Proposals for Adaptive and Link-Tracing Sampling Designs in Health Surveys

Myron J. Katzoff[1], Monroe G. Sirken[1], Steven K. Thompson[2]

[1] National Center for Health Statistics, 6525 Belcrest Road, Hyattsville, MD 20782

[2] Department of Statistics, 314 Thomas Building, Pennsylvania State University,
University Park, PA 16802-2111

Contact: Myron Katzoff

Key Words: Network sampling, Adaptive Sampling, Sampling of Rare Populations

1. Introduction and Background

All surveys employ, implicitly or explicitly, procedures for linking observational units (the individual elements of a population under study) to selection units (collectively, the elements of a sampling frame or universe used to acquire a sample of observational units). These procedures are called counting rules. For the development of estimators, it will be convenient to define a network to be the collection of observation units that share the same linkage pattern. Note that a single selection unit may intersect more than one network and a network can intersect more than one selection unit. The most common situation for sample surveys is the use of unitary counting rules in which there is a one-to-one correspondence between selection units and observation units and every network consists of one unit. In this paper, we examine two adaptive sampling ideas where this selection unit to observation unit bijection is not maintained. We consider only situations where the inclusion of any one observation unit of a network in the sample would lead to the inclusion of every other observation unit in the network. We also limit ourselves to design-based estimators.

For observation unit i, let $\mathbf{y}_i$ be a $p \times 1$ vector denoting values for the variables from which estimates of population characteristics of interest will be computed. Let $\mathbf{x}$ be some known $p \times 1$ vector of coefficients. In adaptive sampling designs, the addition of observational units may depend on the sum of the quantities $\mathbf{x}'\mathbf{y}_i$, where the summation is over the observational units of a selection unit already contained in the sample. In general, in adaptive designs, the inclusion of observational units in the sample (as a consequence of network intersections or because they are elements in an initial sample of selection units) depends on the observed values of link-variables together with the counting rules as well as the $\mathbf{y}_i$. In these designs, the units that are added are those in a "neighborhood" of a selected unit as determined by the counting rules and the values of the link-variables. In spatial sampling, the area or volume units can be both selection and observational units and the definition of a neighborhood may involve the application of a fixed geographical proximity pattern so that, for example, a unit is linked to another unit if they share a common boundary in which case it will be countable under certain conditions. However, in a human population, the selection units (for example, housing units) and observational units (for example, persons) may be different and it may be advantageous to define neighborhoods utilizing link-variables that describe social relationships.

Network sampling procedures which are also adaptive sampling procedures have been studied by Sirken and his colleagues since the early 1960's. A concise but carefully compiled history of network sampling along with a complete listing of references is contained in Sirken [1]. Recent investigations of network sampling at NCHS have focused on population-based establishment surveys (PBES) in which households are the selection units and transactions with medical establishments are the observation units. In this paper, we take an initial look at two adaptive sampling proposals that are not connected with PBES and which conform quite closely to conventional approaches for applying adaptive methods.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the National Center for Health Statistics.

2. Adaptive Designs Using Social and Geographical Links

In this section we give two contrived examples that are intended to show how adaptive designs can help investigators achieve their statistical objectives.

In the first example, we consider the use of social links in the sampling of certain subpopulations of the U.S. population because they are hard-to-reach. This would be the case when such subpopulations are small, only weakly geographically clustered and when the degree of geographical clustering is dependent upon income-level or occupation. This might be of great concern when, for example, one survey objective is to estimate the prevalence of a rare genetic disorder among, say, the three largest subgroups of Asian-Americans.

For the second example, we consider the use of geographical proximity links in producing estimates of characteristics which might be useful in describing the severity of an anthrax contamination of a building. A sampling procedure for this situation could be the only sensible alternative to the testing of every cubic unit of space in the building when the cost of laboratory tests for doing so would be prohibitive and the amount of time require to complete those tests would be intolerable. One might then proceed as follows:

> (1) Define a volumetric grid of spatial units of uniform dimensions for the building and index the grid-units.
>
> (2) Assume that a grid-unit must contain some critical minimum number (CMN) of spores for biological significance (*i.e.*, to be considered contaminated). Draw a "screening" sample by simple random sampling of sufficient size to estimate the proportion of spatial grid-units which contain the CMN with 98% confidence and with an absolute error of 0.1% or less.
>
> (3) If no unit in the screening sample contains the CMN, declare that there is no evidence that the site is not clear and stop. Otherwise, add grid-units adjacent to each contaminated sample grid-unit according to a fixed "neighboring units" pattern defined prior to sampling and continue to add units adjacent to those grid-units in accordance with this pattern until no other contaminated units are found.

Examples of the kinds of information that it could be important to know include the average size of networks of spores, the total spore-count for the building and where to look for additional networks of spores. The problem of where to look for additional networks of spores involves the development of mathematical models from the screening sample and is not within the scope of this paper.

2.1 The First Example: An Adaptive Design Using Social Links

For concreteness and to illustrate certain features of estimators that are expected to be of interest in applications of network sampling which can be supplements to existing national health surveys, let us assume that the purpose of the design is to quantify the prevalence of the genetic disorder in a particular state. Also assume that "state" is a stratification variable for the sample survey for which the supplement has been developed; that the main survey for each state may be treated as having a two-stage design with counties as PSUs and households (HSDs) within counties as the second stage of sampling; and that t PSUs are chosen from a total of T with unequal probabilities, π_i, without replacement but that n_i HSDs from a total of N_i are selected by simple random sampling (SRS) within counties without replacement. Let π_{ij} denote the joint inclusion probability for PSUs i and j for $i \neq j$; and assume that links among the observation units (persons) found in the selection units (HSDs) are defined by the family relationships: parents, parental siblings, siblings, children and children of siblings who are residents of the state.

If one were to use the notation of the second paragraph of the first section, for each observation unit, he would define $\mathbf{y_l}$ so that three of its components: $y_{lj}, y_{l,j+1}$ and $y_{l,j+2}$, are each jointly indicators of membership in an Asian-American subpopulation of interest and the presence of the genetic disorder. With

$$\mathbf{x} = \begin{cases} 1, & \text{in positions } j, j+1 \text{ and } j+2 \\ 0, & \text{otherwise} \end{cases},$$

unit l would then be interviewed when $\mathbf{x}'\mathbf{y}_l \geq 1$. However, it is enough for this example to restrict attention to the univariate indicator

$$y_l = \begin{cases} 1, & \text{if the person has the disorder} \\ & \text{and is a member of one of the} \\ & \text{Asian-American subpopulations of interest} \\ & \\ 0, & \text{otherwise} \end{cases}.$$

With this simplification, we can now give design-based estimators, their variances and variance estimators for the above-described design. Let A_{ik} denote the set of observation units (in the state) that are linked to selection unit k of PSU i and let m_l denote the number of selection units (in the state population) to which observation unit l is linked. If we define

$$w_{ik} = \sum_{l \text{ in } A_{ik}} \frac{y_l}{m_l}$$

then an unbiased estimator of the total number of state residents with the disorder who belong to the subgroups of interest is

$$\hat{\tau} = \sum_{i=1}^{t} \pi_i^{-1} \frac{N_i}{n_i} \sum_{k=1}^{n_i} w_{ik}. \qquad (1)$$

With $W_i \overset{def}{=} \sum_{k=1}^{N_i} w_{ik}$, it can be shown that

$$\begin{aligned} var(\hat{\tau}) &= \sum_{i=1}^{T} \sum_{j>i}^{T} (\pi_i \pi_j - \pi_{ij}) \left(\frac{W_i}{\pi_i} - \frac{W_j}{\pi_j} \right)^2 \\ &\quad + \sum_{i=1}^{T} \frac{N_i(N_i - n_i)}{n_i \pi_i} S_{2i}^2 \qquad (2) \end{aligned}$$

where

$$S_{2i}^2 = \sum_{k=1}^{N_i} (w_{ik} - \overline{W_i})^2 \,/\, (N_i - 1) \text{ and } \overline{W_i} = W_i / N_i.$$

An estimator of this variance is

$$\begin{aligned} \widehat{var}(\hat{\tau}) &= \sum_{i=1}^{t} \sum_{j>i}^{t} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{\widehat{W_i}}{\pi_i} - \frac{\widehat{W_j}}{\pi_j} \right)^2 \\ &\quad + \sum_{i=1}^{t} \frac{N_i(N_i - n_i)}{n_i \pi_i} s_{2i}^2 \qquad (3) \end{aligned}$$

where

$$s_{2i}^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (w_{ik} - \widehat{\overline{W_i}})^2, \; \widehat{\overline{W_i}} = \frac{1}{n_i} \sum_{k=1}^{n_i} w_{ik},$$

and $\widehat{W_i} = N_i \widehat{\overline{W_i}}$.

2.2 The Second Example: An Adaptive Cluster Design with Proximity Links

For $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$ it will be convenient to define

$$y_{i1} = \begin{cases} 0, \text{ if unit } i \text{ does not contain} \\ \quad \text{the critical minimum number of spores} \\ 1, \text{ otherwise} \end{cases}$$

For each unit i, a neighborhood of adjacent units A_i is defined as consisting of the spatial units sharing a common boundary with unit i. These neighborhoods do not depend upon the value of y_{i1} but all the units in A_i are added to the sample if $y_{i1} = 1$ and unit i is a member of the initial sample. If $y_{j1} = 1$ for any different $j \in A_i$, unit j and all the other units in A_j are added to the sample when unit j is not already a member of the sample. Also, if $j \in A_i$, then $i \in A_j$ for $i \neq j$. If unit i is a member of the initial set of selection units, the union of all the units belonging to neighborhoods that contribute units to the sample because of the inclusion of unit i is called a cluster. In this case, the largest proper subcollection of a cluster such that the selection of any its members would yield a covering of all the units of the cluster $\{j : y_{j1} = 1\}$ is then a network. Any unit for which $y_{i1} = 0$ will be considered a network of size one.

Let $y_{i2} = m_i$, the number of units in the network that includes unit i, and let y_{i3} denote the number of spores in unit i. Following Thompson [3], p. 271, let ψ_i denote the network that includes unit i and, for the i^{th} unit of the initial sample, define

$$w_{1i} = \frac{1}{m_i} \sum_{j \in \psi_i} y_{j1}$$

and

$$w_{2i} = \frac{1}{m_i} \sum_{j \in \psi_i} y_{j1} \cdot y_{j2}.$$

w_{1i} is the indicator for a spore-containing network being linked to the i^{th} unit and w_{2i} is the size of the network (*i.e.*, the number of units in the network) when it contains spores. $w_{2i} = 0$ when the network does not contain spores. If the spatial units all have the same dimensions, w_{2i} is an areal or volumetric measure of network size. With these definitions

$$\hat{\mu_1} = \frac{1}{n} \sum_{i=1}^{n} w_{1i}$$

is an estimator of the proportion of population units that are members of spore-containing networks and

$$\hat{\mu_2} = \frac{1}{n} \sum_{i=1}^{n} w_{2i}$$

is an estimator of the average size of a spore-containing network per population unit. Thus, $\hat{\mu_2}/\hat{\mu_1}$ is one measure of the average size of a spore-containing network. In fact, it is a weighted average in which the sizes of the nonempty networks (those for which $m_i > 1$) are the weights. Note that $N\hat{\mu_1}$ is an estimator of the total number of spatial units that contain spores. If we replace the product $y_{j1} \cdot y_{j2}$ in the definition of w_{2i} with y_{j3}, then w_{2i} is the average number of spores in a spatial unit of the i^{th} network and $\hat{\mu_2}$ estimates the average spore-count per spatial unit. Since $N\hat{\mu_2}$ then provides an estimate of the total spore count, we see that $\hat{\mu_2}/\hat{\mu_1}$ would yield an estimate of the average count in nonempty spatial units.

We end this subsection by noting that, for $k = 1, 2$

$$Var(\hat{\mu_k}) = \frac{(N-n)}{Nn(N-1)} \sum_{i=1}^{N} (w_{ki} - \mu_k)^2 \qquad (4)$$

where

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} w_{ki}.$$

Therefore, employing the Taylor-series method for obtaining expressions for approximate variances, it is seen that

$$Var(\hat{\mu_2}/\hat{\mu_1}) \doteq \frac{(N-n)}{Nn(N-1)} \frac{1}{\mu_2^2} \sum_{i=1}^{N} (w_{1i} - \frac{\mu_1}{\mu_2} w_{2i})^2. \qquad (5)$$

3. The Current Research Project

The research project currently underway focuses on the needs of national demographic health surveys. However, it should be of some interest to readers that an adaptive cluster sampling technique like that illustrated by the second example might be used for an entirely different purpose: investigation of an anthrax contamination of a building. In the context of research for demographic surveys, the same adaptive technique could be applied to increase the number of sample households containing Mexican, Puerto Rican or Cuban Americans by adding housing units which conform with a predefined definition of geographic proximity; that is, by adding neighboring units each time such persons are discovered. Since following all the links of a network encountered in such a sampling plan could overwhelm available field resources, it may be necessary to deliberately terminate the inclusion of units at a specific iteration of the process of adding units. When the addition of units is terminated before all the units of a network have been added to the sample, an instance of truncation is said to have occurred.

The objectives of our research are to:

(1) identify and initiate investigations of adaptive designs for health surveys which are likely to be effective in producing subsamples for various ethnic and racial subpopulations, subgroups of those subpopulations and hidden or hard-to-access populations;

(2) characterize the effects of truncation, nonresponse and incomplete response at different phases in sampling utilizing computer simulation models to the extent feasibile; and

(3) establish guidance for collecting data which can be used to create computer models to simulate adaptive designs appropriate for surveys of human populations and which indicate the effects and extent of nonsampling errors.

References

[1] Sirken,Monroe G.(1997). Network Sampling. *Encyclopedia of Statistics*. Wiley & Sons, **4**, 2977-2986.

[2] Sirken,Monroe G. and Shimizu,Iris(1999). Population-based establishment sample surveys: the Horvitz-Thompson estimator. *Survey Methodology*, **25**, 187-191.

[3] Thompson,S.K.(1992). *Sampling*. New York: John Wiley & Sons.

[4] Thompson, Steven K.(1998). Adaptive sampling in graphs. *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp.13-22.

[5] Thompson, S. and Frank, O.(2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, **26**, 87-98.

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