

ESTIMATING VARIANCE IN THE PRESENCE OF NONRESPONSE UNDER UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT

Chugiat Ponkaew and Nuanpan Lawson*

Received: June 20, 2018; Revised: June 04, 2019; Accepted: June 04, 2019

Abstract

In this paper we propose 2 estimators for estimating a population total under unequal probability sampling without replacement, a linear and a ratio estimator. We show in theory that the linear estimator is an unbiased estimator and the latter is an asymptotically unbiased estimator. We propose a general form of the variance, and also an estimator of the variance in this paper considered under the reverse framework and only when the sampling fraction is negligible. We can show in theory that the variance estimators of the proposed estimators are asymptotically unbiased. A simulation study has been conducted to ascertain the performance of the proposed estimators.

Keywords: Probability proportional to size sampling, first and second order inclusion probability, sampling fraction, reverse framework

Introduction

Nonresponse usually occurs in sample surveys due to a failure to collect data. Unit and item nonresponses are currently defined types of the nonresponse. It is called a unit nonresponse if the selected sample elements do not provide the requested information, while in contrast it is called an item nonresponse when sample elements participate in the survey but refuse to

answer specific portions of the survey questions for any reason, known or otherwise. Shao and Steel (1999) stated that variance estimation is a key element in analyzing survey data and subsequently a lot of research has been conducted under the presence of nonresponse with technical countermeasures in place to overcome it using the most

Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800. Thailand. E-mail: nuanpan.n@sci.kmutnb.ac.th

** Corresponding author*

appropriate variance estimators. At present, a two-phase framework provides the basis for development of the most well-known variance of population total estimator (Särndal, 1992; Deville and Särndal, 1995; Kajal, 2014) and it views nonresponse as the second-phase of selection. In literature covering the two-phase framework developments, Särndal and Lundström (2005) proposed an estimator extending the estimator of Horvitz and Thompson (1952) into a two-phase framework, multiplying inverse response probability.

The Särndal and Lundström (2005) estimator requires known response probabilities; however, they are not always known in practice. Therefore, Lawson (2017) proposed an asymptotically unbiased estimator for the population total without any known or estimated response probabilities and it worked under a reverse framework. This framework was first introduced by Fay (1991) who proposed reversing the order of sampling and response. Lawson's (2017) estimator is proposed under unequal probability sampling with probability proportional to size with replacement (PPSWR) and when a sampling fraction is negligible. Usually the variance of a population total estimator under sampling with replacement is larger than sampling without replacement. Therefore, in this paper, we focus on estimating the variance of the population total in the presence of unit nonresponse under unequal probability sampling without replacement. For simplicity we suppose that the response probability is constant. In section 2 we set out a formal framework and proposed population total estimators and their associated variances. In Section 3, the performance of the variance estimators of our proposed estimators are studied in a simulation study. Finally, some concluding comments are given in Section 4.

Materials and Methods

Basic Setup Under Reverse Framework

Let $U = \{1, 2, \dots, i, \dots, N\}$ be a finite population, N be the size of U , and y be a study variable. Subsequently, the population total of

y can be denoted as $Y = \sum_{i \in U} y_i$. Let n denote the total number of elements in sample S , selected according to unequal probability sampling without replacement, and n_r denote the total number of respondents in response sample r .

Under unequal probability sampling without replacement e.g., with probability proportional to the sample size without replacement, probability proportional to size, systematic sampling, or the Midzuno (1952) scheme, let π_i and π_{ij} denote the first and second order of inclusion probabilities defined by $\pi_i = P(i \in s)$ and $\pi_{ij} = P(i \wedge j \in s)$; $i, j = 1, 2, \dots, N$. Let us define a random variable I_i where $I_i = 1$, if $i \in s$; otherwise $I_i = 0$.

Let $E_s(\cdot)$ and $V_s(\cdot)$ denote the expectation and variance operators respective to unequal probability sampling without replacement. Under nonresponse, let subscript R denote the nonresponse mechanism and r_i denote the response indicator variable of y_i where $r_i = 1$, if y_i is observed; otherwise $r_i = 0$.

Let p_i denote the response probability as defined by $p_i = P(r_i = 1)$. We assume that $p_i = p$ for all units in U . Let $E_R(\cdot)$ and $V_R(\cdot)$ denote the expectation and variance operators with respect to the nonresponse mechanism. We assume that the sampled units respond to item y independently of each other with probability p . Hence, $E_R(r_i) = P(r_i = 1) = p$, $V_R(r_i) = p(1-p)$ and $E_R(r_i r_j) = E_R(r_i) E_R(r_j) = p^2$. Let, E, V denote the overall expectation and variance operators. Finally, we assume that the sampling fraction is negligible.

Variance Estimation Under Reverse Framework

Reverse framework was first introduced by Fay (1991), who reversed the order of sampling and response (Shao and Steel, 1999; Lawson, 2017; Lawson and Ponkaew, 2019). In this framework, the variance of the population total estimator $\hat{Y}$ is defined by $V(\hat{Y}) = E_R V_S(\hat{Y} | \mathbf{R}) + V_R E_S(\hat{Y} | \mathbf{R})$, where $\mathbf{R} = (r_1, r_2, \dots, r_N)'$ is the vector of response indicators. However, if the sampling fraction $f = n/N$ is negligible, the second component may be omitted. Therefore, under the reverse

framework with a negligible sampling fraction, the variance of $\hat{Y}$ can be derived by:

$$V(\hat{Y}) \approx E_R V_S(\hat{Y}|\mathbf{R}). \quad (1)$$

To estimate $E_R V_S(\hat{Y}|\mathbf{R})$ in Equation (1), we use the same procedure that we used to estimate $V_S(\hat{Y}|\mathbf{R})$. Let $\hat{V}_S$ be such that $E_S(\hat{V}_S|\mathbf{R}) = V_S(\hat{Y}|\mathbf{R})$ is consequently the overall expectation obtained by $E[\hat{V}_S|\mathbf{R}] = E_R[E_S(\hat{V}_S|\mathbf{R})] = E_R[V_S(\hat{Y}|\mathbf{R})]$.

Therefore, $V(\hat{Y})$ can be estimated by:

$$\hat{V}(\hat{Y}) = \hat{V}_S(\hat{Y}|\mathbf{R}). \quad (2)$$

The Proposed Estimators

In this section, we propose 2 estimators for estimating population totals under unequal probability sampling without replacement. The first one is a linear estimator and the second one is a ratio estimator. The ratio estimator requires neither known nor estimated response probability under a reverse framework with a negligible sampling fraction. In the following development we assume that (A) $p_i = p$ for all $i \in U$ and (B) sampling fraction $f = \frac{n}{N}$ is negligible.

The Linear Estimator

In the case of full response, the Horvitz and Thompson (1952) estimator, as originally discussed by them, has been used to estimate the population total. The Horvitz and Thompson (1952) estimator is defined by

$$\hat{Y}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i}.$$

In the presence of nonresponse, the estimator must deal with responding elements (Bethlehem, 1988). Therefore, Särndal and Lundström (2005) proposed an unbiased estimator of population total based on Horvitz and Thompson's (1952) estimator along with a set of respondents assessed under a two-phase

framework defined by $\hat{Y} = \sum_{i \in S_r} \frac{y_i}{\pi_i p_i}$ where S_r is a set of respondents, under a two-phase framework. They also discussed a method for investigating the variance and the estimated value of the variance of the proposed estimator under a two-phase framework. The variance estimator is also an unbiased estimator.

In this paper, we extended Särndal and Lundström's (2005) estimator to include a reverse framework for unequal probability sampling without replacement and called it a linear estimator. Under the reverse framework, the linear estimator is defined by:

$$\hat{Y}_r^{(1)} = \sum_{i \in S} \frac{r_i y_i}{\pi_i p_i}, \quad (3)$$

where π_i denotes the first order of inclusion probabilities under unequal probability sampling without replacement. For simplicity, we assume that $p_i = p$ for all units in U . Then:

$$\hat{Y}_r^{(1)} = \sum_{i \in S} \frac{r_i y_i}{\pi_i p}, \quad (4)$$

where π_i is defined in Equation (3). However, if p is unknown, we can estimate p by $\hat{p} = \sum_{i \in S} \frac{r_i}{\pi_i} / \sum_{i \in S} \frac{1}{\pi_i}$. Therefore the linear estimator

is defined by $\hat{Y}_r^{(1)} = \sum_{i \in S} \frac{r_i y_i}{\pi_i \hat{p}}$.

Theorem 1. Assume that assumption (A) is correct. Under the reverse framework and unequal probability sampling without replacement, the linear estimator $\hat{Y}_r^{(1)}$ is an unbiased estimator of Y .

Proof. From (4) we have, $\hat{Y}_r^{(1)} = \sum_{i \in S} \frac{r_i y_i}{\pi_i p}$.

The overall expectation of $\hat{Y}_r^{(1)}$ given $\mathbf{R}$, is obtained by:

$$E(\hat{Y}_r^{(1)}) = E_R E_S(\hat{Y}_r^{(1)}|\mathbf{R}) = E_R E_S \left[\sum_{i \in S} \frac{r_i y_i}{\pi_i p} | \mathbf{R} \right]$$

$$= \sum_{i \in U} \frac{E_R(r_i) y_i E_S(I_i)}{\pi_i p} = \sum_{i \in U} \frac{p y_i \pi_i}{\pi_i p} = \sum_{i \in U} \frac{p y_i \pi_i}{\pi_i p} = Y.$$

Then, $E(\hat{Y}_r^{(1)}) = Y$. Therefore, $\hat{Y}_r^{(1)}$ is an unbiased estimator of Y .

2) The Ratio Estimator

From Equation (4) it is obvious that the linear estimator requires a known response probability. However, response probabilities are not always known in practice. Therefore Lawson (2017) proposed an alternative estimator for solving the unknown response probability under PPSWR sampling. In this section, we extend Lawson's (2017) estimator to include unequal probability sampling without replacement and call it a ratio estimator. Under unequal probability sampling without replacement the ratio estimator for estimating the population mean is defined by $\hat{\bar{Y}}_r^{(2)} = \sum_{i \in S} \frac{r_i y_i}{\pi_i} / \sum_{i \in S} \frac{r_i}{\pi_i}$. Consequently the ratio estimator for estimating the population total is:

$$\hat{Y}_r^{(2)} = N \hat{\bar{Y}}_r^{(2)} = N \sum_{i \in S} \frac{r_i y_i}{\pi_i} / \sum_{i \in S} \frac{r_i}{\pi_i}. \quad (5)$$

Theorem 2. Assuming that assumption (A) is correct. Under a reverse framework with unequal probability sampling without replacement, the ratio estimators $\hat{\bar{Y}}_r^{(2)}$ and $\hat{Y}_r^{(2)}$ are asymptotically unbiased of $\bar{Y}$ and Y , respectively.

Proof. The proof is similar to Lawson (2017).

Estimated Variance and the Proposed Variance Estimators

In this section we propose 2 general forms of variance estimator, a linear and ratio estimator both under unequal probability sampling without replacement. First, we investigate the variance of the linear estimator. From Equations (1) and (4) the variance of the linear estimator is obtained by:

$$\begin{aligned} V(\hat{Y}_r^{(1)}) &\approx E_R V_S(\hat{Y}_r^{(1)} | \mathbf{R}) = E_R V_S \left[\sum_{i \in S} \frac{r_i y_i}{\pi_i p} | \mathbf{R} \right] \\ &= E_R V_S \left[\sum_{i \in S} \frac{z_{1i}}{\pi_i} | \mathbf{R} \right], \end{aligned} \quad (6)$$

where $z_{1i} = \frac{r_i y_i}{p}$. Next, we investigate the variance of the ratio estimator. From Equations (1) and (5) the variance of the ratio estimator is obtained by:

$$V(\hat{Y}_r^{(2)}) \approx E_R V_S(\hat{Y}_r^{(2)} | \mathbf{R}) = E_R V_S(N \hat{\bar{Y}}_r^{(2)} | \mathbf{R}). \quad (7)$$

From Equation (7) we can see that $V_S(N \hat{\bar{Y}}_r^{(2)} | \mathbf{R}) = N^2 V_S(\hat{\bar{Y}}_r^{(2)} | \mathbf{R})$. However, $\hat{\bar{Y}}_r^{(2)}$ is a form of ratio estimator; hence, its variance can be derived by using Taylor's linearization approach (Lawson (2017)). Under this method, the linear estimator of $\hat{\bar{Y}}_r^{(2)}$ can be obtained by $\hat{\bar{Y}}_r^{(2)} \approx \bar{Y}_r + \frac{1}{\sum_{i \in U} r_i} \sum_{i \in S} \frac{r_i}{\pi_i} (y_i - \bar{Y}_r)$,

where $\bar{Y}_r = \sum_{i \in U} r_i y_i / \sum_{i \in U} r_i$. Therefore,

$$\begin{aligned} V_S(N \hat{\bar{Y}}_r^{(2)} | \mathbf{R}) &= N^2 V_S(\hat{\bar{Y}}_r^{(2)} | \mathbf{R}) \\ &\approx N^2 V_S \left(\bar{Y}_r + \frac{1}{\sum_{i \in U} r_i} \sum_{i \in S} \frac{r_i}{\pi_i} (y_i - \bar{Y}_r) \middle| \mathbf{R} \right) \\ &= V_S \left(\sum_{i \in S} \frac{r_i}{\pi_i} \frac{N}{\sum_{i \in U} r_i} (y_i - \bar{Y}_r) \middle| \mathbf{R} \right) \\ &= V_S \left(\sum_{i \in S} \frac{z_{2i}}{\pi_i} \middle| \mathbf{R} \right), \end{aligned} \quad (8)$$

where $z_{2i} = \frac{N}{\sum_{i \in U} r_i} r_i (y_i - \bar{Y}_r)$. Substitute Equation (8) into Equation (7),

$$V(\hat{Y}_r^{(2)}) \approx E_R \left[V_S \left(\sum_{i \in S} \frac{Z_{2i}}{\pi_i} \middle| \mathbf{R} \right) \right] \quad (9)$$

From Equations (6) and (9) the variance of the linear estimator and ratio estimator may be written as:

$$V(\hat{Y}_r^{(m)}) \approx E_R V_S \left(\sum_{i \in S} \frac{Z_{mi}}{\pi_i} \middle| \mathbf{R} \right), \quad (10)$$

where $z_{1i} = \frac{r_i y_i}{p}$, $z_{2i} = \frac{N}{N_r} r_i (y_i - \bar{Y}_r)$, and $N_r = \sum_{i \in U} r_i$.

From Equation (10) we see that $\sum_{i \in S} \frac{Z_{mi}}{\pi_i}$ has the same form as Horvitz and Thompson's (1952) estimator. Therefore, the sampling variance of $\sum_{i \in S} \frac{Z_{mi}}{\pi_i}$ given $\mathbf{R}$ is defined by:

$$V_S \left(\sum_{i \in S} \frac{Z_{mi}}{\pi_i} \middle| \mathbf{R} \right) = \sum_{i \in U} D_i z_{mi}^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} z_{mi} z_{mj} \quad (11)$$

Substitute Equation (11) into Equation (10) and we have,

$$\begin{aligned} V(\hat{Y}_r^{(m)}) &\approx E_R V_S \left(\sum_{i \in S} \frac{Z_{mi}}{\pi_i} \middle| \mathbf{R} \right) \\ &= E_R \left[\sum_{i \in U} D_i z_{mi}^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} z_{mi} z_{mj} \right] \\ &= \sum_{i \in U} D_i E_R(z_{mi}^2) + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} E_R(z_{mi}) E_R(z_{mj}) \end{aligned}$$

Then, the variance of linear and ratio estimators can be derived from:

$$\begin{aligned} V(\hat{Y}_r^{(m)}) &\approx \sum_{i \in U} D_i E_R(z_{mi}^2) + \\ &\sum_{i \in U} \sum_{j \neq i \in U} D_{ij} E_R(z_{mi}) E_R(z_{mj}), \end{aligned} \quad (12)$$

where Z_{mi} and $m = 1, 2$ are defined in Equation (10).

Furthermore, by using Equations (2), (10), and (11), the general form of variance estimator of the linear and ratio estimators for estimating population total is obtained by:

$$\hat{V}(\hat{Y}_r^{(m)}) \approx \sum_{i \in S} \hat{D}_i \hat{z}_{mi}^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} \hat{z}_{mi} \hat{z}_{mj}, \quad (13)$$

where $\hat{z}_{1i} = \frac{r_i y_i}{p}$, $\hat{z}_{2i} = \frac{N}{\hat{N}_r} r_i (y_i - \hat{\bar{Y}}_r)$, $\hat{N}_r = \sum_{i \in S} \frac{r_i}{\pi_i}$, $\hat{\bar{Y}}_r = \sum_{i \in S} \frac{r_i y_i}{\pi_i} / \hat{N}_r$, $\hat{D}_i = (1 - \pi_i) \pi_i^{-2}$ and $\hat{D}_{ij} = (\pi_{ij} - \pi_i \pi_j) (\pi_{ij} \pi_i \pi_j)^{-1}$.

Next, we investigate the variances of the proposed estimators and discuss in Theorem 3 and Theorem 4, respectively.

Theorem 3. Assuming that (A) and (B) are correct. Under a reverse framework with unequal probability sampling without replacement:

(a) The variance of the linear estimator is defined by:

$$V(\hat{Y}_r^{(1)}) \approx \sum_{i \in U} D_{ii} y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} y_i y_j, \quad (14)$$

where $D_{ii} = (1 - \pi_i) \pi_i^{-1}$, $D_{ij} = (\pi_{ij} - \pi_i \pi_j) (\pi_i \pi_j)^{-1}$, and $D_{ii} = (1 - \pi_i) (p \pi_i)^{-1}$.

(b) The variance of the ratio estimator is defined by:

$$V(\hat{Y}_r^{(2)}) \approx \sum_{i \in U} D_{ii} (y_i - \bar{Y})^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} (y_i - \bar{Y}) (y_j - \bar{Y}) \quad (15)$$

(c) The general form of variance of the linear estimator and the ratio estimator is defined by:

$$\begin{aligned} V(\hat{Y}_r^{(m)}) &\approx \frac{1}{p} \sum_{i \in U} \frac{(1 - \pi_i)}{\pi_i} (y_i - \bar{Y}_m)^2 + \\ &\sum_{i \in U} \sum_{j \neq i \in U} D_{ij} (y_i - \bar{Y}_m) (y_j - \bar{Y}_m) \end{aligned} \quad (16)$$

where $\bar{Y}_m = \bar{Y}$, if $m = 2$; otherwise $\bar{Y}_m = 0$.

The proof of Theorem 3 is given in the Appendix.

Theorem 4. Assume that (A) and (B) remain correct. Under a reverse framework with unequal probability sampling without replacement:

The asymptotically unbiased estimator of variance of the linear estimator, $V(\hat{Y}_r^{(1)})$, is obtained by:

$$\hat{V}(\hat{Y}_r^{(1)}) = \begin{cases} \frac{1}{p^2} \left[\sum_{i \in s} \hat{D}_i r_i y_i^2 + \sum_{i \in s} \sum_{j \neq i \in s} \hat{D}_{ij} r_i r_j y_i y_j \right], & \text{when } p \text{ is known} \\ \left(\frac{\sum_{i \in s} \frac{1}{\pi_i}}{\sum_{i \in s} \frac{r_i}{\pi_i}} \right)^2 \left[\sum_{i \in s} \hat{D}_i r_i y_i^2 + \sum_{i \in s} \sum_{j \neq i \in s} \hat{D}_{ij} r_i r_j y_i y_j \right], & \text{when } p \text{ is unknown,} \end{cases} \quad (17)$$

where

$$\hat{D}_i = (1 - \pi_i) \pi_i^{-2} \text{ and } \hat{D}_{ij} = (\pi_{ij} - \pi_i \pi_j) (\pi_{ij} \pi_i \pi_j)^{-1}.$$

(b) The asymptotically unbiased estimator of variance of the ratio estimator, $V(\hat{Y}_r^{(2)})$, is obtained by:

$$\hat{V}(\hat{Y}_r^{(2)}) = \left(\frac{N}{\sum_{i \in s} \frac{r_i}{\pi_i}} \right)^2 \left[\sum_{i \in s} \hat{D}_i r_i (y_i - \hat{\bar{Y}}_r)^2 + \sum_{i \in s} \sum_{j \neq i \in s} \hat{D}_{ij} r_i (y_i - \hat{\bar{Y}}_r) r_j (y_j - \hat{\bar{Y}}_r) \right], \quad (18)$$

$$\text{where } \hat{N}_r = \sum_{i \in s} \frac{r_i}{\pi_i} \text{ and } \hat{\bar{Y}}_r = \sum_{i \in s} \frac{r_i y_i}{\pi_i} / \hat{N}_r.$$

(c) The general form of variance estimator of the linear estimator and the ratio estimator is defined by:

$$\hat{V}(\hat{Y}_r^{(m)}) = \left(\frac{\hat{N}_m}{\hat{N}_{rm}} \right)^2 \left[\sum_{i \in s} \hat{D}_i r_i (y_i - \hat{\bar{Y}}_{rm})^2 + \sum_{i \in s} \sum_{j \neq i \in s} \hat{D}_{ij} r_i (y_i - \hat{\bar{Y}}_{rm}) r_j (y_j - \hat{\bar{Y}}_{rm}) \right] \quad (19)$$

where $\hat{N}_m = 1$, if $m = 1$ and p is known, and

$$\hat{N}_m = \sum_{i \in s} \frac{1}{\pi_i}, \text{ if } m = 1 \text{ and } p \text{ is unknown;}$$

otherwise $\hat{N}_m = N$. $\hat{N}_{rm} = 1$, if $m = 1$ and p is

known; otherwise $\hat{N}_{rm} = \sum_{i \in s} \frac{r_i}{\pi_i}$. $\hat{\bar{Y}}_{rm} = \hat{\bar{Y}}_r$, if

$m = 2$; otherwise $\hat{\bar{Y}}_{rm} = 0$.

The proof of Theorem 4 is given in the Appendix.

Simulation Study and Results

In this section we compared the efficiency of the variance estimator of the linear estimator, as defined in Equation (17), against the variance estimator of the ratio estimator, as defined in Equation (18), using a simulation study. We considered 2 different unequal probability samplings without replacement. The first one is Midzuno's (1952) scheme and the second is a probability proportional to size systematic scheme. The sampling methods are described below:

Midzuno's (1952) Scheme

Midzuno (1952) proposed a simple scheme for unequal probability sampling without replacement. Let k_i , $i \in U$ denote the size of unit i and where $P_i = k_i / \sum_{i \in U} k_i$ is the probability of selection of unit i . We note that, if P_i is the probability of selection of unit i , then $0 \leq P_i \leq 1$, i.e., $k_i \geq 0$ for all $i \in U$. Berger (2005) stated that, under unequal probability sampling, π_i is a function of k_i , i.e. $\pi_i = f(k_i)$ for all $i \in U$ where k_i is a non-negative value of size variable. Then, a sample s with size n is selected using 2 different sampling methods. A unit sample is selected by probability proportional to size and then the rest of the samples with size $n-1$ are selected from the population size $N-1$ by simple random sampling without replacement (SRSWOR). In this paper the simulation of Midzuno's (1952) sample is selected by program R through the "UPmidzuno" command (Tillé and Matei, 2009). Finally the first and second order of inclusion probabilities for this sampling scheme are given, respectively, by

$$\pi_i = P_i \left(\frac{N-n}{N-1} \right) + \frac{n-1}{N-1}$$

$$\pi_{ij} = (P_i + P_j) \frac{N-n}{N-1} \frac{n-1}{N-2} + \frac{n-1}{N-1} \frac{n-2}{N-2}$$

Probability Proportional to Size Systematic Scheme

Let $T_i = \sum_{j \leq i \in U} k_j$ be the cumulative total and $\{r+lc\}$ be corresponding units where $j = 1, 2, \dots, N$; $l = 0, 1, 2, \dots, (n-1)$, $c = \sum_{i \in U} k_i / n$ and r represents a random number from 1 to c . The sample is included with the unit i if $T_{i-1} < r+lc \leq T_i$ for some value of $l = 0, 1, 2, \dots, (n-1)$. In this paper the simulation of a systematic sample is selected by program R through an "UP systematic" command (Tillé and Matei, 2009). The first order of inclusion probabilities must be also obtained by $\pi_i = nP_i$ where $P_i = k_i / \sum_{i \in U} k_i$. In addition, Hájek (1981) can solve the second order of inclusion probabilities as follows:

$$\pi_{ij} \approx \pi_i \pi_j \left[1 - (1 - \pi_i)(1 - \pi_j) \left(\sum_{k \in U} \pi_k (1 - \pi_k) \right)^{-1} \right]$$

Finally the variable y_i is generated following Thompson *et al.* (2010) defined by $y_i = k'_i \beta + \varepsilon_i$. In simulation the following steps must be considered:

Step 1. In order to generate y_i , we generate $k_i \sim \text{Gam}(4, 5)$, $\varepsilon_i \sim N(0, 0.7)$, and $\beta_0 = 0$, $\beta_1 = 1$, $i = 1, 2, \dots, N$ where $N = 1,000$.

Step 2. The samples S with size $n = 100$ and 300 are selected according to Midzuno's (1952) scheme and the probability proportional to size systematic scheme.

Step 3. Sample size n and non-response are generated using a uniform response mechanism with a probability of response $p = 0.7$.

Step 4. Compute the estimator of variance of the linear estimator as Equation (17) and of the ratio estimator as Equation (18).

Step 5. Repeat steps (2) to (4) for 5,000 times ($M = 5,000$).

Step 6. The simulation relative bias of the variance estimator $\hat{V}(Y_r^{(m)})$ is given by:

$$RB_{SIM}[\hat{V}(\hat{Y}_r^{(m)})] = \frac{E_{SIM}[\hat{V}(\hat{Y}_r^{(m)})] - V_{SIM}(\hat{Y}_r^{(m)})}{V_{SIM}(\hat{Y}_r^{(m)})}$$

where $E_{SIM}[\hat{V}(\hat{Y}_r^{(m)})] = \frac{1}{M} \sum_{j=1}^M \hat{V}[\hat{Y}_{r(j)}^{(m)}]$ is

the simulated expectation of the variance estimator and

$$V_{SIM}(\hat{Y}_r^{(m)}) = \frac{1}{M-1} \sum_{j=1}^M [\hat{Y}_{r(j)}^{(m)} - E_{SIM}(\hat{Y}_r^{(m)})]^2$$

is the simulated variance of the estimator $\hat{Y}_r^{(m)}$,

$m = 1, 2, \dots$, $E_{SIM}(\hat{Y}_r^{(m)})$ is defined by

$$E_{SIM}(\hat{Y}_r^{(m)}) = \frac{1}{M} \sum_{j=1}^M \hat{Y}_{r(j)}^{(m)}$$

Table 1. The simulation relative bias of $\hat{V}(Y_r^{(m)})$ under Midzuno (1952) and a probability proportional to size systematic schemes

Sampling design	n	Simulation Relative Bias	
		$\hat{V}(\hat{Y}_r^{(1)})$	$\hat{V}(\hat{Y}_r^{(2)})$
Midzuno (1952) scheme	100	113.44	-23.59
	300	87.55	-14.12
Probability proportional to size systematic scheme	100	164.80	19.40
	300	92.40	12.66

Table 1 shows the simulation relative bias of the variance estimators for both the linear and ratio estimators under Midzuno (1952) and the probability proportional to size systematic schemes. The results show the relative bias in the variance of the ratio estimator to be a lot less than the variance of the linear estimator for both schemes.

Conclusions

In this paper, linear and ratio estimators were proposed for estimating population totals in the presence of unit nonresponse. We have also proposed a general form of the variance and the estimated variance of the proposed estimators. We find a new way of estimating variance under a reverse framework when unit nonresponse occurs in the study. It is assumed that the sampling fraction is negligible under unequal probability sampling without replacement while response probability is equal for all units leading to neither known nor estimated response probabilities for the variance of the ratio estimator. In theory, we proved that the estimator of variance of the linear and ratio estimators are asymptotically unbiased while the simulation's relative bias showed that the ratio estimator was more efficient than the linear estimator.

Acknowledgments

We are grateful to Dr. Yves Berger (University of Southampton, UK) for suggesting this topic and to the referees for helpful comments.

Appendix

Proof of Theorem 3

Recall from Equation (12) that the variance of the linear and the ratio estimators can be derived from:

$$V(\hat{Y}_r^{(m)}) \approx \sum_{i \in U} D_i E_R(z_{mi}^2) + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} E_R(z_{mi}) E_R(z_{mj}), \quad (A1)$$

$$\text{where } z_{1i} = \frac{r_i y_i}{p}, \quad z_{2i} = \frac{N}{N_r} r_i (y_i - \bar{Y}_r), \quad \text{and} \\ N_r = \sum_{i \in U} r_i.$$

First of all, we note that $r_i^2 = r_i$. From Equation (A1), if $m = 1$ we have the variance of the linear estimator which is given by:

$$V(\hat{Y}_r^{(1)}) \approx \sum_{i \in U} D_i E_R(z_{1i}^2) + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} E_R(z_{1i}) E_R(z_{1j}). \quad (A2)$$

$$\text{Recall from Equation (A1) that we have} \\ z_{1i} = \frac{r_i y_i}{p}; \text{ therefore, } E_R(z_{1i}) = y_i \text{ and } E_R(z_{1i}^2) = \frac{y_i^2}{p}.$$

Substitute $E_R(z_{1i})$ and $E_R(z_{1i}^2)$ into Equation (A2) and we have:

$$V(\hat{Y}_r^{(1)}) \approx \sum_{i \in U} D_i \frac{y_i^2}{p} + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} y_i y_j \\ = \sum_{i \in U} D_{1i} y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} y_i y_j, \quad (A3)$$

where $D_{1i} = D_i p^{-1}$.

(b) The proof of (b) is similar to (a) and it is obtained by:

$$V(\hat{Y}_r^{(2)}) \approx \sum_{i \in U} D_{1i} (y_i - \bar{Y})^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} (y_i - \bar{Y})(y_j - \bar{Y}). \quad (A4)$$

(c) From Equations (A3) and (A4) we may write,

$$V(\hat{Y}_r^{(m)}) \approx \sum_{i \in U} D_{1i} (y_i - \bar{Y}_m)^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} (y_i - \bar{Y}_m)(y_j - \bar{Y}_m), \quad (A4)$$

where $\bar{Y}_m = \bar{Y}$, if $m = 2$; otherwise $\bar{Y}_m = 0$.

Proof of Theorem 4

By using Equations (2) and (10) under unequal probability sampling without replacement, $V(\hat{Y}_r^{(m)})$ can be estimated by:

$$\hat{V}(\hat{Y}_r^{(m)}) \approx \sum_{i \in S} \hat{D}_i \hat{z}_{mi}^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} \hat{z}_{mi} \hat{z}_{mj}, \quad (B1)$$

$$\text{where, } \hat{z}_{1i} = \frac{r_i y_i}{p}, \quad \hat{z}_{2i} = \frac{N}{\hat{N}_r} r_i (y_i - \hat{Y}_r), \quad \hat{N}_r = \sum_{i \in S} \frac{r_i}{\pi_i},$$

$$\hat{Y}_r = \sum_{i \in S} \frac{r_i y_i}{\pi_i} / \hat{N}_r, \quad \hat{D}_i = (1 - \pi_i) \pi_i^{-2}, \quad \text{and}$$

$$\hat{D}_{ij} = (\pi_{ij} - \pi_i \pi_j) (\pi_{ij} \pi_i \pi_j)^{-1}.$$

(b) Assume p is known. From Equation (B1),

if $m = 1$ we have $\hat{z}_{1i} = \frac{r_i y_i}{p}$; therefore,

$$\begin{aligned} \hat{V}(\hat{Y}_r^{(1)}) &\approx \sum_{i \in S} \hat{D}_i \hat{z}_{1i}^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} \hat{z}_{1i} \hat{z}_{1j} \\ &= \sum_{i \in S} \hat{D}_i \frac{r_i^2 y_i^2}{p^2} + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} \left(\frac{r_i y_i}{p} \right) \left(\frac{r_j y_j}{p} \right) \\ &= \frac{1}{p^2} \left[\sum_{i \in S} \hat{D}_i r_i^2 y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \right]. \end{aligned} \quad (B2)$$

From Equation (B2), the overall expectation of $\hat{V}(\hat{Y}_r^{(1)})$ is obtained by:

$$\begin{aligned} E[\hat{V}(\hat{Y}_r^{(1)})] &\approx E_R E_S [\hat{V}(\hat{Y}_r^{(1)}) | \mathbf{R}] \\ &= E_R E_S \left[\frac{1}{p^2} \left(\sum_{i \in S} \hat{D}_i y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \right) \middle| \mathbf{R} \right] \\ &= \frac{1}{p^2} E_R E_S \left[\sum_{i \in S} \hat{D}_i y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \middle| \mathbf{R} \right]. \end{aligned} \quad (B3)$$

$$\text{To find } E_R E_S \left[\sum_{i \in S} \hat{D}_i y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \middle| \mathbf{R} \right]$$

in Equation (B3),

$$E_R E_S \left[\sum_{i \in S} \hat{D}_i y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \middle| \mathbf{R} \right]$$

$$= E_R E_S \left[\sum_{i \in U} \frac{(1 - \pi_i) I_i}{\pi_i^2} r_i y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} I_i I_j r_i r_j y_i y_j \right]$$

$$\begin{aligned} &\approx \sum_{i \in U} \frac{(1 - \pi_i) E_S(I_i)}{\pi_i^2} E_R(r_i) y_i^2 + \\ &\sum_{i \in U} \sum_{j \neq i \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} E_S(I_i I_j) E_R(r_i r_j) y_i y_j \\ &= \sum_{i \in U} \frac{(1 - \pi_i) \pi_i}{\pi_i^2} p y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \pi_{ij} p^2 y_i y_j \\ &= \sum_{i \in U} \frac{(1 - \pi_i)}{\pi_i} p y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} p^2 y_i y_j. \end{aligned}$$

Therefore,

$$\begin{aligned} E_R E_S \left[\sum_{i \in S} \hat{D}_i y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \middle| \mathbf{R} \right] \\ = \sum_{i \in U} \frac{(1 - \pi_i)}{\pi_i} p y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} p^2 y_i y_j. \end{aligned} \quad (B4)$$

Substitute Equation (B4) into Equation (B3) and we have,

$$\begin{aligned} E[\hat{V}(\hat{Y}_r^{(1)})] &\approx \frac{1}{p^2} \left[\sum_{i \in U} \frac{(1 - \pi_i) \pi_i}{\pi_i^2} p y_i^2 + \right. \\ &\left. \sum_{i \in U} \sum_{j \neq i \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \pi_{ij} p^2 y_i y_j \right] \\ &= \sum_{i \in U} \frac{D_i}{p} y_i^2 + \sum_{i \in U} \sum_{j \neq i \in U} D_{ij} y_i y_j = \sum_{i \in U} D_{1i} y_i^2 + \\ &\sum_{i \in U} \sum_{j \neq i \in U} D_{ij} y_i y_j \\ &= V(\hat{Y}_r^{(1)}). \end{aligned}$$

Therefore, $E[\hat{V}(\hat{Y}_r^{(1)})] \approx V(\hat{Y}_r^{(1)})$; in

the other words, $\hat{V}(\hat{Y}_r^{(1)})$ is an asymptotically unbiased estimator of $V(\hat{Y}_r^{(1)})$.

(c) The proof of (c) is similar to (b) and it is obtained by:

$$\hat{V}(\hat{Y}_r^{(1)}) = \left(\frac{\sum_{i \in S} \frac{1}{\pi_i}}{\sum_{i \in S} \frac{r_i}{\pi_i}} \right)^2 \left[\sum_{i \in S} \hat{D}_i r_i y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i r_j y_i y_j \right]. \quad (B5)$$

(d) The proof of (d) is similar to (b) and it is given by:

$$\hat{V}(\hat{Y}_r^{(2)}) = \left(\frac{\frac{N}{\sum_{i \in S} \frac{r_i}{\pi_i}}}{\sum_{i \in S} \frac{r_i}{\pi_i}} \right)^2 \left[\sum_{i \in S} \hat{D}_i r_i (y_i - \hat{Y}_r)^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i (y_i - \hat{Y}_r) r_j (y_j - \hat{Y}_r) \right]. \quad (B6)$$

(e) From Equations (B3), (B5), and (B6), the general form of variance estimator of the linear estimator and the ratio estimator are defined by:

$$\hat{V}(\hat{Y}_r^{(m)}) = \left(\frac{\hat{N}_m}{\sum_{i \in S} \frac{r_i}{\pi_i}} \right)^2 \left[\sum_{i \in S} \hat{D}_i r_i (y_i - \hat{Y}_{rm})^2 + \sum_{i \in S} \sum_{j \neq i \in S} \hat{D}_{ij} r_i (y_i - \hat{Y}_{rm}) r_j (y_j - \hat{Y}_{rm}) \right], \quad (B7)$$

where $\hat{N}_m = 1$ if $m = 1$ and p are known, and

$\hat{N}_m = \sum_{i \in S} \frac{1}{\pi_i}$ if $m = 1$ and p are unknown;

otherwise $\hat{N}_m = N$. $\hat{N}_{rm} = 1$ if $m = 1$ and p are

known; otherwise $\hat{N}_{rm} = \sum_{i \in S} \frac{r_i}{\pi_i}$. $\hat{Y}_{rm} = \hat{Y}_r$ if m

$= 2$; otherwise $\hat{Y}_{rm} = 0$.

References

Berger, Y.G. (2005). Variance estimation with Chao's sampling scheme. *J. Stat. Plan. Infer.*, 127:253-277.
 Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *J. Off. Stat.*, 3:251-260.

Deville, J. C. and Särndal, C. E. (1995). Variance estimation for the regression imputed Horvitz-Thompson estimator. *J. Off. Stat.*, 4:381-394.
 Fay, R.E. (1991). A design-based perspective on missing data variance. Proceedings of the 1991 US Bureau of the Census Annual Research Conference; March 17-20, 1991; Arlington, VA, USA, p. 447-450.
 Hájek, J. (1981). Sampling in Finite Population. Marcel Dekker, New York, NY, USA, 264p.
 Horvitz, D.F. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 260:663-685.
 Kajal, D. (2014). Estimating population mean with missing data in unequal probability sampling. *Statistics in Transition*, 3:369-388.
 Lawson, N. (2017). Variance estimation in the presence of nonresponse under probability proportional to size sampling. Proceedings of 6th Annual International Conference on Computational Mathematics, Computational Geometry & Statistics; March 6-7, 2017; Singapore; p. 116:119.
 Lawson, N. and Ponkaew, C. (2019). New generalized regression estimator in the presence of nonresponse under unequal probability sampling. *Comm. Stat. A-Theor.*, 48(10):2,483-2,498,
 Midzuno, H. (1952). On sampling system with probability proportional to sum of sizes. *Ann. I. Stat. Math.*, 3:99-107.
 Särndal, C. E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Surv. Methodol.*, 2:241-252.
 Särndal, C. E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. John Wiley & Sons, Chichester, UK, 194p.
 Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *J. Am. Stat. Assoc.*, 94:54-265.
 Thompson, K.J., Yung, W., and Haziza, D. (2010). The effect of nonresponse adjustments on variance estimation. *Surv. Methodol.*, 36:35-43.
 Tillé, Y. and Matei, A. (2009). The package "sampling". R contributed package. Retrieved from <https://cran.r-project.org/web/packages/sampling/sampling.pdf>.