ADAPTIVE CLUSTER SAMPLING WITHOUT REPLACEMENT OF CLUSTERS

Arthur Dryver, and Steven K. Thompson
Arthur Dryver, Department of Statistics,
The Pennsylvania State University,
University Park, PA 16802 U.S.A.
(dryver@lycos.com)

Key Words: Adaptive sampling; Network; Cluster.

Abstract:

In adaptive cluster sampling, although units may be randomly selected without replacement, it is possible to select a cluster more than once. In many situations it may be more desirable to sample clusters without replacement. Estimators where sampling is done without replacement of clusters are described.

1. Introduction

In the simplest form of adaptive cluster sampling an initial sample of n units is selected by random sampling without replacement and, whenever the variable of interest for a unit in the sample satisfies a prespecified condition, neighboring units are added to the sample until no more units are found that meet the criterion (Thompson 1990). Even though the initial sample is selected without replacement, some units in the sample may be selected more than once because the initial sample may contain more than one unit in a given network of units satisfying the condition. 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, called a cluster. In the figure, the variable of interest for a spatial unit is the number of point-objects within the unit, the neighborhood of a unit is defined as including that unit and the four spatially adjacent units, and the criterion for extra sampling is defined as the condition that the variable of interest is greater than or equal to three. Units that do not meet the criterion, including edge units, are considered networks of size one.

An adaptive cluster sample can also be taken without replacement of networks (Salehi and Seber 1997), by selecting each unit of the initial sample at random from the population exclusive of networks already containing an initially selected unit. Even with this procedure, however, a unit which was previously added as an edge unit may be subsequently selected in the initial sample. In this paper we describe a design in which clusters, rather than just networks, are selected without replacement.

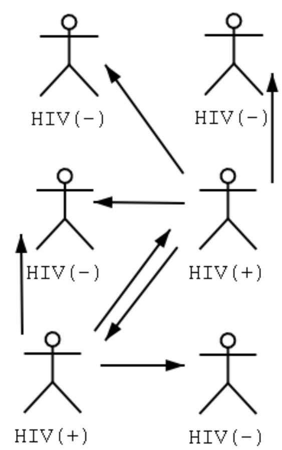

Figure 1: A neighborhood is sexual relations, if a person i has had sexual intercourse with person j then those two people as said to be in the neighborhood of one another. The criterion to adaptively add units (people) was a person had to be HIV positive, if so, their sexual partners were added to the sample. The two people that make are HIV (+) form a network, the remaining neighboring HIV (-) are edge units, and all the units in this example are connected and form a cluster. In this example the size of the network is much smaller than that of the entire cluster.

2. Designs and Terminology

Sampling without replacement of networks (Salehi and Seber 1997). In this design, an initial unit is selected at random from the population and , if its y-value satisfies the condition, its associated network and edge units are observed. The second initial unit is selected at random from the population exclusive of the units in the network already observed. In turn, each of the n units is selected from the population exclusive of previously observed networks of units.

Thus the sample contains exactly n distinct networks, along with their associated edge units. Although sampling is without replacement of networks, repeat observations may occur in the data. A unit observed as an edge unit may be selected more than once. Also, a unit selected in the initial sample may subsequently turn up as an edge unit when an adjacent network is selected.

Sampling without replacement of clusters. In this design, the previous procedure is modified so that each initial unit is selected at random from the population exclusive of all previously observed units, including edge units as well as network units. The sample thus contains n distinct networks.

Repeat observations occur only when a unit in the initial sample is subsequently added as an edge unit of a network or when a unit appears as an edge unit of more than one network. But, in contrast to the previous design, a unit observed first as an edge unit will not be subsequently selected as an initial unit.

3. Estimators

A modified estimator of the Raj type. The Raj estimator, used with the design in which units are selected without replacement such that the i-th draw is performed with probabilities proportional to the size of the remaining units, is

$$\hat{\mu}_{DR} = \frac{1}{Nn}\sum_{i=1}^{n} z_i, \qquad (1)$$

where $z_1 = y_1/p_1$, and, for $i = 2, 3, ..., n$,

$$z_i = \sum_{j=1}^{i-1} y_j + \frac{(1-\sum_{j=1}^{i-1} p_j)y_i}{p_i} \qquad (2)$$

The p_i represent the first-draw probabilities for each unit, so that $p_i/(1-\sum_{j=1}^{i-1} p_j)$ is the conditional i-th draw selection probability for the i-th unit in the sample given the first $i-1$ selections. To avoid double subscripts, the observations are indexed by the order of selection rather than by their population labels.

As shown by Raj (1956), the conditional expectation of each z_i given the inital $i-1$ observations is the population total, so that, unconditionally, z_i/N is an unbiased estimator of the population mean, for $i = 1, ..., N$. Thus their average, the Raj estimator, is an unbiased estimator of the population mean. Raj showed that this strategy would always have variance less than or equal to the variance of the Hansen-Hurwitz estimator used with sampling with replacement.

With the adaptive designs of this paper, the conditional selection probabilities are not known for every sampled unit, and so a modification of the Raj estimator is needed.

Let $p_i^* = m_i/N$, where m_i is the number of units in the network which includes unit i.

For the design in which networks are selected without replacement, let $z_1^* = y_{i.}/p_i^*$, and, for $i = 2, 3, ..., n$,

$$z_i^* = \sum_{j=1}^{i-1} y_{j.} + \frac{(1-\sum_{j=1}^{i-1} p_j^*)y_{i.}}{p_i^*}, \qquad (3)$$

where $y_{i.}$ is the total of the m_i y-values in the network which includes the i-th selection. Note that the edge units are not included in these totals, and hence units not satisfying the criterion are incorporated in the estimate only if they are selected as initial units. Also, the probabilities p_i^* are calculated exclusive of selection as an edge unit. Thus to estimate the population mean and variance without replacement of networks the following equations are used (Salehi and Seber 1997).

$$\hat{\mu}_{Net} = \frac{1}{Nn}\sum_{i=1}^{n} z_i^*, \qquad (4)$$

The variance of the estimator $\hat{\mu}_{Net}$ (Salehi and Seber 1997 based on Raj 1956) is:

$$var(\hat{\mu}_{Net}) = \frac{1}{N^2 n^2}\sum_{i=1}^{n} var(z_i^*) \qquad (5)$$

An unbiased estimate of the variance is:

$$\widehat{var}(\hat{\mu}_{Net}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{z_i^*}{N} - \hat{\mu}_{Net})^2 \qquad (6)$$

When the design is selection without replacement of clusters, the relevant probabilities p_i^* are unchanged, but the z_i are $z_1 = y_{i.}/p_i^*$, and, for $i = 2, 3, ..., n$,

$$z_i = \sum_{j\in s_{i-1}} y_{j.} + \frac{(1-\sum_{j\in s_{i-1}} p_j^*)y_{i.}}{p_i^*}, \qquad (7)$$

where s_{i-1} denotes the collection of all distinct units observed in the first $i-1$ selections.

$$\hat{\mu}_{Clust} = \frac{1}{Nn}\sum_{i=1}^{n} z_i, \tag{8}$$

The variance of the estimator $\hat{\mu}_{Clust}$ (Raj 1956) is:

$$var(\hat{\mu}_{Clust}) = \frac{1}{N^2n^2}\sum_{i=1}^{n} var(z_i) \tag{9}$$

An unbiased estimate of the variance is:

$$\widehat{var}(\hat{\mu}_{Clust}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{z_i}{N} - \hat{\mu}_{Clust})^2 \tag{10}$$

Rao-Blackwell improvement of the modified Raj type estimator - a modified Murthy estimator. The Raj estimator depends on order of selection and so it can be improved by the Rao-Blackwell method. The resulting expected value over all possible reorderings is called the Murthy estimator (Murthy 1957).

The modified Raj type estimators for use with the adaptive designs can similarly be improved by the Rao-Blackwell method. One need in practice consider only the reorderings of the sequence in which the networks (and edge units) in the data are selected, since that is what determines the values of the statistics. Note that different possible reorderings of the data do not all have the same probability. Rather, the probability of each reordering must be computed as the product of the conditional selection probabilities. With the design without replacement of clusters, some reorderings of the data have zero probability, since an edge unit may be selected before a cluster containing it but not after. For the unordering of the estimators we will use a sufficient statistic, d^*, that is not minimal. Let d^* equal the units in the sample, their associated y-values and which networks are intersected in the initial sample. Define the indicator variable e_i as

$$e_i = \begin{cases} 1 & \text{if } i \text{ is in the initial sample.} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$d^* = \{(i, e_i, y_i) : i \in s\} \tag{12}$$

Define s_o to be an ordered sample of n networks. Since the initial sample determines the final sample and every value of the statistic d^*, let $g(s_o^{'})$ denote the function that maps an initial sample into a value of d^* resulting from its selection. For any two values of $s_o^{'}$ and d^* let

$$I(s_o^{'}, d^*) = \begin{cases} 1 & \text{if} \quad g(s_o^{'}) = d^* \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Let $\mathcal{S}$ be the set of all possible ordered initial samples that can be obtained. The following equation can be used for the "unordering" (Murthy 1957) of either estimator, $\hat{\mu}_{Clust}$ or $\hat{\mu}_{Net}$:

$$\begin{aligned} \hat{\mu}_{Murthy} &= \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*) \hat{\mu}(s_o) \div \\ & \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*) \end{aligned} \tag{14}$$

Unfortunately, for without replacement of clusters equation 14 can not be reduced but for without replacement of networks it reduces to the estimator given in Salehi and Seber (1997):

$$\hat{\mu}_{MNet} = \frac{1}{N}\sum_{i=1}^{n}\frac{P(s|i)}{P(s)} y_{i.} \tag{15}$$

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

$$w_i = \frac{1}{m_i}\sum_{j \in \psi_i} y_j \tag{16}$$

The variance for $\hat{\mu}_{MNet}$ (Salehi and Seber 1997) is

$$\begin{aligned} var(\hat{\mu}_{MNet}) &= \frac{1}{N^2}\sum_{i=1}^{K}\sum_{j<i}^{K} m_i m_j \times \Big[1 - \\ & \sum_{s \ni i,j}\frac{P(s|i)P(s|j)}{P(s)}\Big](w_i - w_j)^2 \end{aligned} \tag{17}$$

Let $P(s|i,j)$ denote the probability of sample s given that networks i, j are in the sample regardless of order. Then an unbiased estimator of the variance is

$$\widehat{var}(\hat{\mu}_{MNet}) = \frac{1}{N^2P(s)^2}\Big[\sum_{i=1}^{n}\sum_{j<i}^{n} m_i m_j \times$$

$$(w_i - w_j)^2 \times \{P(s)P(s|i,j) - P(s|i)P(s|j)\}\Big] \quad (18)$$

Equation 14 for without replacement of clusters can also be viewed as

$$\hat{\mu}_{MC} = E[\hat{\mu}_{Clust}|d^*] \quad (19)$$

Thus the variance of the estimator $\hat{\mu}_{MC}$ (Rao 1945, Blackwell 1947) is

$$\begin{aligned} var(\hat{\mu}_{MC}) &= var(\hat{\mu}_{Clust}) - E[(\hat{\mu}_{Clust} - \hat{\mu}_{MC})^2] \\ &= \frac{1}{N^2 n^2}\sum_{i=1}^{n} var(z_i) - \sum_{s_o \in \mathcal{S}} P(s_o)\Big[\hat{\mu}_{Clust}(s_o) - \hat{\mu}_{MC}(s_o)\Big]^2 \end{aligned} \quad (20)$$

An unbiased estimate of the variance is:

$$\begin{aligned} \widetilde{var}(\hat{\mu}_{MC}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{z_i}{N} - \hat{\mu}_{Clust})^2 \\ &\quad - \frac{1}{\sum_{s_o \in \mathcal{S}} P(s_o)I(s_o', d^*)} \times \sum_{s_o \in \mathcal{S}} P(s_o)I(s_o^{'}, d^*) \times \Big[\hat{\mu}_{Clust}(s_o) - \hat{\mu}_{MC}\Big]^2 \end{aligned} \quad (21)$$

A more efficient estimator of the variance is

$$\begin{aligned} \widehat{var}(\hat{\mu}_{MC}) &= E[\widetilde{var}(\hat{\mu}_{MC})|d^*] \\ &= \frac{1}{\sum_{s_o \in \mathcal{S}} P(s_o)I(s_o', d^*)} \times \sum_{s_o \in \mathcal{S}} P(s_o)I(s_o^{'}, d^*) \times \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{z_i}{N} - \hat{\mu}_{Clust})^2 \\ &\quad - \frac{1}{\sum_{s_o \in \mathcal{S}} P(s_o)I(s_o', d^*)} \times \sum_{s_o \in \mathcal{S}} P(s_o)I(s_o^{'}, d^*) \times \Big[\hat{\mu}_{Clust}(s_o) - \hat{\mu}_{MC}\Big]^2 \end{aligned} \quad (22)$$

The estimators $\hat{\mu}_{MC}$ and $\hat{\mu}_{MNet}$ may be hard to compute for large n. The number of different permutations we have to compute is $n!$ for equation 14. The number of calculations can be reduced by noting that certain sets of permutations give the same value of the estimator. Permutations that switch the order of networks that do not meet the condition, are not in the neighborhood of a network selected, that does meet the condition, and have the same y-value give the same value of the estimators. One such situation is when the condition is $y_i > 0$ and two networks of value zero that aren't edge units of an intersected network can be interchanged. Although, this can reduce the number of permutations to be looked at considerably, $\hat{\mu}_{MC}$ may still be complicated to compute. For calculating $\hat{\mu}_{MNet}$ it is possible to reduce the number of calculations still further (Salehi and Seber 1997).

Acknowledgements

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.

4. Bibliography

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

2. Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā A* **18** 379-390.

3. Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association* **51** 269-284.

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

5. Salehi, M.M. and Seber, G.A.F. (1997) Adaptive cluster sampling with networks selected without replacement *Biometrika* **84** 209-219.

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