Chiang Mai J. Sci. 2014; 41(3) : 704-713
http://epg.science.cmu.ac.th/ejournal/
Contributed Paper

Model-Assisted Estimation in Inverse Sampling

Sureeporn Sungsuwan*[a] and Prachoom Suwattee [b]

[a] Department of Mathematics, Faculty of Science, Mahanakorn University of Technology, Nong Chok, Bangkok 10530, Thailand.

[b] School of Applied Statistics, National Institute of Development Administration, Bangkapi, Bangkok 10240, Thailand.

*Author for correspondence; e-mail: sureepor@mut.ac.th

Received: 22 August 2011
Accepted: 6 November 2012

ABSTRACT

The purpose of this study is to propose estimators of the population total and the population mean using a model-assisted approach in inverse simple random sampling with and without replacement. It was found that the proposed estimators are biased, so the mean squared errors of the proposed estimators were investigated. The precision of the estimators is compared with those of the unbiased estimators given by Greco and Naddeo [2]. The simulation results indicate that the absolute relative biases of the proposed estimators decrease when the correlation between *X*, the auxiliary or independent variable and *Y*, the study variable increase. The absolute relative biases are small for all situations. As for the mean squared error estimates of the model-assisted estimate of the population total and the variance estimates of the unbiased estimate, it can be seen for sampling both with and without replacement that if the population prevalence increases then the variance estimates and the mean squared errors increase. They also decrease when the value of *m*, the number of units satisfying specified conditions in the samples increases. At low correlation between the auxiliary variable *X* and the study variable *Y*, the model-assisted estimator is as efficient as the unbiased estimator, whereas at high correlation between *X* and *Y*, the model-assisted estimator is considerably more efficient than the unbiased estimator.

Keywords: inverse simple random sampling, model assisted estimator, sampling with replacement, sampling without replacement

1. INTRODUCTION

For some situations in survey sampling, it is found that many of the units satisfying some conditions may not be present in the sample if a traditional sampling design such as simple random sampling is used. This situation leads a problem in estimating some population parameters. In such a case, an efficient sampling design may be inverse sampling which is a method of sampling that requires continue drawing of units until certain number of units satisfying specified conditions dependent on the results of those drawings has been fulfilled. In other words, the drawing of units is continued until a certain number of units with a characteristic of interest is attained. Hence, the sample size becomes a random variable which has negative binomial distribution for inverse sampling with replacement. For inverse sampling without replacement, the sample size has negative hypergeometric distribution [4].

Many papers on inverse sampling with and without replacement have been presented in various statistical journals. Haldane [3] used inverse simple random sampling with replacement in the estimation of the unknown population prevalence of units with some characteristic of interest. Recently, Christman and Lan [1] considered inverse simple random sampling with and without replacement and gave an unbiased estimator of the population total and its variance. Salehi and Seber [6,7] considered inverse simple random sampling without replacement. They gave an unbiased estimator of the population total and its variance based on Murthy's estimator. Greco and Naddeo [2] considered inverse sampling with replacement when the population units were drawn with unequal probabilities. They derived unbiased estimators of the population total, their variances and an unbiased variance estimator in inverse sampling with replacement. For inverse simple random sampling with and without replacement, they gave an estimator of the population total for each case as well as their variances, which are equivalent to the expressions given by Christman and Lan [1].

In some situations, estimators of certain parameters can be derived from information on auxiliary variable. Särndal, *et. al.* [5] proposed model assisted estimators to improve its precision. In this study, we used model assisted approach to improve the traditional estimators under the inverse simple random sampling with and without replacement.

2. Traditional Estimators

Let $U=\{u_1, u_2,..., u_N\}$ be a population of N distinguishable units. For simplicity, we denote the i^{th} unit by its label i, so the set of N population units can be written as $U = \{1,..., i,..., N\}$ with study values$\{y_1, y_2,..., y_N\}$, respectively. Divide U into 2 disjoint subsets, C and $\bar{C}$ according to the $y-$ values. Let C be the set of M units satisfying a condition and $\bar{C}$ the set of $N–M$ units not satisfying the condition. It is assumed that a unit satisfies the condition if it has the value of y greater than or equal to a constant c. We can write $C=\{i_1,i_2,...,i_M\}$ and $\bar{C} =\{i_{M+1},i_{M+2},...,i_N\}$ where $U = C\cup\bar{C}$ and $C\cap\bar{C} = \varnothing$.

Consider inverse simple random sampling with replacement from a population of size N when the sampling continues until a prespecified number m of units from the set C are obtained in the variable sample of size n. This sample can also be divided into 2 disjoint subsets, the first is of m units from C denoted by s_C and the other of $n–m$ units from $\bar{C}$ denoted by $s_{\bar{C}}$ and $s = s_C\cup s_{\bar{C}}$ with $s_C\cap s_{\bar{C}} = \varnothing$.

Let y_i be the value of y from unit i, $i\in U$, and let $T_y = \Sigma_{i\in U} y_i$ be the unknown total of this study variable. Greco and Naddeo [2] gave an unbiased estimator of T_y as

$$\hat{T}_{y,GN} = N[\hat{P}\,\bar{y}_C + (1-\hat{P})\,\bar{y}_{\bar{C}}] \qquad (1)$$

where $\hat{P} = (m-1)/(n-1)$ is an unbiased estimator of the prevalence of units in the set C, where $\bar{y}_C$, $\bar{y}_{\bar{C}}$ are the sample means of the study variable in C and $\bar{C}$, respectively. For sampling with replacement, the variance of the estimator (1) is given as

$$V(\hat{T}_{y,GN}) = N^2\,\Big[\,(\bar{Y}_C - \bar{Y}_{\bar{C}})^2 V_n(\hat{P}) + \frac{\sigma^2_{y_C}}{m} E_n(\hat{P}^2) + \frac{\sigma^2_{y_{\bar{C}}}}{m-1} E_n[\hat{P}(1-\hat{P})]\,\Big], \qquad (2)$$

where $(\bar{Y}_C,\sigma^2_{y_C})$ and $(\bar{Y}_{\bar{C}},\sigma^2_{y_{\bar{C}}})$ are the means and the variances of the study variable in C and $\bar{C}$, respectively, $E_n(\cdot)$, $V_n(\cdot)$ are expectation and variance with respect to the distribution of n and $V(\cdot)$ is variance with respect to the sampling design. They also gave an unbiased estimator of $V(\hat{T}_{y,GN})$ (2) in the form

$$\hat{V}(\hat{T}_{y,GN}) = N^2\,[\,(\bar{y}_C - \bar{y}_{\bar{C}})^2 \frac{\hat{P}(1-\hat{P})}{n-2} + \frac{s^2_{y_C}}{m}\hat{P}_2 + \frac{s^2_{y_{\bar{C}}}}{m-1}[\hat{P} - \frac{m-1}{m-2}\hat{P}_2]\,], \tag{3}$$

where $\hat{P}_2 = [(m–1)(m–2)]/[(n–1)(n–2)]$, and $S^2_{y_C}$, $S^2_{y_{\bar{C}}}$ the unbiased sample variances of the study variable in C and $\bar{C}$, respectively.

For sampling without replacement, the variance of the estimator (1) was given as

$$V(\hat{T}_{y,GN}) = N^2\left\{(\bar{Y}_C - \bar{Y}_{\bar{C}})^2 V_n(\hat{P}) + \frac{S^2_{y_C}}{m}\left(1-\frac{m}{M}\right)E_n(\hat{P}^2) + \frac{S^2_{y_{\bar{C}}}}{m-1}E_n\left[\hat{P}(1-\hat{P})\left(1-\frac{n-m}{N-M}\right)\right]\right\} \tag{4}$$

They also gave an unbiased estimator of the variance in (4) as

$$\hat{V}(\hat{T}_{y,GN}) = N^2\,[\,(\bar{y}_C - \bar{y}_{\bar{C}})^2 \hat{V}_n(\hat{P}) + \frac{s^2_{y_C}}{m}\left(\hat{P}_3 - \frac{m}{N}\hat{P}\right) + \frac{s^2_{y_{\bar{C}}}}{n-m}[\hat{P}_3 - \frac{n-m}{N}(1-\hat{P})]\,] \tag{5}$$

where $\hat{P}_3 = \frac{m-1}{n-1}\cdot\frac{m-2}{n-1} + \frac{1}{N}\cdot\frac{m-1}{n-1}\left(1-\frac{m-2}{n-1}\right)$..

For given n, the selection procedure under inverse sampling is the same as the selection procedure under stratified sampling with 2 strata where m m and $n–m$ are selected from the first and the second stratum. The sample results in the two groups are independent [2]. Let $T_{y_C} = \Sigma_{i\in C}\, y_i$ be the total of the study variable y in the set C. An unbiased estimator of T_{y_C} is given by the first part of expression in (1) with variance from sampling with replacement as

$$V(\hat{T}_{y_{C,GN}}) = N^2\,[\,\bar{Y}^2_C V_n(\hat{P}) + \frac{\sigma^2_{y_C}}{m}E_n(\hat{P}^2)\,] \tag{6}$$

An unbiased estimator of the variance in (6) is

$$\hat{V}(\hat{T}_{y_{C,GN}}) = N^2\,[\,\bar{y}^2_C \frac{\hat{P}(1-\hat{P})}{n-2} + \frac{s^2_{y_C}}{m}\hat{P}_2\,] \tag{7}$$

For sampling without replacement, the variance of the unbiased estimator of T_{y_C} is

$$V(\hat{T}_{y_C,GN}) = N^2\,[\,\frac{S^2_{y_C}}{m}\left(1-\frac{m}{M}\right)E_n(\hat{P}^2) + \bar{Y}^2_C V_n(\hat{P})\,] \tag{8}$$

An unbiased estimator of the variance in (8) is given as

$$\hat{V}(\hat{T}_{y_C,GN}) = N^2\,[\frac{s^2_{y_C}}{m}[\hat{P}_3 - m\hat{P}] + \bar{y}^2_C \hat{V}_n(\hat{P})\,] \tag{9}$$

3. Model-Assisted Estimators for Inverse Simple Random Sampling with Replacement

Suppose that (x_i, y_i), $i\in s$ are observed where y_i is the value of study variable of unit i and x_i is auxiliary value. From a finite population of size N, assume that $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$. An estimator of β_1 is

$$b_1 = \frac{\left(\frac{\hat{M}}{m}\right)\sum_{i\in s_C}(x_i - \bar{x}_C)(y_i - \bar{y}_C) + \left(\frac{N-\hat{M}}{n-m}\right)\sum_{i\in s_{\bar{C}}}(x_i - \bar{x}_{\bar{C}})(y_i - \bar{y}_{\bar{C}})}{\left(\frac{\hat{M}}{m}\right)\sum_{i\in s_C}(x_i - \bar{x}_C)^2 + \left(\frac{N-\hat{M}}{n-m}\right)\sum_{i\in s_{\bar{C}}}(x_i - \bar{x}_{\bar{C}})^2}$$

where $\hat{M} = N\hat{P}$ is an unbiased estimator of M, $\bar{x}_C = m^{-1}\Sigma_{i\in s_C} x_i$.

The proposed model-assisted estimator of the population total T_y can be written as

$$\tilde{T}_y = \hat{M}\,[\bar{y}_C + b_1(\bar{X} - \bar{x}_C)] + (N-\hat{M})[\bar{y}_{\bar{C}} + b_1(\bar{X} - \bar{x}_{\bar{C}})] \tag{10}$$

where $\bar{X} = N^{-1}\Sigma_{i\in U} x_i$ is the population mean of auxiliary variable x_i, $i\in U$.

Theorem 3.1 A model assisted estimator of T_y in (10) is biased.

Proof $E(\tilde{T}_y) = E_n E[\tilde{T}_y \mid n]$

$$\begin{aligned}
&= E_n\{E[(\hat{M}\bar{y}_C + (N-\hat{M})\bar{y}_{\bar{C}}) \mid n]\} \\
&\quad - E[b_1(\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}} - T_x) \mid n] \\
&= E_n[\hat{M}\bar{Y}_C + (N-\hat{M})\bar{Y}_{\bar{C}}] \\
&\quad - E_n[E[b_1(\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}}) \mid n] \\
&\quad - E(b_1 T_x) \mid n] \\
&= M\bar{Y}_C + (N-M)\bar{Y}_{\bar{C}} \\
&\quad - \{E[b_1(\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}})] \\
&\quad - E_n E(b_1 \mid n) T_x]\} \\
&= T_y - \{E[b_1(\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}})] \\
&\quad - E(b_1) E_n[E(\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}}) \mid n]\} \\
&= T_y - Cov(b_1, (\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}})) \\
&= T_y - Cov(b_1, \hat{T}_x)
\end{aligned}$$

where $\hat{T}_x = \hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}}$.
Therefore, the bias of $\tilde{T}_y$ is
$B(\tilde{T}_y) = E(\tilde{T}_y) - T_y = -Cov(b_1, \hat{T}_x)$.

Since the properties of a model-assisted estimator with respect to the sampling design are usually of a complex form ([5]: 235), they cannot be studied exactly, so approximation expressions are used to study them.

Theorem 3.2 An estimator $\tilde{T}_y$ in (10) is approximately.

$$\begin{aligned}
\tilde{T}_y^* &= M\bar{y}_C + (N-M)\bar{y}_{\bar{C}} \\
&\quad - B_1[(M\bar{x}_C + (N-M)\bar{x}_{\bar{C}}) - N\bar{X}] \\
&\quad + [(\bar{Y}_C - B_1\bar{X}_C) - (\bar{Y}_{\bar{C}} - B_1\bar{X}_{\bar{C}})](\hat{M} - M) \qquad (11)
\end{aligned}$$

with $MSE(\tilde{T}_y) \approx V(\tilde{T}_y^*)$

$$\begin{aligned}
&= \frac{M^2}{m}(\sigma^2_{D_C} - \sigma^2_{D_{\bar{C}}}) + MN\frac{(m+1)}{m^2}\sigma^2_{D_{\bar{C}}} \\
&\quad + (\bar{D}_C - \bar{D}_{\bar{C}})^2 V_n(\hat{M}). \qquad (12)
\end{aligned}$$

Proof For a given n,

$$\begin{aligned}
\tilde{T}_y &= \hat{M}[\bar{y}_C + b_1(\bar{X} - \bar{x}_C)] \\
&\quad + (N-\hat{M})[\bar{y}_{\bar{C}} + b_1(\bar{X} - \bar{x}_{\bar{C}})] \\
&= \hat{M}\bar{y}_C + (N-\hat{M})\bar{y}_{\bar{C}} \\
&\quad + b_1[N\bar{X} - (\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}})] \\
&= h(\bar{y}_C, \bar{y}_{\bar{C}}, b_1, \bar{x}_C, \bar{x}_{\bar{C}}, M)
\end{aligned}$$

Thus, $\tilde{T}_y$ is a nonlinear function of the estimators. From the application of a Taylor linearization technique, this function is approximated by a linear function. The estimator (10) becomes

$$\begin{aligned}
\tilde{T}_y \approx \tilde{T}_y^* &= M\bar{y}_C + (N-M)\bar{y}_{\bar{C}} \\
&\quad - B_1[(M\bar{x}_C + (N-M)\bar{x}_{\bar{C}}) - N\bar{X}] \\
&\quad + [(\bar{Y}_C - B_1\bar{X}_C) - (\bar{Y}_{\bar{C}} - B_1\bar{X}_{\bar{C}})](\hat{M} - M)
\end{aligned}$$

and

$$E(\tilde{T}_y^*) = E_n[E(\tilde{T}_y^* \mid n)] = T_y.$$

Thus, $\tilde{T}_y^*$ is an unbiased estimator of T_y. Consider

$$\begin{aligned}
\tilde{T}_y^* &= NB_1\bar{X} + \frac{M}{m}\sum_{i\in s_C}(y_i - B_1x_i) \\
&\quad + \frac{N-M}{n-m}\sum_{i\in s_{\bar{C}}}(y_i - B_1x_i) \\
&\quad + [(\bar{Y}_C - B_1\bar{X}_C) - (\bar{Y}_{\bar{C}} - B_1\bar{X}_{\bar{C}})](\hat{M} - M) \\
&= NB_1\bar{X} + \frac{M}{m}\sum_{i\in s_C} D_i + \frac{N-M}{n-m}\sum_{i\in s_{\bar{C}}} D_i \\
&\quad + [\bar{D}_C - \bar{D}_{\bar{C}}](\hat{M} - M)
\end{aligned}$$

where $D_i = y_i - B_1x_i$,

$$\begin{aligned}
\bar{D}_C &= \frac{1}{M}\sum_{i\in C}(y_i - B_1x_i) = \bar{Y}_C - B_1\bar{X}_C, \\
\bar{D}_{\bar{C}} &= \frac{1}{N-M}\sum_{i\in \bar{C}}(y_i - B_1x_i) = \bar{Y}_{\bar{C}} - B_1\bar{X}_{\bar{C}}.
\end{aligned}$$

$$\begin{aligned}
V(\tilde{T}_y^*) &= E_n[V(\tilde{T}_y^* \mid n)] + V_n[E(\tilde{T}_y^* \mid n)] \\
&= E_n\Big[M^2\frac{\sigma^2_{D_C}}{m} + (N-M)^2\frac{\sigma^2_{D_{\bar{C}}}}{n-m}\Big] \\
&\quad + (\bar{D}_C - \bar{D}_{\bar{C}})^2 V_n(\hat{M}) \\
&= M^2\frac{\sigma^2_{D_C}}{m} + (N-M)^2\sigma^2_{D_{\bar{C}}} E_n\Big(\frac{1}{n-m}\Big) \\
&\quad + (\bar{D}_C - \bar{D}_{\bar{C}})^2 V_n(\hat{M})
\end{aligned}$$

where $E(\frac{1}{m}\sum_{i\in s_C} D_i) = \frac{1}{M}\sum_{i\in C} D_i = \bar{D}_C$,

$$E(\frac{1}{n-m}\sum_{i\in s_{\bar{C}}} D_i) = \frac{1}{N-M}\sum_{i\in \bar{C}} D_i = \bar{D}_{\bar{C}},$$

$\sigma^2_{D_C} = \frac{1}{M}\sum_{i\in C}(D_i - \bar{D}_C)^2$ and

$$\sigma^2_{D_{\bar{C}}} = \frac{1}{N-M}\sum_{i\in\bar{C}}(D_i - \bar{D}_{\bar{C}})^2$$

By Taylor's expansion,

$$E_n(\frac{1}{n-m}) \approx \frac{mM(N-M)+MN}{m^2(N-M)^2},$$

$$\begin{aligned} V(\tilde{T}^*_y) &= M^2\frac{\sigma^2_{D_C}}{m} + \frac{mM(N-M)+MN}{m^2}\sigma^2_{D_{\bar{C}}} \\ &\quad +(\bar{D}_C - \bar{D}_{\bar{C}})^2 V_n(\hat{M}) \\ &= \frac{M^2}{m}(\sigma^2_{D_C} - \sigma^2_{D_{\bar{C}}}) + MN\frac{(m+1)}{m^2}\sigma^2_{D_{\bar{C}}} \\ &\quad +(\bar{D}_C - \bar{D}_{\bar{C}})^2 V_n(\hat{M}). \end{aligned}$$

The theorem is proved.

An estimator of $MSE(\tilde{T}_y)$ in (12) is in the form

$$\begin{aligned} M\hat{S}E(\tilde{T}_y) = {} & \frac{s^2_{d_C}}{m}\left(N^2\frac{m-1}{n-1}\cdot\frac{m-2}{n-2}\right) + \frac{s^2_{d_{\bar{C}}}}{n-m}(N-\hat{M})^2 \\ & + \left((\bar{d}_C - \bar{d}_{\bar{C}})^2 - \frac{s^2_{d_C}}{m} - \frac{s^2_{d_{\bar{C}}}}{n-m}\right)\frac{\hat{M}(N-\hat{M})}{n-2} \end{aligned} \qquad (13)$$

where $d_i = y_i - b_1 x_i$, $\bar{d}_C = \frac{1}{m}\sum_{i\in s_C} d_i$,

$\bar{d}_{\bar{C}} = \frac{1}{n-m}\sum_{i\in s_{\bar{C}}} d_i$, $s^2_{d_C} = \frac{1}{m-1}\sum_{i\in s_C}(d_i - \bar{d}_C)^2$ and

$$s^2_{d_{\bar{C}}} = \frac{1}{n-m-1}\sum_{i\in s_{\bar{C}}}(d_i - \bar{d}_{\bar{C}})^2.$$

Let $\bar{Y} = N^{-1}\sum_{i\in U} y_i$ be the mean of the study variable Y. A model-assisted estimator of the population mean is

$$\begin{aligned} \tilde{\bar{y}} = \frac{\tilde{T}_y}{N} = \frac{1}{N}\{ & \hat{M}[\bar{y}_C + b_1(\bar{X} - \bar{x}_C)] \\ & +(N-\hat{M})[\bar{y}_{\bar{C}} + b_1(\bar{X} - \bar{x}_{\bar{C}})]\} \end{aligned} \qquad (14)$$

The bias of $\tilde{\bar{y}}$ is equal to $-Cov(b_1,\bar{x})$. Since $\approx \frac{\tilde{T}^*_y}{N}$ where $\tilde{T}^*_y$ is as in theorem 3.2. So the mean squared error of $\tilde{\bar{y}}$ is

$$\begin{aligned} MSE(\tilde{\bar{y}}) \approx V\left(\frac{\tilde{T}^*_y}{N}\right) = \frac{1}{N^2}\Bigg[& \frac{M^2}{m}(\sigma^2_{D_C} - \sigma^2_{D_{\bar{C}}}) \\ & +MN\frac{(m+1)}{m^2}\sigma^2_{D_{\bar{C}}} + (\bar{D}_C - \bar{D}_{\bar{C}})^2 V_n(\hat{M})\Bigg] \end{aligned} \qquad (15)$$

An estimator of $MSE(\tilde{\bar{y}})$ is

$$\begin{aligned} M\hat{S}E(\tilde{\bar{y}}) = {} & \\ \frac{1}{N^2}\Bigg[& \frac{s^2_{d_C}}{m}\left(N^2\frac{m-1}{n-1}\cdot\frac{m-2}{n-2}\right) + \frac{s^2_{d_{\bar{C}}}}{n-m}(N-\hat{M})^2 \\ & + \left((\bar{d}_C - \bar{d}_{\bar{C}})^2 - \frac{s^2_{d_C}}{m} - \frac{s^2_{d_{\bar{C}}}}{n-m}\right)\frac{\hat{M}(N-\hat{M})}{n-2}\Bigg] \end{aligned} \qquad (16)$$

Let C be the set of M units satisfying certain given conditions. Then a model-assisted estimator of the population total of units in C can be written as

$$\tilde{T}_{y_C} = \hat{M}\bar{y}_C + b_1[N\bar{X} - (\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}})] \qquad (17)$$

The model-assisted estimator $\tilde{T}_{y_C}$ in (17) is also biased. The bias of $\tilde{T}_{y_C}$ is

$B(\tilde{T}_{y_C}) = -\{Cov(b_1,\hat{T}_{x_C}) + Cov(b_1,\hat{T}_{x_{\bar{C}}})\}$,

where $\hat{T}_{x_C} = \hat{M}\bar{x}_C$ and $\hat{T}_{x_{\bar{C}}} = (N-\hat{M})\bar{x}_{\bar{C}}$.

The mean squared error of $\tilde{T}_{y_C}$ is approximately

$$\begin{aligned} MSE(\tilde{T}_{y_C}) \approx {} & \frac{M^2}{m}(\sigma^2_{D_C} - B_1^2\sigma^2_{x_{\bar{C}}}) \\ & +MN\frac{(m+1)}{m^2}B_1^2\sigma^2_{x_{\bar{C}}} \\ & +[\bar{Y}_C - B_1(\bar{X}_C - \bar{X}_{\bar{C}})]^2 V_n(\hat{M}). \end{aligned} \qquad (18)$$

where $\sigma^2_{D_C} = \frac{1}{M}\sum_{i\in C}(D_i - \bar{D}_C)^2$ and

$$\sigma^2_{x_{\bar{C}}} = \frac{1}{N-M}\sum_{i\in\bar{C}}(x_i - \bar{X}_{\bar{C}})^2.$$

An estimator of $MSE(\tilde{T}_{y_C})$ is given by

$$\begin{aligned} M\hat{S}E(\tilde{T}_{y_C}) = {} & \frac{s^2_{d_C}}{m}\left(N^2\frac{m-1}{n-1}\cdot\frac{m-2}{n-2}\right) \\ & +b_1^2\frac{s^2_{x_{\bar{C}}}}{n-m}(N-\hat{M})^2 + \Bigg\{[(\bar{y}_C - b_1(\bar{x}_C - \bar{x}_{\bar{C}})]^2 - \frac{s^2_{y_C}}{m} \\ & -b_1^2\frac{s^2_{x_C}}{m} - b_1^2\frac{s^2_{x_{\bar{C}}}}{n-m} + 2b_1\,\mathrm{cov}(\bar{x}_C,\bar{y}_C)\Bigg\}\frac{\hat{M}(N-\hat{M})}{n-2} \end{aligned} \qquad (19)$$

where $s^2_{x_{\bar{C}}} = \frac{1}{n-m-1}\sum_{i\in s_{\bar{C}}}(x_i - \bar{x}_{\bar{C}})^2$,

$s^2_{y_C} = \frac{1}{m-1}\sum_{i\in s_C}(y_i - \bar{y}_C)^2$ and

$$\operatorname{cov}(\bar{x}_C, \bar{y}_C) = \frac{1}{m(m-1)}\sum_{i\in s_C}(x_i - \bar{x}_C)(y_i - \bar{y}_C).$$

Let $\bar{Y}_C = M^{-1}\sum_{i\in C} y_i$ be the population mean of units in *C*. A model-assisted estimator of $\bar{Y}_C$ is given as

$$\tilde{\bar{y}}_C = \bar{y}_C + b_1\left[\bar{X} - \frac{1}{N}(\hat{M}\bar{x}_C + (N-\hat{M})\bar{x}_{\bar{C}})\right] \qquad (20)$$

The bias of $\tilde{\bar{y}}_C$ is

$$B(\tilde{\bar{y}}_C) = -\left\{Cov\left(b_1, \frac{\hat{M}}{N}\bar{x}_C\right) + Cov\left(b_1, \frac{(N-\hat{M})}{N}\bar{x}_{\bar{C}}\right)\right\}.$$

The mean squared error of $\tilde{\bar{y}}_C$ is approximately

$$MSE(\tilde{\bar{y}}_C) \approx \frac{\sigma^2_{y_C}}{m} + \frac{B_1^2}{N^2}\frac{M^2}{m}(\sigma^2_{x_C} - \sigma^2_{x_{\bar{C}}})$$

$$+\frac{M}{N}\frac{(m+1)}{m^2}B_1^2\sigma^2_{x_{\bar{C}}} + \frac{B_1^2}{N^2}(\bar{X}_C - \bar{X}_{\bar{C}})^2 V_n(\hat{M}) \qquad (21)$$

where $\bar{X}_C = \frac{1}{M}\sum_{i\in C} x_i$, $\sigma^2_{x_C} = \frac{1}{M}\sum_{i\in C}(x_i - \bar{X}_C)^2$,

$\bar{X}_{\bar{C}} = \frac{1}{N-M}\sum_{i\in\bar{C}} x_i$, $\sigma^2_{x_{\bar{C}}} = \frac{1}{N-M}\sum_{i\in\bar{C}}(x_i - \bar{X}_{\bar{C}})^2$,

$\bar{Y}_C = \frac{1}{M}\sum_{i\in C} y_i$, $\sigma^2_{y_C} = \frac{1}{M}\sum_{i\in C}(y_i - \bar{Y}_C)^2$.

An estimator of $MSE(\tilde{\bar{y}}_C)$ is in the form

$$M\hat{S}E(\tilde{\bar{y}}_C) = \frac{s^2_{y_C}}{m} + b_1^2\frac{s^2_{x_C}}{m}\left(\frac{m-1}{n-1}\cdot\frac{m-2}{n-2}\right)$$

$$+\frac{b_1^2}{N^2}\frac{s^2_{x_{\bar{C}}}}{n-m}(N-\hat{M})^2 + \left\{\frac{b_1^2}{N^2}\left((\bar{x}_C - \bar{x}_{\bar{C}})^2\right.\right.$$

$$\left.\left.-\frac{s^2_{x_C}}{m} - \frac{s^2_{x_{\bar{C}}}}{n-m}\right)\right\}\frac{\hat{M}(N-\hat{M})}{n-2} \qquad (22)$$

where $\bar{x}_C = \frac{1}{m}\sum_{i\in s_C} x_i$, $\bar{x}_{\bar{C}} = \frac{1}{n-m}\sum_{i\in s_{\bar{C}}} x_i$,

$\bar{y}_C = \frac{1}{m}\sum_{i\in s_C} y_i$, $s^2_{x_C} = \frac{1}{m-1}\sum_{i\in s_C}(x_i - \bar{x}_C)^2$.

4. MODEL-ASSISTED ESTIMATORS FOR INVERSE SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

For inverse simple random sampling without replacement. The estimator of the population parameters is similar to the with replacement case but their variances are different. Consider the estimator of the population total in (10). For sampling without replacement, we have

$$E_n\left(\frac{1}{n-m}\right) \approx \frac{M+1}{m(N-M)} + \frac{(N+1)(M+1)(M-m+1)}{m^2(N-M)^2(M+2)}.$$

The mean squared error of the estimator (10) is

$$MSE(\tilde{T}_y) \approx$$

$$M^2\frac{S^2_{D_C}}{m}\left(1-\frac{m}{M}\right) + S^2_{D_{\bar{C}}}\left[\frac{(N-M)(M-m+1)}{m}\right.$$

$$\left.+\frac{(N+1)(M+1)(M-m+1)}{m^2(M+2)}\right] + [\bar{D}_C - \bar{D}_{\bar{C}}]^2 V_n(\hat{M}) \qquad (23)$$

where $S^2_{D_C} = \frac{1}{M-1}\sum_{i\in C}(D_i - \bar{D}_C)^2$ and

$S^2_{D_{\bar{C}}} = \frac{1}{N-M-1}\sum_{i\in\bar{C}}(D_i - \bar{D}_{\bar{C}})^2$.

An estimator of $MSE(\tilde{T}_y)$ is

$$M\hat{S}E(\tilde{T}_y) =$$

$$s^2_{d_C}\left[\frac{m-1}{n-1}\left(\frac{m-2}{n-2}\cdot\frac{N(N-1)}{m} + \frac{N}{m} - N\right)\right]$$

$$+s^2_{d_{\bar{C}}}\left(\frac{(N-\hat{M})^2}{n-m} - (N-\hat{M})\right)$$

$$+\left[(\bar{d}_C - \bar{d}_{\bar{C}})^2 - s^2_{d_C}\left(\frac{1}{m} - \frac{1}{\hat{M}}\right)\right.$$

$$\left.-s^2_{d_{\bar{C}}}\left(\frac{1}{n-m} - \frac{1}{N-\hat{M}}\right)\right]\frac{\hat{M}(N-\hat{M})}{n-2}\left(1 - \frac{n-1}{N}\right) \qquad (24)$$

Consider an estimator of the population mean. For inverse simple random sampling without replacement, the estimator given by equation (14) has a mean squared error as

$$MSE(\tilde{\bar{y}}) \approx V\left(\frac{\tilde{T}^*_y}{N}\right) = \frac{1}{N_2}\left\{M^2\frac{S^2_{D_C}}{m}\left(1 - \frac{m}{M}\right)\right.$$

$$+S_{D_{\overline{C}}}^{2}\left[\frac{(N-M)(M-m+1)}{m}+\frac{(N+1)(M+1)(M-m+1)}{m^{2}(M+2)}\right]$$

$$\left.+[\overline{D}_{C}-\overline{D}_{\overline{C}}]^{2}V_{n}(\hat{M})\right\} \qquad (25)$$

An estimator of $MSE(\tilde{\bar{y}})$ is $\frac{1}{N^{2}}(M\hat{S}E(\tilde{T}_{y}))$.

Consider at a model-assisted estimator of the population total of units in C. For sampling without replacement, the mean squared error of the estimator (17) is

$$MSE(\tilde{T}_{y_C})\approx\frac{M^{2}}{m}\left(1-\frac{m}{M}\right)S_{D_C}^{2}$$

$$+B_{1}^{2}S_{x_{\overline{C}}}^{2}\left[\frac{(N-M)(N-m+1)}{m}\right.$$

$$\left.+\frac{(N+1)(M+1)(M-m+1)}{m^{2}(M+2)}\right]$$

$$+[\overline{Y}_{C}-B_{1}(\overline{X}_{C}-\overline{X}_{\overline{C}})]^{2}V_{n}(\hat{M}) \qquad (26)$$

where $S_{x_{\overline{C}}}^{2}=\frac{1}{N-M-1}\sum_{i\in\overline{C}}(x_{i}-\overline{X}_{\overline{C}})^{2}$.

An estimator of $MSE(\tilde{T}_{y_C})$ is

$$M\hat{S}E(\tilde{T}_{y_C})=$$

$$s_{d_C}^{2}\left[\frac{m-1}{n-1}\left(\frac{m-2}{n-2}\cdot\frac{N(N-1)}{m}+\frac{N}{m}-N\right)\right]$$

$$+b_{1}^{2}s_{x_{\overline{C}}}^{2}\left(\frac{(N-\hat{M})^{2}}{n-m}-(N-\hat{M})\right)$$

$$+\left\{[\bar{y}_{C}-b_{1}(\bar{x}_{C}-\bar{x}_{\overline{C}})]^{2}-\left(\frac{1}{m}-\frac{1}{\hat{M}}\right)(s_{y_C}^{2}+b_{1}^{2}s_{x_C}^{2})\right.$$

$$-\left(\frac{1}{n-m}-\frac{1}{N-\hat{M}}\right)b_{1}^{2}s_{x_C}^{2}$$

$$\left.+2b_{1}\operatorname{cov}(\bar{x}_{C},\bar{y}_{C})\right\}\frac{\hat{M}(N-\hat{M})}{n-2}\left(1-\frac{n-1}{N}\right) \qquad (27)$$

A model-assisted estimator of the population mean of units in C in (20) has the mean squared error as

$$MSE(\tilde{\bar{y}}_{C})\approx$$

$$\left(1-\frac{m}{M}\right)\frac{S_{y_C}^{2}}{m}+\frac{B_{1}^{2}}{N^{2}}M^{2}\left(1-\frac{m}{M}\right)\frac{S_{x_C}^{2}}{m}$$

$$+\frac{B_{1}^{2}}{N^{2}}S_{x_{\overline{C}}}^{2}\left[\frac{(N-M)(N-m+1)}{m}\right.$$

$$\left.+\frac{(N+1)(M+1)(M-m+1)}{m^{2}(M+2)}\right]$$

$$+\frac{B_{1}^{2}}{N^{2}}(\overline{X}_{C}-\overline{X}_{\overline{C}})^{2}V_{n}(\hat{M}) \qquad (28)$$

where $S_{x_C}^{2}=\frac{1}{M-1}\sum_{i\in C}(x_{i}-\overline{X}_{C})^{2}$,

$S_{x_{\overline{C}}}^{2}=\frac{1}{N-M-1}\sum_{i\in\overline{C}}(x_{i}-\overline{X}_{\overline{C}})^{2}$ and

An estimator of $MSE(\tilde{\bar{y}}_{C})$ is given by

$$M\hat{S}E(\tilde{\bar{y}}_{C})=$$

$$s_{y_C}^{2}\left(\frac{1}{m}-\frac{1}{\hat{M}}\right)+s_{x_C}^{2}\frac{b_{1}^{2}}{N^{2}}\left[\frac{m-1}{n-1}\left(\frac{m-2}{n-2}\cdot\frac{N(N-1)}{m}\right.\right.$$

$$\left.\left.+\frac{N}{m}-N\right)\right]+\frac{b_{1}^{2}}{N^{2}}s_{x_{\overline{C}}}^{2}\left(\frac{(N-\hat{M})^{2}}{n-m}-(N-\hat{M})\right)$$

$$+\left\{\frac{b_{1}^{2}}{N^{2}}\left[(\bar{x}_{C}-\bar{x}_{\overline{C}})^{2}-s_{x\,C}^{2}\left(\frac{1}{m}-\frac{1}{\hat{M}}\right)\right.\right.$$

$$\left.\left.-s_{x_{\overline{C}}}^{2}\left(\frac{1}{n-m}-\frac{1}{(N-\hat{M})}\right)\right]\right\}\frac{\hat{M}(N-\hat{M})}{n-2}\left(1-\frac{n-1}{N}\right)$$

(29)

5. COMPARISON OF ESTIMATORS

The model-assisted estimators are compared to the unbiased estimator obtained by Greco and Naddeo [2]. Since the precision of model-assisted and unbiased estimators cannot be compared directly from expressions of their variances and mean squared errors, the comparison was carried out by a simulation study.

The simulation is based on repeated sampling from each of the 5 generated populations, each of size 5,000. The population values of (x_i, y_i) are generated using the linear relationship between the auxiliary X and the study variable Y. The population prevalence of units in $C=\{u_i : y_i \geq c\}$ was set at 0.05, 0.10 and 0.2. The number m of units from C in the samples were 5, 10 and 15. For each situation, 5,000 samples were drawn from the population and, for each sample, the estimate of the population total and the estimate of the total in C were calculated along with the average of the estimates of the total, $\bar{\hat{\theta}}=\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_{l}$. The absolute relative bias (ARB), $ARB(\hat{\theta}) = \dfrac{|\bar{\hat{\theta}} - \theta|}{\theta}$. The variance estimates is obtained from $\hat{V}(\hat{\theta}) = \dfrac{1}{L-1}\sum_{l=1}^{L}(\hat{\theta}_l - \bar{\hat{\theta}})^2$. The mean squared error estimates is given by $M\hat{S}E(\hat{\theta}) = \hat{V}(\hat{\theta}) + [Bi\hat{a}s(\hat{\theta})]^2$ where $Bi\hat{a}s(\hat{\theta}) = \bar{\hat{\theta}} - \theta$. The mean squared errors of the two model-assisted estimators are compared to the variance estimates of the unbiased estimators given by Greco and Naddeo [2]. The results from sampling with and without replacement are shown in Table 1 and Table 2.

Table 1. Average Sample Size, Absolute Relative Bias (ARB) and Relative Efficiencies by *P*, *m* and *ρ* for Sampling with Replacement.

ρ	P	m	n	ARB($\tilde{T}_y$)	$\dfrac{\hat{V}(\hat{T}_{y,GN})}{M\hat{S}E(\tilde{T}_y)}$	ARB($\tilde{T}_{yc}$)	$\dfrac{\hat{V}(\hat{T}_{yC,GN})}{M\hat{S}E(\tilde{T}_{yC})}$
.1	.05	5	120.118	0.006	0.993	0.200	0.986
		10	219.976	0.003	1.000	0.098	1.000
		15	320.896	0.002	1.007	0.069	1.000
	10	5	60.125	0.012	0.999	0.206	0.995
		10	110.058	0.005	1.002	0.101	0.998
		15	159.466	0.003	1.003	0.063	1.001
	20	5	30.216	0.020	0.978	0.206	0.994
		10	55.185	0.011	1.000	0.104	1.001
		15	79.811	0.006	1.004	0.063	1.003
.3	.05	5	119.665	0.005	1.081	0.191	1.016
		10	219.547	0.002	1.068	0.094	1.015
		15	321.155	0.001	1.090	0.069	1.023
	.10	5	60.006	0.008	1.055	0.198	1.021
		10	109.509	0.004	1.065	0.095	1.021
		15	160.168	0.003	1.068	0.066	1.022
	.20	5	30.023	0.015	1.037	0.199	1.015
		10	54.767	0.007	1.066	0.092	1.021
		15	79.946	0.005	1.064	0.067	1.022
.5	.05	5	121.510	0.003	1.284	0.196	1.056
		10	219.727	0.002	1.341	0.091	1.070
		15	319.453	0.001	1.310	0.060	1.055
	.10	5	60.136	0.006	1.273	0.194	1.068
		10	111.104	0.003	1.281	0.103	1.072
		15	159.737	0.002	1.303	0.062	1.085
	.20	5	30.197	0.012	1.218	0.201	1.065
		10	54.886	0.005	1.210	0.095	1.060
		15	80.041	0.004	1.237	0.065	1.060
.7	.05	5	119.711	0.002	1.792	0.168	1.091
		10	220.538	0.001	1.844	0.092	1.098
		15	317.922	0.000	1.880	0.053	1.102
	.10	5	60.210	0.004	1.745	0.178	1.159
		10	109.670	0.002	1.793	0.085	1.158
		15	159.480	0.001	1.773	0.059	1.141
	.20	5	29.935	0.007	1.587	0.172	1.139
		10	55.381	0.004	1.686	0.098	1.151
		15	79.786	0.002	1.663	0.057	1.149
.9	.05	5	118.989	0.001	5.278	0.157	1.177
		10	221.067	0.000	5.045	0.087	1.187
		15	320.234	0.000	5.199	0.052	1.198
	.10	5	60.029	0.001	4.694	0.162	1.311
		10	109.989	0.000	4.951	0.083	1.322
		15	160.312	0.001	4.778	0.057	1.325
	.20	5	30.254	0.003	4.134	0.168	1.356
		10	54.577	0.001	4.509	0.078	1.356
		15	79.860	0.001	4.680	0.055	1.351

Table 2. Average Sample Size, Absolute Relative Bias (ARB) and Relative Efficiencies by P, m and ρ for Sampling without Replacement.

ρ	P	m	n	ARB($\tilde{T}_y$)	$\frac{\hat{V}(\hat{T}_{y,GN})}{\hat{MSE}(\tilde{T}_y)}$	ARB($\tilde{T}_{yc}$)	$\frac{\hat{V}(\hat{T}_{yC,GN})}{\hat{MSE}(\tilde{T}_{yC})}$
.1	.05	5	118.965	0.006	1.001	0.193	0.991
		10	219.133	0.004	1.007	0.107	1.002
		15	319.410	0.002	1.005	0.066	1.003
	.10	5	59.792	0.011	0.995	0.196	0.994
		10	109.393	0.006	1.011	0.097	0.999
		15	158.685	0.004	1.003	0.066	0.998
	.20	5	29.830	0.019	0.984	0.198	0.995
		10	54.858	0.009	0.996	0.096	1.000
		15	79.097	0.007	1.000	0.068	1.000
.3	.05	5	119.583	0.005	1.074	0.194	1.012
		10	218.127	0.002	1.097	0.092	1.025
		15	319.368	0.001	1.086	0.064	1.017
	.10	5	59.183	0.008	1.064	0.197	1.025
		10	109.459	0.004	1.075	0.092	1.022
		15	159.626	0.003	1.084	0.065	1.023
	.20	5	29.064	0.015	1.050	0.199	1.018
		10	54.137	0.007	1.058	0.098	1.023
		15	79.838	0.005	1.063	0.062	1.021
.5	.05	5	118.841	0.003	1.316	0.183	1.077
		10	219.432	0.002	1.321	0.094	1.051
		15	318.903	0.001	1.320	0.071	1.056
	.10	5	59.694	0.005	1.285	0.181	1.080
		10	109.800	0.003	1.257	0.096	1.071
		15	158.988	0.002	1.262	0.066	1.063
	.20	5	29.935	0.011	1.229	0.188	1.060
		10	54.374	0.005	1.255	0.085	1.067
		15	79.885	0.004	1.235	0.062	1.068
.7	.05	5	119.556	0.002	1.857	0.175	1.090
		10	218.667	0.001	1.848	0.083	1.107
		15	317.425	0.001	1.829	0.057	1.104
	.10	5	59.346	0.004	1.793	0.170	1.147
		10	109.607	0.002	1.813	0.086	1.157
		15	158.746	0.002	1.786	0.059	1.143
	.20	5	29.847	0.007	1.625	0.178	1.142
		10	54.816	0.004	1.726	0.088	1.162
		15	79.129	0.002	1.739	0.059	1.158
.9	.05	5	118.632	0.001	5.251	0.157	1.191
		10	220.210	0.000	5.002	0.088	1.194
		15	319.284	0.000	5.217	0.053	1.195
	.10	5	59.899	0.001	4.790	0.160	1.315
		10	109.685	0.000	4.994	0.084	1.308
		15	160.263	0.001	4.707	0.058	1.312
	.20	5	30.206	0.003	4.261	0.155	1.378
		10	54.547	0.001	4.451	0.079	1.359
		15	79.805	0.001	4.694	0.053	1.363

The average sample size increases as the m value increases. However, if the population prevalence P increases, smaller average sample sizes are obtained. The average sample sizes for sampling with and without replacement are not much different. For sampling with replacement, slightly larger average sample sizes are obtained.

The results in inverse simple random sampling both with and without replacement indicate that the ARB's of $\tilde{T}_y$ and $\tilde{T}_{yC}$ decrease when the correlation between X and Y or the m value increases, but the ARB's of $\tilde{T}_y$ are close to zero for all situations. The ARB's of $\tilde{T}_{yC}$ are larger than those of $\tilde{T}_y$.

As for the mean squared errors estimates of the model-assisted estimate of the population total and the variance estimates of the unbiased estimate, it can be seen that, for sampling both with and without replacement, if the population prevalence increases then the variance estimates and the mean squared errors increase. They also decrease when the value of *m* increases.

For low correlation between *X* and *Y* (less than 0.5), there is little difference between the mean squared error estimates and the variance estimates for any level of prevalence and any number *m* in the samples. This means that the model-assisted estimate is as efficient as the unbiased estimate. For high correlation between *X* and *Y* ($0.5 \leq \rho \leq 0.7$), the model-assisted estimator is more precise for any level of prevalence and the number *m* in the samples. The relative efficiencies of the model-assisted estimates compared to those of unbiased estimates vary from 120.8% to 188%. For very high correlation between *X* and *Y* ($\rho > 0.7$), it is seen that the model-assisted estimate has very large gain, especially when the population prevalence is equal to 0.05. The efficiency does not depend on the number *m.*

For the estimate of the total in *C*, it is seen that the variance estimates and the mean squared error estimates increase if the population prevalence increases. When $\rho < 0.7$, the variance estimates of the unbiased estimates and the mean squared error estimates of the model-assisted estimates are not much different for any values of the population prevalence *P* and the number *m*. At $\rho = 0.7$, the model-assisted estimate is more precise than the unbiased estimate for any level of the population prevalence and the number *m*. The relative efficiency of the model-assisted estimate remains around 114%. For very high correlation between *X* and *Y* ($\rho = 0.9$), the model-assisted estimate has much smaller mean squared error estimates, especially at the population prevalence *P* greater than or equal to 0.10. The relative efficiencies do not depend on the number *m*.

REFERENCES

[1] Christman M.C. and Lan F., Inverse adaptive cluster sampling, *Biom.*, 2001; **57**: 1096-1105.

[2] Greco L. and Naddeo S., Inverse sampling with unequal selection probabilities, *Commun. Stat.-Theory Methods*, 2007; **36**: 1039-1048.

[3] Haldane J.B.S., On a method of estimating frequencies, *Biom.*, 1945; **33**: 222-225.

[4] Lan F., *Sequential Adaptive Sampling Designs to Estimate Abundance in Rare Populations*. PhD Thesis, The American University, USA, 1999.

[5] Särndal C.E., Swensson B. and Wretman J., *Model Assisted Survey Sampling*, Springer-Verlag, New York; 1992.

[6] Salehi M.M. and Seber G.A.F., A new proof of Murthy's estimator with applies to sequential sampling, *Aust. NZ J. Stat.*, 2001; **43**: 281-286.

[7] Salehi M.M. and Seber G.A.F., A general inverse sampling scheme and its application to adaptive cluster sampling, *Aust. NZ J. Stat.,* 2004**; 46**: 483-494.