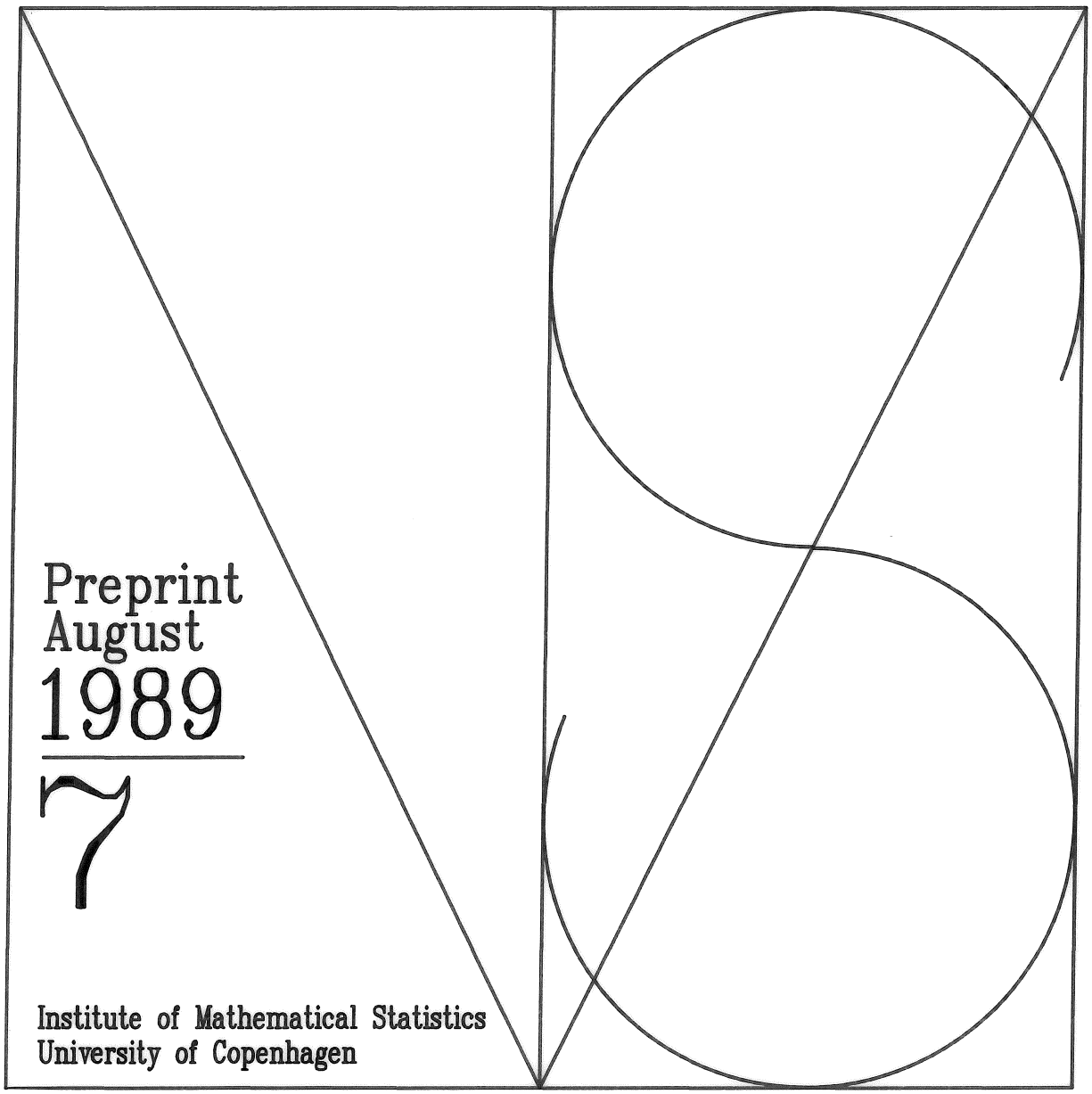
Steven K. Thompson

Stratified Adaptive Cluster Sampling



Preprint
August
1989
7

Institute of Mathematical Statistics
University of Copenhagen

ISSN 0902-8846

Steven K. Thompson*

STRATIFIED ADAPTIVE CLUSTER SAMPLING

Preprint 1989 No. 7

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

August 1989

*University of Alaska Fairbanks, Alaska.

Stratified Adaptive Cluster Sampling

Steven K. Thompson*

Department of Mathematical Sciences
University of Alaska Fairbanks
Fairbanks, Alaska 99775

Abstract. Stratified adaptive cluster sampling refers to designs in which, following an initial stratified sample, additional units are added to the sample from the neighborhood of any selected unit with an observed value that satisfies a condition of interest. For example, in surveys of animal populations, additional observations may be made in the vicinity of any site at which sufficiently high abundance is observed. If any of the added units in turn satisfies the condition, still more units are added to the sample. Such designs differ markedly from conventional stratified sampling designs, in which the entire sample may be selected prior to the survey, selection probabilities do not depend on the values of the variable of interest, and selections in separate strata are independent. Since conventional estimators such as the stratified sample mean are biased with the adaptive designs of this paper, several types of estimators are developed which are unbiased for the population mean or total with stratified adaptive cluster sampling. The variances of the estimators and unbiased estimators of these variances are also obtained. Formulae for optimal allocation of the initial sample among strata apply to some of the estimators. Estimation of the population mean or total with the stratified adaptive cluster designs is complicated by the possibility that a selection in one stratum may result in the addition of units from other strata to the sample, so that observations in separate strata are not independent. The estimators given in this paper differ partly in respect to the weightings given observations resulting from selections in other strata. Improvement of each of the types of estimators through the Rao-Blackwell method is discussed. An example illustrates the use of the different estimators with stratified adaptive cluster sampling.

KEY WORDS: Adaptive sampling; optimal allocation, Rao-Blackwell method, stratified sampling, unbiased estimation.

---

*Department of Mathematical Sciences, University of Alaska Fairbanks, Fairbanks, AK 99775. This research was supported by National Science Foundation Grant DMS-8705812. The paper was written while the author was a sabbatical visitor at the Institute of Mathematical Statistics, University of Copenhagen, Denmark.

1. INTRODUCTION

In stratified adaptive cluster sampling, an initial stratified sample is selected from a population and, whenever the value of the variable of interest for any unit is observed to satisfy a specified condition, additional units from the neighborhood of that unit are added to the sample. Still more units may be added to the sample if in turn any of the subsequently added units satisfies the condition.

The objective of such a design is to obtain the most precise possible estimate of the mean or total of the population, and possibly of individual strata as well. Known patterns of variation are taken into account with the initial stratification of the population, collecting units which tend to be similar into a single stratum. The adaptive addition of neighboring units to the sample whenever a selected unit satisfies the specified condition is designed to take advantage of characteristics such as aggregation tendencies in a population, when the locations and shapes of the aggregations can not be predicted prior to the survey.

Sampling situations in which such adaptive designs apply include surveys of spatially distributed populations such as animal and plant species and geological, mineral, and fossil fuel resources. Whenever sufficiently high abundance is encountered during the survey, neighboring units may be added to the sample. In such surveys, the neighborhood of a unit would typically be defined in terms of spatial proximity. The adaptive procedures also apply to such situations as the study of infectious diseases, in which, following an initial random sample of individuals, additional individuals may be added to the sample whenever an infected person is identified. The neighborhood in such a case can be defined in terms of social or kinship relationships as well as physical proximity.

Stratified adaptive cluster sampling differs markedly from classical stratified sampling, in which the sample selection probabilities do not depend in any way on the variable of interest, the entire sample may be selected prior to the survey, and selections in separate strata are made independently. With the adaptive designs considered in this paper—described in detail in Section 2—the initial selection is made by classical methods, but the subsequent additions to the sample depend on values of the variable of interest associated with selected units. Since the clusters of units added through the adaptive procedure may cross stratum boundaries, selections in one stratum can influence selections in others, and so the final selections are not independent between strata.

Conventional estimators, such as the stratified sample mean, which are

unbiased with classical stratified random sampling, are not unbiased with the adaptive designs. In Section 3, several estimators which are unbiased with the adaptive selection procedures are given, together with formulae for their variances and unbiased estimators of their variances. The unbiased estimators are of five basic types: 1. the stratified sample mean of the initial observations, ignoring all subsequent observations; 2. an estimator using subsequent observations but ignoring any obtained as a result of initial selections in other strata; 3. an estimator related to the "multiplicity estimator" of the network sampling literature; 4. an estimator using expected numbers of intersections between the initial sample and networks of associated units; and 5. an estimator based on probabilities of such intersections. Optimal allocation formulae are obtained for the estimators of the first four types. In addition, since none of the above unbiased estimators is necessarily a function of the minimal sufficient statistic, improvements in each of the estimators through the Rao-Blackwell method are considered. A small example in Section 4 illustrates some of the properties and computations involved with the different estimators.

Adaptive cluster sampling designs without stratification are described in Thompson (1989). Adaptive sampling designs in which the sample size of a simple random sample within each stratum is based on initial observations within the stratum are discussed in Kremers (1987), and Francis (1984). Adaptive designs in which sample size in each stratum or primary unit is based on observations in neighboring strata or primary units are described in Thompson and Ramsey (1983) and Thompson (1988). The importance of adaptive sampling designs for ecological populations with spatial aggregation tendencies is discussed in Seber (1986) and Cormack (1988).

2. DESIGNS

For the adaptive cluster sampling designs of this paper, the population is partitioned into L strata, of which stratum h is comprised of N_h units, for $h = 1, \ldots, L$. The number of units in the population is $N = \sum_{h=1}^{L} N_h$. Associated with unit u_{hi}, the i-th unit of stratum h, is a variable of interest y_{hi}. The object of sampling is estimation of the population mean $N^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}$.

For any unit u_{hi} of the population, the *neighborhood* of unit u_{hi} is defined as a collection of units which includes u_{hi} and with the property that, if unit $u_{h'i'}$ is in the neighborhood of unit u_{hi}, then unit u_{hi} is in the neighborhood of unit $u_{h'i'}$. In spatial sampling situations, the neighborhood of a unit is typically a collection of contiguous or systematically

arranged units centered about that unit. The neighborhood of a unit may include units from more than one stratum.

A unit u_{hi} is said to satisfy the *condition* of interest if the y-value associated with that unit is in a specified set C. In many applications, the condition is specified so that unit u_{hi} satisfies the condition if $y_{hi} \geq c$, for some constant c.

In the designs considered in this paper, an initial sample of units is selected from a population using stratified random sampling; that is, within stratum h, a simple random sample of n_h units is selected without replacement, the selections for separate strata being made independently. Whenever a selected unit satisfies the condition, all units in its neighborhood not already in the sample are added to the sample. Still more units may be added to the sample whenever any of the additionally-added units satisfies the condition, so that the final sample contains every unit in the neighborhood of any sample unit satisfying the condition.

The population may be partitioned into K sets of units, termed *networks*, such that selection in the initial sample of any unit in a network will result in inclusion in the final sample of all units in that network. Any two units satisfying the condition, with one of them in the neighborhood of the other, belong to the same network. A unit not satisfying the condition belongs to a network consisting just of itself. Note that initial selection of a unit satisfying the condition will typically result also in addition to the sample of units not in its network, that is, units not satisfying the condition but in the neighborhood of one or more members of the network.

In spite of the fact that the initial sample is selected without replacement, a unit may be selected more than once. The number of times a unit is selected equals the number of units from its network selected in the initial sample. Let r_{hi} represent the number of times unit u_{hi} is selected. Let m_{khi} denote the number of units in the intersection of stratum k with the network which contains unit u_{hi}. For a unit u_{hi} not satisfying the condition, let a_{khi} be the total number of units in the intersection of stratum k with the collection of distinct networks, exclusive of u_{hi} itself, which intersect the neighborhood of unit u_{hi}. Initial selection of any of these a_{khi} units will result in the addition of unit u_{hi} to the sample. Define a_{khi} to be zero for any unit u_{hi} satisfying the condition. Because of the initial stratified random sampling, the expected number of times unit

u_{hi} is selected is

$$E(r_{hi}) = \sum_{k=1}^{L} n_k \frac{m_{khi} + a_{khi}}{N_k}.$$

The unit u_{hi} will be included in the sample if one or more units from the network to which u_{hi} belongs is included in the initial selection or, for a unit u_{hi} not satisfying the condition, if one or more units from any network which intersects the neighborhood of unit u_{hi} is included in the initial sample. Because of the initial stratified random sampling, the inclusion probability α_{hi} for unit u_{hi} is

$$\alpha_{hi} = 1 - \prod_{k=1}^{L} \binom{N_k - m_{khi} - a_{khi}}{n_k} \Big/ \binom{N_k}{n_k}.$$

The expected sample size ν—that is, the expected number of distinct units in the final sample—is the sum of the above inclusion probabilities (see Godambe, 1955, and Cassel, Särndal, and Wretman, 1977, p. 11):

$$E(\nu) = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \alpha_{hi}.$$

3. ESTIMATORS

Conventional estimators such as the stratified sample mean, although unbiased for the population mean with classical stratified random sampling, are not unbiased with the adaptive designs described in this paper (see the example in the next section). In this section several estimators which are unbiased with stratified adaptive cluster sampling are introduced and evaluated. The estimators and the estimators of variance given are design-unbiased, so that the unbiasedness does not depend on any assumptions about the population itself.

3.1 The Initial Stratified Sample Mean

The stratified mean of the observations in the initial sample,

$$t_1 = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi},$$

is an unbiased estimator of the population mean, since the initial sample is selected by stratified random sampling. Expressions for the variance of

t_1 are well-known, and an unbiased estimator of its variance is available provided that the sample sizes n_h are at least two in each stratum. Optimal allocation of the initial sample among the strata to minimize the variance of t_1 is given by the classical formula. With the estimator t_1, all observations adaptively added to the sample are ignored.

3.2 An Estimator Ignoring Crossover Between Strata

An unbiased estimator which makes use of observations added to the sample subsequent to the initial sample, but using only those added in the same stratum as the initial selection, can be formed as follows. Define the indicator variable $I_{(hi,kj)}$ to be one if unit u_{hi} and unit u_{kj} belong to the same network and zero otherwise. For unit u_{hi}, define ξ_{khi} to be the total of the y-values in the intersection of stratum k with the network that includes unit u_{hi}; that is, $\xi_{khi} = \sum_{j=1}^{N_k} y_{kj} I_{(hi,kj)}$. The number of units in this intersection is $m_{khi} = \sum_{j=1}^{N_k} I_{(hi,kj)}$.

For unit u_{hi}, let the new variable w_{hi} be the total of the y-values in the intersection of the stratum and network of unit u_{hi}, divided by the number of units in that interesction, that is,

$$w_{hi} = \frac{\xi_{hhi}}{m_{hhi}}.$$

The estimator of the population mean is

$$t_2 = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} w_{hi}.$$

To see that t_2 is unbiased for the population mean, let the random variable r_{khi} represent the number of units in the initial sample which are in the intersection of stratum k with the network to which unit u_{khi} belongs. Since the y-value of a unit appears in the estimator in as many terms as there are initially selections in the intersection of the stratum and network to which that unit belongs, the estimator t_2 can be written

$$t_2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi} \frac{N_h}{n_h} \frac{r_{hhi}}{m_{hhi}}.$$

Since r_{hhi} has a hypergeometric destribution with expected value $E(r_{hhi}) = n_h m_{hhi}/N_h$, it follows that the expected value of t_2 equals the population mean.

The variance of t_2 is readily obtained from classical results if one notes that, with the variable w_{hi} replacing the y-variable of unit u_{hi} for each unit of the population, t_2 is the stratified sample mean of a stratified random sample from the population. Hence, the variance of t_2 is

$$var(t_2) = \frac{1}{N^2} \sum_{h=1}^{L} N_h(N_h - n_h)\frac{\sigma_{2h}^2}{n_h},$$

where

$$\sigma_{2h}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (w_{hi} - \mu)^2 .$$

An unbiased estimator $\widehat{var}(t_2)$ of this variance is obtained by replacing σ_{2h}^2 in the above expression with the sample variance s_{2h}^2, for $h = 1, \ldots, L$, where

$$s_{2h}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (w_{hi} - t_2)^2 .$$

3.3 An Estimator Related to the Multiplicity Estimator

The stratified multiplicity estimator for network sampling, as introduced by Birnbaum and Sirken (1965) and investigated in Sirken (1972) and Levy (1977), applies the total weight of the y-values in a network only to the stratum in which the initial selection intersecting that network was made. In the multiplicity estimator, each observation is divided by the number of units—called the "multiplicity"—which if initially selected result in inclusion of the given observation in the sample. With adaptive cluster sampling designs, these multiplicities are not known for every unit in the sample, and so an estimator analagous to the multiplicity estimator must use only those aspects of the multiplicities which can be determined from the data.

For unit u_{hi}, define the new variable w_{hi} to be the total of the y-values in the entire network to which unit u_{hi} belongs, divided by the total number of units in that network; that is,

$$w'_{hi} = \sum_{k=1}^{L} \xi_{khi} \Bigg/ \sum_{k=1}^{L} m_{khi} .$$

The modified stratified multiplicity estimator, for use with stratified adaptive cluster sampling, is

$$t_3 = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} w'_{hi}.$$

For every time any unit of a network is selected in the initial sample, the estimator includes a term with the total of the y-values for that network, divided by the network size and weighted by N_k/n_k for the stratum from which the unit was selected. Thus each individual y-value occurs in the estimator every time any unit from the network to which it belongs is selected in the initial sample, but with weightings depending on the strata from which the initial selections came. Thus, the estimator t_3 can be written in the alternate form

$$t_3 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \left(y_{hi} \sum_{k=1}^{L} \frac{N_k}{n_k} r_{khi} \Big/ \sum_{k=1}^{L} m_{khi} \right).$$

Unbiasedness of t_3 for the population mean follows from the fact that $E(r_{khi}) = n_k m_{khi}/N_k$.

Associating the variable w'_{hi} with unit u_{hi}, the estimator t_3 is a stratified sample mean of a stratified random sample. Hence, the variance of t_3 is

$$var(t_3) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{\sigma_{3h}^2}{n_h},$$

where

$$\sigma_{3h}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (w'_{hi} - \mu)^2 .$$

An unbiased estimator $\widehat{var}(t_3)$ of the variance of t_3 is obtained by replacing in the above expression the variance σ_{3h}^2 with the sample variance

$$s_{3h}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (w'_{hi} - t_3)^2 .$$

3.4 An Estimator Using Expected Numbers of Initial Intersections

Originally designed for sampling with replacement with known, unequal draw-by-draw selection probabilities, the Hansen-Hurwitz estimator achieved unbiasedness by dividing the y-value of each unit by the draw-by-draw selection probability of that unit. In extending the idea of this estimator to other types of designs, it is perhaps more to the point to observe that in the Hansen-Hurwitz estimator each observation is divided by the expected number of times it is selected in the sample and multiplied by the number of times it is selected (or, equivalently, is included in

the estimator as many times as selected). In this sense, the multiplicity estimator used with an unstratified population is of the Hansen-Hurwitz type, but in its stratified form is not of that type.

With stratified adaptive cluster sampling, the selection probabilities and hence the expected number of times selected are not known for every unit in the sample, so that an unbiased estimator must be based only on the aspects of the expected selection numbers that can be determined from the data.

For the unit u_{hi}, define the new variable w''_{hi} to be the total of the y-values of the network to which u_{hi} belongs, divided by a weighted sum of the network-stratum intersection sizes as follows:

$$w''_{hi} = \frac{n_h}{N_h} \sum_{k=1}^{L} \xi_{khi} \Bigg/ \sum_{k=1}^{L} \frac{n_k}{N_k} m_{khi} \, .$$

The estimator of the population mean is

$$t_4 = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} w''_{hi}.$$

With this estimator, a y-value of a unit receives a weight that depends on how many units of the network to which it belongs are selected in the initial sample, but does not depend on the strata from which those units were selected. The estimator can be written in the alternative form

$$t_4 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \left[y_{hi} \sum_{k=1}^{L} r_{khi} \Bigg/ \left(\sum_{k=1}^{L} \frac{n_k}{N_k} m_{khi} \right) \right].$$

Since $E(r_{khi}) = n_k m_{khi}/N_k$, it follows that t_4 is an unbiased estimator of the population mean.

With w''_{hi} as the variable of interest for unit u_{hi} for each unit in the population, t_4 is the stratified sample mean of a stratified random sample and hence has variance

$$var(t_4) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{\sigma^2_{4h}}{n_h},$$

where

$$\sigma^2_{4h} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (w''_{hi} - \mu)^2 \, .$$

An unbiased estimator $\widehat{var}(t_4)$ of the variance of t_4 is obtained by replacing σ_{4h}^2 in the above formula with the sample variance

$$s_{4h}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (w''_{hi} - t_4)^2 .$$

For the estimators t_2, t_3, and t_4, the variance of the estimator depends on the sample size n_h in each stratum through a term which is inversely proportional to n_h. By the usual derivation, therefore, the optimal allocation of the total initial sample size n among the strata is given by

$$n_h = n \frac{N_h \sigma_{ih}}{\sum_{k=1}^{L} N_k \sigma_{ik}},$$

where σ_{ih} designates the square root of the variance term σ_{ih}^2 associated with estimator t_i, for $i = 2, \ldots, 4$.

3.5 An Estimator Using Initial Intersection Probabilities

Unbiasedness in the Horvitz-Thompson estimator is achieved by dividing the y-value for each unit in the sample by the probability that unit is included in the sample. With adaptive cluster sampling, these inclusion probabilities can not be determined from the data for every unit in the sample. An estimator close to the Horvitz-Thompson type can be formed using for each unit the probability that the initial sample intersects the network to which that unit belongs, and giving zero weight to any observation not satisfying the condition that was not included in the initial sample. Since these intersection probabilities are constant for every unit within a network, it will be convenient in this section ut work directly in terms of the networks of the population.

Let the K distinct networks of the population be labelled $1, 2, \ldots, K$, without regard to stratum boundaries. Let y_i denote the total of the y-values in the i-th network of the population. Let x_{hi} be the number of units in stratum h which intersect network i. The probability π_i that the initial sample intersects network i is

$$\pi_{hi} = 1 - \prod_{k=1}^{L} \binom{N_k - x_{ki}}{n_k} \Big/ \binom{N_k}{n_k} .$$

Letting $q_i = 1 - \pi_i$, the probability π_{ij} that the initial sample intersects both networks i and j is

$$\pi_{ij} = 1 - q_i - q_j + \prod_{k=1}^{L} \binom{N_k - x_{ki} - x_{kj}}{n_k} \Big/ \binom{N_k}{n_k} .$$

Let the indicator variable z_i be one if the initial sample intersects network i and zero otherwise. The stratified estimator of modified Horvitz-Thompson type is

$$t_5 = \frac{1}{N} \sum_{i=1}^{K} \frac{y_i z_i}{\pi_i}.$$

For $i = 1, \ldots, K$, z_i is a Bernoulli random variable with $E(z_i) = \pi_i$, $var(z_i) = \pi_i(1 - \pi_i)$, and $cov(z_i, z_j) = \pi_{ij} - \pi_i \pi_j)$, for $i \neq j$. It follows that t_5 is an unbiased estimator of the population mean, and, with the convention that $\pi_{ii} = \pi_i$,

$$var(t_5) = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{K} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right).$$

An unbiased estimator of this variance, since $E(z_i z_j) = \pi_{ij}$, is

$$\widehat{var}(t_5) = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{y_i y_j z_i z_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right),$$

provided that the joint intersection probability π_{ij} is not zero for any pair of networks.

The estimator t_5 is not a true Horwitz-Thompson estimator because the π_i are not the inclusion probabilities for every unit, but give rather the probabilities of intersection of the initial sample with the networks to which units belong. Statistical properties of the sampling strategy such as expected sample size depend on the actual inclusion probabilities as given in Section 2.

3.6 Improvement of the Estimators Using the Rao-Blackwell Method

None of the above five unbiased estimators is a function of the minimal sufficient statistic, and so each may be improved by the Rao-Blackwell method of taking its conditional expectation, given the minimal sufficient statistic. The minimal sufficient statistic D in the finite population setting is the unordered set of distinct, labelled observations (Basu, 1969). The initial stratified sample mean t_1 depends on order—i.e. which of the observations were initial. The estimators t_2, t_3, and t_4 depend on the number of times units are selected in addition to order. The estimator t_5 does not depend on the numbers of times selected, but does depend on order in that a sample unit not satisfying the condition is utilized in the estimator only if it was selected in the initial sample.

Starting with any of the unbiased estimators t, one may obtain the Rao-Blackwell version $t_{RB} = E(t|D)$. Consider a given sample, with minimal sufficient statistic D consisting of the y-values and labels of ν distinct sample units. Consider a selection from this sample with n_1 of the sample units from stratum 1, n_2 from stratum 2,..., and n_L of the sample units from stratum L. Let ν_h denote the number of distinct units in the sample from stratum h. Of the $\prod_{h=1}^{L} \binom{\nu_h}{n_h}$ possible combinations from the data, only some will be *compatible* with D in the sense that an initial sample consisting of that combination of units would lead through the adaptive design to a final sample of precisely the units in D. A combination is compatible with D if and only if it contains at least one unit from each of the distinct networks in D, exclusive of units not satisfying the condition which are in the neighborhood of one or more units in the sample which do satisfy the condition. The estimator t_{RB} is the average of the values of t obtained over all such combinations compatable with D.

Decomposing the variance of any of the estimators t as

$$var(t) = E[var(t|D)] + var[E(t|D)],$$

the variance of the the Rao-Blackwell version of t can be written

$$var(t_{RB}) = var(t) - E[var(t|D)].$$

An improved estimator of $var(t)$ can be obtained by the Rao-Blackwell method with $E[\widehat{var}(t)|D)]$, the average over all the compatible combinations of the unbiased estimators $\widehat{var}(t)$. The term $var(t|D)$ can be computed exactly from the sample as the variance of the values of t over all the compatible combinations. Thus an unbiased estimator of $var(t_{RB})$ is provided by the difference of the above two unbiased estimators; this difference can, however, take on negative values.

Computational aspects of the Rao-Blackwell estimators are deserving of further study. The Rao-Blackwell version of the ordinary Hansen-Hurwitz estimator was obtained in Basu (1958) and Pathak (1962), but has not been a popular estimator because of its computational complexity (see Cassel, et. al., 1977, p. 42 and Chaudhuri and Vos, 1988, p. 259). The ordinary Horvitz-Thompson estimator is a function of the minimal sufficient statistic and hence can not be improved by the Rao-Blackwell method, unlike the closely related estimator t_5. The estimator t_5 gives some observations zero weight based on the order in which they appeared in the sample, whereas the Rao-Blackwell version t_{RB5} may utilize these

same observations with positive weight (as in the example of Section 4). Kremers (1987) provides some computationally simplifying results for the Rao-Blackwell estimator based on the initial mean of an adaptive design.

The Rao-Blackwell version t_{RB1} of the initial stratified sample mean t_1 is identical, based on a result in Thompson (1989), with the Rao-Blackwell version t_{RB2} of t_2. The Rao-Blackwell versions t_{RB2}, t_{RB3}, t_{RB4}, and t_{RB5} are, however, distinct estimators, as demonstrated in the example of the following section.

4. EXAMPLE

The computational differences between the estimators of Section 3 can be illustrated with a very small example. Consider a population of five units with y-values {1,2,10,1000,3} divided into two strata so that the first stratum contains the units with the values {1,2,10} and the second stratum contains the units with the values {1000,3}. Let the condition of interest be specified by $C = \{y : y \geq 5\}$, so that, whenever a value greater than or equal to five is observed, the units in the neighborhood of that observation are added to the sample. The neighborhood of each units is defined to include its immediately adjacent units. Thus, for example, if the unit with value 10 is observed, the adjacent units, having values 2 and 1000, are added to the sample; then, since 1000 also exceeds five, the adjacent unit with value 3 is also added to the sample. The two units with values {10, 1000}, the only units in the population which satisfy the condition, form a network which crosses the boundary between strata.

Consider a stratified adaptive cluster sampling design with an initial sample size of one in each stratum, i.e., $n_1 = n_2 = 1$. The six possible samples obtainable under this design, each with equal probability, are listed in Table 1. The initial observations for each of the six possible samples is followed, after the semicolon, by the observations subsequently added to the sample with the adaptive procedure. For each possible sample, the value of each of the unbiased estimators, other than the initial stratified sample mean, is computed. At the bottom of the table are given the means (equal in each case to 203.2, the population mean) and variances of the estimators under the adaptive design.

In the third row of the table, for illustration, the initial sample selected the unit with value 2 from the first stratum and the unit with value 1000 from the second stratum, resulting in the addition to the sample of the units with values 10 and 3. The computations for each of the estimators are as follow: The intersections of the network of {10, 1000} with each stratum have only one unit each, so $t_2 = (1/5)[3(2) + 2(1000)] = 401.2$.

The sample unit with value 3 does not satisfy the condition and was not intersected by the initial sample, so $t_3 = (1/5)[2+(10+1000)/2] = 203.2$. The expected number of times the unit with value 2 is intersected by an initial sample is 1/3. The expected number for the unit with value 10, as well as for the unit with value 1000, is 1/3+1/2=5/6. Thus, $t_4 = (1/5)[2/(1/3) + 10/(5/6) + 1000/(5/6)] = 243.6$. The intersection probability π_i for the unit with value 2 is 1/3. For the units with values 10 and 1000, the intersection probability is 1-(2/3)(1/2)=2/3. Thus, $t_5 = (1/5)[2/(1/3) + 10/(2/3) + 1000/(2/3) = 304.2$.

The conventional stratified sample mean for the sample of the third row would be (1/5)[3(2+10)/2+3(1000+3)/2]=204.2. The mean of these estimates, over the six possible samples is 136.67. Hence, the conventional stratified sample mean is biased when used with the adaptive design.

Three of the possible samples in the table—those in the third, fifth, and sixth rows—have the same set of four distinct observations, and hence have the same value of the minimal sufficient statistic. The Rao-Blackwell version of any of the estimators for each of these samples is obtained by averaging the value of the corresponding estimator over the three samples. The values of the Rao-Blackwell versions of each of the estimators is listed in Table 3 for each of the six possible samples, and the variances of these improved unbiased estimators are given at the bottom of the table.

[August 8, 1989]

REFERENCES

Basu, D. (1969), "Role of the Sufficiency and Likelihood Principles in Sample Survey Theory," *Sankhyā,* Ser. A., 31, 441-454.

Birnbaum, Z.W., and Sirken, M.G. (1965), "Design of Sample Surveys to Estimate the Prevelence of Rare Diseases: Three Unbiased Estimates," *Vital and Health Statistics,* Ser. 2, No.11, Washington:Government Printing Office.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1977), *Foundations of Inference in Survey Sampling,* New York: Wiley.

Chaudhuri, A. and Vos, J.W.E. (1988), *Unified Theory and Strategies of Survey Sampling,* Amsterdam: North-Holland.

Cormack, R.M. (1988), "Statistical Challenges in the Environmental Sciences: A Personal View," *Journal of the Royal Statistical Society,* Ser. A, 151, 201-210.

Francis, R.I.C.C. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand Journal of Marine and Freshwater Research,* 18, 59-71.

Godambe, V.P. (1955), " A Unified Theory of Sampling from Finite Populations," *Journal of the Royal Statistical Society,* Ser. B, 17, 269-278.

Kremers, W.K. (1987), "Adaptive Sampling to Account for Unknown Variability Among Strata," Preprint No. 128, Institut für Mathematik, Universitat Augsburg, Federal Republic of Germany.

Levy, P.S. (1977), "Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevelence of Attributes in Rare Populations," *Journal of the American Statistical Association,* 72, 758-763.

Seber, G.A.F. (1986), "A Review of Estimating Animal Abundance," *Biometrics,* 42, 267-292.

Sirken, M.G. (1972), "Stratified Sample Surveys with Multiplicity," *Journal of the American Statistical Association,* 67, 224-227.

Thompson, S.K. and Ramsey, F.L. (1983), "Adaptive Sampling of Animal Populations," Technical Report 82, Dept. of Statistics, Oregon State University, Corvallis.

Thompson, S.K. (1988), "Adaptive Sampling," *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 784-786.

—— (1989), " Adaptive Cluster Sampling," Preprint no. 5, Institute of Mathematical Statistics, University of Copenhagen.

Table 1. Values of estimators for the six possible samples in the example.

observations	t_2	t_3	t_4	t_5
$1, 1000; 10, 2, 3$	400.6	202.6	243.0	303.6
$1, 3$	1.8	1.8	1.8	1.8
$2, 1000; 10, 3$	401.2	203.2	243.6	304.2
$2, 3$	2.4	2.4	2.4	2.4
$10, 1000; 2, 3$	406.0	505.0	484.8	303.0
$10, 3; 2, 1000$	7.2	304.2	243.6	304.2
	– – – –	– – – –	– – – –	– – – –
mean :	203.2	203.2	203.2	203.2
variance :	39,766.2	30,361.2	27,504.9	20,220.8

Table 2. Values of estimators improved by the Rao-Blackwell method.

observations	t_{RB2}	t_{RB3}	t_{RB4}	t_{RB5}
$1, 1000; 10, 2, 3$	400.6	202.6	243.0	303.6
$1, 3$	1.8	1.8	1.8	1.8
$2, 1000; 10, 3$	271.47	337.47	324.0	303.8
$2, 3$	2.4	2.4	2.4	2.4
$10, 1000; 2, 3$	271.47	337.47	324.0	303.8
$10, 3; 2, 1000$	271.47	337.47	324.0	303.8
	– – – –	– – – –	– – – –	– – – –
mean :	203.2	203.2	203.2	203.2
variance :	22,305.1	22,494.3	21,040.8	20,220.6

PREPRINTS 1988

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33.

No. 1 Jacobsen, Martin: Discrete Exponential Families: Deciding when the Maximum Likelihood Estimator Exists and Is Unique.

No. 2 Johansen, Søren and Juselius, Katarina: Hypothesis Testing for Cointegration Vectors - with an Application to the Demand for Money in Denmark and Finland.

No. 3 Jensen, Søren Tolver, Johansen, Søren and Lauritzen, Steffen L.: An Algorithm for Maximizing a Likelihood Function.

No. 4 Bertelsen, Aksel: On Non-Null Distributions Connected with Testing that a Real Normal Distribution Is Complex.

No. 5 Tjur, Tue: Statistical Tables for Personal Computer Users.

No. 6 Tjur, Tue: A New Upper Bound for the Efficiency of a Block Design.

No. 7 Bunzel, Henning, Høst, Viggo and Johansen, Søren: Some Simple Non-Parametric Tests for Misspecification of Regression Models Using Sign Changes of Residuals.

No. 8 Brøns, Hans and Jensen, Søren Tolver: Maximum Likelihood Estimation in the Negative Binomial Distribution.

No. 9 Andersson, S.A. and Perlman, M.D.: Lattice Models for Conditional Independence in a Multivariate Normal Distribution.

PREPRINTS 1989

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33 .

No. 1 Bertelsen, Aksel: Asymptotic Expansion of a Complex Hypergeometric Function.

No. 2 Davidsen, Michael and Jacobsen, Martin: Weak Convergence of Twosided Stochastic Integrals, with an Application to Models for Left Truncated Survival Data.

No. 3 Johansen, Søren: Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models.

No. 4 Johansen, Søren and Juselius, Katarina: The Full Information Maximum Likelihood Procedure for Inference on Cointegration - with Applications.

No. 5 Thompson, Steven K.: Adaptive Cluster Sampling.

No. 6 Thompson, Steven K.: Adaptive Cluster Sampling: Designs with Primary and Secondary Units.

No. 7 Thompson, Steven K.: Stratified Adaptive Cluster Sampling.