0.29]$ and $v_{1k} = I[u_{2k} = 1]$Exponential(1), $v_{2k} = I[u_{2k} = 2]$Exponential(1/2) and $v_{3k} = 40I[u_{2k} = 3]$, $u_{1k} \sim$ lognormal$(10, 1)/100\ 000$, u_{2k} follows a discrete uniform distribution of $\{1, 2, 3\}$. As a note, the 60^{th} percentile of u_1 is 0.29. Each ϵ_{1k} is independently and identically (iid) distributed as Normal $(0, 300)$. Each ϵ_{2k} is iid and follow Normal $(0, 15)$. Errors in both phases (ϵ_{1k} and ϵ_{2k}) are mutually independent. Variances of ϵ_{1k} and ϵ_{2k} are set to control the correlations between the study variable and its linear predictor in Phases 1 and 2 to about 0.68 and 0.74, respectively.

In Phase 1, the non-probability sample $S_{\text{NP},1}$ is selected by Poisson sampling with the true selection probability model $\pi_{1k} = \{1 + \exp[-(\eta_1 + 0.08a_{1k} + 0.05a_{2k} + 0.05a_{3k})]\}^{-1}$. We use $\eta_1 = -3.5$ in π_{1k} to control the expected Phase 1 sampling fraction to be 5%. As for the probability sample S_{P}, each population unit is selected according to the Bernoulli sampling with a constant inclusion probability of 5%.

In Phase 2, we consider two different selection mechanisms $\pi_{2k}(\boldsymbol{I}_1)$ under Poisson sampling. In Scenario 1, the true $\pi_{2k}(\boldsymbol{I}_1)$ is $\{1 + \exp[-(-0.2 + 0.25a_{1k} + 0.05a_{2k} - 0.05a_{3k} + 0.00025y_{1k})]\}^{-1}$. In Scenario 2, the true $\pi_{2k}(\boldsymbol{I}_1)$ is $\{1 + \exp[-(-1.5 + 100r_{1k} + 100r_{2k} + 100r_{3k} + 100r_{4k})]\}^{-1}$ where $r_{1k} = a_{1k}/\sum_{j\in U} a_{1j}I_{1j}$, $r_{2k} = a_{1k}/\sum_{j\in U} a_{2j}I_{1j}$, $r_{3k} = a_{1k}/\sum_{j\in U} a_{3j}I_{1j}$ and $r_{4k} = |y_{1k}|/\sum_{j\in U} |y_{1j}|I_{1j}$. Scenarios 1 and 2 differ in whether Phase 2 selection depends on Phase 1 selection (Beaumont and Haziza, 2016). In Scenario 1 (invariance), Phase 2 selection is independent of Phase 1 selection, while in Scenario 2 (non-invariance), Phase 2 selection depends on Phase 1 selection. Across all $R = 12\ 000$ repetitions, the average Phase 2 response

rate in Scenario 1 is around 35%. As for Scenario 2, around 25% of Phase 1 units respond to Phase 2.

In each repetition of the simulation, we use auxiliary variables $\boldsymbol{x}_{1k} = (a_{1k}, a_{2k}, a_{3k})$ to estimate parameters for π_1 and $\boldsymbol{x}_{2k} = (a_{1k}, a_{2k}, a_{3k}, y_{1k})$ to estimate $\pi_2(\boldsymbol{I}_1)$. Under the non-probability two-phase weighting with calibration, we use the known population size N and population totals of auxiliary variables a_1, a_2 and a_3 for the first-phase calibration, where $\boldsymbol{z}_{1k} = (1, a_{1k}, a_{2k}, a_{3k})$. The calibration at the end of Phase 2 matches the four totals from Phase 1 calibration and the estimated total of the Phase 1 study variable y_1, where $\boldsymbol{z}_{2k} = (1, a_{1k}, a_{2k}, a_{3k}, y_{1k})$. In the end, we use $B = 32$ for the pseudo-population bootstrap, equivalent to $C = B^2 = 1\ 024$ different pairs of bootstrap samples in each $R = 12\ 000$ repetitions.

4.2 Result and Discussion

Table 1 summarizes the simulation results for estimated totals, means and medians. As expected, the biases of all these estimators are quite small across the two scenarios. In terms of efficiency, we find that the mean squared errors (MSEs) are lower when using the weighting system with calibration, as the calibration variables are correlated with y_2. This observation aligns with the findings of Särndal (2007). However, our improvement is based on the non-probability two-phase setup.

Table 2 compares variance estimators from Section 3. A key observation is that all variance estimators and the associated confidence intervals perform excellently. The biases of the variance estimators are all small. The coverage probabilities of the 95% confidence intervals are close to the nominal value. Among the three variance estimators of $\widehat{t}_2^*$, $\widehat{\mathrm{Var}}^{\mathrm{L}}[\widehat{t}_2^*]$ has a smaller bias than $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*]$ and $\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widehat{t}_2^*]$ across the two scenarios (Panel (A) of Table

Table 1: Summary of Simulation Results, Estimated Totals, Means and Medians for Phase 2, Non-Probability Two-Phase Weighting without ($\widehat{w}_2^*$) and with Calibration ($\widetilde{w}_2$)

	Total				**Mean**				**Median**			
	Without Calibration		**With Calibration**		**Without Calibration**		**With Calibration**		**Without Calibration**		**With Calibration**	
	%RB	RE	%RB	RE	%RB	RE	%RB	RE	%RB	RE	%RB	RE
Scenario 1: Invariance	0.04	343	0.01	100	-0.01	161	0.01	100	-0.44	115	-0.46	100
Scenario 2: Non-Invariance	0.07	427	0.01	100	0.01	251	0.01	100	-0.59	132	-0.65	100

2). This observation is expected because $\widehat{\text{Var}}^{\text{Alt}}[\widehat{t}_2^*]$ and $\widehat{\text{Var}}^{\text{PPB}}[\widehat{t}_2^*]$ have ignored the extra sampling variability from the Phase 2 selection $\boldsymbol{I}_2$. In Panel (B) of Table 2, we focus solely on the PPB variance estimators for mean and median. This is because we do not need to derive the Taylor linearizations for mean, median and calibration adjustment. Once again, these PPB variance estimators perform well: low biases and coverage probabilities near 95%.

Table 2: Summary of Simulation Results, Estimated Variances of Estimated Totals, Means and Medians for Phase 2, Non-Probability Two-Phase Weighting without ($\widehat{w}_2^*$) and with Calibration ($\widetilde{w}_2$)

Panel (A): Variance Estimation for Estimated Totals

	Without Calibration						**With Calibration**	
	$\widehat{\text{Var}}^{\text{L}}[\widehat{t}_2^*]$		$\widehat{\text{Var}}^{\text{Alt}}[\widehat{t}_2^*]$		$\widehat{\text{Var}}^{\text{PPB}}[\widehat{t}_2^*]$		$\widehat{\text{Var}}^{\text{PPB}}[\widetilde{t}_2]$	
	%RB	%CP	%RB	%CP	%RB	%CP	%RB	%CP
Scenario 1: Invariance	-1.96	94.69	-2.56	94.59	-3.45	94.67	-1.93	94.86
Scenario 2: Non-Invariance	3.97	95.55	-4.81	94.58	-4.03	94.66	-1.93	94.86

Panel (B): Pseudo-Population Bootstrap Variance Estimation for Means and Medians

	Mean				**Median**			
	Without Calibration		**With Calibration**		**Without Calibration**		**With Calibration**	
	$\widehat{\text{Var}}^{\text{PPB}}[\widehat{\mu}_2^*]$		$\widehat{\text{Var}}^{\text{PPB}}[\widetilde{\mu}_2]$		$\widehat{\text{Var}}^{\text{PPB}}[\widehat{m}_2^*]$		$\widehat{\text{Var}}^{\text{PPB}}[\widetilde{m}_2]$	
	%RB	%CP	%RB	%CP	%RB	%CP	%RB	%CP
Scenario 1: Invariance	-2.44	94.91	-1.93	94.86	-0.87	94.62	-0.28	94.76
Scenario 2: Non-Invariance	-5.24	94.35	-1.93	94.86	0.04	95.42	2.35	95.55

5 An Application to the 2020 November Cash Alternative Survey of the Bank of Canada

In this section, we apply our proposed method to the Bank of Canada 2020 November Cash Alternative Survey (Chen et al., 2021). This survey is a non-probability two-phase sample, which consists of a survey questionnaire (SQ) as Phase 1 ($n_1 = 3,893$) and a three-day diary survey instrument (DSI) as Phase 2 ($n_2 = 2,084$). Key questions from Phase 1 address respondents' cash holdings and ownership of various payment instruments. Phase 2 is a payment diary in which respondents record their purchase information (i.e., the payment method used and the amount of each purchase) over three days. For the individual-level probability sample, we chose the Canadian Perspectives Survey Series 5 (CPSS 5), administered by Statistics Canada, for the following reasons. First, the CPSS sample ($n_{\mathrm{P}} = 3,961$) comes from rotation groups of the Labour Force Survey (LFS), a reliable social probability survey. Second, CPSS 5 was collected from September 14, 2020, to September 20, 2020, similar to our non-probability survey's field operation period. In the end, both CPSS 5 and the Bank of Canada 2020 November Cash Alternative Survey were conducted in the online mode.[7]

We estimate the population cash volume share based on the Phase 2 cash and non-cash transactions. For each $k \in S_{\mathrm{NP},2}$, let y_{2k}^{cash} and $y_{2k}^{non-cash}$ be the numbers of cash usage and non-cash usage in the three-day diary, respectively. Thus, the estimated cash volume share is the ratio of the total estimated cash usage to the total estimated transactions (the sum of cash and non-cash transactions). The volume share of cash transactions can be computed

[7]Kim et al. (2021) discuss the importance of survey mode in explaining gaps between non-probability and probability samples' estimates.

using either non-calibrated Phase 2 weights $\widehat{w}^*_{2k}$ from Section 2.1:

$$\widehat{\tau}^*_2 := \frac{\sum_{k \in S_{\text{NP},2}} \widehat{w}^*_{2k} y^{cash}_{2k}}{\sum_{k \in S_{\text{NP},2}} \widehat{w}^*_{2k} (y^{cash}_{2k} + y^{non-cash}_{2k})},$$

or calibrated Phase 2 weights $\widetilde{w}_{2k}$ from Section 2.2:

$$\widetilde{\tau}_2 := \frac{\sum_{k \in S_{\text{NP},2}} \widetilde{w}_{2k} y^{cash}_{2k}}{\sum_{k \in S_{\text{NP},2}} \widetilde{w}_{2k} (y^{cash}_{2k} + y^{non-cash}_{2k})}.$$

Besides the estimated cash volume across the entire population, we also compute disaggregated estimates for three age groups: the 18–34, the 35–54 and the 55+. Results of point estimates and standard errors are reported in Table 3.

The following is a list of auxiliary variables $\boldsymbol{x}_{1k}$ and $\boldsymbol{x}_{2k}$ used to estimate selection probabilities $\pi_{1k} = \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k})$ and $\pi_{2k}(\boldsymbol{I}_1) = \pi_2(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k})$, and calibration variables $\boldsymbol{z}_{1k}$ for the SQ (Phase 1) and $\boldsymbol{z}_{2k}$ for the DSI (Phase 2):

- $\boldsymbol{x}_{1k}$: sex/gender, age group, household size, marital status, highest education attainment and whether the respondent has shopped online during COVID-19;
- $\boldsymbol{x}_{2k}$: $\boldsymbol{x}_{1k}$ plus Phase 1 variables, such as cash on hand and other cash holdings at three age groups;
- $\boldsymbol{z}_{1k}$: population size, sex/gender and age group;
- $\boldsymbol{z}_{2k}$: $\boldsymbol{z}_{1k}$ plus Phase 1 variables, such as cash on hand and other cash holdings at three age groups.

The estimated overall cash volume share is 23.6 percent without using calibrated weights, or 23.5 percent with calibrated weights. Estimates $\widehat{\tau}^*_2$ and $\widetilde{\tau}_2$ at each age group are also close. Consistent with Chen et al. (2021), people in each age group use cash to pay for at least 20 percent of their purchases. Similarly, older people use relatively more cash than the younger

groups. For example, the cash volume share of the 55+ group estimated from uncalibrated Phase 2 weights is 26.4 percent, about 5 percentage points higher than the 18–34 group.

Table 3: Cash Volume Share (percentage) Estimated from Non-Probability Two-Phase Weighting without calibration ($\widehat{w}_2^*$) and with Calibration ($\widetilde{w}_2$)

	Cash Volume Share	
	Without Calibration	**With Calibration**
Overall	23.6 (1.17)	23.5 (1.17)
18–34	21.0 (2.75)	21.2 (2.69)
35–54	21.5 (2.05)	21.6 (2.03)
55+	26.4 (1.70)	26.4 (1.69)

Note: The standard error of each estimate is in brackets. All of them are estimated from the PPB approach outlined in Section 3.2 with $B = 120$ ($C = 14\ 400$ pairs of bootstrap samples in total). We approximate the multistage sampling of CPSS 5 with the Poisson sampling design to avoid computing second-order inclusion probabilities.

In terms of estimating standard errors, we compute them via the PPB approach outlined in Section 3.2. As we have discussed, the edge of the resampling PPB approach over analytical plug-in methods is that PPB is straightforward to apply. In particular, we do not need to derive the Taylor linearizations of complex statistics, the Hájek mean and calibration adjustment. The uncertainty for both weighting schemes is quite small, with standard errors less than 3 percent. Such small standard errors indicate the stability of our estimates, which can provide enough statistical power to detect differences between groups. In the end, we only observe a slight efficiency gain from using the calibrated weights over the uncalibrated ones, especially in the 18–34 age group. This slim difference suggests that future research should explore other calibration variables (i.e., z_1 and z_2) more strongly related to the variables of interest.

6 Conclusion

We have discussed how to make statistical inferences for two-phase survey samples whose weights for Phase 1 and (conditional) Phase 2 are unknown. The starting point is to use the pseudo maximum likelihood method for weight estimation. Each unit's weight is the inverse of the selection probability estimated from this method. For Phase 1, we rely on auxiliary variables from a probability sample for estimation. This leads to a weighting system parallel to the empirical double expansion (EDE) used in probability sampling for totals. As for variance estimation, our preferred way is a pseudo-population bootstrap (PPB) approach from Sections 3.2 and 3.3. This method can account for the estimation of probabilities from both phases as well as calibrations.

Our approach is the same as mainstream survey literature in that it assumes the selection mechanisms for both phases are ignorable (Assumption A1) (Little and Rubin, 2019). However, it may be worthwhile to revisit this assumption. For instance, methods that do not rely on this assumption can help address concerns about bias from unobserved confounders. One such method is the sensitivity analysis proposed by Hartman and Huang (2024).

A Appendix

We use the same setup for asymptotics as Chen et al. (2020), which adapts the framework from Isaki and Fuller (1982) for non-probability sampling. In this framework, there is a sequence τ of finite populations. Each population in this sequence has a Phase 1 non-probability sample, a Phase 2 sample as a subsample of Phase 1 and a probability sample. As the sequence $\tau \to \infty$, the population size and the sample sizes also approach infinity. In this paper, we suppress the index τ and use the notion $N \to \infty$ to indicate this limiting process.

A.1 Regularity Conditions

Regularity conditions for Phase 1 come from Chen et al. (2020). Those for Phase 2 conditional on Phase 1 come from Kim and Kim (2007).

- **C1**: The population size N and the sample sizes $n_1 := |S_{\text{NP},1}|$, $n_2 := |S_{\text{NP},2}|$ and $n_P = |S_{\text{P}}|$ satisfy $n_1 \to \infty$, $n_2 \to \infty$ and $n_2 \leq n_1$ with $\lim_{N\to\infty} \frac{n_1}{N} \in [0, 1)$ and $\lim_{N\to\infty} \frac{n_{\text{P}}}{N} \in (0, 1)$.
- **C2**: The finite population and the sampling design for S_{P} satisfy $\frac{1}{N}\sum_{k\in S_{\text{P}}} d_k \boldsymbol{x}_{1k} - \frac{1}{N}\sum_{k\in U} \boldsymbol{x}_{1k} = O_p(n_{\text{P}}^{-1/2})$.
- **C3**: There exists positive constants γ_1 and γ_2 such that $0 < \gamma_1 \leq \frac{N\pi_{1k}}{n_1} \leq \gamma_2$, $0 < \gamma_1 \leq \frac{n_1\pi_{2k}(\boldsymbol{I}_1)}{n_2} \leq \gamma_2$ and $0 < \gamma_1 \leq \frac{N}{d_k n_{\text{P}}} \leq \gamma_2$ for all units k.
- **C4**: The finite population and the selection probabilities satisfy $\frac{1}{N}\sum_{k\in U} y_{2k}^2 = O(1)$, $\frac{1}{N}\sum_{k\in U} ||\boldsymbol{x}_{1k}||^3 = O(1)$ and $\frac{1}{N}\sum_{k\in U} ||\boldsymbol{x}_{2k}||^3 = O(1)$. Both $\frac{1}{N}\sum_{k\in U} \pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal$ and $\frac{1}{N}\sum_{k\in U} \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal$ are positive definite matrices.

C2 assumes that first moments of auxiliary variables $\boldsymbol{x}_{1k}$ from S_{P} and $S_{\text{NP},1}$ are asymptotically equivalent. C3 prevents extreme weights by restricting the probabilities of being selected into $S_{\text{NP},1}$, $S_{\text{NP},2}$ and S_{P} do not differ in terms of order of magnitude from simple random sampling without replacement. C4 gives us finite moment conditions for valid Taylor series expansions.

A.2 Proof for Theorem 1

For simplicity, we assume there is only one Phase 2 study variable. Suppose that $\boldsymbol{\eta}^\intercal := [\mu_2 \ \boldsymbol{\alpha}^\intercal \ \boldsymbol{\beta}^\intercal]^\intercal$ where $\mu_2 := t_2/N$ is the population mean of the Phase 2 study variable y_2, $t_2 := \sum_{k\in U} y_{2k}$ is the population total of y_2, $\boldsymbol{\alpha}$ is a q-vector of the true parameters of the propensity score model $\pi_{1k} := \pi_1(\boldsymbol{\alpha}; \boldsymbol{x}_{1k})$ for Phase 1 with a q-vector of auxiliary variables $\boldsymbol{x}_{1k}$ and

$\boldsymbol{\beta}$ is a r-vector of the true parameters of the propensity score model $\pi_{2k}(\boldsymbol{I}_1) := \pi_{2k}(\boldsymbol{\beta}; \boldsymbol{I}_1, \boldsymbol{x}_{2k})$ with an r-vector of auxiliary variables $\boldsymbol{x}_{2k}$. I_k^{P} is an indicator of selection into S_{P}.

Let $\widehat{\boldsymbol{\eta}} = (\widehat{\mu}_{2,N}^*, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ be the solution to the following system of estimating equations:

$$\Phi_n(\boldsymbol{\eta}) = \frac{1}{N} \begin{bmatrix} \sum_{k \in U} \left(I_{1k} I_{2k} \dfrac{y_{2k}}{\pi_{1k} \pi_{2k}(\boldsymbol{I}_1)} - \mu_2 \right) \\ \sum_{k \in U} \left(I_{1k} \boldsymbol{x}_{1k} - I_k^{\mathrm{P}} d_k \pi_{1k} \boldsymbol{x}_{1k} \right) \\ \sum_{k \in U} (I_{1k} I_{2k} \boldsymbol{x}_{2k} - I_{1k} \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{x}_{2k}) \end{bmatrix} = \mathbf{0}_{(1+q+r)\times 1}, \tag{16}$$

where I_k^{P} is an indicator of whether unit $k \in U$ is selected into S_{P}.

Under the joint randomization of the selection probability models for Phase 1 and (conditional) Phase 2 and the sampling design of S_{P}, we have $E[\Phi_n(\boldsymbol{\eta}_0)] = \mathbf{0}$ because

$$E[\Phi_n(\boldsymbol{\eta}_0)] = E_p \left\{ E_1[E_2(\Phi_n(\boldsymbol{\eta}) \mid \boldsymbol{I}_1)] \right\} = \frac{1}{N} \begin{bmatrix} \sum_{k \in U} \left\{ y_{2k} E_1 \left[\dfrac{I_{1k}}{\pi_{1k}} E_2 \left(\dfrac{I_{2k}}{\pi_{2k}(\boldsymbol{I}_1)} \;\middle|\; \boldsymbol{I}_1 \right) \right] - \mu_2 \right\} \\ \sum_{k \in U} \left\{ \boldsymbol{x}_{1k} E_1[I_{1k}] - E_p[I_k^{\mathrm{P}}] d_k \pi_{1k} \boldsymbol{x}_{1k} \right\} \\ \sum_{k \in U} \left\{ E_1 \left[I_{1k} E_2(I_{2k} \mid \boldsymbol{I}_1) \right] \boldsymbol{x}_{2k} - E_1[I_{1k}] \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{x}_{2k} \right\} \end{bmatrix}$$

$$= \frac{1}{N} \begin{bmatrix} \sum_{k \in U} \left[y_{2k} \cdot \dfrac{\pi_{1k}}{\pi_{1k}} \cdot \dfrac{\pi_{2k}(\boldsymbol{I}_1)}{\pi_{2k}(\boldsymbol{I}_1)} - \mu_2 \right] \\ \sum_{k \in U} \left\{ \boldsymbol{x}_{1k} \pi_{1k} - \dfrac{\pi_k^{\mathrm{P}}}{\pi_k^{\mathrm{P}}} \pi_{1k} \boldsymbol{x}_{1k} \right\} \\ \sum_{k \in U} \left\{ \pi_{1k} \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{x}_{2k} - \pi_{1k} \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{x}_{2k} \right\} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} t_2 - N\mu_2 \\ \sum_{k \in U} \boldsymbol{x}_{1k} \pi_{1k} - \sum_{k \in U} \boldsymbol{x}_{1k} \pi_{1k} \\ \sum_{k \in U} \pi_{1k} \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{x}_{2k} - \sum_{k \in U} \pi_{1k} \pi_{2k}(\boldsymbol{I}_1) \boldsymbol{x}_{2k} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0}_{q\times 1} \\ \mathbf{0}_{r\times 1} \end{bmatrix}$$

$$= \mathbf{0}_{(1+q+r)\times 1},$$

where $\boldsymbol{\eta}_0 = (\mu_2, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and the design weight $d_k = 1/\pi_k^{\mathrm{P}}$ is the inverse of the inclusion probability π_k^{P}. Consistency of the estimator $\widehat{\boldsymbol{\eta}}$ follows from Newey and McFadden (1994).

Under regularity conditions C1 – C4, we have $\Phi_n(\widehat{\boldsymbol{\eta}}) = \mathbf{0}$ and $\Phi_n(\boldsymbol{\eta}_0) = O_p(n_1^{-1/2})$. By the first-order Taylor series approximation of $\Phi_n(\widehat{\eta})$ around $\boldsymbol{\eta}_0$ (Proposition 6.1.5 in Brockwell (1991)), we have

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = [\phi_n(\widehat{\eta}_0)]^{-1}\Phi_n(\boldsymbol{\eta}_0) + o_p(n_1^{-1/2}) = [E(\phi_n(\widehat{\eta}_0))]^{-1}\Phi_n(\boldsymbol{\eta}_0) + o_p(n_1^{-1/2}), \tag{17}$$

where $\phi_n(\boldsymbol{\eta})$ is the Jacobian matrix of $\Phi_n(\boldsymbol{\eta})$:

$$\phi_n(\boldsymbol{\eta}) = \frac{1}{N}\begin{bmatrix} -1 & -\sum_{k\in U} I_{1k}I_{2k}\frac{y_{2k}}{\pi_{2k}(\boldsymbol{I}_1)}\cdot\frac{1-\pi_{1k}}{\pi_{1k}}\boldsymbol{x}_{1k}^\intercal & -\sum_{k\in U} I_{1k}I_{2k}\frac{y_{2k}}{\pi_{1k}}\cdot\frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{2k}(\boldsymbol{I}_1)}\boldsymbol{x}_{2k}^\intercal \\ \mathbf{0}_{q\times 1} & \sum_{k\in U} I_k^{\mathrm{P}} d_k \pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal & \mathbf{0}_{q\times r} \\ \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times q} & \sum_{k\in U} I_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal \end{bmatrix}.$$

It follows that $\widehat{\mu}_{2,N}^* = \mu_2 + O_p(n_1^{-1/2})$ and hence $\widehat{t}_2^* = N\widehat{\mu}_{2,N}^* = t_2 + O_p(Nn_1^{-1/2})$. We also have

$$\mathrm{Var}[\widehat{\boldsymbol{\eta}}] = [E(\phi_n(\widehat{\eta}_0))]^{-1}\cdot\mathrm{Var}[\Phi_n(\eta_0)]\cdot\{[E(\phi_n(\widehat{\eta}_0))]^{-1}\}^\intercal + o(n_1^{-1/2}).$$

Moreover,

$$E[\phi_n(\boldsymbol{\eta})] = \frac{1}{N}\begin{bmatrix} -1 & -\sum_{k\in U}(1-\pi_{1k})y_{2k}\boldsymbol{x}_{1k}^\intercal & -\sum_{k\in U}(1-\pi_{2k}(\boldsymbol{I}_1))y_{2k}\boldsymbol{x}_{2k}^\intercal \\ \mathbf{0}_{q\times 1} & \sum_{k\in U}\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal & \mathbf{0}_{q\times r} \\ \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times q} & \sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal \end{bmatrix} \text{ and}$$

$$\{E[\phi_n(\boldsymbol{\eta})]\}^{-1} = N\begin{bmatrix} -1 & -\boldsymbol{b}_1^\intercal & -\boldsymbol{c}_2^\intercal \\ \mathbf{0}_{q\times 1} & \left[\sum_{k\in U}\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal\right]^{-1} & \mathbf{0}_{q\times r} \\ \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times q} & \left[\sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal\right]^{-1} \end{bmatrix}, \tag{18}$$

where $\boldsymbol{b}_1^\intercal = \left[N^{-1}\sum_{i\in U}(1-\pi_{1i})y_{2i}\boldsymbol{x}_{1i}^\intercal\right]\left[N^{-1}\sum_{i\in U}\pi_{1i}(1-\pi_{1i})\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^\intercal\right]^{-1}$ and
$\boldsymbol{c}_2^\intercal = \left[N^{-1}\sum_{i\in U}(1-\pi_{2i}(\boldsymbol{I}_1))y_{2i}\boldsymbol{x}_{2i}^\intercal\right]\left[N^{-1}\sum_{i\in U}\pi_{1i}\pi_2(\boldsymbol{I}_1)(1-\pi_{2i}(\boldsymbol{I}_1))\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}^\intercal\right]^{-1}$.

The following finds the total variance $\mathrm{Var}[\Phi_n(\boldsymbol{\eta}_0)]$. We can decompose $\Phi_n(\boldsymbol{\eta}) := \boldsymbol{A}_1 - \boldsymbol{A}_2$ into two independent $(1+q+r)$-vectors $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$:

$$\Phi_n(\boldsymbol{\eta}) = \frac{1}{N}\begin{bmatrix} \sum_{k\in U}\left(I_{1k}I_{2k}\dfrac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} - \mu_2\right) \\ \sum_{k\in U} I_{1k}\boldsymbol{x}_{1k} \\ \sum_{k\in U}(I_{1k}I_{2k}\boldsymbol{x}_{2k} - I_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}) \end{bmatrix} - \frac{1}{N}\begin{bmatrix} 0 \\ \sum_{k\in U}\left(I_k^{\mathrm{P}}d_k\pi_{1k}\boldsymbol{x}_{1k}\right) \\ \mathbf{0}_{r\times 1} \end{bmatrix} =: \boldsymbol{A}_1 - \boldsymbol{A}_2.$$

Because $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are independent, we have $\mathrm{Var}[\Phi_n(\boldsymbol{\eta})] = \mathrm{Var}[\boldsymbol{A}_1] + \mathrm{Var}[\boldsymbol{A}_2]$:

$$\mathrm{Var}[\boldsymbol{A}_1] = \frac{1}{N^2}\begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} & \boldsymbol{V}_{13} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} & \boldsymbol{V}_{23} \\ \boldsymbol{V}_{31} & \boldsymbol{V}_{32} & \boldsymbol{V}_{33} \end{bmatrix} \text{ and } \mathrm{Var}[\boldsymbol{A}_2] = \frac{1}{N^2}\begin{bmatrix} 0 & \mathbf{0}_{1\times q} & \mathbf{0}_{1\times r} \\ \mathbf{0}_{q\times 1} & \boldsymbol{D}_{q\times q} & \mathbf{0}_{q\times r} \\ \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times q} & \mathbf{0}_{r\times r} \end{bmatrix},$$

where $\boldsymbol{D} = \mathrm{Var}_p\left[\sum_{k\in U} I_k^{\mathrm{P}} d_k \pi_{1k} \boldsymbol{x}_{1k}\right]$ is the design-based variance-covariance matrix under the probability sampling design for the reference probability sample S_{P}. We derive each element of the variance-covariance matrix $\mathrm{Var}[\boldsymbol{A}_1]$ as follows.

First, we derive matrices $\boldsymbol{V}_{11}$, $\boldsymbol{V}_{22}$ and $\boldsymbol{V}_{33}$ under assumptions A1 – A4.

$$
\begin{aligned}
\boldsymbol{V}_{11} &= V\left[\sum_{k\in U} I_{1k} I_{2k} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] = \sum_{k\in U} \frac{y_{2k}^2}{\pi_{1k}^2 \pi_{2k}(\boldsymbol{I}_1)^2} V\left[I_{1k} I_{2k}\right] \\
&= \sum_{k\in U} \frac{y_{2k}^2}{\pi_{1k}^2 \pi_{2k}(\boldsymbol{I}_1)^2} \left\{V_1[E_2(I_{1k} I_{2k} \mid \boldsymbol{I}_1)] + E_1[V_2(I_{1k} I_{2k} \mid \boldsymbol{I}_1)]\right\} \\
&= \sum_{k\in U} \frac{y_{2k}^2}{\pi_{1k}^2 \pi_{2k}(\boldsymbol{I}_1)^2} \left\{\pi_{1k}(1-\pi_{1k})\pi_{2k}(\boldsymbol{I}_1)^2 + \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\right\} \\
&= \sum_{k\in U} \left[\frac{1-\pi_{1k}}{\pi_{1k}} + \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] y_{2k}^2,
\end{aligned}
$$

where the second line comes from the law of total variance.

$$
\boldsymbol{V}_{22} = V\left[\sum_{k\in U} I_{1k} \boldsymbol{x}_{1k}\right] = \sum_{k\in U} \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}^{\mathsf{T}} V[I_{1k}] = \sum_{k\in U} \pi_{1k}(1-\pi_{1k}) \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}^{\mathsf{T}}
$$

$$
\begin{aligned}
\boldsymbol{V}_{33} &= V\left[\sum_{k\in U} \left(I_{1k} I_{2k} \boldsymbol{x}_{2k} - I_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right)\right] = V\left[\sum_{k\in U} I_{1k} \boldsymbol{x}_{2k} \left(I_{2k} - \pi_{2k}(\boldsymbol{I}_1)\right)\right] \\
&= E_1\left[V_2\left(\sum_{k\in U} I_{1k} \boldsymbol{x}_{2k} \left(I_{2k} - \pi_{2k}(\boldsymbol{I}_1)\right) \middle| \boldsymbol{I}_1\right)\right] + V_1\left[E_2\left(\sum_{k\in U} I_{1k} \boldsymbol{x}_{2k} \left(I_{2k} - \pi_{2k}(\boldsymbol{I}_1)\right) \middle| \boldsymbol{I}_1\right)\right] \\
&= E_1\left[\sum_{k\in U} I_{1k} \boldsymbol{x}_{2k} \boldsymbol{x}_{2k}^{\mathsf{T}} \pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\right] \\
&= \sum_{k\in U} \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1)) \boldsymbol{x}_{2k} \boldsymbol{x}_{2k}^{\mathsf{T}},
\end{aligned}
$$

where the second line comes from the law of total variance.

For matrices $\boldsymbol{V}_{12} = \boldsymbol{V}_{21}^\intercal$, $\boldsymbol{V}_{13} = \boldsymbol{V}_{31}^\intercal$ and $\boldsymbol{V}_{23} = \boldsymbol{V}_{32}^\intercal$, under assumptions A1 – A4,

$$\boldsymbol{V}_{12} = \mathrm{Cov}\left[\sum_{k\in U} I_{1k}I_{2k}\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}, \sum_{\ell\in U} I_{1\ell}\boldsymbol{x}_{1\ell}\right] = \sum_{k\in U}\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\sum_{\ell\in U}\mathrm{Cov}\left[I_{1k}I_{2k}, I_{1\ell}\right]\boldsymbol{x}_{1\ell}^\intercal = \boldsymbol{V}_{21}^\intercal$$

$$\begin{aligned}\boldsymbol{V}_{13} &= \mathrm{Cov}\left[\sum_{k\in U} I_{1k}I_{2k}\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}, \sum_{\ell\in U}\left(I_{1\ell}I_{2\ell}\boldsymbol{x}_{2\ell} - I_{1\ell}\pi_{2\ell}(\boldsymbol{I}_1)\boldsymbol{x}_{2\ell}\right)\right], \\ &= \sum_{k\in U}\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\sum_{\ell\in U}\mathrm{Cov}\left[I_{1k}I_{2k}, I_{1\ell}I_{2\ell} - I_{1\ell}\pi_{2\ell}(\boldsymbol{I}_1)\right]\boldsymbol{x}_{2\ell}^\intercal \\ &= \sum_{k\in U}\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\sum_{\ell\in U}\left(\mathrm{Cov}\left[I_{1k}I_{2k}, I_{1\ell}I_{2\ell}\right] - \pi_{2k}(\boldsymbol{I}_1)\mathrm{Cov}\left[I_{1k}I_{2k}, I_{1\ell}\right]\right)\boldsymbol{x}_{2\ell}^\intercal = \boldsymbol{V}_{31}^\intercal \text{ and}\end{aligned}$$

$$\begin{aligned}\boldsymbol{V}_{23} &= \mathrm{Cov}\left[\sum_{k\in U} I_{1k}\boldsymbol{x}_{1k}, \sum_{k\in U}\left(I_{1k}I_{2k}\boldsymbol{x}_{2k} - I_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right)\right] = \sum_{k\in U}\boldsymbol{x}_{1k}\sum_{\ell\in U}\mathrm{Cov}\left[I_{1k}, \left(I_{1\ell}I_{2\ell} - I_{1\ell}\pi_{2\ell}(\boldsymbol{I}_1)\right)\right]\boldsymbol{x}_{2\ell}^\intercal \\ &= \sum_{k\in U}\boldsymbol{x}_{1k}\sum_{\ell\in U}\left(\mathrm{Cov}\left[I_{1k}, I_{1\ell}I_{2\ell}\right] - \pi_{2\ell}(\boldsymbol{I}_1)\mathrm{Cov}\left[I_{1k}, I_{1\ell}\right]\right)\boldsymbol{x}_{2\ell}^\intercal = \boldsymbol{V}_{32}^\intercal.\end{aligned}$$

Note that we have

$$\begin{aligned}\mathrm{Cov}[I_{1k}I_{2k}, I_{1\ell}] &= E[I_{1k}I_{2k}I_{1\ell}] - E[I_{1k}I_{2k}]E[I_{1\ell}] \\ &= \begin{cases} E[I_{1k}I_{2k}](1 - E[I_{1k}]) = \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{1k}) & \text{if } k = \ell \\ \pi_{1k}\pi_{1\ell}\pi_{2k}(\boldsymbol{I}_1) - \pi_{1k}\pi_{1\ell}\pi_{2k}(\boldsymbol{I}_1) = 0 & \text{if } k \neq \ell \end{cases} \\ \mathrm{Cov}[I_{1k}I_{2k}, I_{1\ell}I_{2\ell}] &= E[I_{1k}I_{2k}I_{1\ell}I_{2\ell}] - E[I_{1k}I_{2k}]E[I_{1\ell}I_{2\ell}] \\ &= \begin{cases} E[I_{1k}I_{2k}] - E[I_{1k}I_{2k}]^2 = \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1 - \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)) & \text{if } k = \ell \\ \pi_{1k}\pi_{1\ell}\pi_{2k}(\boldsymbol{I}_1)\pi_{2\ell}(\boldsymbol{I}_1) - \pi_{1k}\pi_{1\ell}\pi_{2k}(\boldsymbol{I}_1)\pi_{2\ell}(\boldsymbol{I}_1) = 0 & \text{if } k \neq \ell \end{cases},\end{aligned}$$

where, by independence, $E_2[I_{2k}I_{2\ell}|\boldsymbol{I}_1] = \pi_{2k}(\boldsymbol{I}_1)\pi_{2\ell}(\boldsymbol{I}_1)$ for $k \neq \ell$ and $E_1[I_{1k}I_{1\ell}] = \pi_{1k}\pi_{1\ell}$ for $k \neq \ell$.

It follows that

$$\boldsymbol{V}_{12} = \sum_{k\in U} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{1k})\boldsymbol{x}_{1k}^\intercal = \sum_{k\in U}(1-\pi_{1k})y_{2k}\boldsymbol{x}_{1k}^\intercal = \boldsymbol{V}_{21}^\intercal,$$

$$\begin{aligned}
\boldsymbol{V}_{13} &= \sum_{k\in U} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} \left(\mathrm{Cov}\left[I_{1k}I_{2k}, I_{1k}I_{2k}\right] - \pi_{2k}(\boldsymbol{I}_1)\mathrm{Cov}\left[I_{1k}I_{2k}, I_{1k}\right]\right) \boldsymbol{x}_{2k}^\intercal \\
&= \sum_{k\in U} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} \left(\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)) - \pi_{2k}(\boldsymbol{I}_1)\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{1k})\right) \boldsymbol{x}_{2k}^\intercal \\
&= \sum_{k\in U} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} [\pi_{1k}\pi_{2k}(\boldsymbol{I}_1) - \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)^2]\boldsymbol{x}_{2k}^\intercal = \sum_{k\in U}(1-\pi_{2k}(\boldsymbol{I}_1))y_{2k}\boldsymbol{x}_{2k}^\intercal = \boldsymbol{V}_{31}^\intercal \text{ and}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{V}_{23} &= \sum_{k\in U} \boldsymbol{x}_{1k} \left(\mathrm{Cov}\left[I_{1k}, I_{1k}I_{2k}\right] - \pi_{2k}(\boldsymbol{I}_1)\mathrm{Cov}\left[I_{1k}, I_{1k}\right]\right) \boldsymbol{x}_{2k}^\intercal \\
&= \sum_{k\in U} \boldsymbol{x}_{1k} \left[\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{1k}) - \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{1k})\right] \boldsymbol{x}_{2k}^\intercal = \boldsymbol{0}_{q\times r} = \boldsymbol{V}_{32}^\intercal.
\end{aligned}$$

In summary,

$$\begin{aligned}
&\mathrm{Var}[\boldsymbol{A}_1] \\
=&\frac{1}{N^2}\begin{bmatrix}
\sum_{k\in U}\left[\frac{1-\pi_{1k}}{\pi_{1k}} + \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] y_{2k}^2 & \sum_{k\in U}(1-\pi_{1k})y_{2k}\boldsymbol{x}_{1k}^\intercal & \sum_{k\in U}[1-\pi_{2k}(\boldsymbol{I}_1)]y_{2k}\boldsymbol{x}_{2k}^\intercal \\
\sum_{k\in U}(1-\pi_{1k})\boldsymbol{x}_{1k}y_{2k} & \sum_{k\in U}\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal & \boldsymbol{0}_{q\times r} \\
\sum_{k\in U}[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}y_{2k} & \boldsymbol{0}_{r\times q} & \sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal
\end{bmatrix}.
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathrm{Var}[\Phi_n(\eta_0)] = \mathrm{Var}[\boldsymbol{A}_1] + \mathrm{Var}[\boldsymbol{A}_2] \\
=&\frac{1}{N^2}\begin{bmatrix}
\sum_{k\in U}\left[\frac{1-\pi_{1k}}{\pi_{1k}} + \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] y_{2k}^2 & \sum_{k\in U}(1-\pi_{1k})y_{2k}\boldsymbol{x}_{1k}^\intercal & \sum_{k\in U}[1-\pi_{2k}(\boldsymbol{I}_1)]y_{2k}\boldsymbol{x}_{2k}^\intercal \\
\sum_{k\in U}(1-\pi_{1k})\boldsymbol{x}_{1k}y_{2k} & \sum_{k\in U}\left[\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal\right] + \boldsymbol{D} & \boldsymbol{0}_{q\times r} \\
\sum_{k\in U}[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}y_{2k} & \boldsymbol{0}_{r\times q} & \sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal
\end{bmatrix}.
\end{aligned}$$

Therefore, the asymptotic variance for the linearized $\widehat{t}_2^*$ is the first diagonal element of the variance-covariance matrix $\mathrm{Var}[\widehat{\boldsymbol{\eta}}]$ multiplied by N^2:

$$
\begin{aligned}
&\mathrm{Var}^L[\widehat{t}_2^*] = N^2 \mathrm{Var}^L[\widehat{\mu}_{2,N}^*] \\
&= \sum_{k\in U}\left[\frac{1-\pi_{1k}}{\pi_{1k}} + \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] y_{2k}^2 - \boldsymbol{b}_1^\intercal\left[\sum_{k\in U}(1-\pi_{1k})\boldsymbol{x}_{1k}y_{2k}\right] - \boldsymbol{c}_2^\intercal\left[\sum_{k\in U}[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}y_{2k}\right] \\
&\quad - \boldsymbol{b}_1^\intercal\left[\sum_{k\in U}(1-\pi_{1k})y_{2k}\boldsymbol{x}_{1k}^\intercal\right] + \boldsymbol{b}_1^\intercal\left[\sum_{k\in U}[\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal] + \boldsymbol{D}\right]\boldsymbol{b}_1 \\
&\quad + \boldsymbol{c}_2^\intercal\sum_{k\in U}[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}^\intercal y_{2k} + \boldsymbol{c}_2^\intercal\left[\sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)[1-\pi_{2k}(\boldsymbol{I}_1)]\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal\right]\boldsymbol{c}_2 \\
&= \sum_{k\in U}(1-\pi_{1k})y_{2k}\left(\frac{y_{2k}}{\pi_{1k}} - \boldsymbol{b}_1^\intercal\boldsymbol{x}_{1k}\right) + \sum_{k\in U}(1-\pi_{2k}(\boldsymbol{I}_1))y_{2k}\left(\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} - \boldsymbol{c}_2^\intercal\boldsymbol{x}_{2k}\right) \\
&\quad + \sum_{k\in U}[(1-\pi_{1k})\boldsymbol{b}_1^\intercal(\pi_{1k}\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal\boldsymbol{b}_1 - y_{2k}\boldsymbol{x}_{1k}^\intercal)] + \boldsymbol{b}_1^\intercal\boldsymbol{D}\boldsymbol{b}_1 \\
&\quad + \sum_{k\in U}[(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{c}_2^\intercal(\boldsymbol{x}_{2k}^\intercal + \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal\boldsymbol{c}_2)] \\
&= \sum_{k\in U}\left[(1-\pi_{1k})\pi_{1k}\frac{y_{2k}^2}{\pi_{1k}^2} - \frac{(1-\pi_{1k})\pi_{1k}y_{2k}\boldsymbol{b}_1^\intercal\boldsymbol{x}_{1k}}{\pi_{1k}} + (1-\pi_{1k})\pi_{1k}\boldsymbol{b}_1^\intercal\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal\boldsymbol{b}_1 - \frac{(1-\pi_{1k})\pi_{1k}y_{2k}\boldsymbol{b}_1^\intercal\boldsymbol{x}_{1k}}{\pi_{1k}}\right] \\
&\quad + \sum_{k\in U}\left[(1-\pi_{2k}(\boldsymbol{I}_1))\frac{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)y_{2k}^2}{\pi_{1k}^2\pi_{2k}(\boldsymbol{I}_1)^2} - \frac{(1-\pi_{2k}(\boldsymbol{I}_1))\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)y_{2k}\boldsymbol{c}_2^\intercal\boldsymbol{x}_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right. \\
&\qquad \left. +(1-\pi_{2k}(\boldsymbol{I}_1))\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{c}_2^\intercal\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal\boldsymbol{c}_2 - \frac{(1-\pi_{2k}(\boldsymbol{I}_1))\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)y_{2k}\boldsymbol{c}_2^\intercal\boldsymbol{x}_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] + \boldsymbol{b}_1^\intercal\boldsymbol{D}\boldsymbol{b}_1.
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\mathrm{Var}^L[\widehat{t}_2^*] = &\sum_{k\in U}(1-\pi_{1k})\pi_{1k}\left(\frac{y_{2k}}{\pi_{1k}} - \boldsymbol{b}_1^\intercal\boldsymbol{x}_{1k}\right)^2 + \boldsymbol{b}_1^\intercal\boldsymbol{D}\boldsymbol{b}_1 \\
&+ \sum_{k\in U}(1-\pi_{2k}(\boldsymbol{I}_1))\pi_{2k}(\boldsymbol{I}_1)\pi_{1k}\left(\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} - \boldsymbol{c}_2^\intercal\boldsymbol{x}_{2k}\right)^2,
\end{aligned}
$$

where $\boldsymbol{b}_1^\intercal = \left[N^{-1}\sum_{i\in U}(1-\pi_{1i})y_{2i}\boldsymbol{x}_{1i}^\intercal\right]\left[N^{-1}\sum_{i\in U}\pi_{1i}(1-\pi_{1i})\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^\intercal\right]^{-1}$, $\boldsymbol{D} = \mathrm{Var}_p\left[\sum_{i\in \mathrm{S_P}} d_i\pi_{1i}\boldsymbol{x}_{1i}\right]$ and $\boldsymbol{c}_2^\intercal = \left[N^{-1}\sum_{i\in U}(1-\pi_{2i}(\boldsymbol{I}_1))y_{2i}\boldsymbol{x}_{2i}^\intercal\right]\left[N^{-1}\sum_{i\in U}\pi_{1i}\pi_{2i}(\boldsymbol{I}_1)(1-\pi_{2i}(\boldsymbol{I}_1))\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}^\intercal\right]^{-1}$.

As a note, to obtain the asymptotic variance for $\widehat{\mu}_2^* := \sum_{k\in S_{\text{NP},2}} \widehat{w}_{2k}^* y_{2k} / \sum_{k\in S_{\text{NP},2}} \widehat{w}_{2k}^*$, we can modify the first element in (16) from $\sum_{k\in U}\left(I_{1k}I_{2k}\frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} - \mu_2\right)$ to $\sum_{k\in U}\left(I_{1k}I_{2k}\frac{y_{2k}-\mu_2}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right)$. Following the same procedure, we can obtain the asymptotic variance for $\widehat{\mu}_2^*$ by replacing every y_{2k} in $\text{Var}^L[\widehat{t}_2^*]$ with $y_{2k} - \mu_2$, and then multiply by $1/N^2$.

A.3 Proof of Proposition 2

The proof for Proposition 2 comes from the first-order Taylor series expansion of $\widehat{t}_2^*$ around $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, which essentially adapts the suggested "cookbook" approach from Beaumont et al. (2015) for our setup with unknown Phase 1 and (conditional) Phase 2 inclusion probabilities. The "cookbook" consists of the following steps: (i) linearize $\widehat{t}_2^*$ through a first-order Taylor expansion; (ii) express all sums over $S_{\text{NP},2}$ as sums over $S_{\text{NP},1}$; and (iii) treat I_{2k}, $\widehat{\pi}_{1k}$ and $\widehat{\pi}_{2k}(\boldsymbol{I}_1)$ as fixed and estimate the first-phase variance based on the Poisson sampling with $\widehat{\pi}_{1k}$.

Step 1: Linearize $\widehat{t}_2^*$ through a first-order Taylor expansion

By Equation (17) and $\widehat{t}_2^* = N\widehat{\mu}_{2,N}^*$,

$$
\begin{aligned}
\widehat{t}_2^* &= \sum_{k\in S_{\mathrm{NP},2}} \frac{y_{2k}}{\widehat{\pi}_{1k}\widehat{\pi}_{2k}(\boldsymbol{I}_1)} \\
&= \sum_{k\in S_{\mathrm{NP},2}} \left[\frac{1}{\pi_{1k}} - \frac{1-\pi_{1k}}{\pi_{1k}}\boldsymbol{x}_{1k}^\intercal(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\right]\left[\frac{1}{\pi_{2k}(\boldsymbol{I}_1)} - \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{2k}(\boldsymbol{I}_1)}\boldsymbol{x}_{2k}^\intercal(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] y_{2k} + o_p(Nn_1^{-1/2}) \\
&= \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] - \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{1-\pi_{1k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} y_{2k}\boldsymbol{x}_{1k}^\intercal\right](\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) \\
&\quad - \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} y_{2k}\boldsymbol{x}_{2k}^\intercal\right](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + o_p(Nn_1^{-1/2}) \\
&= \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] - \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{1-\pi_{1k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} y_{2k}\boldsymbol{x}_{1k}^\intercal\right]\left\{\left[\sum_{k\in U}\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal\right]^{-1}\right. \\
&\quad \left.\left[\sum_{k\in U}\left(I_{1k}\boldsymbol{x}_{1k} - I_k^{\mathrm{P}} d_k\pi_{1k}\boldsymbol{x}_{1k}\right)\right]\right\} \\
&\quad - \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{1-\pi_{2k}(\boldsymbol{I}_1)}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)} y_{2k}\boldsymbol{x}_{2k}^\intercal\right]\left\{\left[\sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal\right]^{-1}\left[\sum_{k\in U}\left(I_{1k}I_{2k}\boldsymbol{x}_{2k} - I_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right)\right]\right\} \\
&\quad + o_p(Nn_1^{-1/2}) \\
&= \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] - \left[\sum_{k\in U}(1-\pi_{1k})y_{2k}\boldsymbol{x}_{1k}^\intercal\right]\left\{\left[\sum_{k\in U}\pi_{1k}(1-\pi_{1k})\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}^\intercal\right]^{-1}\left[\sum_{k\in U}\left(I_{1k}\boldsymbol{x}_{1k} - I_k^{\mathrm{P}} d_k\pi_{1k}\boldsymbol{x}_{1k}\right)\right]\right\} \\
&\quad - \left[\sum_{k\in U}(1-\pi_{2k}(\boldsymbol{I}_1))y_{2k}\boldsymbol{x}_{2k}^\intercal\right]\left\{\left[\sum_{k\in U}\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)(1-\pi_{2k}(\boldsymbol{I}_1))\boldsymbol{x}_{2k}\boldsymbol{x}_{2k}^\intercal\right]^{-1}\left[\sum_{k\in U}\left(I_{1k}I_{2k}\boldsymbol{x}_{2k} - I_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right)\right]\right\} \\
&\quad + o_p(Nn_1^{-1/2}) \\
&= \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] - \boldsymbol{b}_1^\intercal\left\{\left[\sum_{k\in U}\left(I_{1k}\boldsymbol{x}_{1k} - I_k^{\mathrm{P}} d_k\pi_{1k}\boldsymbol{x}_{1k}\right)\right]\right\} - \boldsymbol{c}_2^\intercal\left\{\left[\sum_{k\in U}\left(I_{1k}I_{2k}\boldsymbol{x}_{2k} - I_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right)\right]\right\} \\
&\quad + o_p(Nn_1^{-1/2}) \\
&= \left[\sum_{k\in S_{\mathrm{NP},2}} \frac{y_{2k}}{\pi_{1k}\pi_{2k}(\boldsymbol{I}_1)}\right] - \boldsymbol{b}_1^\intercal\left[\sum_{k\in S_{\mathrm{NP},1}}\boldsymbol{x}_{1k}\right] + \boldsymbol{b}_1^\intercal\left[\sum_{k\in S_{\mathrm{P}}} d_k\pi_{1k}\boldsymbol{x}_{1k}\right] - \boldsymbol{c}_2^\intercal\left[\sum_{k\in S_{\mathrm{NP},1}} I_{2k}\boldsymbol{x}_{2k}\right] + \boldsymbol{c}_2^\intercal\left[\sum_{k\in S_{\mathrm{NP},1}}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right] \\
&\quad + o_p(Nn_1^{-1/2}),
\end{aligned}
$$

where we replace $(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ and $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ with the first-order approximation of $\Phi_n(\widehat{\boldsymbol{\eta}})$ around $\boldsymbol{\eta}$ in (18).

Step 2: Express all sums over $S_{\mathbf{NP},2}$ as sums over $S_{\mathbf{NP},1}$

$$\begin{aligned}\widehat{t}_2^* &= \left[\sum_{k\in S_{\mathrm{NP},1}} \frac{1}{\pi_{1k}}\left(\frac{I_{2k}y_{2k}}{\pi_{2k}(\boldsymbol{I}_1)} - \pi_{1k}\boldsymbol{b}_1^\intercal \boldsymbol{x}_{1k} - \pi_{1k}I_{2k}\boldsymbol{c}_2^\intercal \boldsymbol{x}_{2k} + \boldsymbol{c}_2^\intercal \pi_{2k}(\boldsymbol{I}_1)\boldsymbol{x}_{2k}\right)\right] + \boldsymbol{b}_1^\intercal\left[\sum_{k\in S_{\mathrm{P}}} d_k\pi_{1k}\boldsymbol{x}_{1k}\right] + o_p(Nn_1^{-1/2}) \\ &= \left[\sum_{k\in S_{\mathrm{NP},1}} \frac{1}{\pi_{1k}}a_k\right] + \boldsymbol{b}_1^\intercal\left[\sum_{k\in S_{\mathrm{P}}} d_k\pi_{1k}\boldsymbol{x}_{1k}\right] + o_p(Nn_1^{-1/2}),\end{aligned}$$

where

$$a_k := \frac{y_{2k}I_{2k}}{\pi_{2k}(\boldsymbol{I}_1)} - \pi_{1k}\boldsymbol{b}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{1k}} - \pi_{1k}I_{2k}\boldsymbol{c}_{\boldsymbol{2}}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{2k}} + \pi_{1k}\pi_{2k}(\boldsymbol{I}_1)\boldsymbol{c}_{\boldsymbol{2}}^{\boldsymbol{T}}\boldsymbol{x}_{2k}.$$

This concludes the proof of Part (i) in Proposition 2.

Under Assumptions A1 – A4, we define

$$\mathrm{Var}^{\mathrm{Alt}}[\widehat{t}_2^*] := \left(\sum_{k\in U}\frac{1}{\pi_{1k}^2}a_k^2\mathrm{Var}[I_{1k}]\right) + \boldsymbol{b}_1^\intercal\boldsymbol{D}\boldsymbol{b}_1 = \left(\sum_{k\in U}\frac{1-\pi_{1k}}{\pi_{1k}}a_k^2\right) + \boldsymbol{b}_1^\intercal\boldsymbol{D}\boldsymbol{b}_1. \tag{19}$$

Step 3: Treat I_{2k}, π_{1k} and $\pi_{2k}(\boldsymbol{I}_1)$ as fixed and estimate the first-phase variance based on the Poisson sampling with $\widehat{\pi}_{1k}$

Therefore, a consistent estimator for $\mathrm{Var}^{\mathrm{Alt}}[\widehat{t}_2^*]$ is

$$\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*] = \left(\sum_{k\in S_{\mathrm{NP},1}}\frac{1-\widehat{\pi}_{1k}}{\widehat{\pi}_{1k}^2}\widehat{a}_k^2\right) + \widehat{\boldsymbol{b}}_1^\intercal\widehat{\boldsymbol{D}}\widehat{\boldsymbol{b}}_1.$$

Following Section 2.1 of Beaumont et al. (2015) with the Poisson sampling at the second phase (our Assumptions A1–A5) and the consistency of $\widehat{\boldsymbol{b}}_1$, $\widehat{\boldsymbol{c}}_2$, $\widehat{\pi}_1$ and $\widehat{\pi}_2(\boldsymbol{I}_1)$, we have $\widehat{\mathrm{Var}}^{\mathrm{Alt}}[\widehat{t}_2^*] - \mathrm{Var}[\widehat{t}_2^*] = o_p(N^2n_1^{-1}) + O(n_1/N)$. This concludes the proof of Part (ii) in Proposition 2.

A.4 Proof of Proposition 3

The proof in this subsection shows the validity of the pseudo-population bootstrap (PPB) approach for variance estimation in Section 3.2. Specifically, the goal is to show that $\widehat{\text{Var}}^{\text{PPB}}[\widehat{t}_2^*]$ is consistent for $\text{Var}[\widehat{t}_2^*]$ if the sampling fractions n_1/N and n_p/N are negligible.

Let $p^{(\cdot)}$ and $1^{(\cdot)}$ denote the bootstrap sampling mechanisms that lead to $S_{\text{P}}^{(c)}$ and $S_{\text{NP}}^{(c)}$, $c = 1, \ldots, C$, respectively. From (19), we have

$$\begin{aligned} E_{p^{(\cdot)}} E_{1^{(\cdot)}} \left[\widehat{\text{Var}}^{\text{PPB}}[\widehat{t}_2^*] \right] &= \text{Var}^{\text{Alt}}[\widehat{t}_2^* \mid U_{\text{NP}}, U_{\text{P}}] + o(N^{(\cdot)2} n_2^{(\cdot)-1}) \\ &= \left(\sum_{k \in U_{\text{NP}}} \frac{1 - \pi_{1k}^{(\cdot)}}{\pi_{1k}^{(\cdot)}} a_k^{(\cdot)2} \right) + \boldsymbol{b}_1^{(\cdot)\intercal} \boldsymbol{D}^{(\cdot)} \boldsymbol{b}_1^{(\cdot)} + o(N^{(\cdot)2} n_1^{(\cdot)-1}), \end{aligned}$$

where $E_{p^{(\cdot)}}$ and $E_{1^{(\cdot)}}$ denote the expectation with respect to mechanisms $p^{(\cdot)}$ and $1^{(\cdot)}$, respectively. $\pi_{1k}^{(\cdot)}$, $a_k^{(\cdot)}$, $\boldsymbol{b}_1^{(\cdot)}$, $\boldsymbol{D}^{(\cdot)}$ and $n_2^{(\cdot)}$ are defined the same way as their counterparts without the superscript $(\cdot)$, but under pseudo-populations U_{NP} and U_{P}. $N^{(\cdot)}$ is the maximum of N_{P} and N_{NP}. The first term is

$$\begin{aligned} \sum_{k \in U_{\text{NP}}} \frac{1 - \pi_{1k}^{(\cdot)}}{\pi_{1k}^{(\cdot)}} a_k^{(\cdot)2} &= \sum_{k \in S_{\text{NP},1}} \lfloor \pi_{1k}^{(\cdot)-1} \rfloor \frac{1 - \pi_{1k}^{(\cdot)}}{\pi_{1k}^{(\cdot)}} a_k^{(\cdot)2} \\ &= \sum_{k \in S_{\text{NP},1}} [\pi_{1k}^{(\cdot)-1} - (\pi_{1k}^{(\cdot)-1} - \lfloor \pi_{1k}^{(\cdot)-1} \rfloor)] \frac{1 - \pi_{1k}^{(\cdot)}}{\pi_{1k}^{(\cdot)}} a_k^{(\cdot)2} \\ &\approx \sum_{k \in S_{\text{NP},1}} \pi_{1k}^{(\cdot)-1} \frac{1 - \pi_{1k}^{(\cdot)}}{\pi_{1k}^{(\cdot)}} a_k^{(\cdot)2}. \end{aligned}$$

The assumption of a negligible n_1/N justifies the above approximation. Intuitively, if the Phase 1 sampling fraction is negligible, the number of population units a sampled unit represents is large, and hence, the weight attached to the sampled unit is large. Therefore, the fractional part of the weight is negligible compared with the integer part of the weight. Indeed, for every k, the magnitude of the fractional part of $\pi_{1k}^{(\cdot)}$, $(\pi_{1k}^{(\cdot)-1} - \lfloor \pi_{1k}^{(\cdot)-1} \rfloor)$, to $\pi_{1k}^{(\cdot)-1}$ is of order $\frac{\pi_{1k}^{(\cdot)-1} - \lfloor \pi_{1k}^{(\cdot)-1} \rfloor}{\pi_{1k}^{(\cdot)-1}} = \frac{O(1)}{O(N^{(\cdot)}/n_1^{(\cdot)})} = O\left(\frac{n_1^{(\cdot)}}{N^{(\cdot)}}\right) = O\left(\frac{n_1}{N}\right)$, where the first equality comes from the fact that $\pi_{1k}^{(\cdot)-1} -$

$\lfloor \pi_{1k}^{(\cdot)-1} \rfloor$ is a fraction and condition C3. It can be shown that $NN^{(\cdot)-1} = 1 + o_p(1) + O(n_1/N)$ and $n_1 n_1^{(\cdot)-1} = 1 + o_p(1) + O(n_1/N)$. Therefore, the fractional part of $\pi_{1k}^{(\cdot)-1}$ is negligible if n_1/N is negligible.

It can also be shown that $a_k^{(\cdot)} - a_k = o_p(1) + O(n_1/N)$, $\boldsymbol{b}_1^{(\cdot)} - \boldsymbol{b}_1 = o_p(1) + O(n_1/N)$, $\boldsymbol{c}_2^{(\cdot)} - \boldsymbol{c}_2 = o_p(1) + O(n_1/N)$ and $\boldsymbol{D}^{(\cdot)} - \boldsymbol{D} = o_p(1) + O(n_{\mathrm{P}}/N)$ when n_1/N and n_{P}/N are negligible. We also have $\pi_{1k}^{(\cdot)} = \widehat{\pi}_{1k}$. Therefore, we have

$$E\left[\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widehat{t}_2^*]\right] = \mathrm{Var}[\widehat{t}_2^*] + o(N^2 n_1^{-1}) + O(n_1/N) + O(n_{\mathrm{P}}/N),$$

where the equality comes from Proposition 2, following Equation (A.30) in Chen et al. (2019). Therefore, we have $\widehat{\mathrm{Var}}^{\mathrm{PPB}}[\widehat{t}_2^*]$ consistent for $\mathrm{Var}[\widehat{t}_2^*]$, if n_1/N and n_{P}/N are negligible. This concludes the proof for Proposition 3.

Data Availability Statement

The paper uses data from the Bank of Canada 2020 November 2020 Cash Alternative Survey. Requests related to the Bank of Canada 2020 November 2020 Cash Alternative Survey can be directed to the Bank of Canada's Data Statistics Office Meta Data Repository (MEDR) email address: medrsa@ bankofcanada.ca. In addition, we have prepared detailed replication file that consists of R programs that were used for this manuscript. The programs can be requested from the Bank of Canada's Data Statistics Office Meta Data Repository (MEDR) email address: medrsa@ bankofcanada.ca.

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):524–542.

Beaumont, J.-F. and Haziza, D. (2016). A note on the concept of invariance in two-phase sampling designs. *Survey Methodology*, 42(2):319–323.

Beaumont, J.-F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80(1):127–148.

Bessonneau, P., Brilhaut, G., Chauvet, G., and Garcia, C. (2021). With-replacement bootstrap variance estimation for household surveys: Principles, examples and implementation. *Survey methodology*, 47(2):313–.

Binder, D. A., Babyak, C., Brodeur, M., Hidiroglou, M., and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *Canadian Journal of Statistics*, 28(4):751–764.

Brockwell, P. J. (1991). *Time series: Theory and methods*. Springer-Verlag.

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. Technical report, Bank of Canada.

Chen, S., Haziza, D., Léger, C., and Mashreghi, Z. (2019). Pseudo-population bootstrap methods for imputed survey data. *Biometrika*, 106(2):369–384.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532):2011–2021.

Cloutier, E. C. and Langlet, É. (2014). *Aboriginal Peoples Survey, 2012: Concepts and methods guide*. Statistics Canada.

Cohen, N., Ben-Hur, D., and Burck, L. (2017). Variance estimation in multi-phase calibration. *Survey Methodology*, 43(1):125–.

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, volume 1814184. American Statistical Association, Alexandria, VA.

Hájek, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, part I', by D. Basu. *Foundations of Statistical Inference*, 326.

Hartman, E. and Huang, M. (2024). Sensitivity analysis for survey weights. *Political Analysis*, 32(1):1–16.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75(1):25–43.

Henry, C. S., Shimoda, M., and Zhu, J. (2022). 2021 methods-of-payment survey report. Technical report, Bank of Canada.

Hidiroglou, M. and Särndal, C. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24:11–20.

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

Huber, M. (2014). Treatment evaluation in the presence of sample selection. *Econometric Reviews*, 33(8):869–905.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4):501–514.

Kim, J. K., Navarro, A., and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101(473):312–320.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3):941–963.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10(none).

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21(1):43–52.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83:242–272.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

Wang, L., Valliant, R., and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24):5237–5250.

Wang, Z., Peng, L., and Kim, J. K. (2022). Bootstrap inference for the finite population mean under complex sampling designs. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 84(4):1150–1174.

Welte, A. and Wu, J. (2023). The 2021–22 merchant acceptance survey pilot study. Technical report, Bank of Canada.

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

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3:625–650.