Environmental and Ecological Statistics **6,** 183–195 (1999)

Rao-Blackwell versions of the Horvitz-Thompson and Hansen-Hurwitz in adaptive cluster sampling

MOHAMMAD SALEHI M.

School of Mathematics, Isfahan University of Technology, Isfahan, I.R. of Iran
E-mail: salehi_m@cc.iut.ac.ir

Received March 1998; revised October 1998

Thompson (1990) introduced the adaptive cluster sampling design and developed two unbiased estimators, the modified Horvitz-Thompson (HT) and Hansen-Hurwitz (HH) estimators, for this sampling design and noticed that these estimators are not a function of the minimal sufficient statistics. He applied the Rao-Blackwell theorem to improve them. Despite having smaller variances, these latter estimators have not received attention because a suitable method or algorithm for computing them was not available. In this paper we obtain closed forms of the Rao-Blackwell versions which can easily be computed. We also show that the variance reduction for the HH estimator is greater than that for the HT estimator using Rao-Blackwell versions. When the condition for extra samples is $y > 0$, one can expect some Rao-Blackwell improvement in the HH estimator but not in the HT estimator. Two examples are given.

Keywords: adaptive sampling, clustered population, Rao-Blackwell theorem.

1. Introduction

Adaptive cluster sampling has been shown to be a useful sampling method for parameter estimation in clustered and rare populations (Thompson and Seber, 1996; Smith *et al.*, 1995). Suppose that we have a population of units. An initial sample of units is selected by some conventional sampling design. Whenever the value of the variable of interest (or any associated variable) of a selected unit satisfies a specified condition, say C, its neighboring units are added to the sample. Furthermore, if any other units in these neighboring units satisfy C then they are also added to the sample. This process continues until a cluster of units is formed with a boundary of units, called edge units, which do not satisfy C. A cluster without its edge units forms a network, and a unit not satisfying the condition C also forms a network of size one. The networks are disjoint and form a partition of population units. The condition for extra sampling is defined on the value of a variable which is known only for units in the sample.

The author would like to express his sincere thanks to four anonymous referees for many valuable suggestions which considerably improved the presentation of the paper. Research is partially supported by the Research Council of Isfahan University of Technology.

2. Notation and estimators

Consider a population of N units $(u_1, u_2, \ldots, u_N)$ labeled by $(1, 2, \ldots, N)$. With u_i is associated a variable of interest y_i, for $i = 1, 2, \ldots, N$. A simple random sample of size n_1 is taken without replacement. Further units are then added adaptively using condition C. Following the notation of Thompson and Seber (1996), suppose that the final ordered sample of the labels is $s_o = (i_1, i_2, \ldots, i_n)$ where n is the sample size. We note that repeated y-values can occur in this sample. Let d_o denote the data vector whose components are the unit labels and their corresponding y-values, namely $d_o = ((i, y_i) : i \in s_o)$. Consider the unordered reduced set $s_R = \{i_1, i_2, \ldots, i_\nu\}$ of the ν distinct labels in s_o. Then $D_R = \{(i, y_i) : i \in s_R\}$ is minimal sufficient for $\boldsymbol{\theta} = (y_1, y_2, \ldots, y_N)$ in any adaptive sampling design (see Thompson and Seber, 1996, Chapter 2).

We now present the modified HT and HH estimators introduced by Thompson (1990). Let A_i denote the network containing unit i and let m_i denote the number of units in A_i. Thompson (1990) ignored the edge units and defined the ''partial'' inclusion probability, namely

$$\alpha_i = 1 - \binom{N - m_i}{n_1} \Big/ \binom{N}{n_1}. \tag{1}$$

This is the probability that the initial sample intersects A_i. The modified HT estimator for the mean, μ, based on this probability is

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k}, \tag{2}$$

where y_k^* is the sum of the y-values for the kth network, κ is the number of distinct networks intersected by the initial sample. We have $\alpha_k = \alpha_i$ for every unit i in network k. The probability that both networks j and k are intersected by the initial sample is

$$\alpha_{jk} = 1 - \left[\binom{N - m_j}{n_1} + \binom{N - m_k}{n_1} - \binom{N - m_j - m_k}{n_1} \right] \Big/ \binom{N}{n_1}.$$

The variance of $\hat{\mu}_{HT}$ is

$$\mathrm{var}[\hat{\mu}_{HT}] = \frac{1}{N^2} \left[\sum_{j=1}^{K} \sum_{k=1}^{K} y_j^* y_k^* \left(\frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_j \alpha_k} \right) \right], \tag{3}$$

where K and denote the number of networks in the population. An unbiased estimator of the above variance is

$$v[\hat{\mu}_{HT}] = \frac{1}{N^2} \left[\sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa} \frac{y_j^* y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right], \tag{4}$$

where α_{jj} is interpreted as α_j.

Thompson (1990) introduced another unbiased estimator based on the frequency of selection called the modified Hansen-Hurwitz estimator, namely

$$\hat{\mu}_{HH} = \frac{1}{n_1} \sum_{i=1}^{n_1} w_i, \tag{5}$$

where w_i is the mean of the m_i observations (units) in A_i. Here $\hat{\mu}_{HH}$ can be recognized as the sample mean obtained by taking a simple random sample of size n_1 from a population of w_i values (Thompson and Seber, 1996). Using the theory of simple random sampling we have

$$\mathrm{var}[\hat{\mu}_{HH}] = \frac{N - n_1}{Nn_1(N-1)} \sum_{i=1}^{N} (w_i - \mu)^2, \tag{6}$$

with unbiased estimator

$$v[\hat{\mu}_{HH}] = \frac{N - n_1}{Nn_1(n_1-1)} \sum_{i=1}^{n_1} (w_i - \hat{\mu}_{HH})^2. \tag{7}$$

3. Applying the Rao-Blackwell theorem

We use the notation of Thompson and Seber (1996) and their theory in this section. Let

$$G = \binom{\nu}{n_1},$$

the number of combinations of n_1 distinct units from the ν units in the sample. Suppose that these combinations are indexed in an arbitrary way by the label g $(g = 1, 2, \ldots, G)$. Let t_g denote the value of an estimator T when the initial sample consists of combination g. We define the indicator variable I_g to be 1 if the gth combination could give rise to d_R (i.e., is compatible with d_R), and 0 otherwise. The number of compatible combinations is then

$$\xi = \sum_{g=1}^{G} I_g$$

and, conditional on d_R, each of these is equally likely. Hence, given d_R, $T = t_g$ with probability $1/\xi$ for all compatible g so that an improved estimator is given by

$$\begin{aligned} T_{RB} &= E[T \mid d_R] \\ &= \frac{1}{\xi} \sum_{g=1}^{G} t_g I_g. \end{aligned} \tag{8}$$

Since (8) is based on the samples compatible with d_R, one has to evaluate all ξ estimators from the compatible samples. The number ξ can be very large. In the next sections we apply (8) to an estimator based on the units in d_R. The variance of T_{RB} is given by

$$\mathrm{var}[T_{RB}] = \mathrm{var}[T] - E\{\,\mathrm{var}[T \mid d_R]\},$$

and an unbiased estimator of $\mathrm{var}[T_{RB}]$ is given by

$$\begin{aligned} v[T_{RB}] &= v[T] - \ \mathrm{var}[T \mid d_R] \\ &= v[T] - \frac{1}{\xi} \sum_{g=1}^{\xi} (t_g - T_{RB})^2, \end{aligned} \tag{9}$$

where $v[T]$ is an unbiased estimator of $\mathrm{var}[T]$.

The improvement gained (in the sense of the variance reduction) from using Rao-Blackwell version is $E\{\mathrm{var}[T \mid d_R]\}$ and its estimation is $\mathrm{var}[T \mid d_R]$.

4. A closed form for the RB version of HT estimator

The modified Horvitz-Thompson estimator is based on the network rather than on the units. We therefore define the set E to be the set of network labels in the sample. We partition set E into three sets F_1, F_2 and F_3. The set F_1 denotes the set of network labels in E that their networks sizes are greater than 1. Set F_2 denotes the set of network labels that their networks are eventually selected as edge units (possibly more than once), even if in the initial sample. Set F_3 denotes the set of network labels in E and not in F_2 that their networks are of size 1. If $|\,.\,|$ denotes the cardinality of a set, let $\eta = |F_1|$, $\phi = |F_2|$, $\zeta = |F_3|$, $\nu' = \nu - \zeta$ and $n_1' = n_1 - \zeta$. Since all the networks (units) associated with F_3 must be intersected by an initial sample, we allocate one initial unit to each network associated with F_3. We now have the remaining n_1' initial units to be allocated on ν' units which can be done in

$$\binom{\nu'}{n_1'}$$

ways. Those combinations which do not contain at least one unit from each network associated with F_1 are not compatible with the observation of D_R, say d_R. Let C_k be the set of the combinations which contain no units from network k associated with F_1. The set $\bigcup_{k\in F_1} C_k$ is the set of possible samples which contain no units from at least one of the η network associated with F_1. The number of initial combinations which give rise to d_R (compatible with d_R) is therefore

$$\binom{\nu'}{n_1'} - \Big|\bigcup_{k\in F_1} C_k\Big|,$$

which is given by the ''inclusion-exclusion'' formula for the cardinality of the union of several sets, namely

$$\begin{aligned}\xi &= \binom{\nu'}{n_1'} - \sum_{k\in F_1} |C_K| + \sum_{k,l\in F_1} |C_k \cup C_l| + \cdots + (-1)^{\eta}\Big|\bigcap_{k\in F_1} C_k\Big| \\ &= \binom{\nu'}{n_1'} - \sum_{k\in F_1}\binom{\nu' - m_k}{n_1'} + \sum_{k,l\in F_1}\binom{\nu' - m_k - m_l}{n_1'} \\ &\quad + \cdots + (-1)^{\eta}\binom{\nu' - \sum_{k\in F_1} m_k}{n_1'}, \end{aligned} \tag{10}$$

where m_k is the number of units in network k. Although the last term is actually zero we leave it in for notational convenience. We define an indicator function I_{gk} which takes the value 1 when the gth initial combination contains at least one of the units in network k and the value 0 otherwise. From (8), the Rao-Blackwell version of the HT estimator is given by

$$\begin{aligned}\hat{\mu}_{RBHT} &= \frac{1}{\xi}\sum_{g=1}^{\xi}\frac{1}{N}\sum_{k\in E}\frac{y_k^*}{\alpha_k}I_{gk}\\ &= \frac{1}{\xi N}\sum_{k\in E}\frac{y_k^*}{\alpha_k}\sum_{g=1}^{\xi}I_{gk},\end{aligned}\tag{11}$$

where $\sum_{g=1}^{\xi} I_{gk}$ is the number of initial combinations which contain at least one unit from network k. Since all the combinations contain at least one unit from each of the networks associated with $F_1 \cup F_3$, $\sum_{g=1}^{\xi} I_{gk} = \xi$ for $k \in F_1 \cup F_3$. The sum $\sum_{g=1}^{\xi} I_{gk}$'s are equal, to say ξ_1, for all the networks associated with F_2 because they are of size one and an initial unit must intersect them. To obtain ξ_1, we allocate one initial unit for a particular network of size one giving us $\nu - 1$ units in d_R and $n_1 - 1$ initial units. If we follow a similar approach to the one used for obtaining ξ, we have

$$\begin{aligned}\xi_1 = &\binom{\nu' - 1}{n_1' - 1} - \sum_{k\in F_1}\binom{\nu' - m_k - 1}{n_1' - 1} + \sum_{k,l\in F_1}\binom{\nu' - m_k - m_l - 1}{n_1' - 1}\\ &+ \cdots + (-1)^{\eta}\binom{\nu' - \sum_{k\in F_1} m_k - 1}{n_1' - 1}.\end{aligned}\tag{12}$$

On substitution into (11), we have

$$\begin{aligned}\hat{\mu}_{RBHT} &= \frac{1}{N}\left(\sum_{k\in F_1\cup F_3}\frac{y_k^*}{\alpha_k} + \sum_{k\in F_2}\frac{y_k^*\xi_1}{\alpha_k\xi}\right)\\ &= \frac{1}{N}\sum_{k\in F_1\cup F_3}\frac{y_k^*}{\alpha_k} + \frac{\xi_1}{n_1\xi}\sum_{k\in F_2} y_k^*,\end{aligned}\tag{13}$$

since $\alpha_k = n_1/N$ for networks of size one.

In the Appendix, we show that

$$\text{var}[\hat{\mu}_{HT}|d_R] = \frac{1}{(n_1\xi)^2}\left((\xi_1\xi - \xi_1^2)\sum_{k\in F_2} y_k^{*2} + 2(\xi_{12}\xi - \xi_1^2)\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*\right),$$

where ξ_{12} is the number of initial combinations which contain any two networks (units) associated with F_2. We allocate two particular initial units to two networks, so that we have $\nu - 2$ units in d_R and $n_1 - 2$ initial units to select. Once again we follow a similar approach to the one used for obtaining ξ.

$$\begin{aligned}\xi_{12} = &\binom{\nu' - 2}{n_1' - 2} - \sum_{k\in F_1}\binom{\nu' - m_k - 2}{n_1' - 2} + \sum_{k,l\in F_1}\binom{\nu' - m_k - m_l - 2}{n_1' - 2}\\ &+ \cdots + (-1)^{\eta}\binom{\nu' - \sum_{k\in F_1} m_k - 2}{n_1' - 2},\end{aligned}\tag{14}$$

where the last term may be zero. Finally, an unbiased variance estimator of $\hat{\mu}_{HTRB}$ is

$$v[\hat{\mu}_{RBHT}] = v[\hat{\mu}_{HT}] - \frac{1}{(n_1\xi)^2}\left((\xi_1\xi - \xi_1^2)\sum_{k\in F_2} y_k^{*2} + 2(\xi_{12}\xi - \xi_1^2)\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*\right).\tag{15}$$

From (15), we can clearly see that if the y-values for the (edge) units which do not satisfy condition C are zero, then $\hat{\mu}_{HT}$ could not be improved using the Rao-Blackwell theorem. A frequently used the condition C for extra samples is $y_i > 0$, where the y-values are non-negative. In this case

$$E(\hat{\mu}_{HT}|d_R) = \hat{\mu}_{HT},$$

which is intuitively obvious. It does not matter which of the networks of size 1 not satisfying the condition are in the initial sample, and which are edge units. The y_i^* are zero in both cases. With this condition it seems that $\hat{\mu}_{HT}$ is more efficient than $\hat{\mu}_{HH}$ (Thompson, 1990).

5. A closed form for the RB version of HH estimator

The modified HH estimator is based on the value of w_i for each unit so that we define the set F_{1u} which contains all the unit labels in the networks with size greater than one in E. Here F_{1u} is simply F_1 expressed in terms of unit labels rather than network labels. Since the networks associated with F_2 and F_3 are of size one, we now consider them as sets of unit labels rather than sets of network labels in this section. We re-define the indicator function J_{gi} to take the value 1 when the gth initial combination contains unit i, and the value 0 otherwise. The HH estimator using the gth combination can be expressed as

$$t_g = \frac{1}{n_1}\sum_{i\in d_R} w_i J_{gi}.$$

where w_i is the network mean for the network containing unit i. Since all the unit labels in F_3 must be in every combination g, we have

$$\begin{aligned} t_g &= \frac{1}{n_1}\left[\sum_{i\in F_{1u}} w_i J_{gi} + \sum_{i\in F_2} w_i J_{gi} + \sum_{i\in F_3} w_i\right] \\ &= \frac{1}{n_1}\left[\sum_{i\in F_{1u}\cup F_2} w_i J_{gi} + \sum_{i\in F_3} w_i\right]. \end{aligned} \tag{16}$$

From (8) the Rao-Blackwell version of HH estimator is given by,

$$\begin{aligned} \hat{\mu}_{RBHH} &= \frac{1}{\xi}\sum_{g=1}^{\xi}\frac{1}{n_1}\left[\sum_{i\in F_{1u}} w_i I_{gi} + \sum_{i\in F_2} w_i I_{gi} + \sum_{i\in F_3} w_i\right] \\ &= \frac{1}{n_1}\left[\frac{1}{\xi}\sum_{i\in F_{1u}} w_i \sum_{g=1}^{\xi} I_{gi} + \frac{1}{\xi}\sum_{i\in F_2} w_i \sum_{g=1}^{\xi} I_{gi} + \sum_{i\in F_3} w_i\right], \end{aligned} \tag{17}$$

where $\sum_{g=1}^{\xi} I_{gi}$ is the number of the combinations which contain unit i, say ξ_i. To obtain ξ_i we allocate one initial unit to unit i and ζ initial units to the units associated with F_3: we then have $\nu' - 1$ units in d_R and $n_1{}' - 1$ initial units. The number ξ_i is given by

$$\binom{\nu' - 1}{n_1' - 1}$$

minus the number of those combinations that contain no unit label from at least one of η

networks associated with F_1. If unit i is in one of the networks associated with F_1, say the kth, it is guaranteed that at least one unit from network k is in all the

$$\binom{\nu' - 1}{n'_1 - 1}$$

combinations. Thus, if η is the size of F_1,

$$\xi_i = \begin{cases} \binom{\nu'-1}{n'_1-1} - \sum_{\{l:\, l\neq k;\, l\in F_1\}} \binom{\nu'-m_l-1}{n'_1-1} & k\in F_1 \\ +\sum_{\{l,h:\, l,h\neq k;\, l,h\in F_1\}} \binom{\nu'-m_l-m_h-1}{n'_1-1} & u_i \text{ is in network } k \\ +\cdots+(-1)^{\eta}\binom{\nu'-\sum m_l-1}{n'_1-1} & \\ & \\ \binom{\nu'-1}{n'_1-1} - \sum_{l\in F_1}\binom{\nu'-m_l-1}{n'_1-1} + \sum_{l,h\in F_1}\binom{\nu'-m_l-m_h-1}{n'_1-1} & i\in F_2 \\ +\cdots+(-1)^{\eta}\binom{\nu'-\sum m_l-1}{n'_1-1} & \end{cases}$$

We now have

$$\hat{\mu}_{RBHH} = \frac{1}{n_1}\left[\frac{1}{\xi}\sum_{i\in F_{1u}\cup F_2} w_i\xi_i + \sum_{i\in F_3} w_i\right]. \tag{18}$$

Using a similar argument to that given in the Appendix, we have

$$\begin{aligned} \operatorname{var}[\hat{\mu}_{HH}|d_R] &= \frac{1}{\xi}\sum_g (t_g - \hat{\mu}_{RBHH})^2 \\ &= \frac{1}{(n_1\xi)^2}\left(\sum_{i\in F_{1u}\cup F_2}(\xi_i\xi - \xi_i^2)w_i^2 + 2\sum_{i\in F_{1u}\cup F_2}\sum_{i<j}(\xi_{ij}\xi - \xi_i\xi_j)w_iw_j\right) \end{aligned}$$

here ξ_{ij} is the number of initial combinations which contain units i and j in $F_{1u}\cup F_2$, namely

$$\xi_{ij} = \begin{cases} \Big[\binom{\nu'-2}{n'_1-2} - \sum_{h\neq k,l}\binom{\nu'-m_h-2}{n'_1-2} & \text{if } k,l\in F_1 \\ +\sum_{h,e\neq k,l}\binom{\nu'-m_h-m_e-2}{n'_1-2} & (u_i \text{ is in network } k \text{ and} \\ +\cdots+(-1)^{\eta}\binom{\nu'-\sum m_h-2}{n'_1-2}\Big] & u_j \text{ is in network } l) \\ & \\ \Big[\binom{\nu'-2}{n'_1-2} - \sum_{h\neq k}\binom{\nu'-m_h-2}{n'_1-2} & \text{if } k\in F_1 \\ \sum_{h,l\neq k}\binom{\nu'-m_h-m_e-2}{n'_1-2} & (u_i \text{ is in network } k,\, j\in F_2), \\ +\cdots+(-1)^{\eta}\binom{\nu'-\sum m_h-2}{n'_1-2}\Big] & \text{or } (u_i \text{ and } u_j \text{ are in network } k) \\ & \\ \Big[\binom{\nu'-2}{n'_1-2} - \sum_{h\in F_1}\binom{\nu'-m_h-2}{n'_1-2} & \text{if } i,j\in F_2 \\ +\sum_{h,e\neq k,l}\binom{\nu'-m_h-m_e-2}{n'_1-2} & \\ +\cdots+(-1)^{\eta}\binom{\nu'-\sum m_h-2}{n'_1-2}\Big] & \end{cases} \tag{19}$$

Finally, an unbiased variance estimator of (17) is given by,

$$v[\hat{\mu}_{RBHH}] = v[\hat{\mu}_{HH}] - \frac{1}{(n_1\xi)^2} \sum_{i\in F_{1u}\cup F_2} (\xi_i\xi - \xi_i^2)w_i^2 - \frac{2}{(n_1\xi)^2} \sum_{i\in F_{1u}\cup F_2} \sum_{i<j} (\xi_{ij}\xi - \xi_i\xi_j)w_iw_j. \tag{20}$$

Comparing (15) and (20) shows that the variance reduction for the HH estimator is greater than that for the HT estimator using the Rao-Blackwell version. We gain improvement using the Rao-Blackwell version of the HH estimator when the y-values of edge units are zero in contrast to the HT estimator which does not.

6. Examples

To illustrate the computations associated with $\hat{\mu}_{HTRB}$ and $\hat{\mu}_{HHRB}$ and to clarify the notation, two examples are worked out. Example 1 is a small example which should clarify the notation.

Example 1 (A small population) Consider population

$$\{(1,12),(2,1000),(3,4),(4,0),(5,5),(6,500),(7,30)\}.$$

The neighborhood of each unit includes all adjacent units. The condition is defined by $C = \{y : y > 10\}$ and initial simple random sample of size 3 is chosen. The resulting observations and the values of $\hat{\mu}_{RBHT}$ and $\hat{\mu}_{RBHH}$ are listed in Table 1. As we expected the improvement of $\hat{\mu}_{HH}$ is more than that of $\hat{\mu}_{HT}$. The estimator $\hat{\mu}_{RBHH}$ is also more efficient than $\hat{\mu}_{RBHT}$ in this example. The variances of $\hat{\mu}_{RBHH}$ and $\hat{\mu}_{RBHT}$ are respectively 7040.44 and 8386.58. The percentage reduction of var$[\hat{\mu}_{RBHH}]$ is

$$100 \times \left(\frac{\text{var}[\hat{\mu}_{HH}] - \text{var}[\hat{\mu}_{RBHH}]}{\text{var}[\hat{\mu}_{HH}]}\right)\% = 28.21\%,$$

and the percentage reduction of var$[\hat{\mu}_{RBHT}]$ is 0.004%. In order to illustrate the methods of Sections 3 and 4, we work out those estimators for the fourth row of Table 1. The 1st, 2nd and 6th units, with y-values 12, 1000 and 500, were selected as the initial sample. Since all of them exceed 10, their neighboring units were also added to the sample. This included the 7th unit, whose neighboring units were also added. Thus,

$$d_R = \{(1,12),(2,1000),(3,4),(5,5),(6,500),(7,30)\},$$

giving $F_1 = \{\{(1,12),(2,1000)\},\{(6,500),(7,30)\}\}$, $F_2 = \{(3,4),(5,5)\}$ and $F_3 = \{\}$. Thus $\nu' = 6$, $n'_1 = 3$,

$$\xi = \binom{6}{3} - \binom{6-2}{3} - \binom{6-2}{3} = 12 \quad \text{and} \quad \xi_1 = \binom{5}{2} - \binom{5-2}{2} - \binom{5-2}{2} = 4.$$

From Table 1 the sample numbers associated with ξ are 4, 5, 8, 9, 13, 14, 18, 19, 23, 24, 25 and, with ξ_1 are 8,9,18,19 when unit 3 is selected and 13,14,23,24 when unit 5 is selected. Since $\hat{\mu}_{HT} = 308.40$, we have

$$\hat{\mu}_{RBHT} = 308.40 + \frac{4 \times 4}{3 \times 12} + \frac{4 \times 5}{3 \times 12} = 309.40.$$

The estimation of improvement gained from using the Rao-Blackwell version is $\mathrm{var}[\hat{\mu}_{HT}|d_R] = 0.52$.

To obtain $\hat{\mu}_{RBHH}$, we have $F_{1u} = \{(1,12),(2,1000),(6,500),(7,30)\}$. Thus

$$\xi_i = \binom{5}{2} - \binom{5-2}{2} = 7$$

for $i \in F_{1u}$ and $\xi i = 4$ for $i \in F_2$. From (18), we have

$$\hat{\mu}_{RBHH} = \frac{1}{3 \times 12}((1012 + 530)7 + (4 + 5)4) = 300.83.$$

The estimation of improvement gained from using the Rao-Blackwell version is $\mathrm{var}[\hat{\mu}_{HH}|d_R = 4122.03$.

Table 1. All possible outcomes of adaptive cluster sample for a population of seven units with y-values 12, 1000, 4, 0, 5, 500, 30 in which the neighborhood of each unit consists of itself plus adjacent units.

No.	*Sample*	$\hat{\mu}_{RBHH}$	$\hat{\mu}_{RBHT}$	*No.*	*Sample*	$\hat{\mu}_{RBHH}$	$\hat{\mu}_{RBHT}$
1	1,2,3;	338.67	203.73	19	2,3,6;1,5,6	300.83	309.40
2	1,2,4;3	225.78	203.29	20	2,4,5;1,3	170.33	204.07
3	1,2,5;3	227.44	204.96	21	2,4,6;1,3,5,7	257.00	308.40
4	1,2,6;3,5,7	300.83	309.40	22	2,4,7;1,3,5,6	257.00	308.40
5	1,2,7;3,5,6	300.83	309.40	23	2,5,6;1,3,7	300.83	309.40
6	1,3,4;2	225.78	203.29	24	2,5,7;1,3,6	300.83	309.40
7	1,3,5;2	227.44	204.96	25	2,6,7;1,3,5	300.83	309.40
8	1,3,6;2,5,7	300.83	309.40	26	3,4,5;	3.00	3.00
9	1,3,7;2,5,6	300.83	309.40	27	3,4,6;5,7	89.67	107.33
10	1,4,5;2	170.33	204.07	28	3,4,7;5,6	89.67	107.33
11	1,4,6;2,3,5,7	257.00	308.40	29	3,5,6;7	120.22	108.44
12	1,4,7;2,3,5,6	257.00	308.40	30	3,4,7;5,6	120.22	108.44
13	1,5,6;2,3,7	300.83	309.40	31	3,6,7;5	120.22	108.44
14	1,5,7;2,3,6	300.83	309.40	32	4,5,6;7	118.89	107.11
15	1,6,7;2,3,5	300.83	309.40	33	4,5,7;6	118.89	107.11
16	2,3,4;1	225.78	203.29	34	4,6,7;5	118.89	107.11
17	2,3,5;1	227.44	204.96	35	5,6,7;	178.33	107.67
18	2,3,6;1,5,7	300.83	309.40				
	Mean					221.57	221.57
	Variance					7040.44	8286.58

An initial sample of three units by simple random sampling without replacement. Whenever an observed y-value exceeds 10, the neighboring units are added to the sample. Initial observed labels are separated from subsequent observed labels in the table by a semicolon. For each possible sample, the value of each estimator is given. The bottom line of the table gives the mean and variance for each estimator.

Example 2 (Ring-necked Ducks) To demonstrate the methods in Sections 4 and 5, we give an example based on a real life population from Smith *et al.* (1995). They used a population of three waterfowl species in their simulation to find out about adaptive cluster sampling. An effort was made to count every individual duck of the three species in a 5,000 km^2 area of central Florida by making systematic flights over the entire study region. The study region extended 100 km east and 50 km north from the southwest corner at 0438000, 3056000 (Universal Transverse Mercator coordinate; zone 17). Two biologists counted waterfowl from separate helicopters during 13–15 December, 1992. Ring-necked ducks were the most abundant duck (4.67 km^{-2}), and blue-winged teal (2.82 km^{-2}) were more numerous than green-winged teal (0.48 km^{-2}). This example is introduced to demonstrate the computations involved with ξ, ξ_i and ξ_{ij}. In Fig. 1, the population of Ring-necked Ducks of Smith *et al.* (1995) is partitioned into 200 units and an initial sample of size $n_1 = 15$ is supposedly selected at random without replacement. The neighborhood of a unit is defined to consist of unit itself and the four adjacent units sharing a

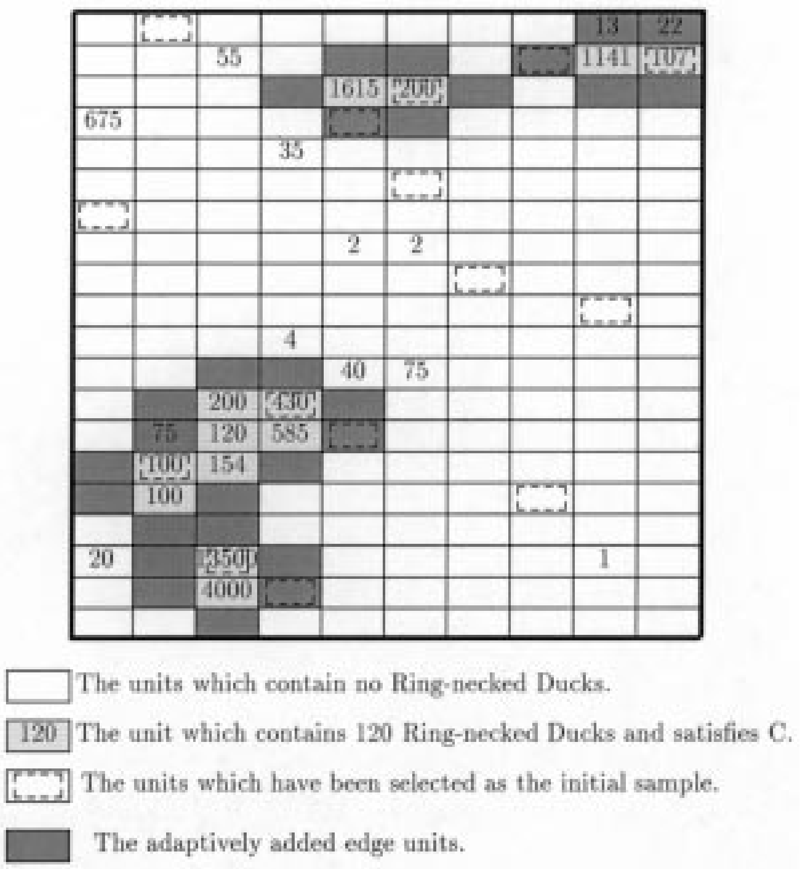
13
22
55
1141
107
1615
200
675
35
2
2
4
40
75
200
430
75
120
585
100
154
100
20
350
1
4000
The units which contain no Ring-necked Ducks.
120 The unit which contains 120 Ring-necked Ducks and satisfies C.
The units which have been selected as the initial sample.
The adaptively added edge units.

Figure 1. Ring-necked Ducks population of Smith *et al.* (1995) is partitioned into 200 units. An adaptive cluster sampling with initial sample of size 15 is selected. The condition for extra sampling is $y_i > 100$.

common boundary. The condition for extra sampling is $\{y : y \geq 100\}$ where y_i is the number of Ring-necked Ducks in unit i. We deliberately consider an initial sample which involves the heaviest computation possible for an initial sample size of 15 from this population.

From Fig. 1, the number of selected units is $\nu = 47$ of which $\zeta = 6$ are non-edge units intersected by the initial sample and which do not satisfy the condition; hence $\nu' = 41$ and $n_1' = 9$. Four networks of size $m_1 = 2$, $m_2 = 7$, $m_1 = 2$ and $m_1 = 2$ were selected. Thus

$$\xi = \binom{41}{9} - 3\binom{39}{9} - \binom{34}{9} + 3\binom{37}{9} + 3\binom{32}{9} - \binom{35}{9} - 3\binom{30}{9} + \binom{28}{9} = 12882163.$$

By similar computation, we find $\xi_1 = 1944684$ and $\xi_{12} = 243257$. We can now substitute them into (13) and (15) to get $\hat{\mu}_{RBHT}$ and its variance estimator. The results are as follows

$$\hat{\mu}_{HT} = 730.38,\ \hat{\mu}_{RBHT} = 731.48;\ v[\hat{\mu}_{HT}] = 308444.65,\ \mathrm{var}[\hat{\mu}_{HT}|d_R] = 3.47.$$

For evaluating $\hat{\mu}_{RBHH}$,

$$\xi_i = \binom{40}{8} - 2\binom{38}{8} - \binom{33}{8} + \binom{36}{8} + 2\binom{31}{8} - \binom{29}{8} = 6959190,$$

for units in the 1st, 3rd and 4th networks, and $\xi_i = 2819025$ for the units in the 2nd network. Here $\xi_i = \xi_1 = 1944684$ for the edge units. On substitution into (17) we have $\hat{\mu}_{HH} = 717.47$, $\hat{\mu}_{RBHH} = 766.32$. Finally, ξ_{ij} is $\binom{39}{7} - \binom{32}{7} - \binom{37}{7} + \binom{30}{7} = 3755409$ when unit i is in network k $(k = 1, 3, 4)$ and unit j in network l $(l \neq k)$ for $k = 1, 3, 4$. It is 1514513 when unit i is in network 2 and unit j in network k $(k = 1, 3, 4)$; it is 659477 when units i and j are in network k $(k = 1, 3, 4)$ or unit i is in network k $(k = 1, 3, 4)$ and j is an edge unit; it is 39603 when units i and j are in network 2, or unit i is in network 2, and j is an edge unit; and is 243257 when i and j are edge units. Using this values we have $v[\hat{\mu}_{HH}] = 309071.23$, $\mathrm{var}[\hat{\mu}_{HT}|d_R] = 4158.18$.

Appendix: *The variance of the modified HT estimator given the sample set*

The modified HT estimator for gth combination can be rewritten as,

$$\begin{aligned} t_g &= \frac{1}{N}\sum_{k\in E}\frac{y_k^*}{\alpha_k}I_{gk} \\ &= \frac{1}{N}\sum_{k\in F_1\cup F_3}\frac{y_k^*}{\alpha_k} + \frac{1}{n_1}\sum_{k\in F_2} y_k^* I_{gk}. \end{aligned} \tag{21}$$

On substitution (13) and (21) into the second term of (9), the first term of (13) and (21) are canceled, we then have

$$\begin{aligned}
\operatorname{var}[\hat{\mu}_{HT}|d_R] &= \frac{1}{\xi}\sum_{g}(t_g - \hat{\mu}_{RBHT})^2 \\
&= \frac{1}{\xi}\sum_{g=1}^{\xi}\left(\frac{1}{n_1}\sum_{k\in F_2} y_k^* I_{gk} - \frac{\xi_1}{n_1\xi}\sum_{k\in F_2} y_k^*\right)^2 \\
&= \frac{1}{n_1^2\xi^3}\sum_{g=1}^{\xi}\left(\sum_{k\in F_2} y_k^*(\xi I_{gk} - \xi_1)\right)^2 \\
&= \frac{1}{n_1^2\xi^3}\sum_{g=1}^{\xi}\sum_{k\in F_2} y_k^{*2}(\xi I_{gk} - \xi_1)^2 \\
&\quad + \frac{2}{n_1^2\xi^3}\sum_{g=1}^{\xi}\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*(\xi I_{gk} - \xi_1)(\xi I_{gl} - \xi_1). \qquad (22)
\end{aligned}$$

Let A be the first term of (22). Since $I_{gk}^2 = I_{gk}$ and $\sum_{g=1}^{\xi} I_{gk} = \xi_1$ for $k \in F_2$,

$$\begin{aligned}
A &= \frac{1}{n_1^2\xi^3}\sum_{k\in F_2} y_k^{*2}\sum_{g=1}^{\xi}(\xi I_{gk} - \xi_1)^2 \\
&= \frac{1}{n_1^2\xi^3}\sum_{k\in F_2} y_k^{*2}\sum_{g=1}^{\xi}(\xi_1^2 + \xi^2 I_{gk}^2 - 2\xi\xi_1 I_{gk}) \\
&= \frac{1}{n_1^2\xi^3}\sum_{k\in F_2} y_k^{*2}\left(\sum_{g=1}^{\xi}\xi_1^2 + \xi^2\sum_{g=1}^{\xi} I_{gk} - 2\xi\xi_1\sum_{g=1}^{\xi} I_{gk}\right) \\
&= \frac{1}{(n_1\xi)^2}\sum_{k\in F_2} y_k^{*2}(\xi_1\xi - \xi_1^2). \qquad (23)
\end{aligned}$$

Let B be the second term of (22). The function $I_{gk}I_{gl}$ is an indicator function which takes value 1 when the gth combination contains at least one unit from network k and one unit from network l and value 0 otherwise. Since networks k and l associated with F_2 each contains 1 unit, $\sum_{g=1}^{\xi} I_{gk}I_{gl} = \xi_{12}$, the number of combinations which contain any two units associated with F_2. We then have

$$\begin{aligned}
B &= \frac{2}{n_1^2\xi^3}\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*\sum_{g=1}^{\xi}(\xi I_{gk} - \xi_1)(\xi I_{gl} - \xi_1) \\
&= \frac{2}{n_1^2\xi^3}\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*\sum_{g=1}^{\xi}(\xi^2 I_{gk}I_{gl} - \xi\xi_1 I_{gk} - \xi\xi_1 I_{gl} - \xi_1^2) \\
&= \frac{2}{n_1^2\xi^3}\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*\left(\xi^2\sum_{g=1}^{\xi} I_{gk}I_{gl} - \xi\xi_1\sum_{g=1}^{\xi} I_{gk} - \xi\xi_1\sum_{g=1}^{\xi} I_{gl} + \sum_{g=1}^{\xi}\xi_1^2\right) \\
&= \frac{2}{(n_1\xi)^2}\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^*(\xi_{12}\xi - \xi_1)^2. \qquad (24)
\end{aligned}$$

On substituting A and B into (22), we have

$$\mathrm{var}[\hat{\mu}_{HT}|d_R] = \frac{1}{(n_{1\xi})^2}\left((\xi_1\xi - \xi_1^2)\sum_{k\in F_2} y_k^{*2} - 2(\xi_{12}\xi - \xi_1^2)\sum_{k\in F_2}\sum_{l<k} y_k^* y_l^* \right).$$

References

Smith, D.R., Conroy, M.J., and Brakhage, D.H. (1995) Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics*, **51**, 777–88.

Thompson, S.K. (1990) Adaptive cluster sampling. *Journal of the American Statistical Association*, **85**, 1050–59.

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

Biographical sketches

Mohammad Salehi Marzijarani was a lecturer in Department of Mathematics, Arak University, Arak, Iran, from 1990 to 1994. He got his Ph.D. in Statistics at Auckland University, New Zealand in 1998. He is an assistant professor in the School of Mathematics, Isfahan University of Technology, Isfahan, Iran.