Improving Unbiased Estimators in Adaptive Cluster Sampling

Arthur L. Dryver and Steven K. Thompson, Pennsylvania State University
Arthur L. Dryver, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802 (dryver@stat.psu.edu)

Key Words: Adaptive Cluster Sampling; Designed-based unbiased; Rao-Blackwell.

Abstract:

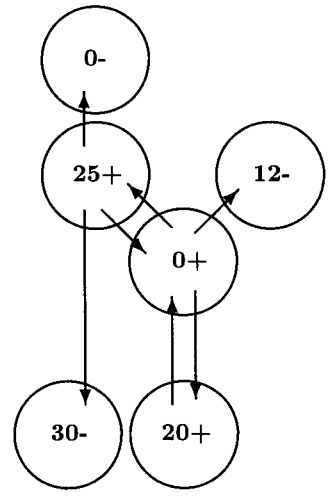
The usual design-unbiased estimators in adaptive cluster sampling can be improved using the Rao-Blackwell method by conditioning on the minimal sufficient statistic. However, the resulting estimators are not commonly used because they are complicated to compute. In this paper easy-to-compute unbiased estimators are presented. These estimators are obtained by conditioning on a statistic that is sufficient but not minimal.

1. Introduction

When dealing with rare or hidden populations, it is often useful after locating a unit that meets a specified criterion to continue sampling in that region. One way of doing so is provided by adaptive cluster sampling. In spatial sampling, adaptive cluster sampling can provide efficient unbiased estimators for the abundance of rare, clustered populations (cf., Thompson and Seber 1996). For sampling hidden human populations, social links play the same role as geographic proximity in spatial sampling and adaptive cluster sampling becomes a type of link-tracing design in a graph or social network (Thompson 1997). In the simplest form of adaptive cluster sampling an initial sample of units is selected by random sampling without replacement (Thompson 1990). Whenever the variable of interest for a unit in the sample satisfies a prespecified condition, neighboring or connected units are added to the sample and observed. This procedure continues until no more units are found that meet the criterion. Conventional estimators, such as a sample mean or expansion estimator, that are unbiased with a conventional design such as simple random sampling are not unbiased with an adaptive design but for adaptive cluster sampling simple design-unbiased estimators of a population mean or total are available.

Support for this research was provided by the National Institutes of Health, National Institute on Drug Abuse, grant RO1 DA09872, and the National Science Foundation, grant DMS-9626102.

Figure 1: The Numbers represent the value of the variable of interest. The network consists of all HIV positive sexually linked people and the edge units are people who are sexually linked and HIV negative.



The usual unbiased estimators in adaptive cluster sampling are very simple but do not necessarily utilize all the information gathered. In particular, the values of edge units are utilized in the estimators only for edge units that were picked in the initial sample. Estimators of higher efficiency can be obtained by taking the expected value of one of the usual estimators conditional on the minimal sufficient statistic. Unfortunately the Rao-Blackwell version of the original estimator, is computationally complex. In this paper new estimators will be presented along with the Rao-Blackwell estimators that can incorparate this previously unused information.

2. Ordinary Estimators in Adaptive Sampling

As in the typical finite population sampling situation, the population consists of N units labeled $1, 2, .., N$ and their associated variables of interest, $\mathbf{y} = \{y_1, y_2, ..., y_N\}$. The population vector $\mathbf{y}$ will be considered fixed but unknown constants. The parameter of interest in this paper is the population mean,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{1}$$

In the simplest form of adaptive cluster sampling an initial sample of units is selected by random sampling without replacement (Thompson 1990). Whenever the variable of interest for a unit in the sample satisfies a prespecified condition, neighboring or connected units are added to the sample and observed. This procedure continues until no more units are found that meet the criterion. Conventional estimators, such as a sample mean or expansion estimator, that are unbiased with a conventional design such as simple random sampling are not unbiased with an adaptive design, but for adaptive cluster sampling simple design-unbiased estimators of a population mean or total are available.

The set of all units meeting the criterion in the neighborhood of one another is called a network. The units that were adaptively sampled that did not meet the criterion are called edge units. Figure 1 illustrates a network and its associated edge units, which together will be called a cluster. In the figure, the neighborhood of a unit is defined as people who are sexually linked, and the criterion for extra sampling is if a person is HIV positive. Units that do not meet the criterion, including edge units, are considered networks of size one.

Two estimators of the population mean which are design-unbiased with adaptive cluster sampling are described below. We call them the ordinary estimators and denote them as $\hat{\mu}_1$ and $\hat{\mu}_2$. Neither of the two estimators is uniformly better than the other, though in empirical studies $\hat{\mu}_2$ is generally more efficient than $\hat{\mu}_1$ (Thompson 1992). These estimators are used when simple random sampling is used to select the initial sample. The units selected in the initial sample are denoted by s_0 and the units in the final sample by s. s_0 is the set of unit labels obtained in the initial sample and s is the set of distinct unit labels in the final sample. Let n denote the initial sample of size and ν the final sample size.

Let ψ_i denote the network which includes unit i and m_i the number of units in that network. w_i represents the average value of a unit in the network which contains unit i, that is

$$w_i = \frac{1}{m_i}\sum_{j\in\psi_i} y_j \tag{2}$$

An unbiased estimator of the population mean is

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} w_i \tag{3}$$

The variance of $\hat{\mu}_1$ is

$$var(\hat{\mu}_1) = \frac{N-n}{Nn(N-1)}\sum_{i=1}^{N}(w_i - \mu)^2 \tag{4}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_1) = \frac{N-n}{Nn(n-1)}\sum_{i=1}^{n}(w_i - \hat{\mu}_1)^2 \tag{5}$$

We next consider how to calculate $\hat{\mu}_2$. Let K equal the number of distinct networks in the population, ψ_k is the set of units in the k^{th} network and x_k denotes the number of units that make up network ψ_k. (Note: x_k is equivalent to m_i except x_k is defined for the distinct networks and m_i the individual units.)

$$y_k^* = \sum_{i\in\psi_k} y_i \tag{6}$$

The sum of the y-values in network k
and the inclusion probability of network k

$$\alpha_k = 1 - \frac{\binom{N-x_k}{n}}{\binom{N}{n}} \tag{7}$$

z_k is an indicator variable which equals one if any unit in the initial sample intersect the k^{th} network.

$$z_k = \begin{cases} 1 & \text{if any unit of the } k^{th} \text{ network is in } s_0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The estimator $\hat{\mu}_2$ is

$$\hat{\mu}_2 = \frac{1}{N}\sum_{k=1}^{K}\frac{y_k^* z_k}{\alpha_k} \tag{9}$$

The joint probability of two networks, k and h being intersected in the initial sample is

$$\alpha_{kh} = 1 - \frac{\left\{\binom{N-x_k}{n} + \binom{N-x_h}{n} - \binom{N-x_k-x_h}{n}\right\}}{\binom{N}{n}} \tag{10}$$

Also $\alpha_{kk} = \alpha_k$. The variance of $\hat{\mu}_2$ is

$$var(\hat{\mu}_2) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^*(\alpha_{kh} - \alpha_k\alpha_h)}{\alpha_k\alpha_h} \tag{11}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_2) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^* z_k z_h(\alpha_{kh} - \alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \tag{12}$$

3. New Estimators

In this section, two new estimators are arrived at by applying the Rao-Blackwell theorem (Rao 1945 and Blackwell 1947) to $\hat{\mu}_1$ and $\hat{\mu}_2$. These estimators are virtually as easy to compute as their ordinary counterparts. When computing the ordinary estimators $\hat{\mu}_1$ and $\hat{\mu}_2$, we incorporate only those edge units which were in the initial sample. The new estimators, which we call $\hat{\mu}_{1+}$ and $\hat{\mu}_{2+}$, are developed considering only how many edge units were initially picked, but not which ones.

3.1 The New Estimator $\hat{\mu}_{1+}$

The final sample s can be partitioned into two parts, a "core" part s_c and the remaining part $s_{\bar{c}}$. The core part s_c is the set of all the distinct units in the sample for which the criterion $y_i \geq c$ is satisfied. The remaining part $s_{\bar{c}}$ consists of all the distinct units in the sample for which $y_i < c$. For unit i, let f_i be the number of times the network to which unit i belongs to is intersected by the initial sample; that is, f_i is the number of units in the initial sample that are in the network to which unit i belongs.

Let the statistic d^+ be defined as

$$d^+ = \{(i, y_i, f_i) : i \in s_c, (j, y_j) : j \in s_{\bar{c}}\} \tag{13}$$

In d^+, the intersection frequency f_i is included only for $i \in s_c$. Let D^+ denote a random variable that takes on possible values of d^+. Also let $\mathcal{D}^+$ denote the sample space for d^+.
For $i \in s$, define the indicator variable e_i as

$$e_i = \begin{cases} 1 & \text{if } y_i < c \text{ and } i \text{ is in} \\ & \text{the neighborhood of some } j \in s_c \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Thus $e_i = 1$ if i is an edge unit and the network that makes it an edge unit is selected in the initial sample. Should $e_i = 1$, we shall refer to that unit as a sample edge unit. Other units picked in the initial sample may be edge units, but sample units are the edge units whose network that classifies them as an edge unit was intersected in the initial sample.

The number of sample edge units in the sample is

$$e_s = \sum_{i=1}^{\nu} e_i = \sum_{i \in s} e_i \tag{15}$$

The number of sample edge units picked in the initial sample s_0 is

$$e_{s_0} = \sum_{i=1}^{n} e_i = \sum_{i \in s_0} e_i \tag{16}$$

The average y-value for the sample edge units in the final sample is

$$\bar{y}_e = \frac{\sum_{i=1}^{\nu} e_i y_i}{e_s} \tag{17}$$

For the ith unit in the sample, define a new variable of interest w_i' by

$$w_i' = w_i(1 - e_i) + \bar{y}_e e_i \tag{18}$$

The variable w_i' is the original w_i when not dealing with sample edge units. When dealing with sample edge units w_i' equals the average of the sample edge units.

The new estimator $\hat{\mu}_{1+}$ is defined by

$$\hat{\mu}_{1+} = E[\hat{\mu}_1 | D^+ = d^+] \tag{19}$$

By the Rao-Blackwell Theorem, $\hat{\mu}_{1+}$ is unbiased for μ, since $\hat{\mu}_1$ is unbiased, and the variance of the new estimator $\hat{\mu}_{1+}$ is less than or equal to the the variance of the ordinary estimator $\hat{\mu}_1$.

$$var(\hat{\mu}_{1RB}) \leq var(\hat{\mu}_{1+}) \leq var(\hat{\mu}_1) \tag{20}$$

Further, unlike $\hat{\mu}_{1RB}$, the new estimator is very easily computed, as shown by the following theorem.

Theorem 1

$$\hat{\mu}_{1+} = \frac{1}{n} \sum_{i=1}^{n} w_i' \tag{21}$$

Since the initial sample determines the final sample and every value of the statistic d^+, let $g(s_0')$ denote the function that maps an initial sample into a value of d^+ resulting from its selection. For any two values of s_0' and d^+ let

$$I(s_0', d^+) = \begin{cases} 1 & \text{if} \quad g(s_0') = d^+ \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Let $L(d^+)$ be the number of initial samples compatible with d^+ and $P(d^+)$ be the probability that $D^+ = d^+$. Also let $\mathcal{S}$ be the sample space containing all possible initial samples. The variance of $\hat{\mu}_{1+}$ is

$$var(\hat{\mu}_{1+}) = \frac{N-n}{Nn(N-1)}\sum_{i=1}^{N}(w_i-\mu)^2 - \frac{1}{n^2}\sum_{d^+\in\mathcal{D}^+}\frac{P(d^+)}{L(d^+)}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d^+)\times \Big(\sum_{i\in s_0^{'},e_i=1} y_i - e_{s_0^{'}}\bar{y}_e\Big)^2 \quad (23)$$

An unbiased estimate of the variance of $\hat{\mu}_{1+}$ when sampling is done without replacement is given by

$$\widehat{var}(\hat{\mu}_{1+}) = \frac{N-n}{Nn(n-1)}\sum_{i=1}^{n}(w_i-\hat{\mu}_1)^2 - \frac{1}{Ln^2}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d^+)\times \Big(\sum_{i\in s_0^{'},e_i=1} y_i - e_{s_0^{'}}\bar{y}_e\Big)^2 \quad (24)$$

However a more efficient estimator is

$$\begin{aligned}\widehat{var}(\hat{\mu}_{1+}) &= E[\widehat{var}(\hat{\mu}_{1+})|d^+] \\ &= \frac{1}{L}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d^+)\frac{N-n}{Nn(n-1)}\times \sum_{i=1}^{n}(w_i-\hat{\mu}_1)^2 \\ &\quad -\frac{1}{Ln^2}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d^+)\times \Big(\sum_{i\in s_0^{'},e_i=1} y_i - e_{s_0^{'}}\bar{y}_e\Big)^2\end{aligned} \quad (25)$$

3.2 The New Estimator $\hat{\mu}_{2+}$

For the kth network in the sample, define the indicator variable

$$e_k^{'} = \begin{cases} 1 & \text{if } y_k^* < c \text{ and } k \text{ is in the} \\ & \text{neighborhood of some } k^{'} \in s_c \\ 0 & \text{otherwise}\end{cases} \quad (26)$$

The variable $e_k^{'}$, for $i = 1,\ldots K$ has meaning similar to e_i but is indexed by network rather than individual unit. Note that all networks of size greater than one must have $e_k^{'} = 0$. Also $e_k^{'} = 0$ for those units not in s.

Let

$$y_k^{'} = \begin{cases} y_k^* & \text{if } e_k^{'} = 0 \\ \bar{y}_e & \text{if } e_k^{'} = 1\end{cases} \quad (27)$$

Thus, for a network of units satisfying the condition, $y_k^{'}$ is the total of the y-values in that network, while for an sample edge unit (a network of size one) $y_k^{'}$ is the average of the y-values for all the sample edge units in the sample.

The new estimator $\hat{\mu}_{2+}$ is defined by

$$\hat{\mu}_{2+} = E[\hat{\mu}_2|D^+ = d^+] \quad (28)$$

By the Rao-Blackwell Theorem $\hat{\mu}_{2+}$ is unbiased for μ since $\hat{\mu}_2$ is unbiased and the variance of $\hat{\mu}_{2+}$ is less than or equal to the variance of $\hat{\mu}_2$. In fact,

$$var(\hat{\mu}_{2RB}) \le var(\hat{\mu}_{2+}) \le var(\hat{\mu}_2) \quad (29)$$

Unlike $\hat{\mu}_{2RB}$, the new estimator is very easily computed, as shown by the following theorem.

Theorem 2

$$\hat{\mu_{2+}} = \frac{1}{N}\sum_{k=1}^{K}\frac{y_k^{'}z_k}{\alpha_k} \quad (30)$$

and the variance of $\hat{\mu}_{2+}$ is

$$var(\hat{\mu}_{2+}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^*y_h^*(\alpha_{kh}-\alpha_k\alpha_k)}{\alpha_k\alpha_k} - \frac{1}{n^2}\sum_{d^+\in\mathcal{D}^+}\frac{P(d^+)}{L(d^+)}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d^+)\times \Big(\sum_{i\in s_0^{'},e_i=1} y_i - e_{s_0^{'}}\bar{y}_e\Big)^2 \quad (31)$$

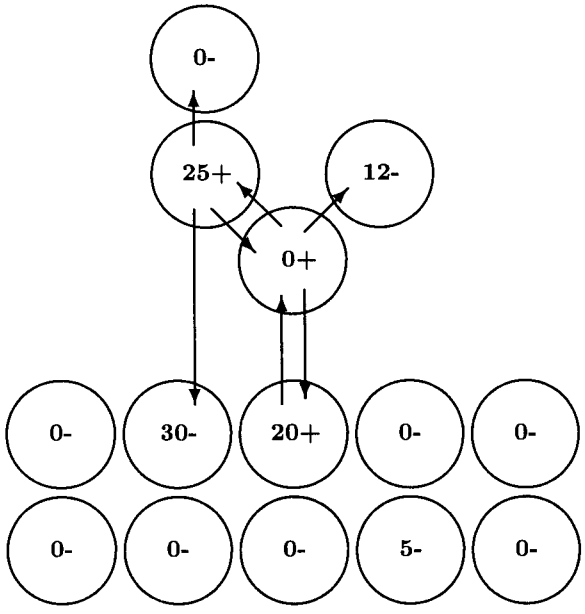
An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_{2+}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^*y_h^*z_kz_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} - \frac{1}{Ln^2}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d^+)\times \Big(\sum_{i\in s_0^{'},e_i=1} y_i - e_{s_0^{'}}\bar{y}_e\Big)^2 \quad (32)$$

A more efficient estimator of the variance is

$$\begin{aligned}\widehat{var}(\hat{\mu}_{2+}) &= E[\widehat{var}(\hat{\mu}_{2+})|d^+] \\ &= \frac{1}{L}\sum_{s_0'\in s} I(s_0', d^+)\frac{1}{N^2}\times \\ &\quad \sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^* z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \\ &\quad -\frac{1}{Ln^2}\sum_{s_0'\in s} I(s_0', d^+)\times \\ &\quad (\sum_{i\in s_0', e_i=1} y_i - e_{s_0'}\bar{y}_e)^2 \end{aligned} \tag{33}$$

4. An Example



The initial sample consists of the $n = 10$ people in the bottom two rows.
For the network intercepted with HIV (+) people
$m_i = 3 \qquad w_i = (20 + 0 + 25)/3 = 15$
Note: There are seven networks in the initial sample with $y_i = 0$
$\hat{\mu}_1 = 1/10 * (7 * 0 + 30 + 15 + 5) = 5$
$\bar{y}_e = 1/3 * (30 + 12 + 0) = 14$
$\hat{\mu}_{1+} = 1/10 * (7 * 0 + 14 + 15 + 5) = 3.4$
Inorder to calculate variance estimates the population size will be $N = 1000000$.

$$\begin{aligned}\widehat{var}(\hat{\mu}_1) &= \frac{1000000-10}{1000000(10)(10-1)}\times \\ &\quad \Big[\ 7\times(0-5)^2+(30-5)^2 \\ &\quad +(15-5)^2+(5-5)^2\Big] \\ &= 9.9999\end{aligned}$$

$$\begin{aligned}\widehat{var}(\hat{\mu}_{1+}) &= 9.9999-\frac{1}{10^2\times\binom{3}{1}}\times \\ &\quad \Big[\ (30-14)^2+(0-14)^2+ \\ &\quad (12-14)^2\Big] \\ &= 8.4799\end{aligned}$$

$$\begin{aligned}\widehat{var}(\hat{\mu}_{1+}) &= (9.99+2.33+3.24)/3-1.52 \\ &= 3.67\end{aligned}$$

5. Bibliography

1. Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics* **18**, 105-110.

2. Rao, C.R. (1945). Information and accuracy attainable in estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* **37**, 81-91.

3. Thompson, S.K. (1990). Adaptive Cluster Sampling. *Journal of the American Statistical Association* **85** 1050-1059.

4. Thompson, S.K. (1992). *Sampling.* New York: Wiley.

5. Thompson, S.K. (1997). Adaptive sampling in behavioral surveys. In Harrison, L., and Hughes, A. eds., The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates. NIDA Research Monograph 167. Rockville, MD: National Institute of Drug Abuse, 296-319.

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