Improved Ratio Estimators in Adaptive Cluster Sampling

Chang-Tai Chao * Feng-Min Lin † Tzu-Ching Chiang ‡

Abstract

For better inference of the population quantity of interest, ratio estimators are often recommended when certain auxiliary variables are available. Two types of ratio estimators, modified for adaptive cluster sampling via transformed population and initial intersection probability approaches, have been studied in Dryver and Chao (2007). Unfortunately, none of them are a function of a minimal sufficient statistic, and therefore can be improved with Rao-blackwellization procedure. The purpose of this paper is to obtain new ratio estimators that are not only more efficient than the original ratio estimators proposed by Dryver and Chao, but simple to calculate. Additionally, explicit formulas for the approximated variance of these easy-to-compute estimators are derived.

Key Words: Auxiliary Variable, Ratio Estimator, Adaptive Cluster Sampling, Rao-blackwell Theorem

1. Introduction

First proposed by Thompson (1990), adaptive cluster sampling is an alternative to estimate the population quantity of interest especially under rare or clustered populations. The basic idea behind this approach is to take a small initial sample by some conventional designs, and then to increase the sampling efficiency in the neighborhoods of the sampling units satisfying a condition previously defined. Under the design-based inferential approach, although the usual unbiased estimators in adaptive cluster sampling are simple to calculate, they do not necessarily utilize all the information provided by the resultant final sample. More efficient estimators, the Rao-blackwell estimators, can be obtained by using Rao-blackwell idea of conditioning on a minimum sufficient statistic. However, Thompson did not present analytical expressions for any of the Rao-blackwell estimators but he computed the value of a Rao-blackwell estimator by averaging the values of the estimator over all the initial samples giving rise to the observed final sample. Clearly, the excessive number of calculations is required and hence it is essential for ordinary applications to achieve simply analytical expressions. A few papers have given some analytical expressions for the Rao-blackwell estimators. For example, Salehi (1999) and Félix-Medina (2000) have provided the colsed-form expressions for the Rao-blackwell estimators based on the modified Hansen-Hurwitz estimator and the modified Horvitz-Thompson estimator. In addition, Dryver and Thompson (2005) have presented alternative mathematical formulas for the two Rao-blackwell estimators derived by taking the expected value of the usual estimators conditional on a sufficient statistic. The alternative Rao-blackwell estimators may not as efficient as the Rao-blackwell estimators obtained by taking the expected value of the usual estimators conditional on a minimum sufficient statistic, but they are rather simple.

The estimators mentioned previously utilize the information provided by the population variable of interest only. Nevertheless, to improve the quality of the estimate in sampling survey, one not only depends on the the information of the primary variable, but reasonably needs to take relevant aspects of data into account. For many sampling survey situations, certain auxiliary variables are often available and it is suggested to make use of the the auxiliary information for better inference. Ratio estimation is a popular and widely used method to take advantage of the data from the variable of interest along with available auxiliary variables. Although design-biased, ratio estimators are more efficient because they can give lower mean-square errors when sufficient correlations between the variable of interest and auxiliary variable exist. Moreover, the performances of the ratio estimators become more apparent as the correlations increase (e.g., Lohr, 1999). Dryver and Chao (2007) used auxiliary information into estimators in adaptive cluster sampling to obtain ratio estimators, which is a straightforward extension of the ratio estimator under unequal probability sampling. None of those ratio estimators , however, is a function of a minimum sufficient statistic and therefore can be improved with the approach similar to that the univariate estimators derived.In this paper the Rao-blackwell ratio estimators are derived by taking expected value of the ordinary ratio estimators conditional on the same sufficient statistic that Dryver and Thompson (2005) utilized. However, the formulas for these Rao-blackwell ratio estimators are not easily computed as those univariate estimators proposed by Dryver and Thompson, but are expressed as an average over the ratios for all edge units in the observed final sample (where edge units are as defined in Section 2). The computation becomes much more intensive as the number of the edge units turns large. In the interest of simplicity, we therefore further propose new, efficient, and easy-to-compute ratio

*Department of Statistics, National Cheng-Kung University
†Department of Statistics, National Cheng-Kung University
‡Institute of Statistical Science, Academia Sinica

estimators. Furthermore, explicit formulas for the approximated variance of the easy-to-compute ratio estimators are derived.

The paper is organized as follows. In Section 2, we briefly describe the ordinary estimators in adaptive cluster sampling, including the design-unbiased estimators and ratio estimators. In Section 3 are two types of the new ratio estimators proposed in this article. The Rao-blackwell ratio estimators, derived with Rao-blackwellization procedure by conditioning on a sufficient statistic, are described in Subsection 3.1; the derivation is given in Appendix A. In Subsection 3.2, the easy-to-compute ratio estimators are illustrated and the formulas for their approximated variance are derived in Appendix B. Section 4 presents concluding comments.

2. Ordinary estimators in adaptive cluster sampling

In adaptive cluster sampling, an initial sample of units can be selected by different types of conventional probability sampling. In this article the simplest form of adaptive cluster sampling, an initial sample is selected by simple random sampling without replacement (SRSWOR), will be considered (Thompson, 1990). Nevertheless, the result can be extended to other various types of adaptive cluster sampling associated with different initial sampling designs. In this section, we will briefly describe the methodology and concepts involved in adaptive cluster sampling, and we will also introduce the notation used throughout this paper. The ordinary estimators in adaptive cluster sampling have been proposed, including the design-unbiased and ratio estimators, are addressed in this section as well.

2.1 Design-unbiased estimators in adaptive cluster sampling

In a basic sampling view, population is a finite set of units consisting of N units with labels $1, \ldots, N$, denoted as $u = \{u_1, \ldots, u_N\}$. Associated with each unit i, the values of the population variable of interest is denoted as y_i. Through this article, the population quantity of interest to be estimated is the population mean of the $y's$, that is,

$$\mu_y = \sum_{i=1}^{N} y_i / N. \tag{1}$$

In adaptive cluster sampling, the sampling procedure is selecting a small initial sample by some conventional designs, and whenever the variable of interest of a unit in the small initial sample satisfies a given condition C, units in some predefined neighborhood will be included into the sample and observed. C is typically a function of the population variable of interest based on the options and the experience of experts for various populations. Neighborhood can be defined by social or institutional relationships between units. The most prevalent, by far, is the neighborhood consisting of the unit itself and the four adjacent units, left, right, top and bottom. In this paper, consider an initial sample $s_0 = (u_1, \ldots, u_{n_0})$ of size n_0 is selected from u via SRSWOR. If any of the units in s_0 satisfy C, for example, $y_i \leqslant c$ where c is a constant, their respective neighborhoods are added to the sample and observed. Furthermore, if any added units satisfy C, the units in the neighborhood are added to the sample and observed as well, and so on. This procedure is iterated until no new units satisfy. The set formed by the original unit in s_0 and together with the units added as a consequence of selecting u_i is called a cluster. The units adaptively selected but not meet C are called edge units. A cluster minus the edge units is called a network. Any unit not satisfying C is, by definition, a network of size 1. Let A_k be the network containing unit i and m_k denote the number of units in A_k. Then the average of the $y-$values in the kth network is

$$w_{y_k} = \frac{1}{m_k} \sum_{i \in A_k} y_i. \tag{2}$$

The population mean of the $y's$ can be written in terms of networks and denote as $\mu_y = \sum_{k=1}^{K} w_{y_k}/N$, where K is the number of the distinct networks in the population.

2.1.1 Ordinary estimators

Adaptive cluster sampling is a case of Unequal Probability Design if networks are considered as sampling units. Thompson (1990) developed two unbiased estimators based on the modifications of the Hansen-Hurwitz and Horvitz-Thompson estimators. With this design, unfortunately, neither the draw-by-draw selection probability, nor the inclusion probability can be determined from the data for the units that do not satisfy C and are not included in the initial sample. Consequently, observations that do not satisfy C are ignore if they are not included in the initial sample.

One of the unbiased estimators in adaptive cluster sampling, the modified Hansen-Hurwitz estimator, is based on the initial draw-by-draw selection probabilities. Let I(.) denote an indicator function equalling 1 when the expression inside is true and 0 otherwise. The number of units selected from the kth network in the initial sample is

$$n_k = \sum_{i \in A_k} I(i \in s_0). \tag{3}$$

The modified Hansen-Hurwitz estimator and its variance are

$$\hat{\mu}_{y \cdot hh} = \frac{1}{n_0} \sum_{k=1}^{\kappa} n_k w_{y_k}, \tag{4}$$

where κ is the number of distinct networks intersected by the initial sample and

$$\mathrm{var}(\hat{\mu}_{y \cdot hh}) = \frac{N - n}{Nn(N-1)} \sum_{k=1}^{K} m_k (w_{y_k} - \mu)^2. \tag{5}$$

Another unbiased estimator using the partial inclusion probabilities is

$$\hat{\mu}_{y \cdot ht} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{u_{y_k}}{\alpha_k}, \tag{6}$$

where $u_{y_k} = m_k \cdot w_{y_k}$ and $\alpha_k = 1 - \left[\binom{N - m_k}{n_0} / \binom{N}{n_0} \right]$ is the initial intersection probability of the kth network. The joint initial intersection probabilities is

$$\alpha_{kh} = 1 - \left[\binom{N - m_k}{n_0} + \binom{N - m_h}{n_0} + \binom{N - m_k - m_h}{n_0} \right] / \binom{N}{n_0}. \tag{7}$$

The variance of $\hat{\mu}_{y \cdot ht}$ is

$$\mathrm{var}(\hat{\mu}_{y \cdot hh}) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} u_{y_k} u_{y_h} \left(\frac{\alpha_{kh} - \alpha_k \alpha_h}{\alpha_k \alpha_h} \right). \tag{8}$$

2.1.2 Rao-Blackwell estimators

Under the design-based inferential approach, the ordinary estimators do not necessarily utilize all the information provided by the resultant final sample. Only the edge units in the initial sample are incorporated when computing them. The Rao-blackwell theorem can be used to improve the efficiency of the ordinary estimators since some variability can be reduced by making use of the observations of the edge units which are not in the initial sample. Dryver and Thompson (2005) utilized a sufficient statistic instead of the minimum sufficient statistic, and obtained the easy-to-compute Rao-blackwell estimators which were developed by considering only how many edge units were initial selected, but not which ones. In this section, only the estimators and their corresponding variances will be introduced. More detailed proofs and descriptions can be found in Dryver and Thompson's paper.

A statistic d^+ is defined as

$$d^+ = \{(i, y_i, f_i), (j, y_j); i \in s_c, j \in s_{\bar{c}}\}. \tag{9}$$

For unit i, f_i is the number of times that the network to which unit i belongs is intersected by the initial sample. The union of a core part s_c and the remaining part $s_{\bar{c}}$ is the final sample s. The core part s_c consists of all the distinct units in the sample which satisfy the condition. The remaining part $s_{\bar{c}}$ is the set of all the distinct units in the sample for which the condition is not met. The statistic d^+ is sufficient for μ so applying by the Rao-blackwell theorem to $\hat{\mu}_{y \cdot hh}$ and $\hat{\mu}_{y \cdot ht}$, the easy-to-compute Rao-blackwell estimators $\hat{\mu}^+_{y \cdot hh}$ and $\hat{\mu}^+_{y \cdot ht}$ are arrived.

One of the Rao-blackwell estimators, $\hat{\mu}^+_{y \cdot hh}$, is defined by

$$\hat{\mu}^+_{y \cdot hh} = E\left(\hat{\mu}_{y \cdot hh} | D^+ = d^+\right) = \frac{1}{n_0} \sum_{k=1}^{\kappa} n_k w^+_{y_k}. \tag{10}$$

where

$$w^+_{y_k} = \begin{cases} w_{y_k} & , \text{if } \sum_{i \in \Psi_k} e_i = 0, \\ \bar{y}_e = \frac{\sum_{i \in s} e_i y_i}{e_s} & , \text{if } \sum_{i \in \Psi_k} e_i = 1. \end{cases} \tag{11}$$

Note $\bar{y}_e$ is the average $y-$value for the sample edge units in the final sample and e_s is the number of sample edge units in s. For the ith unit in the sample, the indicator variable e_i is defined as

$$e_i = \begin{cases} 1, & \text{if } i \text{ does not meet the condition but is in the neighborhood,} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Additionally, for those units which are not in s, $e_i = 0$.

The variance of $\hat{\mu}^+_{y\cdot hh}$ is

$$\begin{aligned} \text{var}(\hat{\mu}_{y\cdot hh}) =& \frac{N-n}{Nn(N-1)} \sum_{k=1}^{K} m_k (w_{y_k} - \mu)^2 \\ & - \frac{1}{n^2} \sum_{d^+ \in D^+} P(d^+) \left(\frac{e_{s_0}}{e_s} \sum_{i \in s, e_i = 1} y_i^2 + \frac{e_{s_0}(e_{s_0}-1)}{e_s(e_s-1)} \sum_{i \in s, e_i = 1} \sum_{j \neq i} y_i y_j - e_{s_0}^2 \bar{y}_e^2 \right), \end{aligned} \tag{13}$$

where $P(d^+)$ is the probability that $D^+ = d^+$ and e_{s_0} is the number of units picked in s_0.

The other estimators $\hat{\mu}^+_{y\cdot ht}$ is defined by

$$\hat{\mu}^+_{y\cdot ht} = E\left(\hat{\mu}_{y\cdot ht} | D^+ = d^+\right) = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{m_k \cdot w^+_{yk}}{\alpha_k} \tag{14}$$

and has variance

$$\begin{aligned} \text{var}(\hat{\mu}_{y\cdot ht}) =& \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} u_{y_k} u_{y_h} \left(\frac{\alpha_{kh} - \alpha_k \alpha_h}{\alpha_k \alpha_h} \right) \\ & - \frac{1}{n^2} \sum_{d^+ \in D^+} P(d^+) \left(\frac{e_{s_0}}{e_s} \sum_{i \in s, e_i = 1} y_i^2 + \frac{e_{s_0}(e_{s_0}-1)}{e_s(e_s-1)} \sum_{i \in s, e_i = 1} \sum_{j \neq i} y_i y_j - e_{s_0}^2 \bar{y}_e^2 \right). \end{aligned} \tag{15}$$

2.2 Ordinary ratio estimators in adaptive cluster sampling

In many applied survey situations of adaptive cluster sampling, auxiliary variable is often collected together with the population variable of interest. Dryver and Chao (2007) utilized the auxiliary information into the estimation, and proposed two ratio estimators in adaptive cluster sampling by taking advantage of the correlation between the variable of interest and the auxiliary variable. In this section these two ordinary ratio estimators and their variances will be briefly described.

Let μ_x be the population mean of the auxiliary variable x. The ordinary ratio estimator related to the modified Hansen-Hurwitz estimator is

$$\hat{\mu}_{r\cdot hh} = \frac{\hat{\mu}_{y\cdot hh}}{\hat{\mu}_{x\cdot hh}} \mu_x, \tag{16}$$

where $\hat{\mu}_{x\cdot hh}$ is the modified Hansen-Hurwitz estimator for μ_x. The approximated variance of $\hat{\mu}_{r\cdot hh}$ is

$$\text{Avar}(\hat{\mu}_{r\cdot hh}) = \frac{N-n}{Nn(N-1)} \sum_{k=1}^{K} m_k (w_{y_k} - R w_{x_k})^2, \tag{17}$$

where R is the population ratio between w_{y_k} and w_{x_k}.

The other ratio estimator of μ_y can be constructed based on the modified Horvitz-Thompson estimator and is defined by

$$\hat{\mu}_{r\cdot ht} = \frac{\hat{\mu}_{y\cdot ht}}{\hat{\mu}_{x\cdot ht}} \mu_x, \tag{18}$$

where $\hat{\mu}_{y\cdot ht}$ and $\hat{\mu}_{x\cdot ht}$ are the modified Horvitz-Thompson estimators for μ_y and μ_x, respectively. The approximated variance is the variance of the modified Horvitz-Thompson estimator of the variable $u'_k = u_{y_k} - R u_{x_k}$.

$$\text{Avar}(\hat{\mu}_{r\cdot ht}) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} u'_k u'_h \left(\frac{\alpha_{kh} - \alpha_k \alpha_h}{\alpha_k \alpha_h} \right). \tag{19}$$

3. New ratio estimators in adaptive cluster sampling

In subsection 3.1 the real Rao-blackwell ratio estimators, derived with Rao-blackwellization procedure by conditioning on the sufficient statistic d^+, are arrived. The derivation of these Rao-blackwell ratio estimators is given in Appendix A. The formulas for them are not as easy as those univariate estimators proposed by Dryver and Thompson (2005), and the computation becomes much more intensive as the number of the edge units turns large. In the interest of simplicity, we therefore further propose two efficient and easy-to-compute ratio estimators and the formulas for them are presented in Subsection 3.2. Furthermore, explicit formulas for the approximated variance of the easy-to-compute ratio estimators are derived and the derivation is given in Appendix B.

3.1 Rao-Blackwell ratio estimators

Mentioned in the previous section, the statistic d^+ is sufficient for μ. So the Rao-blackwell ratio estimators $\hat{\mu}^+_{r\cdot hh}$ and $\hat{\mu}^+_{r\cdot ht}$ are able to be arrived at by applying the Rao-blackwell theorem to $\hat{\mu}_{r\cdot hh}$ and $\hat{\mu}_{r\cdot ht}$. The Rao-blackwell ratio estimator $\hat{\mu}^+_{r\cdot hh}$ is defined by

$$\hat{\mu}^+_{r\cdot hh} = E\left(\hat{\mu}_{r\cdot hh}|D^+ = d^+\right). \tag{20}$$

The estimator is not easily computed as shown by the formula

$$\hat{\mu}^+_{r\cdot hh} = \frac{e_{s_0}}{e_s} \sum_{i\in s, e_i=1} \frac{\sum_{k=1}^{\kappa} n_k w_{y_k}(1 - \sum_{i\in\Psi_k} e_i) + e_i y_i}{\sum_{k=1}^{\kappa} n_k w_{x_k}(1 - \sum_{i\in\Psi_k} e_i) + e_i x_i}\mu_x. \tag{21}$$

The other Rao-blackwell ratio estimator $\hat{\mu}^+_{r\cdot ht}$ is not easily computed as well and the formula is given as

$$\hat{\mu}^+_{r\cdot ht} = E\left(\hat{\mu}_{r\cdot ht}|D^+ = d^+\right) \tag{22}$$

$$= \frac{e_{s_0}}{e_s} \sum_{i\in s, e_i=1} \frac{\sum_{k=1}^{\kappa} \frac{u_{y_k}}{\alpha_k}(1 - \sum_{i\in\Psi_k} e_i) + \frac{n_0}{N} e_i y_i}{\sum_{k=1}^{\kappa} \frac{u_{x_k}}{\alpha_k}(1 - \sum_{i\in\Psi_k} e_i) + \frac{n_0}{N} e_i x_i}\mu_x. \tag{23}$$

The proofs that equations (20) and (21), and (22) and (23) are respectively equivalent are given in Appendix A.

3.2 Easy-to-compute ratio estimators

The formulas for the real Rao-blackwell ratio estimators are too complicated to be calculated in practice. To simplify the calculation, we therefore further construct other improved estimators via a ratio of the Rao-blackwellized univariate estimators conditioning on the sufficient statistic d^+. The new ratio estimators are very easily computed and their approximated variances are less than or equal to the variances of the original ratio estimators proposed by Dryver and Chao (2007). The two easy-to-compute ratio estimators and their variances will be briefly described.

One of the easy-to-compute estimators $\hat{\mu}_{r\cdot hh(+)}$ is defined to be

$$\hat{\mu}_{r\cdot hh(+)} = \frac{\hat{\mu}^+_{y\cdot hh}}{\hat{\mu}^+_{x\cdot hh}}\mu_x = \frac{\sum_{k=1}^{\kappa} n_k w_{y_k}(1 - \sum_{i\in\Psi_k} e_i) + \bar{y}_e}{\sum_{k=1}^{\kappa} n_k w_{x_k}(1 - \sum_{i\in\Psi_k} e_i) + \bar{x}_e}\mu_x, \tag{24}$$

where $\hat{\mu}^+_{y\cdot hh}$ and $\hat{\mu}^+_{x\cdot hh}$ are the Hansen-Hurwitz type Rao-blackwell estimator for μ_y and μ_x. $\bar{x}_e$ is the average $x-$value for the sample edge units in the final sample. The approximated variance of $\hat{\mu}_{r\cdot hh(+)}$ is

$$\begin{aligned} \mathrm{Avar}\left(\hat{\mu}_{r\cdot hh(+)}\right) = \mathrm{Avar}\left(\hat{\mu}_{r\cdot hh}\right) - \frac{1}{n_0^2} \sum_{d^+\in D^+} P(d^+)\Big(\frac{e_{s_0}}{e_s} \sum_{i\in s, e_i=1} (y_i - Rx_i)^2 \\ + \frac{e_{s_0}(e_{s_0}-1)}{e_s(e_s-1)} \sum_{i\in s, e_i=1} \sum_{j\neq i} (y_i - Rx_i)(y_j - Rx_j) - e_{s_0}^2\,(\bar{y}_e - R\bar{x}_e)^2\Big). \end{aligned} \tag{25}$$

The other easy-to-compute estimator $\hat{\mu}_{r\cdot ht(+)}$ is

$$\hat{\mu}_{r\cdot ht(+)} = \frac{\hat{\mu}^+_{y\cdot ht}}{\hat{\mu}^+_{x\cdot ht}}\mu_x = \frac{\sum_{k=1}^{\kappa} \frac{u_{y_k}}{\alpha_k}(1 - \sum_{i\in\Psi_k} e_i) + \frac{n_0}{N}\bar{y}_e}{\sum_{k=1}^{\kappa} \frac{u_{x_k}}{\alpha_k}(1 - \sum_{i\in\Psi_k} e_i) + \frac{n_0}{N}\bar{x}_e}\mu_x. \tag{26}$$

The approximated variance of $\hat{\mu}_{r\cdot ht(+)}$ is

$$\begin{aligned}\operatorname{Avar}\left(\hat{\mu}_{r\cdot ht(+)}\right) = \operatorname{Avar}\left(\hat{\mu}_{r\cdot ht}\right) - \frac{1}{n_0^2}\sum_{d^+\in D^+} P(d^+)\Big(\frac{e_{s_0}}{e_s}\sum_{i\in s, e_i=1}(y_i - Rx_i)^2 \\ + \frac{e_{s_0}(e_{s_0}-1)}{e_s(e_s-1)}\sum_{i\in s,e_i=1}\sum_{j\neq i}(y_i - Rx_i)(y_j - Rx_j) - e_{s_0}^2\left(\bar{y}_e - R\bar{x}_e\right)^2\Big).\end{aligned} \tag{27}$$

4. Conclusions

In order to make the best use of survey data in adaptive cluster sampling, we discuss how to utilize auxiliary information into estimation. Improving ratio estimators with Rao-blackwellization is the main object in this study. We derive the real Rao-blackwell ratio estimators by taking expected value of the ordinary ratio estimators conditional on the same sufficient statistic that Dryver and Thompson (2005) utilized. However, the formulas for these Rao-blackwell ratio estimators are too complicated to be calculated in practice. In the interest of simplicity, we therefore further construct other improved estimators via a ratio of the Rao-blackwellized univariate estimators conditioning on the sufficient statistic. Furthermore, we have been able to obtain the explicit formulas for the approximated variance of those easy-to-compute ratio estimators and therefore guarantee their approximated mean square errors are lower than those of the unimproved ratio estimators proposed by Dryver and Chao (2007).

From the model-based perspective population values are considered to be random variables, and represent just one outcome of many possible outcomes under a specific model. This probability model can be constructed by detailed surveys or experience and may offer more efficient inferences than design-based approach. However, validity of inference depends on the correctness of this assumed model. We did not discuss the inferences via model-based perspective in this research. Similar study under the model-based point of view will be investigated in the future.

Appendix

A. Derivation of the Rao-blackwell ratio estimators

We only derive $\hat{\mu}^+_{r\cdot hh}$ and leave the derivation of $\hat{\mu}^+_{r\cdot ht}$ to the reader because the approach to obtain the formula for $\hat{\mu}^+_{r\cdot ht}$ is not much different from the approach for $\hat{\mu}^+_{r\cdot hh}$. Let $\hat{\mu}^+_{r\cdot hh}(s_0)$ represent $\hat{\mu}^+_{r\cdot hh}$ as a function of the initial sample s_0. Let S_0 be a random variable taking on values from the sample space S and $P(S_0 = s_0|D^+ = d^+)$ is the probability of that initial sample given d^+. Thus the Hansen-Hurwitz type Rao-blackwell ratio estimator is

$$\hat{\mu}^+_{r\cdot hh} = E\left(\hat{\mu}_{r\cdot hh}|D^+ = d^+\right) = \sum_{s_0\in S}\hat{\mu}^+_{r\cdot hh}(s_0)P(S_0 = s_0|D^+ = d^+).$$

The conditional probability $P(S_0 = s_0|D^+ = d^+)$ can be written as

$$I\{g(s_0) = d^+\}/L,$$

where $I\{.\}$ is an indicator function and $L = \sum_{s_0\in S} I\{g(s_0) = d^+\}$ is the total number of combinations compatible with d^+. And out of the L combinations any single unit appears

$$\frac{\binom{e_s - 1}{e_{s_0} - 1}}{\binom{e_s}{e_{s_0}}}L.$$

Thus the estimator can be written as

$$\begin{aligned}\hat{\mu}^+_{r\cdot hh} &= \frac{1}{L}\sum_{s_0\in S} I\{g(s_0) = d^+\}\hat{\mu}^+_{r\cdot hh}(s_0) \\ &= \frac{\binom{e_s - 1}{e_{s_0} - 1}}{\binom{e_s}{e_{s_0}}}\sum_{s_0\in S}\hat{\mu}^+_{r\cdot hh}(s_0) \\ &= \frac{e_{s_0}}{e_s}\sum_{i\in s,e_i=1}\frac{\sum_{k=1}^{\kappa} n_k w_{y_k}(1 - \sum_{i\in\Psi_k} e_i) + e_i y_i}{\sum_{k=1}^{\kappa} n_k w_{x_k}(1 - \sum_{i\in\Psi_k} e_i) + e_i x_i}\mu_x.\end{aligned}$$

B. Derivation of the approximated variance of the easy-to-compute ratio estimators

The improved ratio estimator is

$$\hat{\mu}_r^+ = \frac{\hat{\mu}_y^+}{\hat{\mu}_x^+}\mu_x = \hat{R}\mu_x;$$

hence the estimator can be written

$$\hat{\mu}_r^+ = \hat{\mu}_y^+ + \hat{R}(\mu_x - \hat{\mu}_x^+).$$

The first term in Taylor's formula, expanding about the point (μ_x, μ) gives the approximation

$$\hat{\mu}_r^+ \approx \hat{\mu}_y^+ + R(\mu_x - \hat{\mu}_x^+).$$

Consequently,

$$\hat{\mu}_r^+ - \mu \approx \hat{\mu}_y^+ + R(\mu_x - \hat{\mu}_x^+) - \mu = \hat{\mu}_y^+ - R\hat{\mu}_x^+.$$

The approximation for the variance is

$$\mathrm{Avar}\left(\hat{\mu}_r^+\right) = \mathrm{E}\left(\hat{\mu}_y^+ - R\hat{\mu}_x^+\right)^2 = \mathrm{var}\left(\hat{\mu}_y^+ - R\hat{\mu}_x^+\right) = \mathrm{var}\left(\mathrm{E}(\hat{\mu}_y - R\hat{\mu}_x|D^+)\right).$$

Hence,

$$\begin{aligned}
\mathrm{Avar}\left(\hat{\mu}_r^+\right) =& \mathrm{var}\left(\mathrm{E}(\hat{\mu}_y - R\hat{\mu}_x|D^+)\right)\\
=& \mathrm{var}\left(\hat{\mu}_y - R\hat{\mu}_x\right) - \mathrm{E}\left(\mathrm{var}(\hat{\mu}_y - R\hat{\mu}_x|D^+)\right)\\
=& \mathrm{Avar}\left(\hat{\mu}_r\right) - \mathrm{E}\left((\hat{\mu}_y - R\hat{\mu}_x) - (\hat{\mu}_y^+ - R\hat{\mu}_x^+)\right)^2\\
=& \mathrm{Avar}\left(\hat{\mu}_r\right) - \frac{1}{n^2}\sum_{d^+\in D^+}\frac{P(d^+)}{L(d^+)}\sum_{s_0\in S} I\{g(s_0) = d^+\}\left(\sum_{i\in s_0, e_i=1}(y_i - Rx_i) - e_{s_0}(\bar{y}_e - R\bar{x}_e)\right)^2\\
=& \mathrm{Avar}\left(\hat{\mu}_r\right) - \frac{1}{n^2}\sum_{d^+\in D^+}\frac{P(d^+)}{L(d^+)}\sum_{s_0\in S} I\{g(s_0) = d^+\}\left((\sum_{i\in s_0, e_i=1}(y_i - Rx_i))^2 - e_{s_0}^2\left(\bar{y}_e - R\bar{x}_e\right)^2\right)\\
=& \mathrm{Avar}\left(\hat{\mu}_r\right) - \frac{1}{n^2}\sum_{d^+\in D^+}\frac{P(d^+)}{L(d^+)}\sum_{s_0\in S} I\{g(s_0) = d^+\}\times\\
&\left(\sum_{i\in s_0, e_i=1}(y_i - Rx_i)^2 + \sum_{i\in s_0, e_i=1}\sum_{j\neq i}(y_i - Rx_i)(y_j - Rx_j) - e_{s_0}^2\left(\bar{y}_e - R\bar{x}_e\right)^2\right)\\
=& \mathrm{Avar}\left(\hat{\mu}_r\right) - \frac{1}{n^2}\sum_{d^+\in D^+} P(d^+)\times\\
&\left(\frac{\binom{e_s - 1}{e_{s_0} - 1}}{\binom{e_s}{e_{s_0}}}\sum_{i\in s, e_i=1}(y_i - Rx_i)^2 + \frac{\binom{e_s - 2}{e_{s_0} - 2}}{\binom{e_s}{e_{s_0}}}\sum_{i\in s, e_i=1}\sum_{j\neq i}(y_i - Rx_i)(y_j - Rx_j) - e_{s_0}^2\left(\bar{y}_e - R\bar{x}_e\right)^2\right)\\
=& \mathrm{Avar}\left(\hat{\mu}_r\right) - \frac{1}{n^2}\sum_{d^+\in D^+} P(d^+)\times\\
&\left(\frac{e_{s_0}}{e_s}\sum_{i\in s, e_i=1}(y_i - Rx_i)^2 + \frac{e_{s_0}(e_{s_0} - 1)}{e_s(e_s - 1)}\sum_{i\in s, e_i=1}\sum_{j\neq i}(y_i - Rx_i)(y_j - Rx_j) - e_{s_0}^2\left(\bar{y}_e - R\bar{x}_e\right)^2\right)
\end{aligned}$$

REFERENCES

Dryver, A. L., and Thompson, S. K. (2005), "Improved unbiased estimators in adaptive cluster sampling," *Journal of the Royal Statistical Society B*, 67, 157–166.

Dryver, A. L., and Chao, C. T. (2007), "Ratio estimators in adaptive cluster sampling," *Environmetrics*, 18, 607–620.

Félix Medina, M. H. (2000), "Analytical expressions for Rao-Blackwell estimators in adaptive cluster sampling," *Journal of Statistical Planning and Inference*, 84, 221–236.

Salehi, M. M. (1999), "Rao-Blackwell versions of the Hansen-Hurwitz and Horvitz-Thompson estimators in adaptive cluster sampling," *Ecological and Environmental Statistics*, 6, 183–195.

Thompson, S. K.(1990), "Adaptive cluster sampling," *Journal of the American Statistical Association*, 77, 848–854.

Thompson, S. K., and Seber, G.A.F. (1996), *Adaptive Sampling*, New York: Wiley.