ADAPTIVE SAMPLING IN GRAPHS

Steven K. Thompson, Pennsylvania State University
Department of Statistics, 326 Thomas Bldg., University Park, PA 16802, U.S.A.
skt@stat.psu.edu

Key Words: Adaptive sampling, Design-based approach, Network sampling, Sampling hidden populations, Sampling in graphs, Snowball sampling.

Abstract:

Adaptive sampling designs are those in which the procedure for selecting the units to include in the sample may depend on values of variables of interest observed during the survey. For example, neighboring units may be added to the sample whenever high values are observed. In spatial sampling the neighborhood is defined by geographic proximity. In studies of human populations the neighborhood may also be defined by social relationships.

In studies of hidden and hard-to-reach human populations such as injection drug users and others at risk for HIV transmission, adaptive link-tracing designs in which initial respondents lead investigators through social links to other individuals often provide the only practical way to obtain a sample large enough for the study. Data summaries or inference from such samples can be misleading, however, if the sample-selection procedure is not taken into account. The situation is conceptualized as sampling in a graph, with the nodes of the graph representing people and the arcs or arrows representing social relationships. The problem is that data are observed for only a sample of the nodes and arcs, from which we wish to infer characteristics of the whole graph or population.

Examples of link-tracing designs include network sampling, snowball sampling, chain-referral methods, "random walk" designs, and adaptive cluster sampling. Design-based and model-based methods of inference with such designs will be discussed in this paper.

Support for this research was provided by the National Science Foundation, grant DMS-9626102, and the National Institutes of Health, National Institute on Drug Abuse, grant RO1 DA09872. The author would like to thank Arthur Dryver for the figure.

1. Introduction

Studies of hidden or hard-to-reach groups often rely on link-tracing designs for obtaining a sample containing sufficient numbers of the people of interest (Friedman, et al. 1997, Neaigus 1995, Neaigus et al. 1996, Rothenberg et al. 1995, Thompson 1997). For example, in studies of injection drug users in relation to transmission of the human imunodeficiency virus (HIV), social leads from initial respondents may be traced and the linked individuals added to the sample. In such studies, the social links are not only useful—indeed necessary—in obtaining the sample, but are of inherent interest in themselves, since transmission of the disease is related to sexual and drug-injection relationships. From a sampling and inference point of view, the problem is that we are interested in characteristics of the entire graph—that is, of the larger population with its social structure—but can observe only a sample of nodes and links from the graph.

An adaptive design is one in which the procedure for selecting units to include in the sample may depend on values of the variable of interest observed during the survey. Many of the link-tracing designs used for hidden and hard-to-reach populations are inherently adaptive in that the selection procedure depends on observed link-variables, as well as node variables, the values of which are not known prior to the study. In this paper sampling strategies for graph-structured populations will be briefly reviewed, and some design-based strategies from adaptive cluster sampling and adaptive allocation will be described and illustrated with numerical examples.

Human populations with social structure can be conceptualized as graphs, with the nodes of the graph representing people and the edges or arcs between nodes representing social relationships between people (cf., Frank 1977a, 1988, Wasserman and Faust 1994). In the design-based approach to survey sampling, the variables of interest in the population are viewed as fixed values and inference methods are evaluated in terms of hypothetically repeated selection of the sample. With model-based approaches, the variables of interest in the population are viewed as random variables having some

joint distribution.

In the graph setting, the variables of interest include both those associated with nodes, such as behavioral characteristics of people, and those associated with pairs of nodes, such as the presence, absence, or magnitude of a given social relationship between two people. With a fixed-population, design-based approach in the graph setting, both the characteristics of the people and the social network structure of the population are viewed as fixed, unknown values. An advantage of design-based methods is that properties such as design-unbiasedness do not depend on any assumptions about the population itself. Even when a stochastic population model is used to help in the design or inference choices, design-based methods can ensure certain desirable inference properties even if the model assumptions turn out to be unrealistic (Godambe 1985, Särndal, Swensson, and Wretman 1992). Design-based approaches are emphasized in this paper; a model-based approach to sampling and inference in graphs is given in Thompson and Frank (1998).

The statistical literature on link-tracing designs, some of it explicitly formulated in the graph framework and some not, includes various methods of snowball sampling, network or multiplicity sampling, chain-referral methods, and "targeted sampling." In snowball sampling, as described by Goodman (1961), an initial sample of individuals were asked to identify a fixed number of acquaintances, who in turn were asked to name the same number of acquaintances, for a fixed number of waves. Frank (1971, 1977a,b, 1978a,b, 1979) developed a number of design-based and model-based methods for inference from samples in graphs and considered generalized snowball sampling procedures with varying numbers of links and waves. Frank and Snijders (1994) developed design- and model-based methods for estimating the size of a hidden population, that is, the number of nodes in the population graph. Snijders (1992) described snowball designs in which only a subsample of the links from each individual were traced. In network or multiplicity sampling (Birnbaum and Sirken 1965, Kalton and Anderson 1986, Levy 1977, Levy and Lemeshow 1991, Sirken 1970, 1972a, b, Sirken and Levy 1974, Sudman, Sirken, and Cowan 1988) social, kinship, and administrative links—generally assumed to be symmetric—were used to obtain observations of additional units. Recognizing that conventional estimators were biased with such procedures, design-unbiased methods were developed for use with a variety of initial sampling designs. Klovdahl (1989) used the term "random walk" to describe a link-tracing design in which only one of the linked individuals from each respondent is selected at random to be added to the sample. Situations in which there is inherently at most one link to follow from each respondent have been termed "chains" (Erickson 1979). Additional discussion of practical issues of link-tracing designs are discussed in Granovetter (1976), Morgan and Rytina (1977), Frank (1980, 1988), van Meter (1990), Spreen (1992), Wasserman and Faust (1994), and Spreen and Zwaagstra (1994). The term "targeted sampling" was introduced by Waters and Biernacki (1989) to describe a combination of survey sampling and ethnographic procedures used to obtain a sample of members of a hidden population, including ethnographic mapping that can be used for stratification and allocation of effort as well as link-tracing from one individual to another.

Adaptive cluster sampling is a class of designs in which neighboring units are added to the sample whenever an observed value satisfies a specified condition. In the spatial setting, neighborhood relationships are defined geographically, while in the graph setting the relationships are typically defined by social connections. When used in the graph setting, the strategy provides design-unbiased estimators applicable when the selection procedure is dependent on observed node values as well as link values and when some of the links are asymmetric.

Adaptive cluster sampling in which the initial sample is selected by random sampling, with or without replacement, was described in Thompson (1990). Other adaptive cluster sampling designs described in the literature include initial unequal probability sampling with replacement (Roesch 1993, Smith et al. 1995), initial cluster and systematic designs (Thompson 1991a), initial stratified designs (Thompson 1991b), initial two-stage designs (Salehi and Seber 1997a), strategies in which the condition for adaptive sampling is based on the order statistics of the initial sample (Thompson 1995), initial "Latin square +1" designs (Munholland and Borkowski 1993) and strategies in which the sampling is without replacement of networks (Salehi M and Seber 1997b) and without replacement of clusters (Dryver and Thompson 1998b). Multivariate aspects are discussed in Thompson (1993). Adaptive cluster sampling sequentially stopped when total sample size exceeds a specified limit is described in Brown (1994) and Brown and Manly (1998). Adaptive cluster sampling was applied to household surveys of rare characteristics in Danaher and King (1994). Adaptive cluster sampling without a fixed frame is described in Roesch (1993) and generalized in Mosquin (1998). Properties of adaptive cluster

sampling are further examined in Christman (1997) and in Thompson and Seber (1996).

Adaptive stratification and allocation refer to stratified designs in which stratum boundaries or allocation of sampling effort among strata depends on values of variables of interest observed during the survey. Reviews of the literature on these strategies can be found in Solomon and Zacks (1970) and Thompson and Seber (1996). Design-unbiased adaptive allocation strategies are described in Thompson, Seber, and Ramsey (1992) and Thompson and Seber (1996). An optimal adaptive design in two phases under an assumed model is described in Chow and Thompson (1997).

2. Sampling in Graphs

In the usual setup for finite-population sampling the population consists of N units with associated label set $U = \{1, 2, ..., N\}$. Associated with the ith unit is a variable of interest y_i and auxiliary variable x_i, each of which can be vector valued. In the fixed-population approach the population y-values, denoted $\mathbf{y} = (y_1, \ldots, y_N)$, are viewed as a collection of fixed, unknown values. In the stochastic population or model-based approach, the population vector $\mathbf{Y} = (Y_1, \ldots, Y_N)$ is viewed as a random vector having some probability distribution $F(\mathbf{y}; \phi)$, which may depend on one or more unknown parameters ϕ. A sample s is a subset of units from U or, if order of selection should be distinguished, a sequence of units from U. The collection of possible samples is denoted $\mathcal{S}$. The y values are observed only for units in the sample, while the x values are usually assumed known for all units in the population. The sampling design is the procedure by which the sample is selected and is characterized by a probability function $p(\cdot)$ defined on $\mathcal{S}$. A selection procedure that does not depend on any values of the variable of interest or on any unknown parameter values can be written $p_x(s)$ (or $p_x(s; \delta)$ if any unknown design parameters δ are involved). More generally, the sampling design is $p_x(s \mid \mathbf{y}; \delta)$. Designs $p(s)$ that do not depend on any values of the variable of interest will be termed *conventional*, while designs that depend on observed values of variables of interest will be termed *adaptive*.

In the graph setting, variables are defined on pairs of units as well as on individual units, so that the population consists not only of units in U but pairs of units in U^2. In this paper, the terms "unit" and "node" will be used interchangeably. A variable of interest associated with an individual node i will be denoted y_i, while a variable of interest associated with a pair of nodes (i, j) will be denoted a_{ij}. Often the variable of interest a_{ij} is an indicator variable with $a_{ij} = 1$ indicating an arc or arrow from unit i to unit j and $a_{ij} = 0$ indicating no such arc, but more generally continuous variables such as the size of a transaction can also be defined on pairs of nodes. The $N \times N$ matrix of a-variables is denoted $\mathbf{a}$. A sample from a graph can include both a sample of nodes and a sample of arcs and is denoted $s = (s^{(1)}, s^{(2)})$, where $s^{(1)}$ is the set of labels on which the unit variable of interest is observed and $s^{(2)}$ is the set of label pairs for which linking variables of interest are observed. The design $p(s \mid \mathbf{y}, \mathbf{a})$ can depend on a-values, as when links are followed from nodes in the initial sample, and on y-values, as when the decision to follow links is based on observed characteristics of the initial nodes. In the fixed population, design-based approach, both $\mathbf{y}$ and $\mathbf{a}$ are considered fixed, while in the stochastic population approach $\mathbf{Y}$ and $\mathbf{A}$ are a random vector and matrix respectively, with as assumed joint probability distribution $F(\mathbf{y}, \mathbf{a}; \phi)$.

3. Adaptive Cluster Sampling in Graphs

In adaptive cluster sampling, linked units are added to the sample whenever the variable of interest for a sample unit satisfies a specified condition. In the social network setting, this means that investigators can choose a protocol that makes the decision to add socially linked people dependent on behavioral or other characteristics of the person already in the sample. In the spatial setting, the inherent linkages of units are given by geographically defined neighborhood relationships. In either setting, the linkages can be asymmetric. For example, person A if included in the sample will lead investigators to person B, but person B will not lead investigators back to A, either because person B does not satisfy the specified condition or because person B chooses not to reveal the identity of A. The asymmetric linkages complicate design-unbiased estimation with such designs by making some inclusion probabilities unknown from sample data.

In the simplest form of adaptive cluster sampling, an initial sample of units is selected by random sampling without replacement. Whenever a unit in the sample satisfies the condition, all units linked to it are added, that is, all units to which there is an arc or arrow from the initial unit. If any of these added units satisfies the condition, the units linked to them are added and so on. A network of units is defined as a complete, strongly connected component; that

is, inclusion of any units in the network will result in the other units in the network being added. Inclusion of a unit may also result in units not in its network being added, because there is an arc from the first unit to a second but not an arc back to the first from the second.

The actual probably that unit i in included in the sample depends not only on the other units in its network, but also on units with arcs or paths leading to i but without paths back. The existence of sone of these asymmetric paths leading in to sample units typically can not be determined from the sample data. Unbiased estimation therefore starts with the symmetric network relationships.

The simplest of the unbiased estimators of the population total with adaptive cluster sampling has the form

$$\hat{\tau}_1 = \frac{N}{n}\sum_{i=1}^{N}\frac{y_i f_i}{m_i}$$

where n is the initial sample size, m_i is the number of units in the network that includes unit i, and f_i is the number of units from that network included in the initial sample. The estimator may be written more simply as $\hat{\tau}_1 = (N/n)\sum_{i=1}^{n} w_i$ where the summation is understood to be over the n selections of the initial sample and w_i is the average unit y-value in the network intersected on the ith selection. An unbiased estimator of the variance of $\hat{\tau}_1$ is

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

where $\hat{\mu} = \hat{\tau}/N$.

A second unbiased estimator has the form

$$\hat{\tau}_2 = \sum_{k=1}^{K}\frac{y_k^* J_k}{\alpha_k}$$

where the summation is over the k networks in the population, y_k^* is the total y-value in the kth network, J_k is an indicator variable equal to one if only if the initial sample intersects network k (that is, one or more units of network k are included in the initial sample), and α_k is the probability that the initial sample intersects network k. This estimator has the form of a Horvitz-Thompson estimator but uses intersection probabilities instead of the actual inclusion probabilities and gives no weight to units in the sample that were selected only through asymmetric linkages out from the initial sample, so that their networks were not intersected by the initial sample. An unbiased estimator of variance is

$$\widehat{\text{var}}(\hat{\tau}_2) = \sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^*}{\alpha_{kh}}\left(\frac{\alpha_{kh}}{\alpha_k\alpha_h} - 1\right)J_k J_h$$

where α_{kh} is the probability that both networks k and h are intersected by the initial sample.

With the initial simple random sample, the intersection probability is

$$\alpha_k = 1 - \binom{N-x_k}{n}\Big/\binom{N}{n}$$

where x_k denotes the number of units in the kth network. The joint intersection probability is

$$\alpha_{kh} = 1 - \left[\binom{N-x_k}{n} + \binom{N-x_h}{n} - \binom{N-x_k-x_h}{n}\right]\Big/\binom{N}{n}$$

for $k \neq h$ and $\alpha_{kk} = \alpha_k$. Slightly more complicated expressions give the intersection probabilities with more complex initial designs.

3.1 Bernoulli initial sample

In the literature on hidden human populations, Bernoulli sampling designs have played an important role as an approximation to the natural process by which initial respondents come into the sample (Frank 1971, 1977, Frank and Snijders 1992). With a Bernoulli sampling design, units in the population are selected for inclusion in the sample independently, with possibly unequal probabilities. Properties of such designs are discussed in Hájek (1981) under the term "Poisson sampling." Let Z_i be the indicator random variable associated with unit i, so that $Z_i = 1$ if $i \in s$ and $Z_i = 0$ if $i \notin s$. The inclusion probability for unit i is $\pi_i = \text{E}(Z_i)$. Also, $\text{var}(Z_i) = \pi_i(1-\pi_i)$ and $\text{cov}(Z_i, Z_j) = 0$ for $i \neq j$. With a Bernoulli sample the Horvitz-Thompson estimator $\hat{\tau} = \sum_{i\in s}(y_i/\pi_i)$ is design-unbiased for the population total τ and has variance $\text{var}(\hat{\tau}) = \sum_{i=1}^{N} y_i^2(1-\pi_i)/\pi_i$ and unbiased variance estimator $\widehat{\text{var}}(\hat{\tau}) = \sum_{i=1}^{N} y_i^2(1-\pi_i)/\pi_i^2$.

An adaptive cluster sampling starting with an initial Bernoulli sample and adding connected units whenever a unit satisfies the condition has, for the kth network, intersection probability

$$\alpha_k = 1 - \prod_{i\in A_k}(1-\pi_i)$$

in which π_i is the probability that unit i is included in the initial sample. The joint intersection probability for two distinct networks k and k' is $\alpha_{kk'} = \alpha_k\alpha_{k'}$. Thus the unbiased estimate and its variance for this type of design is $\hat{\tau} = \sum_{k=1}^{K} y_k^*/\alpha_k$, $\text{var}(\hat{\tau}) = \sum_{i=1}^{N} y_i^2(1-\alpha_k)/\alpha_k$, with unbiased variance estimator $\widehat{\text{var}}(\hat{\tau}) = \sum_{i=1}^{N} y_i^2(1-\alpha_k)/\alpha_k^2$.

3.2 Estimating Equation Approach

In empirical studies of the efficiency of adaptive cluster sampling, the estimator $\hat{\mu}_2$ related to the Horvitz-Thompson estimator has performed better than the simpler $\hat{\mu}_1$ related to the multiplicity or Hansen-Hurwitz estimator. Each of these estimators is design-unbiased for the population mean. A different approach starts with an estimating function for the whole population and then uses a design-unbiased estimator of the estimating function (Godambe and Thompson 1986, Thompson (M.E.) 1997). For instance, letting y_k^* denote the total of the variable of interest for the kth network in the population, suppose it is assumed under a population model that $\mathrm{E}(y_k^*) = x_k\theta$, where x_k is the number of units in the kth network and θ is a parameter of the population model (superpopulation). Then an estimating function having expectation zero under the assumed model is

$$\sum_{k=1}^{K}(y_k^* - \theta x_k)$$

Setting this function equal to zero and solving for θ gives the finite population mean $\theta_N = \sum_{k=1}^{K} y_k^* / \sum_{i=1}^{N} x_k = \sum_{i=1}^{N} y_i/N = \mu$. A design-unbiased estimate of the population estimating function is provided by

$$g(d,\theta) = \sum_{k=1}^{K} \frac{(y_k - \theta x_k)J_k}{\alpha_k}$$

Setting $g = 0$ and solving for θ gives the generalized ratio estimator

$$\hat{\mu}_3 = \frac{\sum_{k=1}^{K} y_k J_k/\alpha_k}{\sum_{k=1}^{K} x_k J_k/\alpha_k}$$

This estimator would be at its best if the y value of each network was exactly proportional to the x-value for that network. Estimators of this form were suggested by Hájek (1971) and are given for adaptive cluster sampling in Thompson (1991a) and examined more widely in Félix Medina (1998).

3.3 Improved Unbiased Estimators

Let s_0 represent an original sample, in order selected and possibly including repeat selections, and $r(s_0)$ the reduction function giving the unordered set s of distinct units. Let $\hat{\tau}(s_0)$ be the value of estimator $\hat{\tau}$ with sample s_0. Let $d = \{(i, y_i), i \in s\}$ be the value of the minimal sufficient statistic actually obtained. Starting with any unbiased estimator $\hat{\tau}$ for τ, the Rao-Blackwell method can be used to obtain an improved unbiased estimator $\hat{\tau}^*$ given by

$$\begin{aligned}\hat{\tau}^* &= \mathrm{E}(\hat{\tau} \mid d) \\ &= \sum_{\{s_0 : r(s_0) = s\}} \hat{\tau}(s_0) \frac{p(s_0 \mid \mathbf{y})}{p(s \mid \mathbf{y})}\end{aligned}$$

With the initial design simple random sampling, in which every sample has equal probability, the improved estimator is simply the average value of the original estimator over all initial samples leading to the same final sample. The improved estimators for adaptive cluster sampling, starting with $\hat{\tau}_1$ and $\hat{\tau}_2$, are described in Thompson (1990, 1991b). Computational forms are given in Salehi (1998). An easy to compute improved estimator involving only the averaging of edge units is described in Dryver and Thompson (1998a).

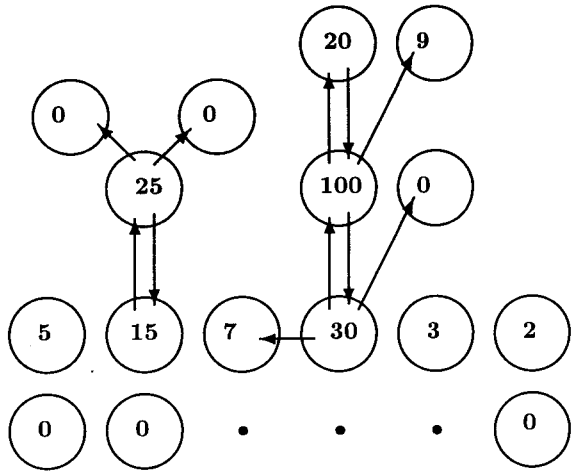
3.4 Example

The following numerical example illustrates a link-tracing strategy in which the design-unbiased estimators of adaptive cluster sampling can be used. The unbiased estimates are contrasted to the conventional sample mean or expansion estimators, which are biased with the link-tracing selection procedure.

Consider a survey of drug use in a population of 1000 people. The variable of interest is amount spent in the last week on the drug and the object of the survey is to estimate the total amount spent during that period by the population or, equivalently, the mean amount spent per person. An initial sample of 100 people is selected using random sampling without replacement. Drug use is relatively rare in the population, and of the 100 people, only 6 people report any drug use at all. The values reported (in dollars) are 5, 15, 7, 30, 3, and 2, with the other 94 initial respondents reporting zero.

Now to obtain a larger sample of users the investigators will follow social links whenever 10 dollars or more is reported spent. So whenever a respondent reports \$10 or more he or she is asked to name close social contacts (not necessarily drug use contacts), and those linked people are added to the sample. The person who reported spending 15 is asked and names one contact who, when interviewed, reports spending 25. This added person in turn reports two additional people. But each of those two people reports spending zero, so they are not questioned on their contacts. The person in the initial sample who spent 30 reports two new people, one who spent 100 and one who spent 0; he also reports the person already in the initial sample who spent 7. The added

Figure 1: The final sample of the example. Links are traced whenever a node has a value of 10 or more.



person who spent 100 reports two new people, reporting 20 and 9. The added person who spent 20 is questioned but reports no contacts other than the person already in the sample who had reported him. The directed graph structure of the sample is shown in Figure 1.

Thus, starting with an initial sample of 100 people, the link-tracing design leads to a final sample of 107 people. The naive sample mean of amount spent per person is

$$\bar{y} = (5+15+7+30+3+2+25+100 +20+9)/107 = 216/107 = 2.019$$

or just over 2 dollars per person. The conventional expansion estimator of the population total is $N\bar{y} = 1000(2.019) = 2019$, so that the conventional estimate of the size of this underground economy that week is over 2019 dollars.

The final sample contains 10 people who reported any use at all, so the ratio of dollars spent to users in the sample is $216/10 = 21.60$, giving almost 22 dollars per user.

However, these conventional data summary statistics are not unbiased estimates of the corresponding quantities for the population, because of the way the sample was selected. Unbiased estimates for this situation are provided by the design-unbiased estimators of adaptive cluster sampling.

Estimation in adaptive cluster sampling uses the network structure in the sample. The person who spent 15 and the person who spent 25 together form one network, because with the design if either one is included in the initial sample both end up in the final sample. The three people reporting 30, 100, and 20 together form another network, because inclusion of any one in the initial sample results in inclusion of all three in the final sample. Each of the other people in the sample forms a network of size one.

The simplest of the design-unbiased estimators simply replaces the original value for each unit in the initial sample with the average of the values in its network. For the network of two units, the average is $(15+25)/2 = 20$. For the network of three units, the average is $(30+100+20)/3 = 50$. The unbiased estimator of the mean amount spent per person on drugs in the population is

$$\hat{\mu}_1 = (5+20+7+50+3+2)/100 = 87/100 = .87$$

so that the unbiased estimate is 87 cents spent per person in contrast to the naive estimate of over two dollars.

An unbiased estimate of the total amount spent in the population is given by the expansion $\hat{\tau}_1 = 1000(.87) = 870$, in contrast to the naive estimate of over 2000 dollars.

There were 6 users in the initial sample, so an unbiased estimate of the number of users in the population is $100(6)/100 = 60$. The ratio of unbiased estimates gives $870/60 = 14.50$ or \$14.50 spent on average by each user in the population, in contrast to the naive estimate of almost \$22.

Another type of design-unbiased estimator from adaptive cluster sampling is only slightly less simple to compute and in empirical studies tends to be more efficient than the first. The second estimator divides the total value of a network by the probability that network was intersected by the initial sample, for each network intersected. In this example, for a network of one person, the intersection probability is simple the probability the person is included in the initial sample, or n/N=0.1. For a larger network, the probability of intersection is the probability that one or more of the units in the network are included in the initial sample. This is readily computed as one minus the probability that the initial sample completely misses the network. The computation is straightforward and involves calculating the number of ways to choose the initial sample from the units not in the network. For the network of two people the intersection probability is .19 and for the network of three people it is .27. The second unbiased estimate of the total amount spent is

$$\hat{\tau}_2 = (5/.1)+(40/.19)+(7/.1)+(150/.27) +(3/.1)+(2/.1) = 936$$

The estimate of total of \$936 in the hidden economic activity is similar to the other unbiased estimate, but again is in contrast to the naive estimate.

The second unbiased estimate of the population mean is $\hat{\mu}_2 = 936/1000 = .934$ or about 94 cents per person.

An unbiased estimate of the number of users in the population is obtained from this method by using as the variable of interest for each person the indicator variable which is one when reported amount spent is greater than zero. The unbiased estimate is $(1/.1) + (2/.19) + (1/.1) + (3/.27) + (1/.1) + (1/.1) = 62$ users in the population. The ratio of unbiased estimates is $936/62 = 15.10$ or about \$15 per user, again in contrast to the naive figure of about \$22.

Table 1. Values of the original estimators for the different types of original samples giving rise to the same final sample.

$f_1, f_2, \ldots, f_7$	$p(s_0 \mid d)$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$
2,1,0,0,0,0,0	3/39	1000	866	873
1,2,0,0,0,0,0	6/39	1300	866	873
1,1,1,0,0,0,0	6/39	870	936	935
1,1,0,1,0,0,0	6/39	890	956	954
1,1,0,0,1,0,0	18/39	800	866	865

The Rao-Blackwell estimates then can be computed as weighted averages of the ordinary estimates using the conditional probabilities in column 2 of the table. The Rao Blackwell estimates are

$$\hat{\tau}_1^* = 917$$

$$\hat{\tau}_2^* = 891$$

$$\hat{\tau}_3^* = 891$$

4. Adaptive Stratification and Allocation

"Targeted sampling" for hidden human populations relies on ethnographic mapping and other means to focus sampling effort in those parts of the study region of most interest to the investigators (Waters and Biernacki 1989, Carlson et al. 1994). To the extent that the ethographic map can be drawn prior to sampling of respondents for the study, the map can be used as auxiliary information for conventional stratification procedures. In reality, however, the mapping depends on talking to respondents who are also part of the ongoing study, so that the use of this information in stratification or allocation is adaptive.

Adaptive stratification refers to designs in which the drawing of stratum boundaries depends on observations made during the survey. Adaptive allocation refers to designs in which the allocation of sampling effort among strata may depend on observations during the survey, even though the stratum boundaries may be fixed. Conventional estimators such as the stratified sample mean that are unbiased with ordinary stratified sampling are typically not unbiased with the adaptive stratification and allocation procedures. In simulation studies, the introduced biases have been small (Francis 1984, 1991), and the conventional estimator has some justification from a model-based viewpoint (Thompson and Seber 1996). It is also possible, however, to use design-unbiased strategies for adaptive stratification and allocation.

Unbiased strategies for adaptive allocation include applying the Rao-Blackwell method to an unbiased estimator based on the initial (conventional) stratified sample (Kremers 1987), basing subsequent allocation on initial observations in different strata (Thompson, Seber, and Ramsey 1992), and multiphase adaptive allocation or stratification strategies with estimators based on fixed-weight averages of the unbiased estimators from each phase (Thompson and Seber 1996, p. 189-191).

When link-tracing or other adaptive designs are used along with stratification, values of observations in one stratum may induce investigators to add linked units not only in the same stratum but in other strata as well, since the inherent graph structure of the population may cross stratum boundaries. Stratified adaptive cluster sampling strategies (Thompson 1991b) provide design-unbiased estimators and estimators of variance for such designs, even though links are followed across stratum boundaries. With the fixed-weight adaptive stratification or allocation strategy, the adaptive cluster sampling strategy or any other design-unbiased strategy may be used at each phase.

In the fixed-weight strategies, the stratification and allocation for each phase after the first can depend adaptively on previous phases, but the weights chosen for averaging the estimators are fixed prior to the survey. Just as the choices of design and sample sizes in a conventional survey can make use of data from past surveys of the same population without biasing the results of the new survey, so the data from previous phases of a single survey can be used in determining stratification boundaries and allocation for the next phase. Let $\hat{\tau}_j$ be a design-unbiased estimator of the population total for the jth phase. Then the estimator

$$\hat{\tau}_w = \sum_{j=1}^{m} w_j \hat{\tau}_j$$

is design-unbiased for τ, where the w_j are any set

of fixed weights with $\sum_{j=1}^{m} w_j = 1$. Since the design and allocation at phase j depend only on the data d_{j-1} from the first $j-1$ phases, the estimators and estimators of variance $\hat{\tau}_j$ and $\widehat{\text{var}}(\hat{\tau}_j)$ are unbiased, over all samples that might be selected in the jth phase, conditional on d_{j-1}. Thus unconditionally the overall variance estimator $\widehat{\text{var}}(\hat{\tau}_w) = \sum_{j=1}^{m} w_j^2 \widehat{\text{var}}(\hat{\tau}_j)$ is design-unbiased as well. Notice that, because the weight that a given data value is given in the estimator depends on the phase and hence the order it was obtained in, the Rao-Blackwell method could also be applied to produce an improved unbiased estimator not depending on order.

4.0.1 Example

A simple numerical example with stratified random sampling at each of two phases illustrates the difference of the unbiased fixed-weight estimator from the biased conventional stratified estimator. Consider a population partitioned into $L = 3$ strata each with $N_h = 10$ units. At the first phase a conventional stratified sample is used, allocating the total first-phase sample size $n_1 = 6$ equally, so that $n_{1h} = 2$ units selected at random without replacement in each stratum. The observed $y - values$ for the three strata are respectively $(10, 5), (0, 2), (3, 6)$ giving first-phase sample means $\bar{y}_{jh}$ for the three strata of $\bar{y}_{11} = 7.5$, $\bar{y}_{12} = 1$, and $\bar{y}_{13} = 4.5$ and sample standard deviations $s_{11} = 3.5$, $s_{12} = 1.4$, and $s + 13 = 2.1$. Allocating a second-phase total sample size of $n_2 = 6$ approximately proportional to the first-phase sample standard deviations gives $n_{21} = 3$, $n_{22} = 1$, and $n_{23} = 2$. The y-values observed at the second phase for the three strata are $(2, 0, 6), (11), (4, 8)$, giving second-phase sample means of $\bar{y}_{21} = 2.7$, $\bar{y}_{22} = 11$, and $\bar{y}_{23} = 6$. The first-phase estimate is $\hat{\tau}_1 = \sum N_h \bar{y}_{1h} = 130$ and the second-phase estimate is $\hat{\tau}_2 = \sum N_h \bar{y}_{2h} = 197$. With equal weights $w_1 = w_2 = 1/2$, the unbiased estimate of the population total is $\hat{\tau} = \sum w_j \hat{\tau}_j = 164$. The conventional but biased estimator on the other hand would use the overall sample means in the three strata of $\bar{y}_1 = 4.6$, $\bar{y}_2 = 4.3$, and $\bar{y}_3 = 5.3$, giving the estimate $\hat{\tau} = \sum N_h \bar{y}_h = 142$.

References

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

Brown, J.A. (1994). The application of adaptive cluster sampling to ecological studies. *In* D.J. Fletcher and B.F.J. Manly (Eds), *Statistics in Ecology and Environmental Monitoring*, pp. 86–97. Otago Conference Series No. 2. Dunedin, New Zealand: University of Otago Press.

Brown, J.A., and Manly, B.F.J. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics,* to appear.

Carlson, R. G., Wang, J., Siegal, H., Falck, R., and Guo, J. (1994). An ethnographic approach to targeted sampling: problems and solutions in AIDS prevention research among injection drug and crack-cocaine users. *Human Organization* **53** 279-286.

Chao, C., and Thompson, S.K. (1997). Optimal sampling design under a spatial model. Technical Report 97-11, Department of Statistics, Pennsylvania State University.

Christman, M. (1997). Efficiency of some sampling designs for spatially clustered populations. *Environmentrics* **8** 145-166.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition. New York: Wiley.

Danaher, P.J., and King, M. (1994). Estimating rare household characteristics using adaptive sampling. *The New Zealand Statistician* **29**, 14–23.

Dryver, A., and Thompson, S.K. (1998a). Improved unbiased estimators for adaptive cluster sampling. Technical Report 98-00, Department of Statistics, Pennsylvania State University.

Dryver, A., and Thompson, S.K. (1998b). Adaptive cluster sampling without replacement of clusters. Technical Report 98-00, Department of Statistics, Pennsylvania State University.

Félix Medina, M.H. (1998). A design-based approach for making inferences from adaptive cluster samples. Technical Report 98-03, Department of Statistics, Pennsylvania State University.

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.

Francis, R.I.C.C. (1991). Statistical properties of two-phase surveys: comment. *Canadian Journal of Fisheries and Aquatic Sciences* **48**, 1228.

Frank, O. (1971). *Statistical Inference in Graphs.* Stockholm: Försvarets forskningsanstalt.

Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference* **1** 235-264.

Frank, O. (1977a). Survey sampling in graphs. Journal of Statistical Planning and Inference, 1, 235-264.

Frank, O. (1977b). Estimation of graph totals. Scandinavian Journal of Statistics, 4, 81-89.

Frank, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. Scandinavian Journal of Statistics, 5, 177-188.

Frank, O. (1978b). Sampling and estimation in large social networks. Social Networks, 1, 91-101.

Frank, O. (1979a). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research,* edited by P.W. Holland and S. Leinhardt. New York: Academic Press, 319-347.

Frank, O. (1979b). Moment properties of subgraph counts in stochastic graphs. *Annals of the New York Academy of Sciences* **319** 207-218. by P.W. Holland and S. Leinhardt. New York: Academic Press, 319-347.

Frank, O. (1988). Random sampling and social networks: a survey of various approaches. *Mathmatiques, Informatique et Sciences humaines* **26** 19-33.

Frank, O. (1997). Composition and structure of social networks. *Mathmatiques, Informatique et Sciences humaines* **35** 11-23.

Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics, 10, 53-67.

Friedman, S.R., Neaigus, A., Jose, B., Curtis, R., Goldstein, M., Ildefonso, G., Rothenberg, R.B., and Des Jarlais, D.C., (1997). Sociometric risk networks and HIV risk. *American Journal of Public Health,* in press.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society B, 17, 269-278.

Godambe, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association* **77**, 393–403.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review* **54** 127-138.

Goodman, L.A. (1961). Snowball sampling. Annals of Mathematical Statistics, 32 148-170.

Granovetter, M. (1976). Network sampling: some first steps. *American Journal of Sociology* **81** 1287-1303.

Hájek, J. (1971). Discussion of An essay on the logical foundations of survey sampling, part one, by D. Basu. *In* V.P. Godambe and D.A. Sprott (Eds), *Foundations of Statistical Inference*, p. 236. Toronto: Holt, Rinehart, Winston.

Hájek, J. (1981). *Sampling From a Finite Population.* New York: Marcel Dekker.

Kalton, G. and Anderson, D.W. (1986). "Sampling Rare Populations," *Journal of the Royal Statistical Society, Ser. A* **149** 65-82.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In M. Kochen, ed. The Small World, Norwood, NJ: Ablex Publishing, 176-210.

Kremers, W.K. (1987). Adaptive sampling to account for unknown variability among strata. Preprint No. 128. Institut für Mathematik, Universitat Augsburg, Germany.

Levy, P.S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association* **72**, 758-763.

Levy, P.S. and Lemeshow, S. (1991). *Sampling of Populations; Methods and Applications.* New York: Wiley.

Morgan, D.L., and Rytina, S. (1977). Comment on "Network sampling: some first steps" by Mark Granovetter. *American Journal of Sociology* **83** 722-727.

Mosquin, P. (1998). Frame-free adaptive sampling designs. Masters research paper. Department of Statistics, Pennsylvania State University.

Munholland, P.L., and Borkowski, J.J. (1993). Adaptive Latin square sampling + 1 designs. Technical Report No. 3-23-93, Department of Mathematical Sciences, Montana State University, Bozeman. 722-727.

Neaigus, A., Friedman, S.R., Goldstein, M.F., Ildefonso, G., Curtis, R., and Jose, B.(1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. In Needle, R.H., Genser, S.G., and Trotter, R.T. II, eds., *Social Networks, Drug Abuse, and HIV Transmission.* NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 20-37.

Neaigus, A., Friedman, S.R., Jose, B., Goldstein, M.F., Curtis, R., Ildefonso, G., and Des Jarlais, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **11** 499-509.

Roesch, F.A. Jr. (1993). Adaptive cluster sampling for forest inventories. *Forest Science* **39**, 655–669.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klovdahl, A.S. 1995. Social networks in disease transmission:

The Colorado Springs study. In Needle, R.H., Genser, S.G., and Trotter, R.T. II, eds., *Social Networks, Drug Abuse, and HIV Transmission.* NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 3-19.

Salehi M, M. (1988). Adaptive Cluster Sampling Designs. Ph.D. Thesis, University of Auckland, New Zealand.

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

Salehi M, M. and Seber G.A.F. (1997b). Two-stage adaptive cluster sampling. *Biometrics* **53** 959-970.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association* **63**, 257-266.

Sirken, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association* **67** 224-227.

Sirken, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics* **28** 869-873.

Sirken, M.G. and Levy, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association* **69** 68-73.

Smith, D.R., Conroy, M.J., and Brakhage, D.H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics* **51** 777-788.

Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin de Methodologie Sociologique* **36** 59-70.

Solomon H., and Zacks, S. (1970). Optimal design of sampling from finite populations: A critical review and indication of new research areas. *Journal of the American Statistical Association* **65**, 653–677.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs; what and why? *Bulletin de Methodologie Sociologique* **36** 34-58.

Spreen, M., and Zwaagstra, R. (1994). Personal network sampling, outdegree analysis and multilevel analysis: introducing the network concept in studies of hidden populations. *International Sociology* **9** 475-491.

Sudman, S., Sirken, M.G., and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science* **240** 991-996.

Thompson, M.E. (1997). *Theory of Sample Surveys.* London: Chapman and Hall.

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

Thompson, S.K. (1991a). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics* **47**, 1103-1115.

Thompson, S.K. (1991b). Stratified adaptive cluster sampling. *Biometrika* **78**, 389-397.

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

Thompson, S.K. (1993). Multivariate aspects of adaptive cluster sampling. *In* G.P. Patil and C.R. Rao (Eds), *Multivariate Environmental Statistics*, pp.561–572. New York: North Holland/Elsevier Science Publishers.

Thompson, S.K. (1996). Adaptive cluster sampling based on order statistics. *Environmetrics* **7** 123-133.

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.

Thompson, S., and Frank, O. (1998). Model-based estimation with link-tracing sampling designs. Technical Report 98-01, Department of Statistics, Pennsylvania State University.

Thompson, S.K., Ramsey, F.L., and Seber, G.A.F. (1992). An adaptive procedure for sampling animal populations. Biometrics, 48, 1195–1199.

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

van Meter, K. M. (1990). Methodological and design issues: techniques for assessing the representatives of snowball samples. In Lambert, E.Y. (1990), ed., *The Collection and Interpretation of Data from Hidden Populations.* NIDA Monograph 98. Rockville, MD: National Institute on Drug Abuse, 31-43.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* New York: Cambridge University Press.

Watters, J.K., and Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems* **36** 416-430.