Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Improved Horvitz-Thompson estimator in survey sampling

by Xianpeng Zong, Rong Zhu and Guohua Zou

Release date: May 7, 2019

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, 2019 (special issue)
Vol. 45, No. 1, pp. 165-184
Statistics Canada, Catalogue No. 12-001-X

Improved Horvitz-Thompson estimator in survey sampling

Xianpeng Zong, Rong Zhu and Guohua Zou[1]

Abstract

The Horvitz-Thompson (HT) estimator is widely used in survey sampling. However, the variance of the HT estimator becomes large when the inclusion probabilities are highly heterogeneous. To overcome this shortcoming, in this paper we propose a hard-threshold method for the first-order inclusion probabilities. Specifically, we carefully choose a threshold value, then replace the inclusion probabilities smaller than the threshold by the threshold. Through this shrinkage strategy, we construct a new estimator called the improved Horvitz-Thompson (IHT) estimator to estimate the population total. The IHT estimator increases the estimation accuracy much although it brings a bias which is relatively small. We derive the IHT estimator's mean squared error and its unbiased estimator, and theoretically compare the IHT estimator with the HT estimator. We also apply our idea to construct an improved ratio estimator. We numerically analyze simulated and real data sets to illustrate that the proposed estimators are more efficient and robust than the classical estimators.

Key Words: Horvitz-Thompson estimator; Inverse probability weighting; Hard-threshold; Robustness; Unequal probability sampling; Sampling without/with replacement; Ratio estimator.

1 Introduction

The Horvitz-Thompson (HT) estimator proposed by Horvitz and Thompson (1952) is widely used in survey sampling. It has also been applied to other fields such as functional data analysis (Cardot and Josserand, 2011) and the treatment effect (Rosenbaum, 2002). The HT estimator is an unbiased estimator constructed via inverse probability weighting. However, when the inclusion probabilities are highly heterogeneous, i.e., inclusion probabilities of some units are relatively tiny, the variance of the HT estimator becomes large due to inverse probability weighting. In this paper, we propose an improved Horvitz-Thompson (IHT) estimator to address this problem.

Our approach is to use a hard-threshold for the first-order inclusion probabilities. Specifically, we carefully choose an inclusion probability as the threshold. The inclusion probabilities that are smaller than the threshold are replaced by the threshold, while the others remain unchanged. In this way, we obtain the modified inclusion probabilities, and construct an estimator based on the modified inclusion probabilities through inverse probability weighting. We call this estimator the IHT estimator. This method looks very easy but is more efficient than the HT estimator. This hard-threshold approach can be explained as a shrinkage method. Shrinkage is very commonly used in statistics, such as ridge regression (Hoerl and Kennard, 1970) and high-dimensional statistics (Tibshirani, 1996). In this paper, we use it to reduce the negative effect of highly heterogeneous inclusion probabilities. Similar to other shrinkage methods, our approach introduces a bias, which is proved to be very small, but reduces the variance to a larger extent, so it improves the estimation efficiency. We will theoretically and numerically show the improvement from using the modified inclusion probabilities. In addition to the population total estimator, we also extend this strategy to the ratio estimator, and accordingly, an improved ratio estimator is obtained.

1. Xianpeng Zong, School of Mathematical Sciences, Capital Normal University, Beijing, 100048, China; Rong Zhu, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. E-mail: rongzhu@amss.ac.cn; Guohua Zou, School of Mathematical Sciences, Capital Normal University, Beijing, 100048, China.

The remainder of the paper is organized as follows. Section 2 introduces the HT estimator and shows its drawback. Section 3 proposes our modified inclusion probabilities and the resultant IHT estimator. We also provide the IHT estimator's properties, and theoretically compare it with the HT estimator in this section. Section 4 extends our idea to obtain an improved ratio estimator and shows that this modification is efficient. Section 5 presents numerical evidence from simulations and a real data analysis. Section 6 concludes. Proofs of theoretical results are given in the Appendix.

2 HT estimator and its drawback

Consider a finite population $U = \{U_1, \ldots, U_N\}$ of size N, where U_k denotes the k^{th} unit. For simplicity, we write $U = \{1, \ldots, k, \ldots, N\}$. For each unit k, suppose that the value y_k of the target characteristic Y is measured. Our aim is to estimate the total, $t_y = \sum_U y_k$, using a sample s of size n which is randomly drawn from the population U. We implement unequal probability sampling without replacement. Denote $\{\pi_k\}_{k=1}^N$ as the first-order inclusion probabilities and $\{\pi_{kl}\}_{k \neq l}$ as the second-order inclusion probabilities.

Horvitz and Thompson (1952) proposed the HT estimator as follows

$$\hat{t}_{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

The HT estimator $\hat{t}_{\text{HT}}$ is an unbiased estimator of t_y and its variance is

$$V(\hat{t}_{\text{HT}}) = \sum_U \frac{\Delta_{kk}}{\pi_k^2} y_k^2 + \sum\sum_{k \neq l}{}_U \frac{\Delta_{kl}}{\pi_k \pi_l} y_l y_k, \tag{2.1}$$

where $\Delta_{kk} = \pi_k - \pi_k^2$ for all k and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ for all $k \neq l$. When the inclusion probabilities are highly imbalanced, i.e., some π_k's are very small, the variance of the HT estimator may be very large.

3 Improved HT estimator

In this section, we improve the HT estimator in the sense of reducing its mean squared error (MSE). The resultant estimator is referenced as the IHT estimator. For doing this, we first propose the modified first-order inclusion probabilities, where the hard-threshold method is used to reduce the effect of those inclusion probabilities with relatively tiny values.

Definition 1. *Let* $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(N)}$ *be the ordered values of the first-oder inclusion probabilities* $\{\pi_1, \pi_2, \ldots, \pi_N\}$. *Assume that there exists an integer* $K \geq 2$ *such that* $\pi_{(K)} \leq (K+1)^{-1}$. *We define the modified first-order inclusion probabilities as follows*

$$\pi_k^* = \begin{cases} \pi_k & \pi_k > \pi_{(K)}, \\ \pi_{(K)} & \pi_k \leq \pi_{(K)}, \end{cases} \qquad 1 \leq k \leq N.$$

From the definition, we partition the finite population into two parts: $U_1 = \{k : \pi_k > \pi_{(K)}\}$ with size $N - K$, and $U_2 = \{k : \pi_k \le \pi_{(K)}\}$ with size K. For U_1, the first-order inclusion probabilities remain unchanged, while all of first-order inclusion probabilities for U_2 are replaced by $\pi_{(K)}$. From this hard-threshold, we get our modified first-order inclusion probabilities $\{\pi_k^*\}_{k=1}^N$. Obviously, the choice of K is very important. In Section 3.2, we shall provide a simple way to choose K.

Remark on existence of K. The assumption in Definition 1 is quite weak. If $\pi_{(2)} > 1/(2+1)$, then the sampling fraction $f > \frac{1}{3} - \frac{1}{3N}$. However that situation that $f > \frac{1}{3}$ rarely happens in practical surveys. Thus, the inequality that $\pi_{(2)} \le 1/(2+1)$ generally holds.

Instead of the original first-order inclusion probabilities $\{\pi_k\}_{k=1}^N$, we use our defined modified first-order inclusion probabilities $\{\pi_k^*\}_{k=1}^N$ to construct an improved Horvitz-Thompson (IHT) estimator by inverse probability weighting.

Definition 2. *The IHT estimator is defined as*

$$\hat{t}_{IHT} = \sum_{k \in s} \frac{y_k}{\pi_k^*}.$$

Unlike the unbiased HT estimator, the IHT estimator is biased. However, this modification leads to much smaller MSE due to reducing the variance. It is worth pointing out that, although we focus on sampling without replacement in this paper, our modification idea is equally applicable to the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) for sampling with replacement.

3.1 Properties of the IHT estimator

In this section, we derive the properties of the IHT estimator. We first provide the expressions of its bias, variance, MSE and an unbiased estimator of MSE in Theorem 1. Then we compare the IHT estimator with the HT estimator in Theorems 2 and 3.

Theorem 1. *The bias and variance of the IHT estimator* $\hat{t}_{IHT}$ *are expressed as*

$$Bias(\hat{t}_{IHT}) = \sum\nolimits_{U_2} \left(\frac{\pi_k}{\pi_{(K)}} - 1 \right) y_k,$$

and

$$Var(\hat{t}_{IHT}) = \sum\nolimits_{U} \frac{\Delta_{kk}}{\pi_k^{*2}} y_k^2 + \sum\sum_{k \ne l} \nolimits_{U} \frac{\Delta_{kl}}{\pi_k^* \pi_l^*} y_k y_l,$$

respectively, where $\Delta_{kk} = \pi_k (1 - \pi_k)$, $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l \ (k \ne l)$ *as defined before. Therefore, its MSE is given by*

$$MSE(\hat{t}_{IHT}) = \left[\sum\nolimits_{U_2} \left(\frac{\pi_k}{\pi_{(K)}} - 1 \right) y_k \right]^2 + \sum\nolimits_{U} \frac{\Delta_{kk}}{\pi_k^{*2}} y_k^2 + \sum\sum_{k \ne l} \nolimits_{U} \frac{\Delta_{kl}}{\pi_k^* \pi_l^*} y_k y_l. \qquad (3.1)$$

An unbiased estimator of the MSE is

$$\widehat{MSE}\left(\hat{t}_{IHT}\right) = \sum_{s_2} \frac{\left(\pi_k - \pi_{(K)}\right)^2}{\pi_{(K)}^2 \pi_k} y_k^2 + \sum\sum_{k \neq l}{}_{s_2} \frac{\left(\pi_k - \pi_{(K)}\right)\left(\pi_l - \pi_{(K)}\right)}{\pi_{(K)}^2 \pi_{kl}} y_k y_l$$
$$+ \sum_s \frac{\breve{\Delta}_{kk}}{\pi_k^{*2}} y_k^2 + \sum\sum_{k \neq l}{}_s \frac{\breve{\Delta}_{kl}}{\pi_k^* \pi_l^*} y_k y_l,$$

where $\breve{\Delta}_{kk} = \frac{\Delta_{kk}}{\pi_k}$, $\breve{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$, s *is the sample set, and* $s_2 = s \cap U_2$.

Proof. See Appendix A.1.

To derive the properties of the IHT estimator, we need the following regularity conditions:

Condition C.1. $\min_{i \in U} \pi_i \geq \lambda > 0$, $\min_{i,j \in U} \pi_{ij} \geq \lambda^* > 0$, *and*

$$\lim \sup_{N arrow \infty} n \max_{i \neq j \in U} \left| \pi_{ij} - \pi_i \pi_j \right| < \infty.$$

Condition C.2. $\max_{i \in U} |y_i| \leq C$ *with* C *a positive constant not depending on* N.

Condition C.1 is a common condition imposed on the first-order and second-order inclusion probabilities. The same conditions are used in Breidt and Opsomer (2000), where further comments on C.1 are provided. Condition C.2 is also a common condition.

Theorem 2. *For the HT estimator* $\hat{t}_{HT}$ *and the IHT estimator* $\hat{t}_{IHT}$, *under the Conditions C.1-C.2, we have*

$$Bias\left(N^{-1}\hat{t}_{HT}\right) = 0, \qquad Bias\left(N^{-1}\hat{t}_{IHT}\right) = O\left(n^{-1}\right);$$

and

$$MSE\left(N^{-1}\hat{t}_{HT}\right) = O\left(n^{-1}\right), \qquad MSE\left(N^{-1}\hat{t}_{IHT}\right) = O\left(n^{-1}\right).$$

Proof. See Appendix A.2.

From Theorem 2, the squared-bias of our IHT estimator is very small compared to its MSE. Although our IHT estimator brings a bias to reduce the variance, the price for this is relatively small. The following theorem theoretically compares the efficiency of the two estimators.

Theorem 3. *Under the Conditions C.1-C.2, we have*

$$MSE\left(N^{-1}\hat{t}_{IHT}\right) \leq MSE\left(N^{-1}\hat{t}_{HT}\right) + o\left(n^{-1}\right). \tag{3.2}$$

Especially, for Poisson sampling, we obtain

$$MSE\left(N^{-1}\hat{t}_{IHT}\right) \leq MSE\left(N^{-1}\hat{t}_{HT}\right),$$

where the strict inequality is true if there exist $k \neq l \in U_2$ *such that* $\left(\pi_k - \pi_{(K)}\right) y_k \neq \left(\pi_l - \pi_{(K)}\right) y_l$.

Proof. See Appendix A.3.

Theorem 3 shows that, under some mild conditions, the proposed IHT estimator is asymptotically more efficient than the HT estimator. From the proof in Appendix A.3, the term $o(n^{-1})$ in equation (3.2) is due to the interaction term from the second-order inclusion probabilities. We theoretically bound the term as $o(n^{-1})$. For Poisson sampling, the term does not exist, so the MSE of the IHT estimator is uniformly not larger than that of the HT estimator. Empirically, we compare the IHT estimator with the HT estimator in Section 5.

3.2 The choice of K

The efficiency of the IHT estimator relies on the choice of K, which provides a control of the variance-and-bias tradeoff. The choice of K needs to satisfy the condition that $\pi_{(K)} < 1/(K+1)$ of Definition 1, since the modified inclusion probabilities would cause large bias when K becomes large. On the other hand, the improvement of the IHT estimator would not be significant if K is small. In the proofs of Theorem 3, equation (A.5) provides a lower bound of the main term of $\operatorname{MSE}(N^{-1}\hat{t}_{\mathrm{HT}}) - \operatorname{MSE}(N^{-1}\hat{t}_{\mathrm{IHT}})$. The lower bound increases as $\pi_{(K)}$ increases. Therefore, denoting the maximum value $K^* = \max\{i : \pi_{(i)} \leq 1/(i+1)\}$, we choose K^* as the threshold. In practice, we propose the following algorithm to find the maximum value K^*.

Algorithm 1	**The choice of K**
Step (i)	Obtain the ordered inclusion probabilities $\{\pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(N)}\}$ by sorting $\{\pi_k\}_{k=1}^N$ from small to large.
Step (ii)	Test and modify. If j satisfies $\pi_{(j)} \leq \frac{1}{j+1}$ and $\pi_{(j+1)} > \frac{1}{j+2}$, the modified first-order inclusion probabilities are defined as $\boldsymbol{\pi}^* = \{\underbrace{\pi_{(j)}, \ldots, \pi_{(j)}}_{j-1}, \pi_{(j)}, \pi_{(j+1)}, \ldots, \pi_{(N)}\},$ and $K = j$.

Note that the choice of K^* based on Algorithm 1 is not optimal in terms of MSE. However, we simulate an example in Section 5 where the performance of Algorithm 1 is very close to that of the theoretically ideal choice.

4 Extension to the ratio estimator

When an auxiliary variable is available, the ratio estimator is usually used to estimate the population total. In this section, we extend the IHT estimator to the case of ratio estimation.

4.1 Improved ratio estimator

Denote the ratio between the population totals of Y and Z as

$$R = t_y \big/ t_z ,$$

where t_y and t_z are the totals of the finite populations Y and Z, respectively. Let $\hat{t}_{y\pi} = \sum_s \frac{y_k}{\pi_k}$, $\hat{t}_{z\pi} = \sum_s \frac{z_k}{\pi_k}$, $\hat{t}^*_{y\pi} = \sum_s \frac{y_k}{\pi^*_k}$, and $\hat{t}^*_{z\pi} = \sum_s \frac{z_k}{\pi^*_k}$. The classical estimator and our modified estimator of R are given by

$$\hat{R} = \hat{t}_{y\pi} \big/ \hat{t}_{z\pi} , \quad \text{and} \quad \hat{R}^* = \hat{t}^*_{y\pi} \big/ \hat{t}^*_{z\pi} .$$

We assume that the population total t_z is known. To estimate the population total t_y of Y, the classical ratio estimator is given by

$$\hat{Y}_R = t_z \cdot \hat{t}_{y\pi} \big/ \hat{t}_{z\pi} .$$

Alternatively, our improved ratio estimator of t_y based on the modified inclusion probabilities is expressed as

$$\hat{Y}^*_R = t_z \cdot \hat{t}^*_{y\pi} \big/ \hat{t}^*_{z\pi} .$$

4.2 Properties of the improved ratio estimator

To show theoretically that the improved ratio estimator $\hat{Y}^*_R$ is more efficient than the classical ratio estimator $\hat{Y}_R$, we need the following regularity conditions:

Condition C.3. $\lim_{N\to\infty} \frac{n}{N} = c$, *where* $c \in (0, 1)$ *is a constant.*

Condition C.4. $\max_{i\neq j\neq k\in U} \left(\pi_{ijk} - \pi_{ij}\pi_k\right) = O\left(n^{-1}\right)$, *and*

$$\max_{i\neq j\neq k\neq l\in U} \left(\pi_{ijkl} - 4\pi_{ijk}\pi_l + 6\pi_{ij}\pi_k\pi_l - 3\pi_i\pi_j\pi_k\pi_l\right) = O\left(n^{-2}\right).$$

Condition C.3 is a common condition. The same condition is used in Breidt and Opsomer (2000). Condition C.4 is a mild assumption on the third-order and fourth-order inclusion probabilities. In Appendix A.5, we present some frequent examples which satisfy Condition C.4.

Comparing our improved estimators with the classical estimators, we have the following result.

Theorem 4. *If Conditions C.1-C.4 are satisfied, and* $c_1 \leq z_k \leq c_2$ *for all* $k \in U$ *with* c_1 *and* c_2 *some positive constants, then*

$$MSE\left(\hat{R}^*\right) \leq MSE\left(\hat{R}\right) + o\left(n^{-1}\right).$$

Furthermore,

$$MSE\left(N^{-1}\hat{Y}^*_R\right) \leq MSE\left(N^{-1}\hat{Y}_R\right) + o\left(n^{-1}\right).$$

Proof. See Appendix A.4.

Like Theorem 3, Theorem 4 shows that the proposed method improves the classical ratio estimators with a tolerance of order $o(n^{-1})$.

5 Numerical studies

In this section, we assess the empirical performance of our IHT estimator using three synthetic examples and one real example. We consider the following two cases: the estimation of a population total and the estimation of a population ratio, where our IHT estimators are compared with the HT estimator. We measure the efficiency improvement in terms of $\text{Re} = \frac{|\text{MSE}^{\text{HT}} - \text{MSE}^{\text{IHT}}|}{\text{MSE}^{\text{HT}}} \times 100\%$, where MSE^{HT} and MSE^{IHT} denote the MSE of the HT estimators and IHT estimators, respectively. We additionally compare the IHT estimator with the HT estimator in the sense of inference performance in the real example.

5.1 Simulations

Example 1: An illustrative example

We generate a finite population Y of size $N = 3{,}000$, where the k^{th} unit value $y_k = |y_{0k}|$ and $y_{0k} \sim N(0, 1)$. Our aim is to estimate the population mean $\bar{Y} = \frac{1}{N}\sum_U y_k$. We perform Poisson sampling according to the inclusion probabilities set as follows

$$\pi_1 = \ldots = \pi_{1{,}000} = 0.2, \quad \pi_{1{,}001} = \ldots = \pi_{2{,}000} = 0.001, \quad \text{and} \quad \pi_{2{,}001} = \ldots = \pi_{3{,}000} = 0.08.$$

In this example, the HT estimator is not efficient since one third of the inclusion probabilities are 0.001, tiny relative to 0.08 or 0.2. From our hard-threshold strategy, we replace these tiny probabilities with 0.08, so the modified inclusion probabilities are given by

$$\pi_1^* = \ldots = \pi_{1{,}000}^* = 0.2, \quad \pi_{1{,}001}^* = \ldots = \pi_{2{,}000}^* = 0.08, \quad \text{and} \quad \pi_{2{,}001}^* = \ldots = \pi_{3{,}000}^* = 0.08.$$

Note that the modified probabilities are not obtained according to Algorithm 1. It is an illustrative example to show that our hard-threshold can bring efficiency improvement. By setting the iteration time $M = 2{,}000$, we get the simulated biases, variances and MSEs of our IHT estimator and the HT estimator. The results are shown in Table 5.1.

Table 5.1
Performance of Example 1

MSE^{HT}	MSE^{IHT}	Bias^{HT}	Bias^{IHT}	Var^{HT}	Var^{IHT}	Re
0.1187	0.0751	5.374×10^{-6}	0.0723	0.1187	0.0029	36.71%

From the table, the variance of the HT estimator is much larger than that of the IHT estimator, so it loses its efficiency in terms of MSE compared to the IHT estimator although the HT estimator is unbiased. Furthermore, in order to show the variations of both estimators, we plot their values among 2,000 iterations in Figure 5.1. It clearly displays that, although there is small bias for the IHT estimator, its variation is much less than that of the HT estimator. These observations empirically verify our theoretical results.

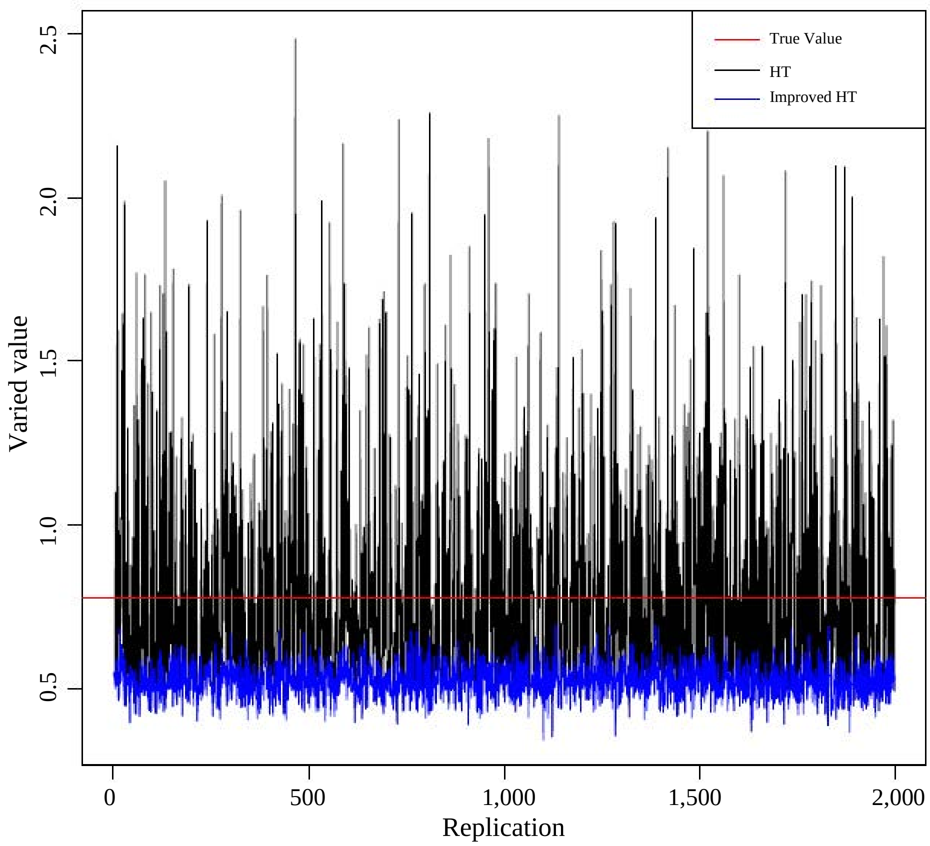


Figure 5.1 The plots of both estimators in Example 1.

Example 2: π_i's depend on an auxiliary variable

We generate the finite population Y of size $N = 3{,}000$ as follows: $y_k = \sqrt{3} \cdot \rho \cdot x_k + \sqrt{3 - 3\rho^2} \cdot |e_k|$, where x_k and e_k are independently generated from $U(0, 2)$ and $N(0, 1)$ respectively, and $0 \le \rho \le 1$ controling the correlation of Y and X. We consider three sampling methods: Poisson sampling, PPS sampling and πPS sampling. The sampling fraction $f = \frac{n}{N} = 0.02, 0.04, 0.06, 0.08, 0.10, 0.15, 0.20, 0.30$. We report the results in Figure 5.2, where $\rho = 0.8$, and list the specific Re values of Figure 5.2 in Table 5.4.

From these results, we get the same observations as Example 1. It indicates that our IHT estimator outperforms the HT estimator in terms of MSE and that the improvement is generally substantial. Comparing with Figures 5.2(a), 5.2(b), and 5.2(c), πPS sampling obtains the biggest advantage of the IHT estimator over the HT estimator, followed by PPS sampling and Poisson sampling. We also show the results for different ρ values under πPS sampling in Table 5.2, where the case of $f = 0.08$ is reported and other cases are ignored because of the similarity. It is observed from the table that, no matter what value ρ takes, the IHT estimator has uniformly much less MSE than the HT estimator.

Table 5.2
The performance of Example 2 for different ρ values, where $f = 0.08$

ρ	MSE^{HT}	MSE^{IHT}	Bias^{HT}	Bias^{IHT}	Var^{HT}	Var^{IHT}	Re
0	3.45×10^{-2}	1.36×10^{-2}	3.43×10^{-5}	5.82×10^{-4}	3.45×10^{-2}	1.30×10^{-2}	60.70%
0.1	2.51×10^{-2}	1.38×10^{-2}	1.16×10^{-5}	8.25×10^{-4}	2.51×10^{-2}	1.30×10^{-2}	44.91%
0.3	2.43×10^{-2}	1.24×10^{-2}	4.65×10^{-6}	8.86×10^{-4}	2.43×10^{-2}	1.15×10^{-2}	48.97%
0.5	2.38×10^{-2}	1.07×10^{-2}	9.83×10^{-6}	8.44×10^{-4}	2.38×10^{-2}	9.88×10^{-3}	54.92%
0.8	9.38×10^{-3}	5.22×10^{-3}	3.04×10^{-7}	3.16×10^{-4}	9.38×10^{-3}	4.91×10^{-3}	44.33%
0.9	4.75×10^{-3}	2.65×10^{-3}	7.98×10^{-6}	2.64×10^{-4}	4.74×10^{-3}	2.38×10^{-3}	44.27%

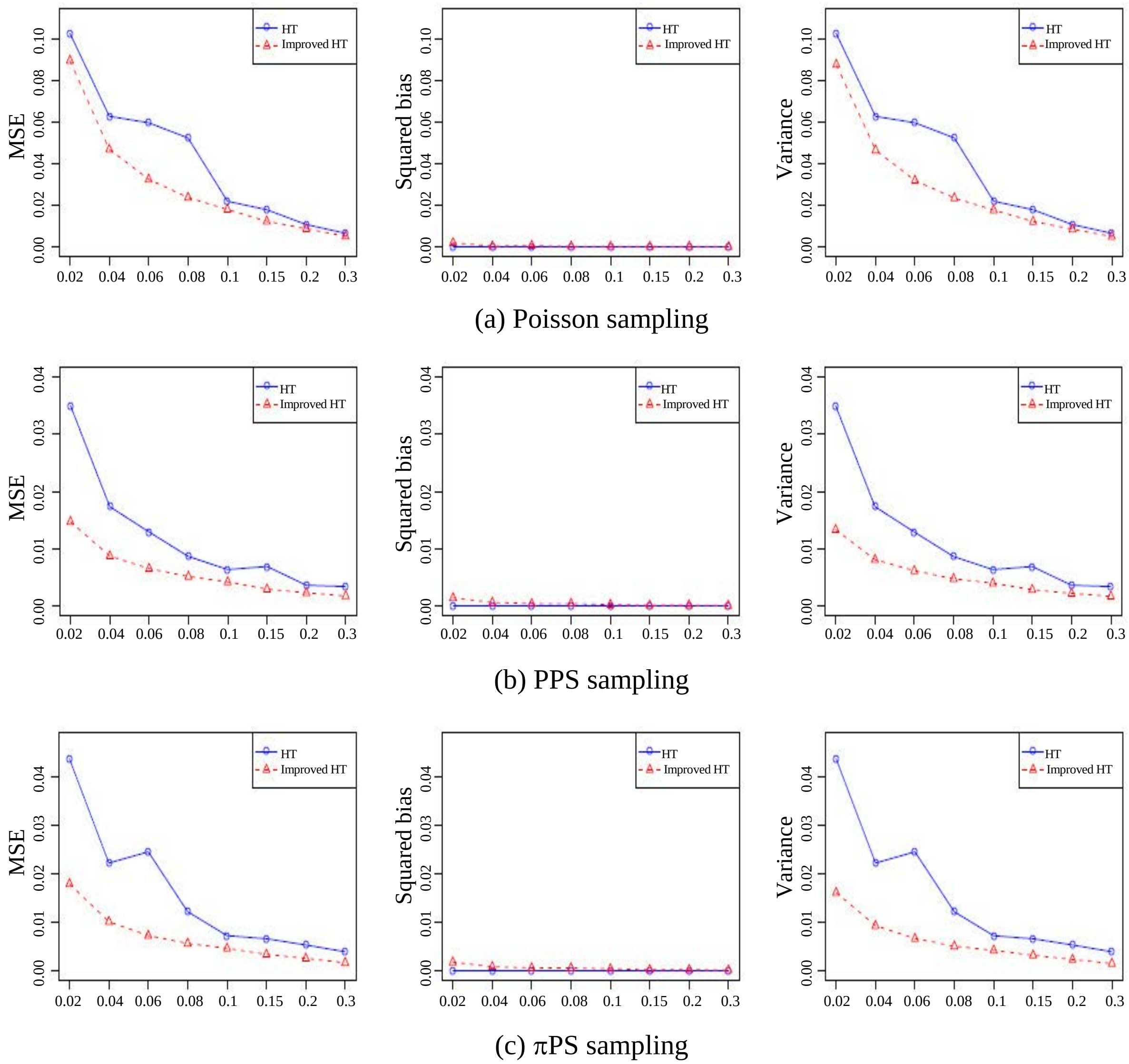


Figure 5.2 The performance of our IHT estimator and the HT estimator in Example 2, where $\rho = 0.8$. From left to right: the MSE performance, the squared-bias performance, and the variance performance.

Example 2 (continued): The performance of Algorithm 1 and the effect of the outcome's coefficient of variation

Here we empirically investigate the performance of Algorithm 1 and the effect of the outcome's coefficient of variation on our IHT estimator. We generate a finite population through a linear model with an intercept: $y_k = \alpha + \sqrt{3} \cdot \rho \cdot x_k + \sqrt{3 - 3\rho^2} \cdot |e_k|$, where x_k and e_k are the same as Example 2. We set $N = 1{,}000$, and control the coefficient of variation of the outcome by varying the intercept term $\alpha = \{-10, 5, 0, 5, 10\}$. Firstly, we study the performance of Algorithm 1. Note that the optimal choice K_{opt} can be derived via minimizing equation (3.1). We compare the MSE values based on K_{opt} and K^* from Algorithm 1, and report the results of $f = 0.03$ in Table 5.3 and ignore other cases because of the similarity. From the table, the MSE values based on K^* are very close to those based on K_{opt}. It indicates that Algorithm 1 provides an efficient choice of K. Secondly, we investigate the effect of the outcome's coefficient of variation. From the table, the IHT estimator always performs much better than the HT estimator when α takes different values. It indicates that our IHT is robust to the outcome's coefficient of variation.

Table 5.3
Performance of Algorithm 1, where $f = 0.03$

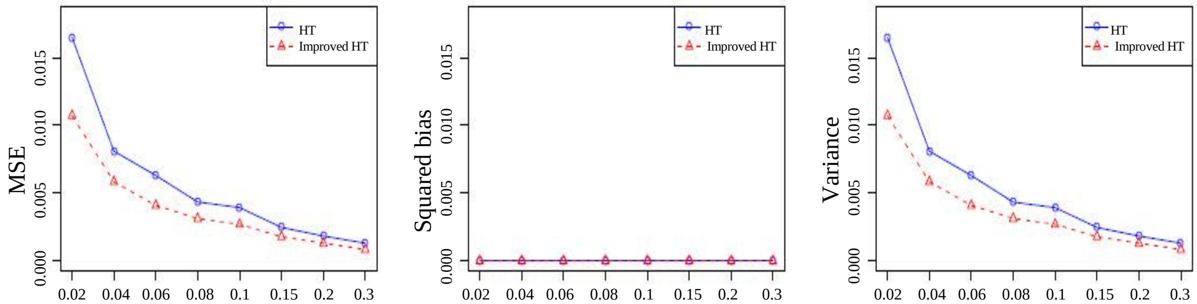
α	$\bar{Y}$	K^*	K_{opt}	MSE_{HT}	MSE_{K^*}	MSE_{opt}	Re
-10	-7.80	125	166	3.3928	1.4130	1.3448	58.35%
-5	-2.81	125	174	0.7097	0.3073	0.2907	56.70%
0	2.19	125	164	0.0623	0.0245	0.0237	60.67%
5	7.20	125	160	1.4056	0.5884	0.5647	58.14%
10	12.24	125	159	4.7510	1.9916	1.9121	58.08%

Example 3: The estimation of population ratio

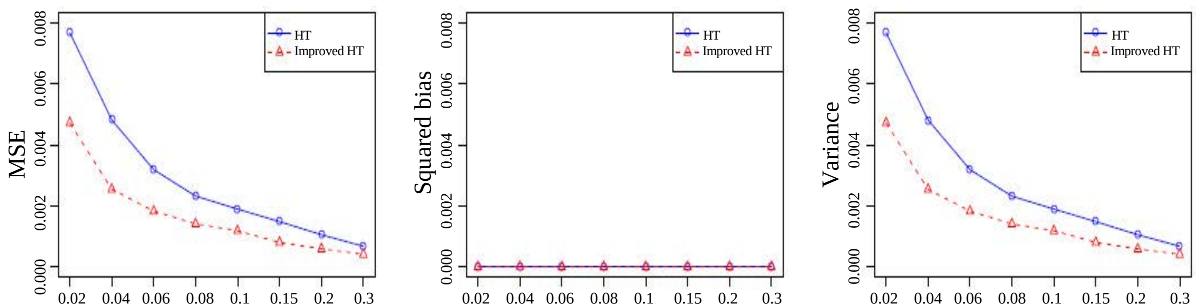
We generate two populations Y and Z of size $N = 3{,}000$: $y_k = \sqrt{12} \cdot \rho_1 \cdot x_k + \sqrt{3 - 3\rho_1^2} \cdot |e_1|$, and $z_k = \sqrt{12} \cdot \rho_2 \cdot x_k + \sqrt{3 - 3\rho_2^2} \cdot |e_2|$, where $x_k \sim U(0, 1)$, $e_1 \sim N(0, 1)$ and $e_2 \sim N(0, 1)$. Our aim is to estimate the ratio $R = t_y / t_z$, where $t_y = \sum_{k=1}^{N} y_k$ and $t_z = \sum_{k=1}^{N} z_k$. We set (ρ_1, ρ_2) as (0.3, 0.4) or (0.7, 0.8), and report the results of two cases in Figures 5.3(a) and 5.3(b), respectively. Similar to the estimation of the population total in examples given above, Figure 5.3 shows that our improved estimator outperforms the classical estimator. We also list the specific Re values of Figure 5.3 in Table 5.4, where the MSEs decrease by 27% to 47%.

Table 5.4
Some specific Re values of Figures 5.2 and 5.3

f	0.02	0.04	0.06	0.08	0.10	0.15	0.20	0.30
Figure 5.2(a)	12.73%	25.33%	45.52%	54.71%	18.15%	30.94%	18.96%	21.99%
Figure 5.2(b)	57.92%	49.78%	49.48%	40.52%	33.81%	57.44%	36.45%	48.70%
Figure 5.2(c)	58.98%	54.41%	70.42%	53.75%	36.05%	48.72%	52.05%	57.65%
Figure 5.3(a)	35.09%	27.92%	35.16%	28.09%	31.50%	28.00%	29.07%	36.31%
Figure 5.3(b)	38.57%	47.18%	42.76%	39.27%	37.49%	46.20%	44.14%	39.55%



(a) πPS sampling (ρ_1 = 0.3, ρ_2 = 0.4)



(b) πPS sampling (ρ_1 = 0.7, ρ_2 = 0.8)

Figure 5.3 Performance of Example 3. From left to right: the MSE performance, the squared-bias performance, and the variance performance.

5.2 Real example

We investigate the data set "Lucy" in the R package "TeachingSampling" (Gutierrez, 2009). This data set includes the variables of 2,396 firms: *ID*, *Level*, *Income*, *Employees*, and *Taxes*. Our aim is to estimate the *Employees* mean $\bar{Y}$ of the 2,300 small or mid-sized firms $(\bar{Y} = 60.59)$. We set the *Income* as the size of the firm, and perform πPS sampling. The sample size n is set among {46, 92, 138, 184, 230, 345, 460, 690}. We list the results in Table 5.5, where the bias, variance, MSE and Re values are reported. We also present the number K^* chosen by Algorithm 1. From Table 5.5, our IHT estimator has much better performance than the HT estimator in terms of MSE. As the sampling fraction f increases, the value of K^* decreases. It means that the number of the modified inclusion probabilities decreases as the sampling fraction increases. This makes sense since the effect of the small inclusion probabilities becomes weak when the sample size increases.

In this real example, we additionally compare the IHT estimator with the HT estimator in the sense of inference performance. Since the squared bias of the IHT estimator is negligible as shown in Theorem 2, the confidence region with 95% coverage is constructed as follow:

$$\left(\hat{t} - 1.96\sqrt{\widehat{\text{MSE}}},\ \hat{t} + 1.96\sqrt{\widehat{\text{MSE}}}\right), \tag{5.1}$$

where $\hat{t}$ is the IHT estimator, and $\widehat{\text{MSE}}$ is its MSE estimator.

Table 5.5
The performance of estimation for the real data set "Lucy"

n	46	92	138	184	230	345	460	690
MSE^{HT}	42.60	20.80	26.87	9.30	6.97	8.01	6.40	2.99
MSE^{IHT}	28.27	14.05	10.18	7.75	5.70	3.77	2.85	1.76
Bias^{HT}	0.0092	0.0002	0.0004	0.0020	0.0041	0.0001	0.0005	0.0112
Bias^{IHT}	0.7520	0.3375	0.2562	0.1093	0.1253	0.0831	0.0539	0.0626
Var^{HT}	42.59	20.80	26.87	9.30	6.97	8.01	6.40	2.97
Var^{IHT}	27.52	13.71	9.92	7.64	5.57	3.68	2.79	1.70
Re ↑	33.64%	32.46%	62.13%	16.75%	18.31%	53.01%	55.49%	41.09%
K^*	166	100	72	59	49	36	29	21

We iteratively simulate $M = 5{,}000$ times and calculate the mean and variance of MSE estimator, and the 95% coverage probabilities. The coverage probabilities (CP) are calculated as CP = $\frac{1}{M}\sum_{m=1}^{M} I\left(\bar{t} \in A_{(m)}\right)$ where $\bar{t}$ is the finite population mean and $A_{(m)}$ is the constructed 95% confidence region of the m^{th} iteration using equation (5.1). The inference performance is reported in Table 5.6. From the table, we have two observations. Firstly, our IHT estimator has smaller MSE than the HT estimator, but it attains almost the same coverage as the HT estimator. Thus, much narrower confidence intervals of the IHT estimator are constructed than those of the HT estimator. Secondly, for the HT estimator, the MSE estimator is much unstable due to the high heterogeneousness of the inclusion probabilities, while our IHT can efficiently overcome this problem. As a summary, our IHT estimator not only increases the estimation accuracy much at the expense of bringing a negligible bias, but also brings much more stable MSE estimator than the HT estimator.

Table 5.6
The inference performance of "Lucy" data set

f	HT				IHT			
	MSE	$\text{E}(\widehat{\text{MSE}})$	$\text{Var}(\widehat{\text{MSE}})$	CP	MSE	$\text{E}(\widehat{\text{MSE}})$	$\text{Var}(\widehat{\text{MSE}})$	CP
0.02	219	76.1	8.28×10^4	91%	48.9	48.4	1.37×10^3	90%
0.04	109	173	2.90×10^7	92%	26.9	26.9	196	92%
0.06	72.7	118	9.11×10^6	91%	18.4	18.2	117	91%
0.08	54.3	67.5	1.94×10^6	93%	14.2	14.1	37.9	92%
0.10	43.2	59.5	1.46×10^6	93%	11.4	11.2	22.8	93%
0.15	28.5	27.2	2.40×10^5	93%	7.47	7.40	17.1	93%

6 Concluding remarks

In this paper, we have proposed a novel and simple method to improve the Horvitz-Thompson estimator in survey sampling. Compared with the HT estimator, the proposed IHT estimator improves the estimation accuracy at the expense of introducing a small bias. Empirical studies show that the improvement is substantial. This new idea has also been used to construct an improved ratio estimator. Naturally, applying it to other estimators, such as the regression estimator and the treatment effect estimator, is of interest as well, and this warrants further study.

The choice of the threshold K is important in our method. Although we have suggested an easy algorithm for the choice and have numerically showed that our choice is very close to the optimal one in terms of MSE, it may not be optimal in terms of MSE. How to choose an optimal threshold is a meaningful topic for future research.

Acknowledgements

The author are grateful to the referees, the associate editor and the editor for their meticulous reading of the manuscript and their invaluable comments. Zhu's work was supported by the National Natural Science Foundation of China (grant nos. 11871459, 71532013 and 71771208). Zou's work was partially supported by the Ministry of Science and Technology of China (Grant no. 2016YFB0502301) and the National Natural Science Foundation of China (Grant nos. 11529101 and 11331011).

Appendix

A.1 Proof of Theorem 1

To obtain the MSE of the IHT estimator, we first define $I_k = 1$ or 0, $k = 1, \ldots, N,$ if the k^{th} unit is drawn or not, then

$$E(I_k) = \pi_k,\ \mathrm{Var}(I_k) = \Delta_{kk},\ \mathrm{Cov}(I_k, I_l) = \Delta_{kl} \quad \text{for} \quad k \neq l,$$

where $\Delta_{kk} = \pi_k(1 - \pi_k),\ \Delta_{kl} = \pi_{kl} - \pi_k\pi_l.$ So the bias of the IHT estimator is

$$\mathrm{Bias}(\hat{t}_{\mathrm{IHT}}) = E\left(\sum_U \frac{y_k}{\pi_k^*} I_k\right) - \sum_U y_k = \sum_{U_2}\left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k. \tag{A.1}$$

The variance of the IHT estimator is given by

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{\mathrm{IHT}}) &= \mathrm{Var}\left(\sum_s \frac{y_k}{\pi_k^*}\right) = \mathrm{Var}\left(\sum_U \frac{y_k}{\pi_k^*} I_k\right) \\
&= \sum_U \left[\left(\frac{y_k}{\pi_k^*}\right)^2 \mathrm{Var}(I_k)\right] + \sum\sum_{k \neq l}{}_U \left(\frac{y_k}{\pi_k^*}\frac{y_l}{\pi_l^*}\mathrm{Cov}(I_k, I_l)\right) \\
&= \sum_{U_1} \frac{\Delta_{kk}}{\pi_k^2} y_k^2 + \sum_{U_2} \frac{\Delta_{kk}}{\pi_{(K)}^2} y_k^2 + \sum\sum_{k \neq l}{}_U \frac{\Delta_{kl}}{\pi_k^*\pi_l^*} y_k y_l. \qquad (A.2)
\end{aligned}$$

Combining (A.1) and (A.2), we obtain

$$\begin{aligned}
\mathrm{MSE}\left(\hat{t}_{\mathrm{IHT}}\right) &= \mathrm{Bias}^2\left(\hat{t}_{\mathrm{IHT}}\right) + \mathrm{Var}\left(\hat{t}_{\mathrm{IHT}}\right) \\
&= \left[\sum\nolimits_{U_2}\left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2 + \sum\nolimits_U \frac{\Delta_{kk}}{\pi_k^{*2}} y_k^2 + \sum\sum\nolimits_{k \neq l}{}_U \frac{\Delta_{kl}}{\pi_k^* \pi_l^*} y_k y_l \\
&= \left\{\sum\nolimits_U \frac{\Delta_{kk}}{\pi_k^{*2}} y_k^2 + \left[\sum\nolimits_{U_2}\left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2\right\} + \sum\sum\nolimits_{k \neq l}{}_U \frac{\Delta_{kl}}{\pi_k^* \pi_l^*} y_k y_l \\
&\triangleq F_1 + F_2.
\end{aligned} \tag{A.3}$$

It is directly verified that $E\left(\widehat{\mathrm{MSE}}\left(\hat{t}_{\mathrm{IHT}}\right)\right) = \mathrm{MSE}\left(\hat{t}_{\mathrm{IHT}}\right)$. Therefore, Theorem 1 is proved.

A.2 Proof of Theorem 2

Using Conditions C.1 and C.2, we see that $\lambda \leq \pi_k \leq \pi_{(K)} \leq 1$ for each $k \in U_2$, and $\max_{k \neq l \in U_2} |\pi_{kl} - \pi_k \pi_l| = O(n^{-1})$. Then, from equation (2.1), we have

$$\begin{aligned}
\left| E\left(\hat{\bar{t}}_{\mathrm{HT}} - \bar{t}\right)^2 \right| &= \left| \frac{1}{N^2} \sum_U \frac{\Delta_{kk}}{\pi_k^2} y_k^2 + \frac{1}{N^2} \sum\sum\nolimits_{k \neq l}{}_U \frac{\Delta_{kl}}{\pi_k \pi_l} y_l y_k \right| \\
&\leq \frac{1}{N^2} \sum_U \frac{1 - \pi_k}{\pi_k} y_k^2 + \frac{1}{N^2} \sum\sum\nolimits_{k \neq l}{}_U \left| \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \right| |y_l y_k| \\
&= O(n^{-1}).
\end{aligned}$$

Similarly, by the MSE of the IHT estimator given in (3.1), we observe

$$\begin{aligned}
\left| E\left(\hat{\bar{t}}_{\mathrm{IHT}} - \bar{t}\right)^2 \right| &= \left| \left[\frac{1}{N} \sum\nolimits_{U_2}\left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2 + \frac{1}{N^2} \sum\nolimits_U \frac{\Delta_{kk}}{\pi_k^{*2}} y_k^2 + \frac{1}{N^2} \sum\sum\nolimits_{k \neq l}{}_U \frac{\Delta_{kl}}{\pi_k^* \pi_l^*} y_k y_l \right| \\
&\leq \left[\frac{K}{N} \frac{1}{K} \sum\nolimits_{U_2}\left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2 + \frac{1}{N^2} \sum\nolimits_U \left| \frac{\pi_k (1 - \pi_k)}{\pi_k^{*2}} \right| y_k^2 \\
&\quad + \frac{1}{N^2} \sum\sum\nolimits_{k \neq l}{}_U \left| \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k^* \pi_l^*} \right| |y_k y_l| \\
&= O(n^{-1}).
\end{aligned}$$

From Conditions C.1 and C.2, it is readily seen that

$$\mathrm{Bias}\left(\hat{\bar{t}}_{\mathrm{IHT}}\right) = \left| \frac{1}{N} \sum\nolimits_{U_2}\left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k \right| \leq \frac{K}{N} \frac{1}{K} \sum\nolimits_{U_2} \left| \frac{\pi_k}{\pi_{(K)}} - 1 \right| |y_k| \leq \frac{1}{N} \sum\nolimits_{U_2} |y_k| = O(n^{-1}),$$

where the third and fourth steps are valid due to $\lambda \leq \pi_k \leq \pi_{(K)} \leq 1$ for each $k \in U_2$ and $K/N = O(n^{-1})$, respectively.

A.3 Proof of Theorem 3

From equation (2.1), since the HT estimator is unbiased, we have

$$\text{MSE}\left(\hat{Y}_{\text{HT}}\right) = \left\{\sum\nolimits_{U_1} \frac{\Delta_{kk}}{\pi_k^2} y_k^2 + \sum\nolimits_{U_2} \frac{\Delta_{kk}}{\pi_k^2} y_k^2\right\} + \sum\sum_{k \neq l}\nolimits_U \frac{\Delta_{kl}}{\pi_k \pi_l} y_k y_l \triangleq F_3 + F_4. \tag{A.4}$$

To illustrate the effectiveness of the new estimator, we compare equation (A.3) and equation (A.4). We prove $F_3 \geq F_1$ at first. It is clear that

$$\begin{aligned}
F_3 - F_1 &= \sum\nolimits_U \frac{\Delta_{kk}}{\pi_k^2} y_k^2 - \left\{\sum\nolimits_{U_1} \frac{\Delta_{kk}}{\pi_k^2} y_k^2 + \sum\nolimits_{U_2} \frac{\Delta_{kk}}{\pi_{(K)}^2} y_k^2 + \left[\sum\nolimits_{U_2} \left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2\right\} \\
&= \sum\nolimits_{U_2} \frac{\Delta_{kk}}{\pi_k^2} y_k^2 - \sum\nolimits_{U_2} \frac{\Delta_{kk}}{\pi_{(K)}^2} y_k^2 - \left[\sum\nolimits_{U_2} \left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2 \\
&= \sum\nolimits_{U_2} \frac{\left(\pi_{(K)}^2 - \pi_k^2\right)(1 - \pi_k)}{\pi_{(K)}^2 \pi_k} y_k^2 - \left[\sum\nolimits_{U_2} \left(\frac{\pi_k}{\pi_{(K)}} - 1\right) y_k\right]^2 \\
&\triangleq D - C.
\end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$C = \left(\sum\nolimits_{U_2} \frac{\pi_k - \pi_{(K)}}{\pi_{(K)}} y_k\right)^2 \leq K \sum\nolimits_{U_2} \frac{\left(\pi_k - \pi_{(K)}\right)^2}{\pi_{(K)}^2} y_k^2 \triangleq E,$$

where the strict inequality holds if there exist $k \neq l \in U_2$ such that $\left(\pi_k - \pi_{(K)}\right) y_k \neq \left(\pi_l - \pi_{(K)}\right) y_l$. Further,

$$\begin{aligned}
F_3 - F_1 \geq D - E &= \sum\nolimits_{U_2} \frac{\left(\pi_{(K)}^2 - \pi_k^2\right)(1 - \pi_k)}{\pi_{(K)}^2 \pi_k} y_k^2 - K \sum\nolimits_{U_2} \frac{\left(\pi_k - \pi_{(K)}\right)^2}{\pi_{(K)}^2} y_k^2 \\
&= \sum\nolimits_{U_2} \frac{\left(\pi_{(K)} - \pi_k\right)\left[(1 - \pi_k - K\pi_k)\pi_{(K)} + (\pi_k - \pi_k^2 + K\pi_k^2)\right]}{\pi_{(K)}^2 \pi_k} y_k^2.
\end{aligned} \tag{A.5}$$

From Definition 1, we have $\pi_k \leq \pi_{(K)} \leq (K+1)^{-1}$ for each $k \in U_2$, thus $D - E \geq 0$. So $F_3 - F_1 = D - C \geq D - E \geq 0$ holds.

For the terms F_2 and F_4, we note that

$$\begin{aligned}
F_2 - F_4 &= \sum\sum_{k \neq l}\nolimits_U \frac{\Delta_{kl}}{\pi_k^* \pi_l^*} y_k y_l - \sum\sum_{k \neq l}\nolimits_U \frac{\Delta_{kl}}{\pi_k \pi_l} y_k y_l \\
&= \sum\sum_{k \neq l}\nolimits_{U_2} \left(\frac{\Delta_{kl}}{\pi_{(K)}^2} - \frac{\Delta_{kl}}{\pi_k \pi_l}\right) y_k y_l + \sum_{k \in U_1} \sum_{l \in U_2} \left(\frac{\Delta_{kl}}{\pi_{(K)} \pi_k} - \frac{\Delta_{kl}}{\pi_k \pi_l}\right) y_k y_l \\
&\quad + \sum_{k \in U_2} \sum_{l \in U_1} \left(\frac{\Delta_{kl}}{\pi_{(K)} \pi_l} - \frac{\Delta_{kl}}{\pi_k \pi_l}\right) y_k y_l \\
&\triangleq \Delta_1 + \Delta_2 + \Delta_3.
\end{aligned}$$

Using Conditions C.1 and C.2, it is seen that

$$\begin{aligned}\frac{|\Delta_1|}{N^2} &= \frac{1}{N^2}\left|\sum\sum_{k\neq l}{}_{U_2}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\left(\frac{\pi_k\pi_l}{\pi_{(K)}^2}-1\right)y_k y_l\right| \\ &\leq \frac{1}{N^2}\sum\sum_{k\neq l}{}_{U_2}\left|\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right|\left|\frac{\pi_k\pi_l}{\pi_{(K)}^2}-1\right||y_k y_l| \\ &\leq \frac{K^2}{N^2}\frac{1}{K^2}\sum\sum_{k\neq l}{}_{U_2}\left|\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right||y_k y_l| = O(n^{-3}),\end{aligned}$$

where the third and fourth steps are valid due to $\lambda \leq \pi_k \leq \pi_{(K)} \leq 1$ for each $k \in U_2$, $K/N = O(n^{-1})$, and $\max_{k\neq l\in U_2}|\pi_{kl}-\pi_k\pi_l| = O(n^{-1})$. Similarly, we obtain

$$\begin{aligned}\frac{|\Delta_2|}{N^2} &= \frac{1}{N^2}\left|\sum_{k\in U_1}\sum_{l\in U_2}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\left(\frac{\pi_l}{\pi_{(K)}}-1\right)y_k y_l\right| \\ &\leq \frac{1}{N^2}\sum_{k\in U_1}\sum_{l\in U_2}\left|\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right|\left|\frac{\pi_l}{\pi_{(K)}}-1\right||y_k y_l| \\ &\leq \frac{1}{N^2}\sum_{k\in U_1}\sum_{l\in U_2}\left|\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right||y_k y_l| = O(n^{-2}),\end{aligned}$$

and $\frac{|\Delta_3|}{N^2} = O(n^{-2})$.

Thus, together with $F_3 \geq F_1$, we have

$$\mathrm{MSE}(N^{-1}\hat{t}_{\mathrm{IHT}}) \leq \mathrm{MSE}(N^{-1}\hat{t}_{\mathrm{HT}}) + o(n^{-1}).$$

For the Poisson sampling case, we have $F_4 = F_2 = 0$. Hence, for Poisson sampling, we obtain

$$\mathrm{MSE}(N^{-1}\hat{t}_{\mathrm{IHT}}) \leq \mathrm{MSE}(N^{-1}\hat{t}_{\mathrm{HT}}).$$

A.4 Proof of Theorem 4

First note that

$$(\hat{R}-R)^2 = \left(\frac{\hat{t}_{y\pi}-R\hat{t}_{z\pi}}{\hat{t}_{z\pi}}\right)^2 = \frac{(\hat{t}_{y\pi}-R\hat{t}_{z\pi})^2}{t_z^2} - \frac{(\hat{t}_{z\pi}^2-t_z^2)(\hat{t}_{y\pi}-R\hat{t}_{z\pi})^2}{t_z^2\,\hat{t}_{z\pi}^2} \triangleq \mathrm{I}+\mathrm{III},$$

and

$$(\hat{R}^*-R)^2 = \left(\frac{\hat{t}_{y\pi}^*-R\hat{t}_{z\pi}}{\hat{t}_{z\pi}^*}\right)^2 = \frac{(\hat{t}_{y\pi}^*-R\hat{t}_{z\pi}^*)^2}{t_z^2} - \frac{(\hat{t}_{z\pi}^{*2}-t_z^2)(\hat{t}_{y\pi}^*-R\hat{t}_{z\pi}^*)^2}{t_z^2\,\hat{t}_{z\pi}^{*2}} \triangleq \mathrm{II}+\mathrm{IV}.$$

Let $u_k = y_k - Rz_k$. By Theorem 3, we have

$$N^{-2}E\left(\hat{t}_u^* - t_u\right)^2 \le N^{-2}E\left(\hat{t}_u - t_u\right)^2 + o\left(n^{-1}\right).$$

Thus, for the terms I and II, we get

$$E\,(\mathrm{I}) \le E\,(\mathrm{II}) + o\left(n^{-1}\right). \tag{A.6}$$

Now, we need to prove that the expectations of III and IV are negligible. Observe that,

$$\begin{aligned}
|\,E\,(\mathrm{III})\,| &= \left| E\,\frac{\left(\hat{t}_{z\pi} + t_z\right)\left(\hat{t}_{z\pi} - t_z\right)\left(\hat{t}_{y\pi} - R\hat{t}_{z\pi}\right)^2}{t_z^2\,\hat{t}_{z\pi}^2} \right| \\
&\le E\,\frac{|\,\hat{t}_{z\pi} + t_z\,|\,|\,\hat{t}_{z\pi} - t_z\,|\left(\hat{t}_{y\pi} - R\hat{t}_{z\pi}\right)^2}{t_z^2\,\hat{t}_{z\pi}^2} \\
&\le \frac{Z^* + |\,t_z\,|}{t_z^2\,Z_*^2}\,E\left(|\,\hat{t}_{z\pi} - t_z\,|\left(\hat{t}_{y\pi} - R\hat{t}_{z\pi}\right)^2\right) \\
&\le \frac{Z^* + |\,t_z\,|}{t_z^2\,Z_*^2}\sqrt{E\left(\hat{t}_{z\pi} - t_z\right)^2 E\left(\hat{t}_{y\pi} - R\hat{t}_{z\pi}\right)^4},
\end{aligned}$$

where $Z^* = \frac{n}{N}\max_{k\in U}\left(\frac{z_k}{\pi_k}\right)$, $Z_* = \frac{n}{N}\min_{k\in U}\left(\frac{z_k}{\pi_k}\right)$. Similarly,

$$|\,E\,(\mathrm{IV})\,| \le \frac{\tilde{Z}^* + |\,t_z\,|}{t_z^2\,\tilde{Z}_*^2}\sqrt{E\left(\hat{t}_{z\pi}^* - t_z\right)^2 E\left(\hat{t}_{y\pi}^* - R\hat{t}_{z\pi}^*\right)^4},$$

where $\tilde{Z}^* = \frac{n}{N}\max_{k\in U}\left(\frac{z_k}{\pi_k^*}\right)$, $\tilde{Z}_* = \frac{n}{N}\min_{k\in U}\left(\frac{z_k}{\pi_k^*}\right)$.

Using Theorem 2 and Lemma 1, we see that $|\,E(\mathrm{III})\,| = O\left(n^{-3/2}\right)$ and $|\,E\,(\mathrm{IV})\,| = O\left(n^{-3/2}\right)$. Combining these and equation (A.6), we get

$$\mathrm{MSE}\left(\hat{R}^*\right) \le \mathrm{MSE}\left(\hat{R}\right) + o\left(n^{-1}\right).$$

It implies that $\mathrm{MSE}\left(N^{-1}\hat{Y}_R^*\right) \le \mathrm{MSE}\left(N^{-1}\hat{Y}_R\right) + o\left(n^{-1}\right)$.

A.5 Discussion on Condition C.4

Case 1: Simple random sampling without replacement

Under the simple random sampling without replacement, we have that $\pi_i = \frac{n}{N}$ for $i \in U$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ for $i \ne j \in U$, $\pi_{ijk} = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}$ for $i \ne j \ne k \in U$, and $\pi_{ijkl} = \frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)}$ for $i \ne j \ne k \ne l \in U$. It follows that

$$\pi_{ijk} - \pi_{ij}\pi_k = -\frac{2n\,(n-1)\,(N-n)}{N^2\,(N-1)\,(N-2)} = O\left(n^{-1}\right),$$

where the last equality is from Condition C.3. We also obtain

$$\begin{aligned}
&\pi_{ijkl} - 4\pi_{ijk}\pi_l + 6\pi_{ij}\pi_k\pi_l - 3\pi_i\pi_j\pi_k\pi_l \\
&\quad = \left(\pi_{ijkl} - \pi_{ijk}\pi_l\right) - 3\left(\pi_{ijk}\pi_l - \pi_{ij}\pi_k\pi_l\right) + 3\left(\pi_{ij}\pi_k\pi_l - \pi_i\pi_j\pi_k\pi_l\right) \\
&\quad = 3\frac{n(n-1)(n-2)(n-N)}{N^2(N-1)(N-2)(N-3)} - 6\frac{n^2(n-1)(n-N)}{N^3(N-1)(N-2)} + 3\frac{n^3(n-N)}{N^4(N-1)} \\
&\quad = O(n^{-2}),
\end{aligned}$$

where the last equality is from Condition C.3. Thus, Condition C.4 holds under the simple random sampling without replacement.

Case 2: Poisson sampling

From the independence of Poisson sampling, $\pi_{ij} = \pi_i\pi_j$ for $i \neq j \in U$, $\pi_{ijk} = \pi_i\pi_j\pi_k$ for $i \neq j \neq k \in U$, and $\pi_{ijkl} = \pi_i\pi_j\pi_k\pi_l$ for $i \neq j \neq k \neq l \in U$. Hence, $\pi_{ijk} - \pi_{ij}\pi_k = 0$, and $\pi_{ijkl} - 4\pi_{ijk}\pi_l + 6\pi_{ij}\pi_k\pi_l - 3\pi_i\pi_j\pi_k\pi_l = 0$. It follows that Poisson sampling satisfies Condition C.4.

A.6 A lemma for proving Theorem 4

Lemma 1. *For the HT estimator $\hat{\bar{t}}_{HT}$ and the IHT estimator $\hat{\bar{t}}_{IHT}$, under the Conditions C.1-C.4, we have*

$$E\left(\hat{\bar{t}}_{HT} - \bar{t}\right)^4 = O(n^{-2}), \quad \textit{and} \quad E\left(\hat{\bar{t}}_{IHT} - \bar{t}\right)^4 = O(n^{-2}).$$

Proof. Noting that

$$\hat{\bar{t}}_{\text{HT}} - \bar{t} = \frac{1}{N}\sum\nolimits_U \frac{I_k - \pi_k}{\pi_k} y_k \triangleq \frac{1}{N}\sum\nolimits_U J_k y_k,$$

we have

$$\begin{aligned}
\left(\hat{\bar{t}}_{\text{HT}} - \bar{t}\right)^4 &= \frac{1}{N^4}\sum_k\sum_l\sum_i\sum_j (J_k y_k)(J_l y_l)(J_i y_i)(J_j y_j) \\
&= \frac{1}{N^4}\sum_U (J_k y_k)^4 + \frac{4}{N^4}\sum_{k\neq l}(J_k y_k)^3 (J_l y_l) + \frac{3}{N^4}\sum_{k\neq l}(J_k y_k)^2 (J_l y_l)^2 \\
&\quad + \frac{6}{N^4}\sum_{i\neq k\neq l}(J_i y_i)^2 (J_k y_k)(J_l y_l) + \frac{1}{N^4}\sum_{i\neq j\neq k\neq l}(J_i y_i)(J_j y_j)(J_k y_k)(J_l y_l) \\
&\triangleq \text{I} + \text{II} + \text{III} + \text{IV} + \text{V}.
\end{aligned}$$

For the first term I, using $\lambda \leq \pi_k \leq 1$ and $|I_k - \pi_k| \leq 1$ for any $k \in U$, we get

$$|E(I)| = E\left(\frac{1}{N^4}\sum_U (J_k y_k)^4\right) = \frac{1}{N^4}\sum_U\left(\frac{y_k}{\pi_k}\right)^4 E(I_k - \pi_k)^4 \leq \frac{1}{N^4}\sum_U\left(\frac{y_k}{\pi_k}\right)^4 = O(n^{-2}).$$

For the terms II and III, we have

$$|E(J_k^3 J_l)| = \left| \frac{1}{\pi_k^3 \pi_l} E\left[(I_k - \pi_k)^3 (I_l - \pi_l)\right] \right| \le \frac{1}{\pi_k^3 \pi_l} E\left[\,|I_k - \pi_k|^3 |I_l - \pi_l|\,\right] \le \frac{1}{\pi_k^3 \pi_l} \le \frac{1}{\lambda^4},$$

and

$$E(J_k^2 J_l^2) = \frac{1}{\pi_k^2 \pi_l^2} E\left[(I_k - \pi_k)^2 (I_l - \pi_l)^2\right] \le \frac{1}{\pi_k^2 \pi_l^2} \le \frac{1}{\lambda^4}.$$

Thus, $|E(\text{II})| = O(n^{-2})$ and $|E(\text{III})| = O(n^{-2})$. For the fourth term IV, we have that

$$\begin{aligned} |E(J_i^2 J_k J_l)| &= \frac{1}{\pi_i^2 \pi_k \pi_l} \left| E\left[(I_i - \pi_i)^2 (I_k - \pi_k)(I_l - \pi_l)\right] \right| \\ &= \frac{1}{\pi_i^2 \pi_k \pi_l} \left| (1 - 2\pi_i)[(\pi_{ikl} - \pi_{ik}\pi_l) - \pi_k(\pi_{il} - \pi_i \pi_l)] + \pi_i^2 (\pi_{kl} - \pi_k \pi_l) \right| \\ &= O(n^{-1}), \end{aligned}$$

where the last step is from Conditions C.1 and C.4. It implies that $|E(\text{IV})| = O(n^{-2})$. For the last term V, we have that

$$\begin{aligned} &\frac{1}{N^4} E\left(\sum_{i \ne j \ne k \ne l} (J_i y_i)(J_j y_j)(J_k y_k)(J_l y_l) \right) \\ &\quad = \frac{1}{N^4} \sum_{i \ne j \ne k \ne l} \frac{\pi_{ijkl} - 4\pi_{ijk}\pi_l + 6\pi_{ij}\pi_k\pi_l - 3\pi_i\pi_j\pi_k\pi_l}{\pi_i \pi_j \pi_k \pi_l} y_i y_j y_k y_l \\ &\quad = O(n^{-2}), \end{aligned}$$

where the last step is from Conditions C.1 and C.4. Thus, $E\left(\hat{\bar{t}}_{\text{HT}} - \bar{t}\right)^4 = O(n^{-2})$ holds.

Next we shall prove $E\left(\hat{\bar{t}}_{\text{IHT}} - \bar{t}\right)^4 = O(n^{-2})$. Noting that

$$\hat{\bar{t}}_{\text{IHT}} - \bar{t} = \frac{1}{N} \sum_U \frac{I_k - \pi_k^*}{\pi_k^*} y_k = \frac{1}{N} \sum_U \frac{I_k - \pi_k}{\pi_k^*} y_k + \frac{1}{N} \sum_U \frac{\pi_k - \pi_k^*}{\pi_k^*} y_k \triangleq A + \Delta,$$

we have

$$E\left(\hat{\bar{t}}_{\text{IHT}} - \bar{t}\right)^4 = E(A + \Delta)^4 = E(A^4) + 4\Delta E(A^3) + 6\Delta^2 E(A^2) + 4\Delta^3 E(A) + \Delta^4. \tag{A.7}$$

Similar as the proofs of the result $E\left(\hat{\bar{t}}_{\text{HT}} - \bar{t}\right)^4 = O(n^{-2})$, using $\lambda \le \pi_k^* \le 1$, it is easy to obtain

$$E(A^4) = E\left(\frac{1}{N} \sum_U \frac{I_k - \pi_k}{\pi_k^*} y_k \right)^4 = O(n^{-2}). \tag{A.8}$$

From equation (A.8), we have that $E(A^2) = O(n^{-1})$ and $E(A^3) = O(n^{-3/2})$. Meanwhile $E(A) = 0$ and

$$\Delta = \frac{1}{N} \sum_U \frac{\pi_k - \pi_k^*}{\pi_k^*} y_k = \frac{K}{N} \left(\frac{1}{K} \sum_{U_2} \frac{\pi_k - \pi_k^*}{\pi_k^*} y_k \right) = O(n^{-1}).$$

Therefore, from equation (A.7), we prove that $E\left(\hat{\bar{t}}_{\text{IHT}} - \bar{t}\right)^4 = O(n^{-2})$.

References

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Cardot, H., and Josserand, E. (2011). Horvitz-thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98, 107-118.

Gutierrez, H.A. (2009). Estrategias de muestreo: Diseno de encuestas y estimacion de parametros. *Editorial Universidad Santo Tomas*.

Hansen, M., and Hurwitz, W. (1943). On the theorey of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hoerl, A., and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

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

Rosenbaum, P. (2002). *Observational Studies*. New York: Springer.

Tibshirani, R. (1996). Regression shrikage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.