Adaptive Web Sampling

To appear in *Biometrics, 2006*

Steven K. Thompson

Department of Statistics and Actuarial Science

Simon Fraser University, Burnaby, BC V5A 1S6 CANADA

email: thompson@stat.sfu.ca

Abstract

A flexible class of adaptive sampling designs is introduced for sampling in network and spatial settings. In the designs, selections are made sequentially with a mixture distribution based on an active set that changes as the sampling progresses, using network or spatial relationships as well as sample values. The new designs have certain advantages compared with previously existing adaptive and link-tracing designs, including control over sample sizes and of the proportion of effort allocated to adaptive selections. Efficient inference involves averaging over sample paths consistent with the minimal sufficient statistic. A Markov chain resampling method makes the inference computationally feasible. The designs are evaluated in network and spatial settings using two empirical populations: A hidden human population at high risk for HIV/AIDS and an unevenly distributed bird population.

Key words: Adaptive sampling, link-tracing designs, Markov chain Monte Carlo, network sampling, Rao-Blackwell, spatial sampling.

1 Introduction

This paper introduces a flexible class of adaptive designs for sampling populations in network and spatial settings. In a network setting, the designs are applied to sampling from a hidden human population at risk for HIV/AIDS. Once an initial sample of people is obtained, social links are followed to add more members of the hidden population to the sample. The decision whether to follow a link from a specific person may depend in part on the assessed risk-related behavior of that person. In a spatial setting, the designs are applied to a survey of migratory waterfowl. The spatial distribution of the birds is highly uneven. Sample plots are observed from aircraft. Whenever high abundance is observed in a sample plot, adjacent plots may be added to the sample.

When sampling in a network, at any point in the sampling there are a certain number of links out from the current sample. One of these can be selected, either at random or as a function of link weight, and followed to add the next unit to the sample. More generally, the next unit or set of units can be selected with a mixture distribution based on an active set that changes as the sampling progresses. The network concept can be applied in spatial and other structured settings as well, to produce adaptive designs more flexible than those hitherto available. The designs are termed *adaptive web sampling* to reflect their ability to reach weblike into interesting areas of the target population.

Because the adaptive web designs direct sampling effort disproportionately into high-valued or otherwise interesting areas of the study population, the samples are not at face value representative of the larger population. Unbiased or consistent estimation therefore involves taking into account initial and conditional selection probabilities under the design. The efficient design-based inference methods described in this paper are computationally intensive, since their direct calculation involves consideration of every possible sample path consistent with the minimal sufficient statistic, with different selection probabilities and estimation values

to be computed for each path. The computations are made feasible, however, with the Markov chain resampling method presented.

2 Sampling setup

The population of interest consists of a set of units labeled $1, 2, \ldots, N$. With each unit i is associated one or more variables of interest, denoted y_i. For a population in the graph or network setting, some additional structure is required. In addition to the set of N units or nodes and variables of interest y_i associated with the ith node, there are variables of interest w_{ij} associated with any pair of nodes i and j and describing relationships between i and j. In many situations w_{ij} is an indicator variable, with $w_{ij} = 1$ if there is a link from node i to node j, and $w_{ij} = 0$ otherwise. More generally, the variable w_{ij} designates a weight on the relationship between i and j. The link variables w_{ij} determine the graph structure of the population and, like the node variables y_i, are variables of interest that are observed only through the sample.

A sample s is a subset or subsequence of units from the population, and in the graph setting the sample consists of both a sample $s^{(1)}$ of nodes, on which the node variables y_i are observed, and a sample $s^{(2)}$ of pairs of nodes, for which the value of the link variables w_{ij} are observed. A design is considered adaptive if it depends on any of the variables of interest in the sample, whether those associated with individual units or those associated with pairs of units.

The original data d_o in sampling consist of the sequence of labels of the sample units, in the order selected, together with their associated y values. In the case of with-replacement designs, the sequence may include repeated selections of some units. The minimal sufficient statistic on the other hand consists only of the set of labels of distinct units selected, together with the associated y values (Godambe 1955, Basu 1969, and for adaptive designs, Thompson

and Seber 1996). In the graph setting, the minimal sufficient statistic consists of the reduced data $d_r = \{(i, y_i), ((j,k), w_{jk}); i \in s^{(1)}, (j,k) \in s^{(2)}\}$, that is, the set of distinct nodes and pairs of nodes in the sample, together with the associated node and relationship variables. Suppose for example that node i but not node j is in $s^{(1)}$, so that y_i is known and y_j is unknown; it could still be the case that it is known whether or not the relationship indicated by w exists from i to j, so that the ordered pair (i, j) would be in $s^{(2)}$. A variable of interest such as the out-degree $w_{i+} = \sum_j w_{ij}$ of unit i, although it depends on the link variables, is considered a node variable of interest associated with unit i.

3 Designs

An adaptive web sample is selected in steps. First, an initial sample s_0 is selected by some design p_0. At the kth step after the initial sample, selection of the next part of the sample s_k depends on values associated with a current *active set* a_k, that is, a subset or subsequence of the sample so far selected, together with any associated variables of interest. Thus, for $k \geq 1$, the selection distribution at step k is $p_k(s_k \mid a_k, y_{a_k}, w_{a_k})$, where a_k is a subset or subsequence of the current sample.

One way to implement such a design is to select the next unit from a mixture distribution, so that with probability d the next unit is selected adaptively using a distribution based on the unit values or graph structure of the active set and with probability $1 - d$ it is selected conventionally using a distribution based on the sampling frame or spatial structure of the population. For example, with probability d one of the links from the active set is selected at random and the unit to which it connects is added to the sample, while with probability $1 - d$ a new unit is selected completely at random from the population or from those units not already in the sample. The probability d may itself depend on the values in the active set.

The adaptive selections can be made unit-by-unit or in waves. Selection can be said to occur in waves if the active set remains constant for several unit selections in a row, so that a whole group of selections is based on a given active set. Snowball-type designs, for example, typically occur in waves, with a whole set of links selected from the previous wave of units or from all the units selected so far.

Designs in the present class have more flexibility than random walk designs by not being confined to only one unit at a time in the active set. They are more flexible than ordinary network, snowball, and adaptive cluster sampling designs by not requiring every link to be followed from a particular wave, nor do connected components intersected by the sample need to be sampled completely. This flexibility can be used to seek a balance between going deep into the population following links for many waves or going wide with only one or a few waves. Flexibility also comes from the conventional selection part of the mixture distribution, which balances spreading the sample out with placing it adaptively in the promising areas. Flexibility also comes from allocating part of the effort to the initial sample, thus controlling how much goes into adaptive effort. The adaptive selection distribution can depend on link values, based for example on node values from which the links originate or distance to the connected units.

Let the current sample at step k be $s_{ck} = \cup_{i=0}^{k-1} s_i$. Let a_k be the current active set at step k. Note that $a_k \subseteq s_{ck}$. Let n_{ak} be the number of units in the current active set a_k, and let n_{ck} be the number of units in the current sample. The next set of units s_k is selected with design probability $\mathrm{q}(s_k \,|\, a_k, y_{a_k}, w_{a_k})$ depending on the current active set a_k and its values. For designs in waves, let s_{ckt} denote the current sample at the time for the tth unit selection in the kth wave, so that s_{ckt} consists of the entire sample that has been selected prior to that time. For designs in which each wave consists of a single unit, the subscript t may be omitted, as will be done for simplicity in the estimation section following this section.

With a nonreplacement design, the next unit or set of units is selected from the collection of units not yet selected. In that way the selection of each unit in the sample depends on every unit selected so far. However, the selection depends on variables of interest y or w only through those units that are in the current active set.

At the time of the tth unit selection in the kth wave, let $w_{a_{kt}+}$ be the total number of links out, or the total of the weight values, from the active set a_k to units not in the current sample s_{ckt}. That is, $w_{a_{kt}+} = \sum_{\{i \in a_k, j \in \bar{s}_{ckt}\}} w_{ij}$. When w is an indicator variable, $w_{a_{kt}+}$ is the total of the net out-degrees of the individual units in the active set a_k, where net out-degree is the out-degree of a unit minus the number of its links to other units already in the current sample.

For each unit i in the sample, the variable of interest y_i and the out-degree (or out-weight) w_{i+} are recorded. In addition, for each pair of units (i, j) for which both i and j are in the sample, the values of the link variables w_{ij} and w_{ji} are observed.

Consider as a candidate for the tth selection in the kth wave a unit i not in the current sample, so $i \notin s_{ckt}$. Suppose the current active set a_k contains one or more units having links or positive weights out to unit i, and let $w_{a_k i} = \sum_{j \in a_k} w_{ij}$ denote their total. The probability that unit i is the next unit selected is

$$q_{kti} = d\frac{w_{a_k i}}{w_{a_{kt}+}} + (1-d)\frac{1}{(N - n_{s_{ckt}})} \tag{1}$$

where d is between 0 and 1. If there are no links at all out from the current active set, then

$$q_{kti} = \frac{1}{(N - n_{ckt})}$$

Thus, with probability d, one of the links out from the current active set is selected at random, or with probability proportional to its weight, and the node to which it leads is added to the sample, while with probability $1 - d$ the new sample unit is selected completely at random from the units not already selected. However, if there are no links or positive

weights out from the active set to any unsampled units, then the next unit is selected from the collection of unsampled units.

Denoting the tth unit selected in the kth wave as i_{kt}, the kth-wave sample in the order selected is $s_k = (i_{k1}, \ldots, i_{kn_k})$, where n_k is the size of the kth wave. The overall sample selection probability for the ordered sample $\mathbf{s}$ is

$$p(\mathbf{s}) = p_0 \prod_{k=1}^{K} \prod_{t=1}^{n_k} q_{kti} \tag{2}$$

where p_0 is the selection probability for the initial sample and K is the number of waves. Calculation of these sample selection probabilities is required for the design based estimators described in the next section.

If the relationship variable w consists of weights, instead of having just 0 or 1 values, then the link-based selection can depend on these weights. For example, link weights can be defined in relation to the y value of an originating node or as a distance measure to the connected node, so that links are followed with higher probability from nodes with higher values or with lower probability to distant nodes. Then a link from the active set can be selected with probability proportional to link weight, or with some other selection probability $\mathrm{p}(i \,|\, s_{ckt}, a_k, y_{a_k}, w_{a_k})$ depending on variables of interest only through the active set. For example, a link out could be selected at random from the links with w_{ij} greater than some constant, or y_i greater than some constant. The selection probability when links are not followed does not have to be uniform over the units not in the current sample, but can be a more general design $\mathrm{p}(i \,|\, s_{ckt})$ such as selecting with probability related to an auxiliary variable or from a spatially defined distribution.

In the more general context with w representing a possibly continuous link weight variable, the probability that unit i is the next unit selected is

$$q_{kti} = d\,\mathrm{p}(i \,|\, s_{ckt}, a_k, y_{a_k}, w_{a_k}) + (1-d)\,\mathrm{p}(i \,|\, s_{ckt})$$

If there are no links or positive weights from a_k to i, then

$$q_{kti} = \mathrm{p}(i \mid s_{ckt}).$$

Once unit i has been selected, it is possible to add an accept/reject step for deciding whether to include it in the active set, for example, accepting with higher probability if unit i has a high value or high degree.

The constant d itself can also be replaced by a probability $d(\, k, t, a_k, y_{a_k}, w_{a_k})$ depending on values related to nodes and links in the active set or changing as sample selection progresses. For example, if the values of the units in a_k are particularly high, we could increase the probability of following links. As for dependence of d on (k, t), the use of an initial conventional sample of size $n_0 > 1$ may be viewed as serving to obtain some information from basic coverage of the population before adaptive sampling is allowed to commence. In principle, one could instead increase d continuously as sampling progresses in order to increase the probability of adaptive sampling as information about the population increases.

The full design may consist of a single adaptive web sample as described above, or of m independently selected samples.

Each of the adaptive web designs described in this paper can also be carried out with replacement. In that case, at each point in the sampling, with probability d a link is selected at random or from some distribution depending on node and link values without regard to whether the unit it leads to has already been selected or not. With probability $1 - d$ the next unit is selected at random or by another distribution, not depending on node or link values, from the entire set of N units in the population.

For a with-replacement design let $w_{a_k+} = \sum_{\{i \in a_k, j=1,\ldots,N\}} w_{ij}$, so that the relevant total of links from the active set includes even links to units in the current sample. With random or weighted selection of links with probability d and random selection from the population at large with probability $1 - d$, the probability that unit i is the next selection, regardless of

whether i has already been selected or not, in the simple with-replacement case is

$$q_{kti} = d\frac{w_{a_k i}}{w_{a_k +}} + (1-d)\frac{1}{N}$$

if there are links or positive weights from the active set to i and

$$q_{kti} = \frac{1}{N}$$

if there are not.

4 Estimation

For the without-replacement designs, several types of design-unbiased estimators of the population mean are described, with modifications for with-replacement designs given later. For any design let $\mathbf{s}$ denote the sample in the original order selected, and let s denote the set of distinct units in the sample. Let r be the reduction function that reduces any sequence $\mathbf{s}$ to the set s of its distinct elements, so that $r(\mathbf{s}) = s$. For a without-replacement design, the only reduction that takes place is the elimination of order-of-selection information, while for a with-replacement design information about numbers of times a unit is selected is eliminated as well. For the estimators in this paper the values of link variables between units in the sample are assumed known, as is the out-degree of each sample unit. The minimal sufficient statistic can then be written $d_r = \{(i, y_i, w_{i+}, w_{ij}); i \in s, j \in s\}$.

4.1 Estimator based on initial sample mean

Suppose a single initial sample unit is selected with probability π_0 and has value y_0. Then $\hat{\mu}_{01} = (1/N)y_0/\pi_0$ is an unbiased estimator of the population mean. More generally, with initial sample size $n_0 \geq 1$, let $\hat{\mu}_{01}$ be an unbiased estimator of the population mean based on the initial sample design, such as the Horvitz-Thompson estimator $\hat{\mu}_{01} = (1/N)\sum_{i \in s_0} y_i/\pi_i$

or, for an initial simple random sample, $\hat{\mu}_{01} = \bar{y}_0$. (Note that the notation p_0 of the previous section refers potentially to the selection of a whole set of initial units, while π_i is used exclusively for the inclusion probability of a single unit.)

A conceptually simple but computationally intense estimator of the population mean is obtained via the Rao-Blackwell approach, finding the conditional expectation of the preliminary estimator given the minimal sufficient statistic. The relevant conditional distribution is given by $p(\mathbf{s} \mid d_r) = p(\mathbf{s})/[\sum_{\{\mathbf{s}:r(\mathbf{s})=s\}} p(\mathbf{s})]$. The improved estimator is

$$\hat{\mu}_1 = \mathrm{E}(\hat{\mu}_{01}|d_r) = \sum_{\{\mathbf{s}:r(\mathbf{s})=s\}} \hat{\mu}_{01}(\mathbf{s}) p(\mathbf{s} \mid d_r)$$

The improved estimator $\hat{\mu}_1$ is the expected value of the initial estimator over all $n!$ reorderings of the sample data. For some designs, for example those with an initial random sample of size n_0, the preliminary estimator has the same value over reordering of the initial n_0 units, so that the expected value is over $\binom{n}{n_0}$ combinations and $(n - n_0)!$ reorderings. In calculating the expectation, each of the reorderings is weighted by the selection probability (2). This estimator is unbiased for $0 \leq d \leq 1$.

4.2 Estimator based on conditional selection probabilities

Another unbiased estimator can be obtained by dividing observed values by conditional selection probabilities depending on the step by step active sets. Des Raj (1956) and Murthy (1957) considered designs in which the selection probability depended on the set, though not the order, of the units already selected. For some of the active set designs, such as those in which the active set consists of the most recently selected units, these probabilities can depend on order of selection as well as the set of units already selected. In addition, the conditional selection probabilities may depend on the y and w values associated with the active sets. Further, since the initial sample may be selected by a different procedure, a composite estimator is needed.

With an initial sample of n_0 units, let $\hat{\tau}_{s_0}$ be an unbiased estimator of the population total $\tau = \sum_{i=1}^{N} y_i$ based on the initial sample s_0. Thus, if the initial sample is selected by simple random sampling, without replacement, $\hat{\tau}_{s_0} = (N/n)\sum_{i\in s_0} y_i = N\bar{y}_0$, while with an initial unequal probability design, $\hat{\tau}_{s_0} = \sum_{i\in s_0} y_i/\pi_i$ could be used. For selections after the initial sample, conditional selection probabilities are utilized. Thus for the ith selection, having value y_i and at which point the current sample is $s_{ck} = (s_0, \ldots, s_{k-1})$, let $z_i = \sum_{j\in s_{ck}} y_j + y_i/q_{ki}$. Note that each z_i is an unbiased estimator of τ. An unbiased estimator of the population mean is

$$\hat{\mu}_{02} = \frac{1}{Nn}\left[n_0\hat{\tau}_{s_0} + \sum_{i=n_0+1}^{n} z_i\right]$$

This estimator is a weighted average of the initial sample estimator $\hat{\tau}_{s_0}$ and the average of the $n - n_0$ subsequent conditional estimators, z_i.

This estimator is unbiased for $0 \leq d < 1$. If $d = 1$, then, unless the population graph is complete, this estimator is not unbiased because not all of the conditional selection probabilities will be strictly greater than zero at each step for every unit not in the current sample.

Since the initial estimator depends on order of selection, applying the Rao-Blackwell method produces the second improved unbiased estimator

$$\hat{\mu}_2 = \mathrm{E}(\hat{\mu}_{02}|d_r) = \sum_{\{\mathbf{s}:r(\mathbf{s})=s\}} \hat{\mu}_{02}(\mathbf{s})p(\mathbf{s}\,|\,d_r)$$

The expectation is again carried out over every sample path of nodes consistent with the sufficient data and using the sequence selection probabilities under the design.

4.3 Composite conditional generalized ratio estimator

Let $\hat{N}_0$ be an estimator of the population size N based on the initial sampling design used to select the first n_0 units. For example, if the initial design is an unequal probability one

and the initial estimator $\hat{\tau}_0$ is a Horvitz-Thompson estimator $\hat{\tau}_0 = \sum_{k \in s_0} y_k/\pi_k$, then $\hat{N}_0 = \sum_{k \in s_0}(1/\pi_k)$.

For $i = n_0 + 1, \ldots, n$, define $\hat{N}_i = \#s_{ck} + 1/q_{ki}$, where $\#s_{ck}$ is the size of the current sample. Note that just as z_i is an unbiased estimator of τ, $\hat{N}_i$ is an unbiased estimator of N. A composite estimator of N combining the initial and subsequent samples is $\hat{N} = (1/n)[n_0\hat{N}_0 + \sum_{i=n_0+1}^{n} \hat{N}_i]$. A generalized ratio estimator is then formed as the ratio of the two conditional probability-based estimators

$$\hat{\mu}_{03} = \frac{N\hat{\mu}_{02}}{\hat{N}}$$

Although the estimator $\hat{\mu}_{03}$ is not precisely unbiased, conditioning on the minimal sufficient statistic produces an improved estimator having the same expected value (hence the same bias) with mean square error as small or smaller than the preliminary estimator:

$$\hat{\mu}_3 = \mathrm{E}(\hat{\mu}_{03}|d_r) = \sum_{\{\mathbf{s}:r(\mathbf{s})=s\}} \hat{\mu}_{03}(\mathbf{s})p(\mathbf{s} \mid d_r)$$

4.4 Composite conditional mean-of-ratios estimator

An alternate way to use the ratios of unbiased estimators in a composite estimator is

$$\hat{\mu}_{04} = \frac{1}{n}\left[n_0\frac{\hat{\tau}_{s_0}}{\hat{N}_0} + \sum_{i=n_0+1}^{n}\frac{z_i}{\hat{N}_i}\right]$$

The improved version of this estimator is

$$\hat{\mu}_4 = \mathrm{E}(\hat{\mu}_{04}|d_r) = \sum_{\{\mathbf{s}:r(\mathbf{s})=s\}} \hat{\mu}_{04}(\mathbf{s})p(\mathbf{s} \mid d_r)$$

None of the estimators $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, or $\hat{\mu}_4$ gives uniformly lower mean square errors than the others, since the minimal sufficient statistic in finite population design-based sampling is not complete. The estimator $\hat{\mu}_2$ though unbiased can produce large values with certain samples having small conditional selection probabilities. With a population in which the y values of

the units are 0-1 valued so that the population mean is a proportion, the estimator $\hat{\mu}_2$, even though unbiased, can produce estimates of the population proportion greater than one, that is, outside the range of possible values. The alternative estimators $\hat{\mu}_3$ and $\hat{\mu}_4$, though not unbiased, do produce estimates strictly between 0 and 1 for such populations.

4.5 Variance estimators

For an estimator $\hat{\mu} = \mathrm{E}(\hat{\mu}_0 \,|\, d_r)$, the conditional decomposition of variances gives $\mathrm{v}ar(\hat{\mu}) = \mathrm{v}ar(\hat{\mu}_0) - \mathrm{E}[\mathrm{v}ar(\hat{\mu}_0 \,|\, d_r)]$. An unbiased estimator of this variance is $\widehat{\mathrm{v}ar}(\hat{\mu}) = \mathrm{E}[\widehat{\mathrm{v}ar}(\hat{\mu}_0) \,|\, d_r] - \mathrm{v}ar(\hat{\mu}_0 \,|\, d_r)$. The first term is the expectation of the initial sample variances over all reorderings of the data, while the second term is the variance of the initial estimator over the reorderings.

For the first estimator, with the initial sample a random sample without replacement of n_0 units, then the variance estimator of the initial sample mean is $\widehat{\mathrm{v}ar}(\hat{\mu}_{01}) = (N - n_0)v_0/(Nn_0)$, where v_0 is the sample variance of the initial sample.

For the second estimator, with $n_0 = 1$ and initial estimator $\hat{\mu}_{02} = \sum_{i=1}^{n} z_i/(Nn)$, the variance estimator $\widehat{\mathrm{v}ar}(\hat{\mu}_{02}) = \sum_{i=1}^{n}(z_i - N\hat{\mu}_0)^2/(n(n-1)N^2)$ can be shown to be unbiased for $\mathrm{v}ar(\hat{\mu}_{02})$, following Des Raj (1956) and Murthy (1957), the useful fact being the z_is are uncorrelated though not independent. Thus $\widehat{\mathrm{v}ar}(\hat{\mu}_2) = \mathrm{E}[\widehat{\mathrm{v}ar}(\hat{\mu}_{02}) \,|\, d_r] - \mathrm{v}ar(\hat{\mu}_{02} \,|\, d_r)$ is an unbiased estimator of the variance of $\hat{\mu}_2$.

With the second estimator and an initial sample with $n_0 > 1$, let v_1 be an unbiased estimator of $\mathrm{v}ar(\hat{\mu}_{s_0})$, where $\hat{\mu}_{s_0} = \hat{\tau}_{s_0}/N$, and let $v_2 = \sum_{i=n_0+1}^{n}(z_i - \bar{z}_2)^2/[(n-n_0)(n-n_0-1)N^2]$, where $\bar{z}_2 = \sum_{i=n_0+1}^{n} z_i/(n-n_0)$. When s_0 is a simple random sample with $n_0 > 1$ then $v_1 = (N-n_0)v_0/(Nn_0)$. An unbiased estimator of the variance of $\hat{\mu}_{02}$ is the composite

$$\widehat{\mathrm{v}ar}(\hat{\mu}_{02}) = \left(\frac{n_0}{n}\right)^2 v_1 + \left(\frac{n-n_0}{n}\right)^2 v_2$$

and an unbiased estimator of the variance of $\mathrm{v}ar(\mu_2)$ is the difference $\widehat{\mathrm{v}ar}(\hat{\mu}_2) = \mathrm{E}[\widehat{\mathrm{v}ar}(\hat{\mu}_{02}) \,|\, d_r] -$

$\mathrm{v}ar(\hat{\mu}_{02} \mid d_r)$.

Variance estimators of this type, although unbiased, have the disadvantage that they can give negative estimates with some samples. An alternative approach that avoids this difficulty, and is therefore recommended, is to select more than one adaptive web sample independently. Let m be the number of independent adaptive web samples, and let $\hat{\mu}_k$ be the estimate of a population quantity with the kth adaptive web sample. Then an unbiased estimate of μ is $\hat{\mu} = \sum_{k=1}^{m} \hat{\mu}_k/m$ and an unbiased estimate of the variance of $\hat{\mu}$ is

$$\widehat{\mathrm{v}ar}(\hat{\mu}) = \sum_{i=1}^{m} (\hat{\mu}_k - \hat{\mu})^2/[m(m-1)]$$

This procedure provides unbiased, invariably positive estimates of variance for any of the estimators.

4.6 Estimators for with-replacement designs

With a with-replacement adaptive web design, the initial sample still provides an unbiased estimator of the population mean or total, and the form of $\hat{\mu}_{01}$ would remain the same. For the estimator $\hat{\mu}_{02}$ the component variables become $z_k = \sum_{i \in s_{ck}} y_k/q_{ki}$. For $\hat{\mu}_{03}$ and $\hat{\mu}_{04}$, the unbiased estimators of N are $\hat{N}_k = 1/q_{ki}$.

For the Rao-Blackwell improvements of each of these estimators, the collection of samples consistent with the minimal sufficient statistic include not only reorderings of the original data, but also recombinations in which repeat selections of different units still give the same set of distinct units.

5 Markov chain resampling estimators

Computation of the estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ and their variance estimators under various adaptive web designs involves in general tabulating the reorderings of the sample selection sequence.

For each reordering, the probability of that ordering under the design is computed, along with the values of the estimators and variance estimators. Direct calculation is very efficient up to sample sizes of ten or so, involving no more than a few million permutations to be enumerated. For larger sample sizes, the numbers of permutations or combinations of potential selection sequences in the conditional sample space become prohibitively large for the exact, enumerative calculation. For this reason, a resampling approach is utilized as a general method for obtaining estimates with designs of these types.

Each of the estimators involves the mean of a function over a conditional distribution. Let x represent a point in the conditional sample space. In this context, x is typically a permutation of the n units selected for the sample and the sample space consists of all possible permutations. The estimator $\hat{\mu} = \mathrm{E}(\hat{\mu}_0 \,|\, d_r)$ can be written $\hat{\mu} = \sum_x \hat{\mu}(x) p(x \,|\, d)$ where x is a point in the sample space, $p(x \,|\, d_r)$ is the probability of selecting x with the given sampling design conditional on the realized value d_r of the minimal sufficient statistic, and the sum is over all points in the sample space. One way to obtain a sample s_r of permutations x from the conditional distribution $p(x \,|\, d_r)$ is through a Markov chain accept/reject procedure (Hastings, 1970).

The resampling procedure used to obtain estimators in the examples in this paper when sample sizes precluded enumerative calculation is as follows. The object is to obtain a Markov chain $x_0, x_1, x_2, \ldots$ having stationary distribution $p(x \,|\, d)$. Suppose that at step $k-1$ the value is $x_{k-1} = j$, so that j denotes the current permutation of the sample data in the chain. A tentative permutation t_k is produced by applying the original sampling design, with sample size n, to the data as if the sample comprised the whole population, that is, as if $N = n$. This resampling distribution, denoted p_t differs from, but has some similarity to, the actual sampling design p. The desired conditional distribution $p(x \,|\, d_r)$ is proportional to the unconditional distribution $p(x)$ under the original design applied to the whole population.

Let $\alpha = \min\{[p(t_k)/p(x_{k-1})][p_t(x_{k-1})/p_t(t_k)], 1\}$. With probability α, t_k is accepted and $x_k = t_k$, while with probability $1 - \alpha$, t_k is rejected and $x_k = x_{k-1}$.

This procedure produces a Markov chain $x_0, x_1, x_2, \ldots$ having the desired stationary distribution $p(x \mid d_r)$. The chain is started with the original sample $\mathbf{s}$ in the order actually selected. Given any value of the minimal sufficient statistic d_r, the chain is thus started in its stationary distribution and so remains in its stationary distribution step by step.

Suppose that n_r resampled permutations are selected by this process and let $\hat{\mu}_{0j}$ denote the value of the initial estimator for the jth permutation. An enumerative estimator of the form $\hat{\mu} = \mathrm{E}(\hat{\mu}_0)$ is replaced by the resampling estimator

$$\tilde{\mu} = \frac{1}{n_r} \sum_{j=0}^{n_r - 1} \hat{\mu}_{0j}$$

Similarly,

$$\tilde{\mathrm{E}}[\widehat{\mathrm{v}ar}(\hat{\mu}_0) \mid d] = \frac{1}{n_r} \sum_{j=0}^{n_r - 1} \widehat{\mathrm{v}ar}(\hat{\mu}_{0j})$$

and

$$\widetilde{\mathrm{v}ar}(\hat{\mu}_0 \mid d) = \frac{1}{n_r} \sum_{j=0}^{n_r - 1} (\hat{\mu}_{0j} - \tilde{\mu})^2$$

To estimate the additional variance $\mathrm{v}ar(\tilde{\mu} \mid d)$, due to resampling, one approach is to divide the Markov chain resampling data into L groups of length K each and use the sample variance of the block means as suggested by Hastings (1970):

$$s_{\bar{y}}^2 = \sum_{i=1}^{L} (\bar{y}_i - \hat{\mu})^2 / [L(L-1)]$$

For the examples motivating this work the expense of resampling, which involves only computation, is small compared to the expense of actual sampling, the recommended approach is to use large resample sizes to make the additional resampling variance negligible.

6 Empirical examples

6.1 HIV/AIDS at-risk population

Figure 1 (top) shows drug-using relationships among at-risk individuals from the Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat et al. 1993, Rothenberg et al. 1995, Darrow et al. 1999). The nodes represent people in the study population, with a solid circle indicating an injection drug user. Links shown represent drug-using relationships between individuals. With hidden human populations of this type, it is generally much more difficult or expensive to select nodes at random or by a conventional probability design than to find new units by following links from units already in the sample.

Simulations of sampling from this population were carried out with a variety of adaptive web designs. The data set used has 595 people and is used as an empirical population from which to sample. Injection drug use was used as the variable of interest and the proportion of injection drug users was the population quantity to be estimated from each sample. Representative properties of the strategies with this population are summerized by the sampling distributions shown in Figure 2. In each case, 4 independent adaptive web samples were selected, each having an initial sample size of 10 and a final sample size of 20, for a total sample size of 80. The initial samples were selected by simple random sampling. The probability d of following links was 0.9. A sample of this type from the population is shown in the lower portion of Figure 1. In Figure 2 the selection of links was at random from those in the active set. Similar results were obtained with a slightly different design, in which links were followed with probability proportional to originating node value, which meant that links were only followed from nodes with $y = 1$, corresponding to high-risk individuals. The number of simulation runs for each design was 2000, and the number of resamples for each of the 2000 samples was 10,000. The estimators having lowest mean square error were $\hat{\mu}_4$ and $\hat{\mu}_1$ for both

designs, though $\hat{\mu}_4$ showed slightly increased bias with the design having weighted selection of links.

Degree distributions, giving the frequency of nodes having each degree value, are of considerable interest in social network analysis and epidemiology. The degree distribution for the Colorado Springs empirical distribution is given in the upper left of Figure 3. The same distribution is shown with logarithmic scales in the upper right. The approximately linear section in the middle range of the scaled distribution indicated a power-law distribution could be used to approximate the degree distribution in that range. The lower left of the figure gives the degree distribution for adaptive web samples of size 20 from the population, with initial random samples of size 10 and random selection of links. The sample degree distribution has a lower frequency of nodes with small degree and a higher frequency of nodes with high degree. The skewness of the sample degree distribution compared to that of the population is especially evident with the rescaled distribution at the lower right.

The sampling distribution of the sample mean degree is shown on the lower left in Figure 3, based on 2000 samples selected with the adaptive web design, each sample consisting of $m = 4$ independent selections having initial sample sizes $n_0 = 10$ and final size $n = 20$, so that each sample contains 80 units in all. Whereas the mean degree in the population is 2.5, the expected value of the sample mean degree is 5.5. In contrast to the highly biased sample degree statistic, an unbiased estimate of the degree of the population network is provided by either of the estimators $\hat{\mu}_1$ or $\hat{\mu}_2$ with node degree as the variable of interest y_i. The sampling distribution of the unbiased estimator $\hat{\mu}_1$ applied to the degree data, with the same 2000 samples, is shown on the lower right in Figure 3. With the design selecting links with probability proportional to the degree of the originating node, the results were very similar and are not shown.

6.2 Wintering waterfowl population

Figure 4 depicts the spatial distribution of a population of blue-winged teal, a migratory waterfowl species, on a wildlife refuge (Smith, Conroy, and Brakhage 1995). The study area has been divided into 50 units or plots, and counts of birds are given for each plot in the population. The variable of interest for each unit is the number of birds in it and the population quantity to be estimated is the total number of birds in the empirical population. The directed graph structure of the population in relation to adaptive sampling is shown at the bottom of the figure.

In each simulation run, an initial simple random sample of n_0 units was selected, for different values of n_0. In subsequent selections links were selected at random with probability $d = 0.9$, while with probability $1 - d = 0.1$ a unit was selected at random from those not already in the sample. Total sample size was fixed at $n = 20$. The design variation in which links were selected with probability proportional to the y value of the originating node was also evaluated, with similar results. The number of simulation runs for each strategy was 2000, and the number of resamples for each of the 2000 samples was 10,000. The gains from improving the initial estimators with Markov chain resampling were substantial.

The issue of how much of the sample to allocate to the random initial selection and how much to the adaptive part is examined in Figure 5. Mean square errors in the figure are standardized by dividing by that of a simple random sample of 20 units, corresponding to the choice $n_0 = 20$. The pattern illustrates a trade-off between focusing on areas near to high encountered values and exploring the study region more widely. A very small random starting sample ($n_0 < 5$) gives little information overall, so that it would be better to explore more widely than to focus. Once a larger initial random sample has been obtained it is more valuable to focus effort in promising areas. Of the values tried, the lowest mean square errors were obtained in the vicinity of $n_0 = 13$ or $n_0 = 14$, depending on the estimator. Thus,

with this population, optimal efficiency was achieved when the initial sample size was 65 to 70 percent of the total sample size, so that by selecting 30-35 percent of the sample units adaptively, a 75 percent gain in efficiency is gained over conventional random sampling with the same sample size. Comparing with the tables in Smith, Brown, and Lo (1995), this design is also somewhat more efficient than the adaptive cluster design having a random sample size with expected final sample size approximately 20.

Different sample allocations with the blue-winged teal population are illustrated in Figure 6, showing two samples each having sample size $n = 20$ and random selections of links. The sample at the top of the figure has $n_0 = 13$ and the sample at the bottom has $n_0 = 1$. Visually, the first sample has better coverage of the study region overall, while still providing some focused exploration of the aggregation areas.

7 Discussion

The new designs have a number of advantages relative to previously available adaptive and link-tracing strategies, as well as providing efficiency gains in some situations over conventional sampling with the same sample size. In comparison with ordinary adaptive cluster sampling (Thompson 1990, Thompson and Seber 1996) and with some types of network or multiplicity sampling (Birnbaum and Sirken 1965), one advantage of the present designs is that no connected component is required to be sampled completely. Relative to some of the standard snowball designs in graphs (Frank 1977a,b, 1978a,b, 1979, Frank and Snijders 1994), an advantage of the present designs is that sample size can be fixed in advance, and depth versus width of penetration into the population can be adjusted. Similarly in adaptive cluster sampling much of the literature has been devoted to containing, even approximately, the random sample size (e.g., Salehi and Seber 1997, Brown and Manley 1998, Christman and Lan 2001, Su and Quinn 2003, and extensive review in Smith et al. 2004), whereas with the present

designs sample size can be fixed exactly. In addition, the allocation between conventional and adaptive sampling efforts can be adjusted as desired. Design-unbiased estimation in adaptive cluster sampling requires that a fixed condition for extra sampling in a neighborhood be fixed in advance, whereas in the present designs the probability of an adaptive selection can depend on a continuous function of sample unit or link values, giving for example higher probability to following a link from a high-valued unit than from a low-valued unit. In contrast to some of the standard methods for network and snowball designs, the strategies discussed in this paper are applicable in directed graph as well as undirected graph situations. Whereas with random walk designs in graphs (Klovdahl 1989, Lovász 1993) selection of the next unit can depend only on the most recently selected unit, the active set of the new designs can take many forms such as the most recent several selections, the whole current sample, or a set of units close in geographic or graph distance to the current selection.

In relation to optimal model-based sampling strategies (Zacks 1969, Chao and Thompson 1999), which can roughly be characterized as adaptively placing new units in proximity to high-valued or "interesting" observations while at the same time striving to spread them out to cover the study region, the proposed designs, while not optimal under any one model, approximate some of the characteristics of optimal strategies while being much simpler to implement and avoiding the dependence on model-based assumptions.

Further study of the choices of initial sample size n_0 and link-tracing probability d would be useful. In the last empirical example of this paper, the optimal choice allocated 65 to 70 percent of the total sample size to the initial sample. If d is set to one, so that there is no possibility of a random jump unless one is stuck with no network links to follow from the current active set, then $\hat{\mu}_1$ still provides an unbiased estimator while $\hat{\mu}_2$ is no longer unbiased. Choices of d less than one and of n_0 greater than one keep the sampling from getting stuck in a single large network component. As d approaches zero the design becomes more like a

conventional one. The choice of n_0 and d reflects the tension between the desire to select each unit from the most promising area in light of the current data and the desire to have wide, representative coverage of the population. More generally, the value of d could depend on the current sample, for example increasing from zero toward one as sample size increases.

8 Acknowledgements

Support for this work has been provided by the National Center for Health Statistics, National Science Foundation grants DMS-9626102 and DMS-0406229, and the Visiting Faculty Program of the Statistical Sciences Group of Los Alamos National Laboratory. The author would like to thank Myron Katzoff, John Potterat, Steve Muth, and Dave Smith for helpful suggestions and access to data.

9 References

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā A* **31** 441–454.

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., and Manley, B.F.J. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics* **5** 49-63.

Chao, T-C., and Thompson, S.K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics* **12** 517-538.

Chow, M. and Thompson, S.K. (2003). Estimation with link-tracing sampling designs—a Bayesian approach. *Survey Methodology* **29** 197-205.

Christman, M.C., and Lan, F. (2001). Inverse adaptive cluster sampling. *Biometrics* **57** 1096-1105.

Darrow, W.W., Potterat, J.J., Rothenberg, R.B., Woodhouse, D.E., Muth, S.Q., and Klovdahl, A.S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs Study. *Sociological Focus* **32** 143-158.

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

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. (1979). 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., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* **10** 53-67.

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

Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika* **57** 97-109.

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.

Lovász, L. (1993). Random walks on graphs: A survey. In Miklós, D., Sós, D., and Szöni, T., eds., *Combinatorics, Paul Erdös is Eighty,*, Vol. 2, pp. 1-46. János Bolyai Mathematical Society, Keszthely, Hungary.

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

Potterat, J. J., Woodhouse, D. E., Rothenberg, R. B., Muth, S. Q., Darrow, W. W., Muth, J. B. and Reynolds, J. U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS* **7** 1517-1521.

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

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.

Smith, D.R, Brown, J.A., and Lo, N.C.H. (2004) Application of Adaptive Cluster Sampling to Biological Populations. In W.L. Thompson, ed., *Sampling Rare and Elusive Species,* Island Press, Covelo, California and Washington, D.C., 75-122.

Su, Z., and Quinn, T.J.II. (2003). Estimator bias and efficiency for adaptive cluster sampling with order statistics and a stopping rule. *Environmental and Ecological Statistics* **10** 17-41.

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

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

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

Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations. *Journal of the American Statistical Association* **64** 1342–1349.

Figure 1: Top: HIV/AIDS at-risk population (Potterat et al. 1993). Dark node indicates injection drug use. Links indicate drug-using relationships. Largest component contains 300 of the 595 individuals. Bottom: Adaptive web sample of 80 nodes and corresponding sample network structure from the population above, with $n_0 = 10$, $n = 20$, $m = 4$, $d = 0.9$, random selection of links.

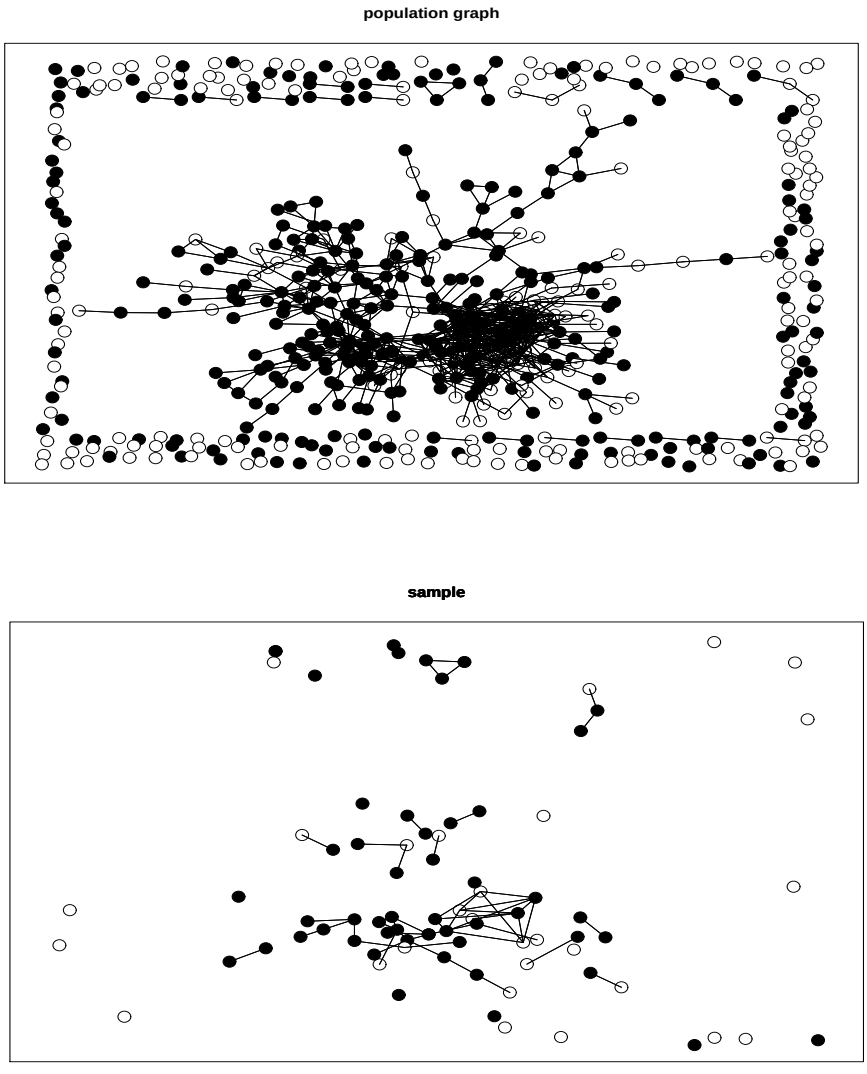

Figure 2: Sampling distributions of estimators of population mean node value for HIV/AIDS at-risk population. Adaptive web design with total sample size 80, $n0 = 10$, $n = 20$, $m = 4$, $d = 0.9$, random selection of links. Based on selection of 2000 samples, 10,000 resamples from each subsample. Population mean=0.5748. The four rows correspond to the four types of estimators presented. Histograms for the preliminary estimators are on the left, while those for the improved estimators are on the right.

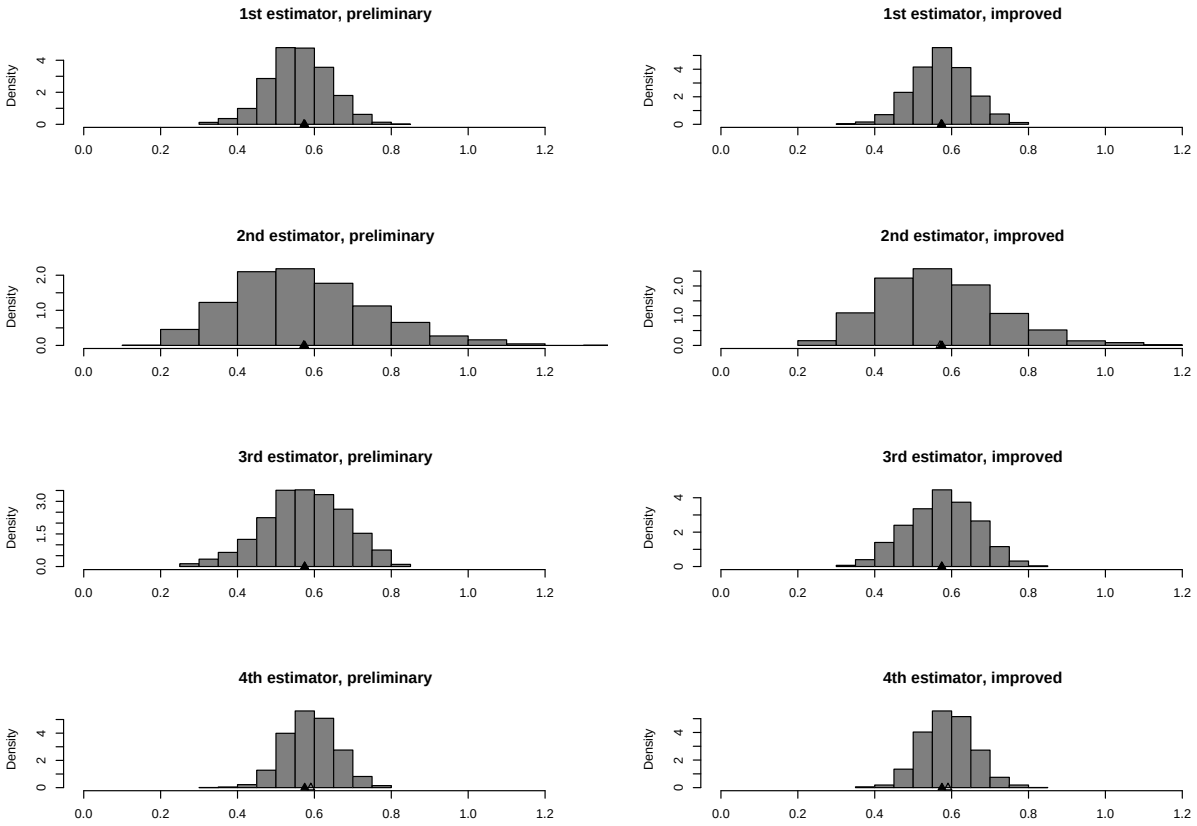

Figure 3: Population (top) and sample (middle) degree distributions for HIV/AIDS at-risk population, natural (left) and logarithmic (right) scales. Distribution of sample mean degree (bottom left) and of unbiased estimator $\hat{\mu}_1$ for mean degree (bottom right).

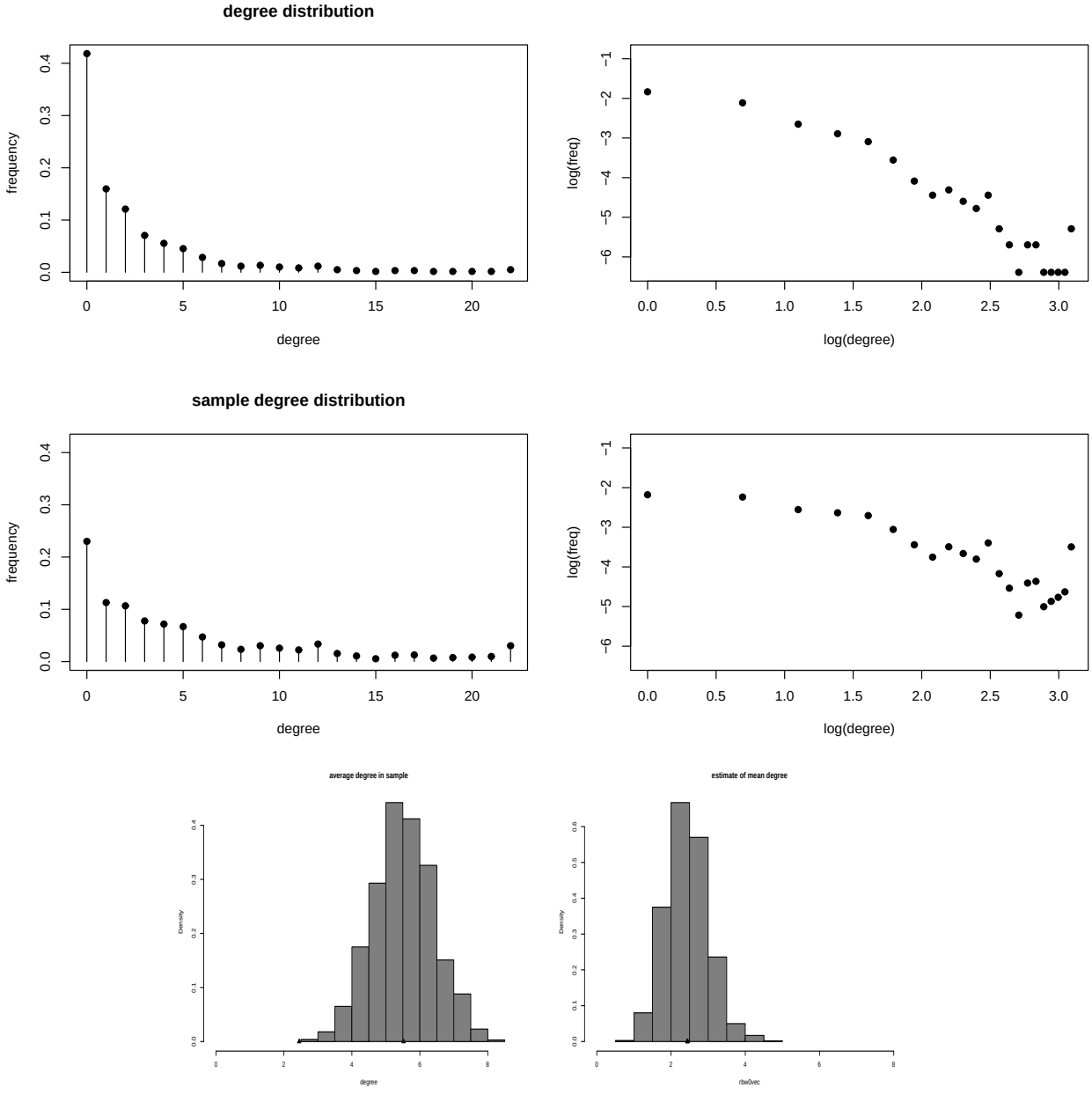

Figure 4: Blue-winged teal population spatial count data (Smith et al. 1995) and population graph structure.

spatial population

0	0	3	5	0	0	0	0	0	0
0	0	0	24	14	0	0	10	103	0
0	0	0	0	2	3	2	0	13639	1
0	0	0	0	0	0	0	0	14	122
0	0	0	0	0	0	2	0	0	177

population graph

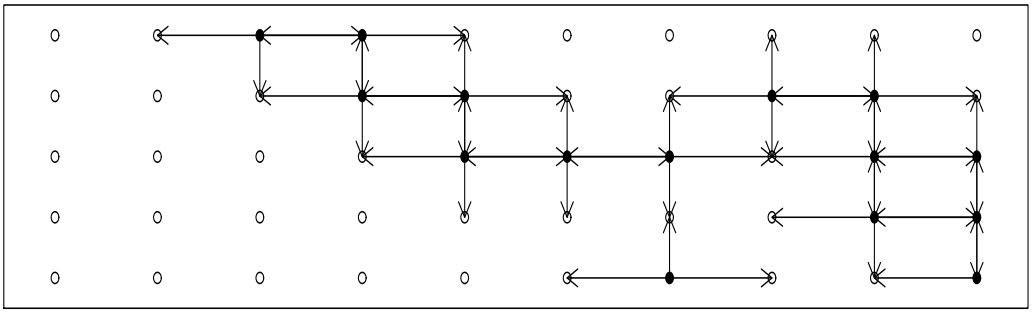

Figure 5: Mean square errors of estimators of population mean for different sample sizes $n0$, with total sample size 20, blue-winged teal population. Adaptive web design used had $d = 0.9$, random selection of links from current sample. Mean square errors are standardized by dividing by that for a random sample of 20 units without replacement.

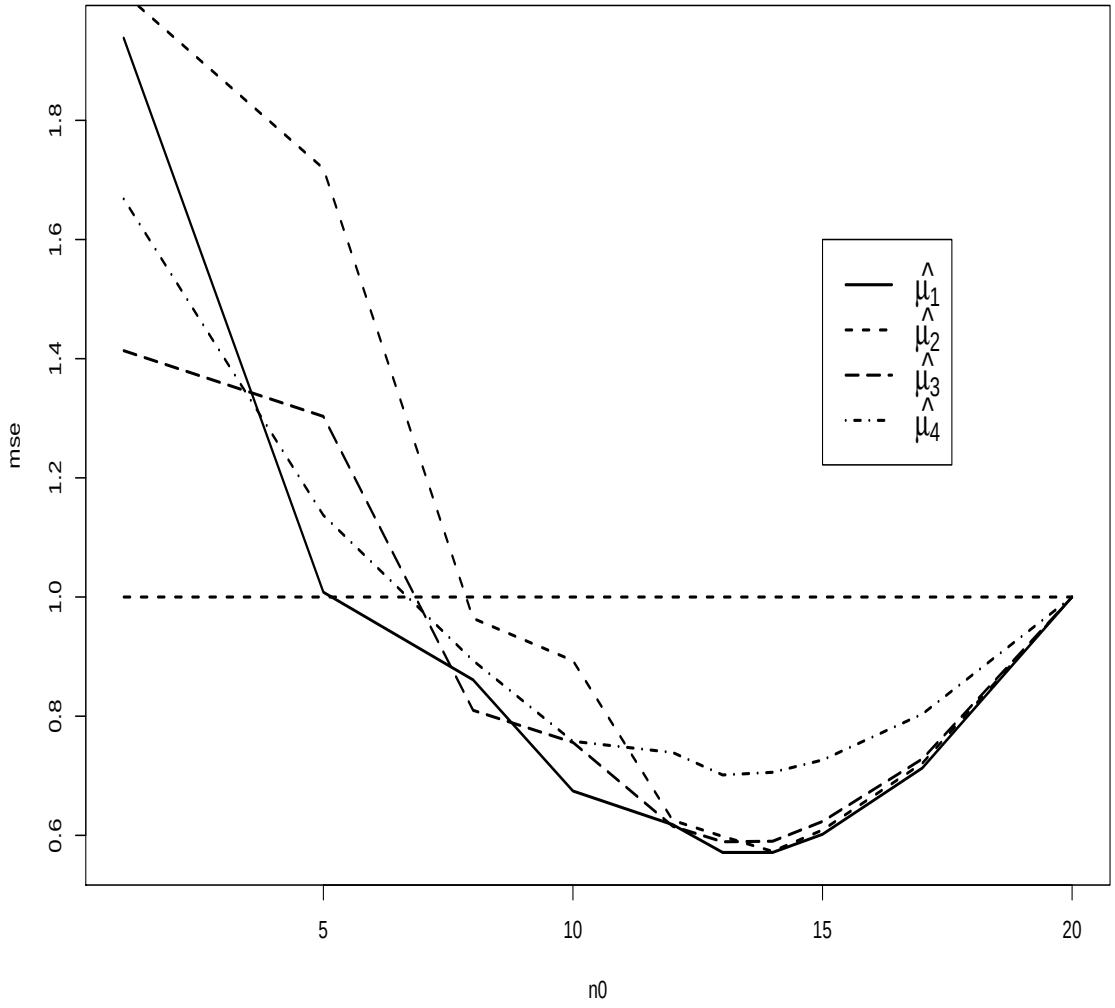

Figure 6: Two types of sample allocations from the blue-winged teal population. Both samples have a total size of 20 and use random selection of links. The first has an initial sample size of 13, while the second has an initial sample size of 1 unit, proceeding adaptively from there. The first design was more efficient than the second.