The Pennsylvania State University

The Graduate School

Department of Statistics

# ADAPTIVE SAMPLING DESIGNS AND ASSOCIATED ESTIMATORS

A Thesis in

Statistics

by

Arthur L. Dryver



Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 1999

We approve the thesis of Arthur L. Dryver.

Date of Signature

| | |
|---|---|
| Steven K. Thompson<br>Associate Professor of Statistics<br>Thesis Adviser<br>Chair of Committee | |
| Jogesh Babu<br>Professor of Statistics | |
| Bing Li<br>Associate Professor of Statistics | |
| Linda Collins<br>Professor of Human Development and Family Studies | |
| Bruce Lindsay<br>Distinguished Professor of Statistics<br>Interim Head of the Department of Statistics | |

# Abstract


## Estimators in Adaptive Sampling

In conventional sampling, we generally use different estimators depending on how the sample was obtained. If a simple random sample was taken the sample mean would be used, assuming the parameter of interest is the population mean. On the other hand, if a sample was taken with probability proportional to size, we would use the Horvitz-Thompson estimator (Horvitz and Thompson 1952). In that case, the sample mean would yield a biased estimate. In adaptive sampling, which estimator is most appropriate to use also depends on how the sample was obtained. This dissertation will cover previous work in adaptive sampling when the initial sample is taken without replacement of units and the case when the sample is taken without replacement of networks. It will also cover new estimators when the initial sample is taken without replacement of units and on adaptive cluster sampling without replacement of clusters.

The latter estimators proposed are design unbiased estimators. Often, a sample cannot be taken in the manner necessary to utilize design unbiased estimators and for this reason model based estimators are important to develop. Maximum likelihood model based estimators for estimating population size utilizing adaptive snowball sampling will be covered.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgments

I would like to thank my advisor, Dr. Steven K. Thompson, for imparting upon me a portion of his wisdom, knowledge, and insight which helped me throughout my graduate career. I would like to thank Dr. Jogesh Babu and Dr. Bing Li for taking time out from their work whenever my research fell into their domain and I asked for their assistance. I would like to thank Dr. Linda Collins for her assistance in gearing my work to a more applied and general audience. Finally, I would like to thank Dr. Martina Morris for working with me this past year and giving me a glimpse into the real world. This, as a result, enabling me to see beyond the theoretical and thus greatly enhancing my research.

Partial support for this research was provided by the National Institutes of Health, National Institute on Drug Abuse, grant RO1 DA09872, and the National Science Foundation, grant DMS-9626102.

# Chapter 1

# Introduction

## 1.1 Background Information

When dealing with rare or hidden populations, it is often useful after locating a unit that meets a specified criterion to continue sampling in that region. One way of doing so is by adaptive cluster sampling. In spatial sampling, adaptive cluster sampling can provide unbiased efficiency for estimating the abundance of rare, clustered populations (cf., Thompson and Seber 1996). For sampling hidden human populations, social links play the same role as geographic proximity in spatial sampling and adaptive cluster sampling becomes a type of link-tracing design in a graph or social network (Thompson 1997).

In the simplest form of adaptive cluster sampling an initial sample of units is selected by random sampling with or without replacement(Thompson 1990). Whenever the variable of interest for a unit in the sample satisfies a prespecified condition, neighboring or connected units are added to the sample and observed. This procedure continues until no more units are found that meet the criterion. The set of all units meeting the criterion in the neighborhood of one another is called a network. The units that were adaptively sampled that did not meet the criterion are called edge units. Figure 1.1 illustrates a network and its associated edge units, which together will be called

a cluster. In the figure, the variable of interest for a spatial unit is the number of point-objects within the unit, the neighborhood of a unit is defined as including that unit and the four spatially adjacent units, and the criterion for extra sampling is defined as the condition that the variable of interest is greater than or equal to three. Units that do not meet the criterion, including edge units, are considered networks of size one.

Conventional estimators, such as a sample mean or expansion estimator, that are unbiased with a conventional design such as simple random sampling are not unbiased with an adaptive design, but for adaptive cluster sampling simple design-unbiased estimators of a population mean or total are available.

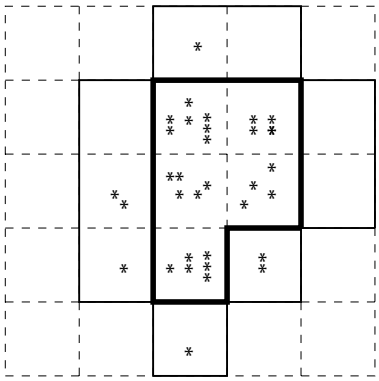


Fig. 1.1. Inside the dark line is the network, the adjacent units, inside the lighter solid line are edge units and the dashed lines represent the grid. The criterion to adaptively add units was the number of observations in a unit being greater than or equal to three.

## 1.2 Content

The usual unbiased estimators in adaptive cluster sampling are very simple but do not necessarily utilize all the information gathered. In the case where an initial sample is taken with replacement repeat selections can occur. The usual unbiased estimators do not take this into account. A more efficient estimator that utilizes this information of a repeat selection is discussed in chapter 2. Various estimators of variance are also covered. The estimators covered can be viewed two different ways, first as conditioning on the number of distinct units, see Raj and Khamis (1958), and secondly as conditioning on a sufficient statistic. Both methods yield the same estimator for the population mean but different estimators for the variance.

Improvements have also been made for when an initial sample is taken without replacement. In particular, the values of edge units are utilized in the estimators only for edge units that were picked in the initial sample. Estimators that can incorporate this information can be obtained using the Rao-Blackwell method conditioning the minimal sufficient statistic (Thompson 1990). These estimators can be much more computationally difficult. For computing the Rao-Blackwell estimators Salehi (1998) derived expressions based on inclusion-exclusion formulas. In chapter 3 new easy-to-compute estimators of higher efficiency are derived by taking the expected value of the usual estimators conditional on a sufficient statistic which is not minimally sufficient. Also, the Rao-Blackwell estimators are computed using computationally efficient new formulas based on the multivariate hypergeometric distribution, derived in Section 3.4. In

addition, the unbiased variance estimators for the Rao-Blackwell estimators are further improved using the Rao-Blackwell method.

Adaptive cluster sampling is not limited to only two types of initial samples, random sampling with or without replacement. An initial sample can be taken by any of the conventional means such as bernoulli, systematic, or stratified sampling. It may be more desirable to use a more complex sampling strategy in order to take more advantage of adaptive sampling. In chapter 4 a new sampling strategy, random sampling without replacement of clusters, will be introduced. Sampling without replacement of clusters is an extension of the work done by Salehi and Seber (1997) sampling without replacement of networks. Other pertinent work in this area was done by Raj (1956) and Murthy (1957).

A Horvitz-Thompson and a Hansen-Hurwitz type estimator are typically used with adaptive cluster sampling. The selection probabilities generally cannot be determined for all the units in the final sample and that is why a modified version of the Horvitz-Thompson and Hansen-Hurwitz must be used (Thompson 1990). Unfortunately, there is not one estimator which is uniformly better than another. Generally the Horvitz-Thompson type estimator is more efficient than the Hansen-Horvitz type estimator in the univariate case. Most studies are done with multivariate data though. Thus an important question that must be answered is,"How will the two types of estimators perform when the data is collected according to one variable but another variable is being estimated?" This question will be addressed in chapter 5.

Another important issue involving adaptive cluster sampling is when it is not possible to follow the specified design. For example, in many situations it is not feasible

to view the entire network. Sometimes there are too many units involved or as in a link tracing design it may not be possible survey all nodes necessary for various reasons. For this reason it is essential to examine model based approaches for analyzing adaptively sampled data. In chapter 6 maximum likelihood model based estimators for estimating population size when an adaptive snowball sample is taken are derived. Simulations are used to test the efficiency of these estimators. Then one of these estimators is used to analyze the Colorado Springs 1990 network data.

# Chapter 2

# Sampling With Replacement of Units

## 2.1 Introduction

As in the typical finite population sampling situation, the population consists of $N$ units labeled $1, 2, .., N$ and their associated variables of interest, $\mathbf{y} = \{y_1, y_2, ..., y_N\}$. The population vector $\mathbf{y}$ will be considered fixed but unknown constants. The parameter of interest in this paper is the population mean,

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{2.1}$$

In the simplest form of adaptive cluster sampling an initial sample of units is selected by random sampling with or without replacement (Thompson 1990). The usual design unbiased estimators for adaptive cluster sampling with an initial sample taken by with or without replacement are of a Hansen-Hurwitz and Horvitz-Thompson type. The estimators are similar for with and without replacement of units. In this chapter only the with replacement case will be covered.

### 2.1.1 The Hansen Hurwitz Type Estimator

An estimator of the population mean which is design-unbiased with adaptive cluster sampling is described below (Thompson 1992). This estimator is used when

simple random sampling with replacement is used to select the initial sample. The units selected in the initial sample are denoted by $s_0$ and the units in the final sample by $s$. $s_0$ is the set of unit labels obtained in the initial sample and $s$ is the set of distinct unit labels in the final sample. Let $n$ denote the initial sample size and $\nu$ the final sample size. Let $\psi_i$ denote the network which includes unit $i$ and $m_i$ the number of units in that network. $w_i$ represents the average value of a unit in the network which contains unit $i$, that is

$$w_i = \frac{1}{m_i} \sum_{j \in \psi_i} y_j \tag{2.2}$$

An unbiased estimator of the population mean is

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} w_i \tag{2.3}$$

The variance of $\hat{\mu}_1$ for a sample taken with replacement is

$$var(\hat{\mu}_1) = \frac{1}{nN} \sum_{i=1}^{N} (w_i - \mu)^2 \tag{2.4}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_1) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu}_1)^2 \tag{2.5}$$

## 2.2 The Horvitz Thompson Type Estimator

The estimator will be denoted $\hat{\mu}_2$. Let $K$ equal the number of distinct networks in the population, $\psi_k$ is the set of units in the $k^{th}$ network and $x_k$ denotes the number of units that make up network $\psi_k$. (Note: $x_k$ is equivalent to $m_i$ except $x_k$ is defined for the distinct networks and $m_i$ the individual units.)

The sum of the y-values in network $k$

$$y_k^* = \sum_{i \in \psi_k} y_i \tag{2.6}$$

and the inclusion probability of network $k$

$$\alpha_k = 1 - (\frac{N - x_k}{N})^n \tag{2.7}$$

Let $z_k$ be the indicator variable which equals one if any unit in the initial sample intersect the $k^{th}$ network.

$$z_k = \begin{cases} 1 & \text{if any unit of the } k^{th} \text{ network is in } s_0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

The estimator $\hat{\mu}_2$ is

$$\hat{\mu}_2 = \frac{1}{N} \sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k} \tag{2.9}$$

The joint probability of two distinct networks, $k$ and $h$ being intersected in the initial sample is

$$\alpha_{kh} = 1 - (\frac{N - x_k}{N})^n - (\frac{N - x_h}{N})^n + (\frac{N - x_k - x_h}{N})^n \tag{2.10}$$

Let $\alpha_{kk} = \alpha_k$. The variance of $\hat{\mu}_2$ is

$$var(\hat{\mu}_2) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h} \tag{2.11}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_2) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* z_k z_h (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h \alpha_{kh}} \tag{2.12}$$

## 2.3 The New Estimator

### 2.3.1 Introduction

When sampling is done with replacement it is possible to sample the same unit more than once, perhaps even many times. Thus an initial sample of size $n$ may have a much smaller effective initial sample size. It is important to take this into account when estimating the parameter of interest. The Hansen-Hurwitz based estimator can be greatly affected by repeat selections unlike the Horvitz-Thompson type estimator which is unaffected by repeat selections. In some situations it is preferred to use the Hansen-Hurwitz type estimator, and for this reason it is worth looking into the more efficient estimator presented in this section. Two different approaches to repeat selections of a unit give the same estimator for the mean, but different estimators for the variance. The estimator along with the associated estimators of the variance will be compared in an example. A comparison of this improved estimator with the usual estimator will also be illustrated. The new estimator in this section can be viewed and derived in more than one way. The different derivations lead to different interpretations of the estimator and also different estimators for the variance. The estimators illustrated in this section will be relatively easy to compute. Considerably more complicated but more efficient estimators derived by taking the conditional expectation of the usual estimators given the minimal sufficient statistic exist.

### 2.3.2 The Improved Estimator

Let $n_1$ denote the number of distinct units in the initial sample and $s_1$ the distinct unit labels.

$$\hat{\mu}_{1v} = \frac{1}{n_1} \sum_{i \in s_1} w_i \tag{2.13}$$

### 2.3.3 First Approach

The expectation of $\hat{\mu}_{1v}$ is unbiased given $n_1$ and thus it is an unbiased estimator of $\mu$ (Raj and Khamis 1958).

### 2.3.4 Second Approach

For unit $i$, let $f_i$ be the number of times the network to which unit $i$ belongs is intersected by distinct units in the initial sample; that is, $f_i$ is the number of different units in the initial sample that are in the network to which unit $i$ belongs. Let $n_1$ and $s_1$ be as defined earlier. Let the statistic $d_v$ be defined as

$$d_v \quad = \quad \{(i, y_i, f_i) : \quad i \in s\} \tag{2.14}$$

The minimal sufficient statistic

$d = \{(i, y_i) : \quad i \in s\}$ is a function of the statistic $d_v$. Thus $d_v$ is a sufficient statistic. The new estimator $\hat{\mu}_{1v}$ is defined by

$$\hat{\mu}_{1v} = E[\hat{\mu}_1 \mid d_v] \tag{2.15}$$

By the Rao-Blackwell Theorem, $\hat{\mu}_{1v}$ is unbiased for $\mu$, since $\hat{\mu}_1$ is unbiased, and the variance of the new estimator $\hat{\mu}_{1v}$ is less than or equal to the the variance of the estimator $\hat{\mu}_1$.

$$E[\hat{\mu}_{1v}] \ = E[E[\hat{\mu}_1 \mid d_v] \ ] \ = \mu \tag{2.16}$$

This estimator can be improved by conditioning on the minimal sufficient statistic $d$. The estimator derived given the minimal sufficient statistic will be denoted $\hat{\mu}_{1RB}$. In fact, since the minimal sufficient statistic $d$, is a function of the sufficient statistic $d_v$,

$$\text{var}(\hat{\mu}_{1RB}) \leq \text{var}(\hat{\mu}_{1v}) \leq \text{var}(\hat{\mu}_1) \tag{2.17}$$

Unlike $\hat{\mu}_{1RB}$, the new estimator is very easily computed, as shown by the following theorem.

THEOREM 2.1.

$$\hat{\mu}_{1v} = \frac{1}{n_1} \sum_{i \in s_1} w_i \tag{2.18}$$

#### 2.3.4.1 Proof of Theorem 2.1

The expected number of repeat selections in the initial sample of a specific unit, $i \in s_1$, given $d_v$ is $\frac{n}{n_1}$ thus the expected value of $E[\hat{\mu}_1 \mid d_v]$ is

$$\begin{aligned} E[\hat{\mu}_1 \mid d_v] &= E[\frac{1}{n}\sum_{i=1}^{n} w_i \mid d_v] \\ &= E[\frac{1}{n}\sum_{i\in s_0} w_i \mid d_v] \\ &= \frac{1}{n}\sum_{i\in s_1} w_i(\frac{n}{n_1}) \\ &= \frac{1}{n_1}\sum_{i\in s_1} w_i \end{aligned} \tag{2.19}$$

## 2.4 The Variance

### 2.4.1 The Variance as Viewed Using the First Approach

The following formulas on how to compute $var(\hat{\mu}_{1v})$ and $E(\frac{1}{n_1})$ follow from Raj and Khamis (1958)

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w_i - \mu)^2 \tag{2.20}$$

Then

$$var(\hat{\mu}_1) = \frac{1}{n}(1 - \frac{1}{N})\sigma^2 = (Q - \frac{1}{N})\sigma^2 \tag{2.21}$$

where

$$Q = \frac{1}{n}(1 + \frac{n-1}{N}) \tag{2.22}$$

$$var(\hat{\mu}_{1v}) = [E(\frac{1}{n_1}) - \frac{1}{N}]\sigma^2 \tag{2.23}$$

Then $var(\hat{\mu}_{1v}) < var(\hat{\mu}_1)$ if $E(\frac{1}{n_1}) < Q$

For the calculation of the variance we need the $E(\frac{1}{n_1})$ and the formulas to do this are,

$$P(n_1) = N^{-n}\binom{N}{n_1}\epsilon^{n_1}O^n \tag{2.24}$$

$$\epsilon^k O^t = \sum_{r=1}^{k}(-1)^{k-r}\binom{k}{r}r^t \tag{2.25}$$

and

$$E(\frac{1}{n_1}) = N^{-n}\sum_{n_1=1}^{n}\frac{1}{n_1}\binom{N}{n_1}\epsilon^{n_1}O^n \tag{2.26}$$

An unbiased estimator of $\sigma^2$ for a given $n_1 \geq 2$ is

$$s^2_{n_1} = \frac{1}{n_1 - 1}\sum_{i \in s_1}(w_i - \hat{\mu})^2 \tag{2.27}$$

and the unbiased estimator for the variance of $\hat{\mu}_{1v}$ is,

$$\widehat{var}(\hat{\mu}_{1v}) = [(\frac{1}{n_1} - \frac{1}{N}) + N^{1-n}(1 - \frac{1}{n_1})]s^2_{n_1} \tag{2.28}$$

$$E[\widehat{var}(\hat{\mu}_{1v}) \mid n_1 \geq 2] = var(\hat{\mu}) \tag{2.29}$$

Another unbiased estimator of the variance, which is not conditional on $n_1 \geq 2$ is

$$\widetilde{var}(\hat{\mu}_{1v}) = [(\frac{1}{n_1} - \frac{1}{N}) + \frac{N-1}{N^n - N}]s^2_{n_1} \tag{2.30}$$

where

$$s^2_{n_1} = \frac{1}{n_1 - 1} \sum_{i \in s_1} (w_i - \hat{\mu}_{1v})^2 \quad \text{for } n_1 \geq 2 \tag{2.31}$$

and

$$s^2_{n_1} = 0 \quad \text{for } n_1 = 1 \tag{2.32}$$

To understand where equations 2.28 and 2.30 come from we must think about

$$E[\frac{1}{n_1} \mid n_1 \geq 2] \tag{2.33}$$

$$E[\frac{1}{n_1} \mid n_1 \geq 2]P(n_1 \geq 2) + E[\frac{1}{n_1} \mid n_1 = 1]P(n_1 = 1) =$$

$$\frac{1}{n_1}\frac{N^n - 1}{N^{n-1}} + (\frac{1}{1})\frac{1}{N^{n-1}} =$$

$$\frac{1}{n_1}\left[\frac{N^{n-1} - 1 + n_1}{N^{n-1}}\right] =$$

$$N^{1-n_1}\left[\frac{N^{n-1}}{n_1}+1+\frac{1}{n_1}\right] \tag{2.34}$$

In the next subsection we will cover another easy to compute estimator for the variance of $\hat{\mu}_{1v}$.

### 2.4.2 Variance as Viewed by Rao-Blackwell Method

The variance of $\hat{\mu}_{1v}$ can also be viewed as

$$\begin{aligned} var(\hat{\mu}_{1v}) &= [E(\frac{1}{n_1})-\frac{1}{N}]\sigma^2 \quad \text{(equation 2.23)} \\ &= var(\hat{\mu}_1) - E[(\hat{\mu}_{1v}-\hat{\mu}_1)^2] \end{aligned} \tag{2.35}$$

Equation 2.35 and 2.23 yield the same value since equation 2.35 is the variance of the same estimator, even though it was derived through a different approach. Equation 2.23 in many cases will be easier to compute, but for estimating the variance (equation 2.23 or 2.35) it is sometimes better to view the variance as equation 2.35. An unbiased easy-to-compute estimator of the variance is,

$$\widetilde{var}(\hat{\mu}_{1v}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(w_i-\hat{\mu}_1)^2-(\hat{\mu}_{1v}-\hat{\mu}_1)^2 \tag{2.36}$$

A more efficient, unbiased, but more complicated to compute estimator of the variance of $\hat{\mu}_{1v}$ is,

$$\begin{aligned}\widehat{var}(\hat{\mu}_{1v}) &= E[\widehat{var}(\hat{\mu}_1) \mid d_v] - E[(\hat{\mu}_{1v} - \hat{\mu}_1)^2 \mid d_v] \\ &= \frac{1}{\sum_{s_0 \in S_{d_v}}} \sum_{s_0 \in S_{d_v}} \widehat{var}(\hat{\mu}_1) - \frac{1}{\sum_{s_0 \in S_{d_v}}} \sum_{s_0 \in S_{d_v}} (\hat{\mu}_{1v} - \hat{\mu}_1)^2\end{aligned} \tag{2.37}$$

where $S_{d_v}$ is the sample space for $s_0$ compatible with the sufficient statistic $d_v$. Therefore $s_0 \in S_{d_v}$ should be interpreted as all initial samples of size $n$ containing $n - n_1$ repetitions and all the distinct units in $s_1$.

### 2.4.3 Which Estimator of the Variance to Use and When

For unbiased estimation of the variance of $\hat{\mu}_{1v}$ a decision of which estimator of the variance should be used can be made by comparison after data collection. In situations where the probability of repeat selections is low the first approach should be taken, for it reduces the variance regardless of whether or not a repeat selection was made, unlike the second approach. In the situation where the expected number of repeat selections is high, however, the second approach is best.

## 2.5 An Example

Table 2.1: The first line is the unit labels and the second line is their associated values. The population consists of the four units. The following lines of the table are necessary components for calculating various estimators in adaptive cluster sampling, with n=3 and a condition of $y_i \geq 5$.

| unit $i$ | = | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $y_i$ | = | 6 | 10 | 0 | 2 |
| $m_i$ | = | 2 | 2 | 1 | 1 |
| $w_i$ | = | $\frac{6+10}{2} = 8$ | $\frac{6+10}{2} = 8$ | 0 | 2 |

Table 2.2: The Calculation of the Estimators. This table consists of all the possible selections of units in the initial samples and a few possible associated estimates of $\mu_{1v}$. In the first column numbers before the semi-colon represent the units in the initial sample. The second column is the set of $y-$values in the final sample.

| $i \in s_0$ | $P(s_0)$ | $n_1$ | $\frac{1}{n_1}$ | Final Sample | $\hat{\mu}_1$ | $\hat{\mu}_{1v}$ |
|---|---|---|---|---|---|---|
| 1,1,1 | 1/64 | 1 | 1.000 | 6,10,0 | 8.000 | 8.000 |
| 1,1,2 | 3/64 | 2 | 0.500 | 6,10,0 | 8.000 | 8.000 |
| 1,2,2 | 3/64 | 2 | 0.500 | 6,10,0 | 8.000 | 8.000 |
| 2,2,2 | 1/64 | 1 | 1.000 | 6,10,0 | 8.000 | 8.000 |
| 1,1,3 | 3/64 | 2 | 0.500 | 6,10,0 | 5.333 | 4.000 |
| 1,3,3 | 3/64 | 2 | 0.500 | 6,10,0 | 2.667 | 4.000 |
| 3,3,3 | 1/64 | 1 | 1.000 | 0 | 0.000 | 0.000 |
| 1,1,4 | 3/64 | 2 | 0.500 | 6,10,0,2 | 6.000 | 5.000 |
| 1,4,4 | 3/64 | 2 | 0.500 | 6,10,0,2 | 4.000 | 5.000 |
| 4,4,4 | 1/64 | 1 | 1.000 | 2 | 2.000 | 2.000 |
| 2,2,3 | 3/64 | 2 | 0.500 | 6,10,0 | 5.333 | 4.000 |
| 2,3,3 | 3/64 | 2 | 0.500 | 6,10,0 | 2.667 | 4.000 |
| 2,2,4 | 3/64 | 2 | 0.500 | 6,10,0,2 | 6.000 | 5.000 |
| 2,4,4 | 3/64 | 2 | 0.500 | 6,10,0,2 | 4.000 | 5.000 |
| 3,3,4 | 3/64 | 2 | 0.500 | 0,2 | 0.667 | 1.000 |
| 3,4,4 | 3/64 | 2 | 0.500 | 0,2 | 1.333 | 1.000 |
| 1,2,3 | 6/64 | 3 | 0.333 | 6,10,0 | 5.333 | 5.333 |
| 1,2,4 | 6/64 | 3 | 0.333 | 6,10,0,2 | 6.000 | 6.000 |
| 1,3,4 | 6/64 | 3 | 0.333 | 6,10,0,2 | 3.333 | 3.333 |
| 2,3,4 | 6/64 | 3 | 0.333 | 6,10,0,2 | 3.333 | 3.333 |
| MEAN | | 2.313 | 0.469 | | 4.500 | 4.500 |
| BIAS | | | | | 0.000 | 0.000 |
| MSE | | | | | 4.250 | 3.719 |

Table 2.3: The Calculation of the Variance Estimators. This table consists of all the possible selections of units in the initial samples and a few possible associated estimates of $\mu_{1v}$. In the first column numbers before the semi-colon represent the units in the initial sample. The second column is the set of $y-$values in the final sample.

| $i \in s_0$ | $P(s_0)$ | $n_1$ | $\widehat{var}(\hat{\mu}_1)$ | $E[\widehat{var}(\hat{\mu}_1)\|n_1 \geq 2]$ | $\widetilde{var}(\hat{\mu}_1)\|n_1$ | $\widehat{var}(\hat{\mu}_{1v})$ |
|---|---|---|---|---|---|---|
| 1,1,1 | 1/64 | 1 | 0.000 | $\cdots$ | 0.000 | 0.000 |
| 1,1,2 | 3/64 | 2 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1,2,2 | 3/64 | 2 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2,2,2 | 1/64 | 1 | 0.000 | $\cdots$ | 0.000 | 0.000 |
| 1,1,3 | 3/64 | 2 | 7.111 | 9.000 | 9.600 | 5.333 |
| 1,3,3 | 3/64 | 2 | 7.111 | 9.000 | 9.600 | 5.333 |
| 3,3,3 | 1/64 | 1 | 0.000 | $\cdots$ | 0.000 | 0.000 |
| 1,1,4 | 3/64 | 2 | 4.000 | 5.063 | 5.400 | 3.000 |
| 1,4,4 | 3/64 | 2 | 4.000 | 5.063 | 5.400 | 3.000 |
| 4,4,4 | 1/64 | 1 | 0.000 | $\cdots$ | 0.000 | 0.000 |
| 2,2,3 | 3/64 | 2 | 7.111 | 9.000 | 9.600 | 5.333 |
| 2,3,3 | 3/64 | 2 | 7.111 | 9.000 | 9.600 | 5.333 |
| 2,2,4 | 3/64 | 2 | 4.000 | 5.063 | 5.400 | 3.000 |
| 2,4,4 | 3/64 | 2 | 4.000 | 5.063 | 5.400 | 3.000 |
| 3,3,4 | 3/64 | 2 | 0.444 | 0.563 | 0.600 | 0.333 |
| 3,4,4 | 3/64 | 2 | 0.444 | 0.563 | 0.600 | 0.333 |
| 1,2,3 | 6/64 | 3 | 7.111 | 2.667 | 2.844 | 7.111 |
| 1,2,4 | 6/64 | 3 | 4.000 | 1.500 | 1.600 | 4.000 |
| 1,3,4 | 6/64 | 3 | 5.778 | 2.167 | 2.311 | 5.778 |
| 2,3,4 | 6/64 | 3 | 5.778 | 2.167 | 2.311 | 5.778 |
| MEAN | | | 4.250 | $\cdots$ | 3.719 | 3.719 |
| BIAS | | | 0.000 | $\cdots$ | 0.000 | 0.000 |

Table 2.4: The Calculation of the Variance Estimators. This table consists of all the possible selections of units in the initial samples and a few possible associated estimates of $\mu_{1v}$. In the first column numbers before the semi-colon represent the units in the initial sample. The second column is the set of $y-$values in the final sample.

| $i \in s_0$ | $P(s_0 \mid n_1)$ | $n_1$ | $E[\widehat{var}(\hat{\mu}_1) \mid n_1 \geq 2]$ |
|---|---|---|---|
| 1,1,2 | 1/20 | 2 | 0.000 |
| 1,2,2 | 1/20 | 2 | 0.000 |
| 1,1,3 | 1/20 | 2 | 9.000 |
| 1,3,3 | 1/20 | 2 | 9.000 |
| 1,1,4 | 1/20 | 2 | 5.063 |
| 1,4,4 | 1/20 | 2 | 5.063 |
| 2,2,3 | 1/20 | 2 | 9.000 |
| 2,3,3 | 1/20 | 2 | 9.000 |
| 2,2,4 | 1/20 | 2 | 5.063 |
| 2,4,4 | 1/20 | 2 | 5.063 |
| 3,3,4 | 1/20 | 2 | 0.563 |
| 3,4,4 | 1/20 | 2 | 0.563 |
| 1,2,3 | 2/20 | 3 | 2.667 |
| 1,2,4 | 2/20 | 3 | 1.500 |
| 1,3,4 | 2/20 | 3 | 2.167 |
| 2,3,4 | 2/20 | 3 | 2.167 |
| MEAN | | | 3.719 |
| BIAS | | | 0.000 |

## 2.6 Simulations

Simulations were performed on a real data set, Blue-Winged Teal Data and all teal data combined (Smith *et al* 1995.) For each estimator 100,000 iterations were performed. A large number of iterations were used to most accurately estimate the true population variances of the estimators. Varying initial sample sizes were used. The same sample is used to calculate $\hat{\mu}_1$, $\hat{\mu}$. The formulas used to estimate the variance are,

$$\widehat{var}(\hat{\mu}) = \frac{1}{100000 - 1} \sum_{i=1}^{100000} (\hat{\mu}_i - \bar{\mu}_h)^2, \tag{2.38}$$

$\hat{\mu}_i$ is the value for the relevant estimator for sample $i$;

$$\bar{\mu}_h = \frac{1}{100000} \sum_{i=1}^{100000} \hat{\mu}_i; \tag{2.39}$$

and

$$eff(\hat{\mu}_{1v}) = \frac{var(\hat{\mu}_1)}{var(\hat{\mu}_{1v})} \tag{2.40}$$

Table 2.5: blue-winged teal data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

Table 2.6: All three combined winged teal data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20 | 203 | 80 | 0 | 0 | 0 | 0 | 675 | 0 |
| 4000 | 13500 | 234 | 1359 | 18 | 0 | 0 | 80 | 178 | 55 |
| 0 | 0 | 0 | 0 | 99 | 3 | 6 | 0 | 17709 | 14 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 122 |
| 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 1484 |

Table 2.7: Results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1v})$ | $eff(\hat{\mu}_{1v})$ |
|---|---|---|---|---|
| 2 | 9.408540 | 239431.199039 | 239431.199039 | 1.000000 |
| 5 | 18.902630 | 96593.852872 | 93921.103018 | 1.028457 |
| 8 | 24.897700 | 60722.208370 | 57186.900046 | 1.061820 |
| 10 | 27.766470 | 48561.937733 | 44840.513882 | 1.082992 |
| 12 | 29.957820 | 40477.689187 | 36503.950776 | 1.108858 |
| 15 | 32.490670 | 32277.723180 | 28304.051394 | 1.140392 |
| 30 | 38.828650 | 16139.558235 | 11994.332909 | 1.345599 |

Table 2.8: Results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 100000$

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1v})$ | $eff(\hat{\mu}_{1v})$ |
|---|---|---|---|---|
| 2 | 1.979980 | 1829568.863462 | 1829568.863462 | 1.000000 |
| 5 | 4.805120 | 724179.843872 | 708215.737339 | 1.022541 |
| 8 | 7.462080 | 450691.262877 | 428585.736581 | 1.051578 |
| 10 | 9.149730 | 362617.371987 | 337071.914752 | 1.075786 |
| 12 | 10.760140 | 303981.322596 | 275533.076251 | 1.103248 |
| 15 | 13.070180 | 243978.165835 | 212089.205625 | 1.150356 |
| 30 | 22.734690 | 120465.937008 | 90242.995239 | 1.334906 |

Table 2.9: Results for the simulations on combined-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1v})$ | $eff(\hat{\mu}_{1v})$ |
|---|---|---|---|---|
| 2 | 14.972870 | 485415.581370 | 485415.581370 | 1.000000 |
| 5 | 27.403770 | 194157.930293 | 188478.855701 | 1.030131 |
| 8 | 33.615140 | 121325.702864 | 114223.601748 | 1.062177 |
| 10 | 36.106800 | 97761.427784 | 90167.053706 | 1.084226 |
| 12 | 37.826400 | 81100.231847 | 73277.408661 | 1.106756 |
| 15 | 39.609680 | 65197.180696 | 57142.216474 | 1.140963 |
| 30 | 43.595080 | 32456.393513 | 24221.569903 | 1.339979 |

### 2.6.1 Comments on Results

In the case where $n = 2$ the estimators yield the same estimates for the parameter of interest regardless of repetition of units or not. For $n \geq 2$ the following property holds. The greater the proportion of $\frac{n}{N}$ the more the reduction in variance by using the new estimator. The reduction in variance for $\hat{\mu}_{1v}$ in relation to $\hat{\mu}_1$ is dependent upon $E(\frac{1}{n_1})$, as can be seen from equations 2.21 and 2.23.

# Chapter 3

# Sampling Without Replacement of Units

## 3.1 Introduction

In this chapter ordinary estimators that can be used when an initial sample is taken without replacement which are very similar to the usual estimators in chapter 2 will be covered. In addition to, more efficient estimators will be derived from the Rao-Blackwell theorem.

The usual estimators in adaptive cluster sampling can be improved by incorporating more of the information obtained in the final sample. In particular, the values of edge units are utilized in the estimators only for edge units that were picked in the initial sample. Estimators that can incorporate this information can be obtained using the Rao-Blackwell method conditioning the minimal sufficient statistic (Thompson 1990). These estimators can be much more computationally difficult. For computing the Rao-Blackwell estimators Salehi (1998) derived expressions based on inclusion-exclusion formulas.

In this chapter new, easy-to-compute estimators of higher efficiency are derived by taking the expected value of the usual estimators conditional on a sufficient but not minimal sufficient statistic. In Section 3.6 empirical comparisons of efficiencies are made among the ordinary estimators, the new estimators, and the Rao-Blackwell estimators.

The Rao-Blackwell estimators are computed using computationally efficient new formulas based on the multivariate hypergeometric distribution, derived in Section 3.4. A comparison of this new approach and Salehi's approach for calculating the Rao-Blackwell estimators is also included in the latter section. In addition, the unbiased variance estimators for the Rao-Blackwell estimators are further improved using the Rao-Blackwell method.

## 3.2 Ordinary Estimators in Adaptive Sampling

Two estimators of the population mean which are design-unbiased with adaptive cluster sampling are described below (Thompson 1990). We call them the ordinary estimators and denote them as $\hat{\mu}_1$ and $\hat{\mu}_2$. Neither of the two estimators is uniformly better than the other, though in empirical studies $\hat{\mu}_2$ is generally more efficient than $\hat{\mu}_1$ (Thompson 1992). These estimators are used when simple random sampling is used to select the initial sample. The units selected in the initial sample are denoted by $s_0$ and the units in the final sample by $s$. $s_0$ is the set of unit labels obtained in the initial sample and $s$ is the set of distinct unit labels in the final sample. Let $n$ denote the initial sample of size and $\nu$ the final sample size. Let $\psi_i$ denote the network which includes unit $i$ and $m_i$ the number of units in that network. $w_i$ represents the average value of a unit in the network which contains unit $i$, that is

$$w_i = \frac{1}{m_i} \sum_{j \in \psi_i} y_j \tag{3.1}$$

An unbiased estimator of the population mean is

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} w_i \tag{3.2}$$

The variance of $\hat{\mu}_1$ is

$$var(\hat{\mu}_1) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} (w_i - \mu)^2 \tag{3.3}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_1) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu}_1)^2 \tag{3.4}$$

Now how to calculate $\hat{\mu}_2$. Let $K$ equal the number of distinct networks in the population, $\psi_k$ is the set of units in the $k^{th}$ network and $x_k$ denotes the number of units that make up network $\psi_k$. (Note: $x_k$ is equivalent to $m_i$ except $x_k$ is defined for the distinct networks and $m_i$ the individual units.)

The sum of the y-values in network $k$

$$y_k^* = \sum_{i \in \psi_k} y_i \tag{3.5}$$

and the inclusion probability of network $k$

$$\alpha_k = 1 - \frac{\binom{N-x_k}{n}}{\binom{N}{n}} \tag{3.6}$$

$z_k$ is an indicator variable which equals one if any unit in the initial sample intersect the $k^{th}$ network.

$$z_k = \begin{cases} 1 & \text{if any unit of the } k^{th} \text{ network is in } s_0 \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

The estimator $\hat{\mu}_2$ is

$$\hat{\mu}_2 = \frac{1}{N} \sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k} \tag{3.8}$$

The joint probability of two networks, $k$ and $h$ being intersected in the initial sample is

$$\alpha_{kh} = 1 - \frac{\{\binom{N-x_k}{n} + \binom{N-x_h}{n} - \binom{N-x_k-x_h}{n}\}}{\binom{N}{n}} \tag{3.9}$$

Also $\alpha_{kk} = \alpha_k$. The variance of $\hat{\mu}_2$ is

$$var(\hat{\mu}_2) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h} \tag{3.10}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_2) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* z_k z_h (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h \alpha_{kh}} \tag{3.11}$$

## 3.3 New Estimators

In this section, two new estimators are arrived at by applying the Rao-Blackwell theorem to $\hat{\mu}_1$ and $\hat{\mu}_2$. The new estimators are virtually as easy to compute as $\hat{\mu}_1$ and $\hat{\mu}_2$. When computing $\hat{\mu}_1$ and $\hat{\mu}_2$, we incorporate only those edge units in the initial sample. The new estimators, which we call $\hat{\mu}_{1+}$ and $\hat{\mu}_{2+}$, are developed considering only how many edge units were initially picked, but not which ones. The estimators presented in this section and their associated variances are derived in the appendix.

### 3.3.1 The New Estimator $\hat{\mu}_{1+}$

The final sample $s$ can be partitioned into two parts: a "core" part $s_c$ and the remaining part $\bar{s}_c$. The core part $s_c$ is the set of all the distinct units in the sample for which the criterion $y_i \geq c$ is satisfied. The remaining part $\bar{s}_c$ consists of all the distinct units in the sample for which $y_i < c$. For unit $i$, let $f_i$ be the number of times the

network to which unit $i$ belongs is intersected by the initial sample; that is, $f_i$ is the number of units in the initial sample that are in the network to which unit $i$ belongs.

Let the statistic $d^+$ be defined as

$$d^+ \quad = \quad \{(i, y_i, f_i) : \; i \in s_c, \; (j, y_j) : \; j \in \bar{s}_c\} \tag{3.12}$$

In $d^+$, the intersection frequency $f_i$ is included only for $i \in s_c$. Let $D^+$ denote a random variable that takes on possible values of $d^+$. Also let $\mathcal{D}^+$ denote the sample space for $d^+$.

For $i \in s$, define the indicator variable $e_i$ as

$$e_i = \begin{cases} 1 & \text{if} \quad y_i < c \quad \text{and } i \text{ is in the neighborhood of some } j \quad \in \quad s_c \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

Thus $e_i = 1$ if $i$ is an edge unit and the network that makes it an edge unit is selected in the initial sample. Should $e_i = 1$, we shall refer to that unit as a sample edge unit. Other units picked in the initial sample may be edge units, but sample units are the edge units whose network that classifies them as an edge unit was intersected in the initial sample.

The number of sample edge units in the sample is

$$e_s = \sum_{i=1}^{\nu} e_i = \sum_{i \in s} e_i \tag{3.14}$$

The number of sample edge units picked in the initial sample $s_0$ is

$$e_{s_0} = \sum_{i=1}^{n} e_i = \sum_{i \in s_0} e_i \tag{3.15}$$

The average $y$-value for the sample edge units in the final sample is

$$\bar{y}_e = \frac{\sum_{i=1}^{\nu} e_i y_i}{e_s} \tag{3.16}$$

For the $i$th unit in the sample, define a new variable of interest $w_i^{'}$ by

$$w_i^{'} = w_i(1 - e_i) + \bar{y}_e e_i \tag{3.17}$$

The variable $w_i^{'}$ is the original $w_i$ when not dealing with sample edge units. When dealing with sample edge units $w_i^{'}$ equals the average of the sample edge units.

The new estimator $\hat{\mu}_{1+}$ is defined by

$$\hat{\mu}_{1+} = E[\hat{\mu}_1 | D^+ = d^+] \tag{3.18}$$

By the Rao-Blackwell Theorem, $\hat{\mu}_{1+}$ is unbiased for $\mu$, since $\hat{\mu}_1$ is unbiased, and the variance of the new estimator $\hat{\mu}_{1+}$ no more than the variance of the ordinary estimator $\hat{\mu}_1$. In fact, since the minimal sufficient statistic $d$, defined later by equation 3.33, is a

function of the sufficient statistic $d^+$,

$$\text{var}(\hat{\mu}_{1RB}) \leq \text{var}(\hat{\mu}_{1+}) \leq \text{var}(\hat{\mu}_1) \tag{3.19}$$

Further, unlike $\hat{\mu}_{1RB}$, the new estimator is very easily computed, as shown by the following theorem.

THEOREM 3.1.

$$\hat{\mu}_{1+} = \frac{1}{n}\sum_{i=1}^{n} w_i^{'} \tag{3.20}$$

The proof of the previous theorem and some of the following equations in this subsection are given in the next subsection.

Since the initial sample determines the final sample and every value of the statistic $d^+$, let $g(s_0^{'})$ denote the function that maps an initial sample into a value of $d^+$ resulting from its selection. For any two values of $s_0^{'}$ and $d^+$ let

$$I(s_0^{'}, d^+) = \begin{cases} 1 & \text{if} \quad g(s_0^{'}) = d^+ \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

Let $L(d^+)$ be the number of initial samples compatible with $d^+$ and $P(d^+)$ be the probability that $D^+ = d^+$. Also let $\mathcal{S}$ be the sample space containing all possible initial samples. The variance of $\hat{\mu}_{1+}$ is

$$\begin{aligned} var(\hat{\mu}_{1+}) &= \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} (w_i - \mu)^2 \\ &\quad - \frac{1}{n^2} \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d^+) \Big( \sum_{i \in s_0^{'}, e_i = 1} y_i - e_{s_0^{'}} \bar{y}_e \Big)^2 \end{aligned} \tag{3.22}$$

An unbiased easy-to-compute estimator of the variance of $\hat{\mu}_{1+}$ when sampling is done without replacement is given by

$$\widehat{var}(\hat{\mu}_{1+}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu}_1)^2 - (\hat{\mu}_1 - \hat{\mu}_{1+})^2 \tag{3.23}$$

However, a more efficient estimator is

$$\begin{aligned} \widehat{var}(\hat{\mu}_{1+}) &= E[\widehat{var}(\hat{\mu}_{1+}) | d^+] \\ &= \frac{1}{L} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d^+) \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu}_1)^2 \\ &\quad - \frac{1}{Ln^2} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d^+) \Big( \sum_{i \in s_0^{'}, e_i = 1} y_i - e_{s_0^{'}} \bar{y}_e \Big)^2 \end{aligned}$$

(3.24)

### 3.3.2 The New Estimator $\hat{\mu}_{2+}$

For the $k$th network in the sample, define the indicator variable

$$e_k^{'} = \begin{cases} 1 & \text{if} \quad y_k^* < c \quad \text{and } k \text{ is in the neighborhood of some } k^{'} \in s_c \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

The variable $e_k^{'}$, for $i = 1, \ldots K$ has meaning similar to $e_i$ but is indexed by network rather than individual unit. Note that all networks of size greater than one must have $e_k^{'} = 0$. Also $e_k^{'} = 0$ for those units not in $s$.

Let

$$y_k^{'} = \begin{cases} y_k^* & \text{if} \quad e_k^{'} = 0 \\ \bar{y}_e & \text{if} \quad e_k^{'} = 1 \end{cases} \tag{3.26}$$

Thus, for a network of units satisfying the condition, $y_k^{'}$ is the total of the $y$-values in that network, while for an sample edge unit (a network of size one) $y_k^{'}$ is the average of the $y$-values for all the sample edge units in the sample.

The new estimator $\hat{\mu}_{2+}$ is defined by

$$\hat{\mu}_{2+} = E[\hat{\mu}_2 | D^+ = d^+] \tag{3.27}$$

By the Rao-Blackwell Theorem $\hat{\mu}_{2+}$ is unbiased for $\mu$ since $\hat{\mu}_2$ is unbiased and the variance of $\hat{\mu}_{2+}$ is less than or equal to the variance of $\hat{\mu}_2$. In fact,

$$\text{var}(\hat{\mu}_{2RB}) \leq \text{var}(\hat{\mu}_{2+}) \leq \text{var}(\hat{\mu}_2) \tag{3.28}$$

Unlike $\hat{\mu}_{2RB}$, the new estimator is very easily computed, as shown by the following theorem.

THEOREM 3.2.

$$\hat{\mu_{2+}} = \frac{1}{N} \sum_{k=1}^{K} \frac{y_k^{'} z_k}{\alpha_k} \tag{3.29}$$

and the variance of $\hat{\mu}_{2+}$ is

$$\begin{aligned} var(\hat{\mu}_{2+}) &= \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* (\alpha_{kh} - \alpha_k \alpha_k)}{\alpha_k \alpha_k} \\ &\quad -\frac{1}{n^2} \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d^+) \Big( \sum_{i \in s_0^{'}, e_i = 1} y_i - e_{s_0^{'}} \bar{y}_e \Big)^2 \end{aligned} \tag{3.30}$$

An unbiased estimator of this variance is

$$\widetilde{var}(\hat{\mu}_{2+}) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* z_k z_h (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h \alpha_{kh}} - (\hat{\mu}_2 - \hat{\mu}_{2+})^2 \tag{3.31}$$

A more efficient estimator of the variance is

$$\begin{aligned}\widehat{var}(\hat{\mu}_{2+}) &= E[\widetilde{var}(\hat{\mu}_{2+})|d^{+}] \\ &= \frac{1}{L}\sum_{s_0^{'}\in s} I(s_0^{'}, d^{+})\frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^{*}y_h^{*}z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \\ &\quad -\frac{1}{Ln^2}\sum_{s_0^{'}\in s} I(s_0^{'}, d^{+})\Big(\sum_{i\in s_0^{'}, e_i=1} y_i - e_{s_0^{'}}\bar{y}_e\Big)^2\end{aligned} \tag{3.32}$$

## 3.4 Rao-Blackwell Estimators

Each of the estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ can be improved by the Rao-Blackwell method (Rao 1945, Blackwell 1947), by conditioning on the minimal sufficient statistic. The minimal sufficient statistic $d$ is the unordered set of distinct units in the sample and their $y$-values. That is,

$$d = \{(i, y_i) : i \in s\} \tag{3.33}$$

Let $D$ be a random variable that takes on possible values of $d$ from the sample sace $\mathcal{D}$ of all possible values of the statistic $d$. The Rao-Blackwell estimators are

$$\hat{\mu}_{1RB} = E[\hat{\mu}_1 \,|\, d] \tag{3.34}$$

and

$$\hat{\mu}_{2RB} = E[\hat{\mu}_2 \,|\, d] \tag{3.35}$$

The initial sample can be thought of as one of many possible ways that will yield that particular final sample. Although different initial samples can produce the same final sample, they do not always generate the same estimate for the parameter of interest. The Rao-Blackwell estimators can be obtained by averaging the values of the ordinary estimators over all initial samples that produce the same final sample.

Since there can be a great number of such samples, straightforward computation of these estimators can be difficult(Thompson 1990). Salehi (1998) derived a method for computing the Rao-Blackwell estimators using inclusion-exclusion formulas. This method is computationally efficient provided that relatively few networks meeting the condition are intersected, regardless of the size of the networks and the number of times the networks were intersected. Unfortunately, as the number of intersected networks meeting the condition increases, the number of computations involved can increase dramatically. The approach that we derive in this section to compute the Rao-Blackwell estimators, based on the multivariate hypergeometric distribution, is affected mostly by the number of times a network is intersected and the size of the networks. Thus, for data with a large number of networks of small size intersected, the new formulas would tend to be more computationally efficient. Thus one approach can have a significantly fewer number of computations than the other depending on the circumstances, and it is best to consider both approaches when calculating the Rao-Blackwell estimators.

### 3.4.1 Rao-Blackwell Applied To $\hat{\mu}_1$ and $\hat{\mu}_2$

$$E[\hat{\mu}|D=d] = \sum_{s_0^{'} \in \mathcal{S}} \hat{\mu}(s_0^{'}) P(S_0 = s_0^{'} | D = d) \tag{3.36}$$

If $s_0^{'}$ can yield $d$ the unordered set of distinct units, and their labels, in the final sample, the probaility $S_0 = s_0^{'}$ is equal to one divided by the number of different samples that can yield $d$. This makes sense in that, each initial sample is equally likely to occur, thus each initial sample that yields $d$ is equally weighted. Otherwise if $s_0^{'}$ does not yield $d$, the probability $S_0 = s_0^{'}$ is zero, (i.e. that initial sample can not produce that final sample.) Let $u(s_0^{'})$ denote the function that maps an initial sample into a value of $d$ resulting from its selection. For any two values of $s_0^{'}$ and $d$ let

$$I(s_0^{'}, d) = \begin{cases} 1 & \text{if} \quad u(s_0^{'}) = d \\ 0 & \text{otherwise} \end{cases} \tag{3.37}$$

Let $H$ be the total number of combinations compatible with $d$:

$$H = \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d) \tag{3.38}$$

Another way of thinking of $H$ is the number of initial samples of size $n$ that yield the final sample $s$. Let $\kappa$ equal the number of distinct networks minus the number of networks that are also sample edge units. The first $\kappa$ terms in the equations to come will represent the networks in the sample not including sample edge units. The $\kappa + 1$ term will represent the group of sample edge units. Let $x_{\kappa+1} = e_s$ and $y^*_{\kappa+1} = \bar{y}_e$. Then:

$$\begin{aligned} H &= \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d) \\ &= \sum_{f_1} \cdots \sum_{f_{\kappa+1}} \prod_{k=1}^{\kappa+1} \binom{x_k}{f_k} \end{aligned} \tag{3.39}$$

The following constraints on $f_k$ exist:

$$\sum_{k=1}^{\kappa+1} f_k = n \tag{3.40}$$

$$1 \leq f_k \leq x_k \; \forall k \leq \kappa \tag{3.41}$$

$$0 \leq f_{\kappa+1} \leq e_s \tag{3.42}$$

**Note:**

$$I(s_0^{'}, d) = \begin{cases} 1 & \text{if the constraints 3.40, 3.41, and 3.42 are met} \\ 0 & \text{otherwise} \end{cases} \tag{3.43}$$

Let $\hat{\mu}_{1RB} = E[\hat{\mu}_1 | D]$.

THEOREM 3.3.

$$\hat{\mu}_{1RB} = \frac{1}{Hn}\{\sum_{f_1}\cdots\sum_{f_{\kappa+1}}[\prod_{k=1}^{\kappa+1}\binom{x_k}{f_k}]\sum_{k=1}^{\kappa}\frac{f_k(y_k^*)}{x_k} + f_{\kappa+1}(y_{\kappa+1}^*)\} \tag{3.44}$$

Given the constraints in equations 3.40, 3.41, and 3.42

$H_i$ will be defined as the number of ways $i$ sample edge units can be selected given $D$ and zero if $i$ is not possible given $D$. Then:

$$H_i = I(i)\binom{x_{\kappa+1}}{i}\sum_{f_1}\cdots\sum_{f_\kappa}\prod_{k=1}^{\kappa}\binom{x_k}{f_k} \tag{3.45}$$

Given the constraints in equations 3.40, 3.41, and 3.42

Let the indicator variable $I(i)$ equal one if it is possible to have an initial sample containing $i$ sample edge units and obtain the final sample and zero otherwise. So,

$$I(i) = \begin{cases} 1 & \text{if the constraints on } f_k \text{ for } 1 \le k \le \kappa \\ & \text{can be met given } f_{\kappa+1} = i \\ 0 & \text{otherwise} \end{cases} \tag{3.46}$$

**Note:** $\sum_{i=0}^{e_s} H_i = H$

Let $\hat{\mu}_{2RB} = E[\hat{\mu}_2|D]$.

THEOREM 3.4.

$$\hat{\mu}_{2RB} = \frac{1}{N}\sum_{k=1}^{\kappa}\frac{y_k^*}{\alpha_k} + \sum_{i=0}^{e_s}\frac{i\bar{y}_e H_i}{nH} \tag{3.47}$$

### 3.4.2 Variance Formulae for Any Unbiased Estimators

Let $\hat{\mu}$ represent either $\hat{\mu}_1$ or $\hat{\mu}_2$ and $\hat{\mu}_{RB}$ represent either $\hat{\mu}_{1RB}$ or $\hat{\mu}_{2RB}$.

$$var(\hat{\mu}_{RB}) = var(\hat{\mu}) - E[(\hat{\mu} - \hat{\mu}_{RB})^2] \tag{3.48}$$

Let $H(d)$ be the number of initial samples compatible with $d$ and $P(d)$ be the probability that $D = d$. The variance of $\hat{\mu}_{1RB}$ is

$$var(\hat{\mu}_{1RB}) = \frac{N-n}{Nn(N-1)}\sum_{i=1}^{N}(w_i - \mu)^2$$

$$
\begin{aligned}
&- \sum_{d \in \mathcal{D}} \frac{P(d)}{H(d)} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d) \times \\
&\Big[ \sum_{i \in s_0^{'}} \frac{1}{n} w_i - \frac{1}{Hn} \{ \sum_{f_1} \cdots \sum_{f_{\kappa+1}} \Big[ \prod_{k=1}^{\kappa+1} \binom{x_k}{f_k} \Big] \sum_{k=1}^{\kappa} \frac{f_k(y_k^*)}{x_k} + f_{\kappa+1}(y_{\kappa+1}^*) \} \Big]^2
\end{aligned}
\tag{3.49}
$$

An unbiased estimate of the variance of $\hat{\mu}_{1RB}$ when sampling is done without replacement is given by

$$
\begin{aligned}
\widetilde{var}(\hat{\mu}_{1RB}) &= \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu}_1)^2 \\
&\quad - \frac{1}{H} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d) \times \\
&\quad \Big[ \sum_{i \in s_0^{'}} \frac{1}{n} w_i - \frac{1}{Hn} \{ \sum_{f_1} \cdots \sum_{f_{\kappa+1}} \Big[ \prod_{k=1}^{\kappa+1} \binom{x_k}{f_k} \Big] \sum_{k=1}^{\kappa} \frac{f_k(y_k^*)}{x_k} + f_{\kappa+1}(y_{\kappa+1}^*) \} \Big]^2
\end{aligned}
\tag{3.50}
$$

However a more efficient estimator is

$$
\begin{aligned}
\widehat{\widetilde{var}}(\hat{\mu}_{1RB}) &= E[\widetilde{var}(\hat{\mu}_1)|d] \\
&= \frac{1}{H} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d) \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu}_1)^2 \\
&\quad - \frac{1}{H} \sum_{s_0^{'} \in \mathcal{S}} I(s_0^{'}, d) \times \\
&\quad \Big[ \sum_{i \in s_0^{'}} \frac{1}{n} w_i - \frac{1}{Hn} \{ \sum_{f_1} \cdots \sum_{f_{\kappa+1}} \Big[ \prod_{k=1}^{\kappa+1} \binom{x_k}{f_k} \Big] \sum_{k=1}^{\kappa} \frac{f_k(y_k^*)}{x_k} + f_{\kappa+1}(y_{\kappa+1}^*) \} \Big]^2
\end{aligned}
$$

(3.51)

The variance of $\hat{\mu}_{2RB}$ is

$$\begin{aligned} var(\hat{\mu}_{2RB}) &= \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^*(\alpha_{kh}-\alpha_k\alpha_k)}{\alpha_k\alpha_k} \\ &\quad -\sum_{d\in\mathcal{D}}\frac{P(d)}{H(d)}\sum_{s_0^{'}\in\mathcal{S}}^{'} I(s_0^{'},d)\times \\ &\quad \Big[\sum_{i\in s_0^{'},e_i=1}^{'}\frac{y_i}{n}-\sum_{i=1}^{e_s}\frac{i\bar{y}_e H_i}{nH}\Big]^2 \end{aligned} \tag{3.52}$$

An unbiased estimator of this variance is

$$\begin{aligned} \widetilde{var}(\hat{\mu}_{2RB}) &= \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^* z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \\ &\quad -\frac{1}{H}\sum_{s_0^{'}\in\mathcal{S}}^{'} I(s_0^{'},d)\times \\ &\quad \Big[\sum_{i\in s_0^{'},e_i=1}^{'}\frac{y_i}{n}-\sum_{i=1}^{e_s}\frac{i\bar{y}_e H_i}{nH}\Big]^2 \end{aligned} \tag{3.53}$$

A more efficient estimator of the variance is

$$\begin{aligned} \widehat{var}(\hat{\mu}_{2RB}) &= E[\widetilde{var}(\hat{\mu}_{2RB})|d] \\ &= \frac{1}{H}\sum_{s_0^{'}\in\mathcal{S}}^{'} I(s_0^{'},d)\frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y_k^* y_h^* z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \end{aligned}$$

$$-\frac{1}{H}\sum_{s_0^{'}\in\mathcal{S}} I(s_0^{'},d)\times$$

$$\Big[\sum_{i\in s_0^{'},e_i=1}\frac{y_i}{n}-\sum_{i=1}^{e_s}\frac{i\bar{y}_e H_i}{nH}\Big]^2$$

(3.54)

## 3.5 An Illustrative Example

The following example serves to illustrate the computation of the new estimators for given samples and shows, for a small population, the relative properties of the different types of estimators. The population consists of $N = 6$ units. The initial sample is a simple random sample of $n = 2$ units. Neighboring (adjacent) units are added whenever the condition $y_i \geq 150$ is satisfied. Table 1 lists basic quantities for every unit and network in the population. Table 2 lists every possible sample and the value of each estimator for that sample with variances of the estimators given at the bottom of the table.

Table 3.1: The first line is the unit labels and the second line is their associated values. The population consists of the six units. The following lines of the table are necessary components for calculating various estimators in adaptive cluster sampling, with n=2 and a condition of $y_i \geq 150$.

| unit $i$ | = | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $y_i$ | = | 2 | 150 | 151 | 146 | 1 | 0 |
| $m_i$ | = | 1 | 2 | 2 | 1 | 1 | 1 |
| $w_i$ | = | 2 | $\frac{150+151}{2} = 150.5$ | $\frac{150+151}{2} = 150.5$ | 146 | 1 | 0 |
| network #k | = | 1 | 2 | 2 | 3 | 4 | 5 |
| $x_k$ | = | 1 | 2 | 2 | 1 | 1 | 1 |
| $y_k^*$ | = | 2 | 150+151=301 | 150+151=301 | 146 | 1 | 0 |
| $\alpha_k$ | = | 1/3 | $1 - \binom{4}{2} \div \binom{6}{2} = 3/5$ | $1 - \binom{4}{2} \div \binom{6}{2} = 3/5$ | 1/3 | 1/3 | 1/3 |

Table 3.2: This table consists of all possible initial samples and a few possible associated etimates of $\mu$. In the first column numbers before the semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units.

| The Sample | $\hat{\mu}_1$ | $\hat{\mu}_{1+}$ | $\hat{\mu}_{1RB}$ | $\hat{\mu}_2$ | $\hat{\mu}_{2+}$ | $\hat{\mu}_{2RB}$ |
|---|---|---|---|---|---|---|
| 2,150;151,146 | 76.250 | 112.250 | 119.900 | 84.61111 | 120.61111 | 113.21111 |
| 2,151;150,146 | 76.250 | 112.250 | 119.900 | 84.61111 | 120.61111 | 113.21111 |
| 2,146 | 74.000 | 74.000 | 74.000 | 74.00000 | 74.00000 | 74.00000 |
| 2,1 | 1.500 | 1.500 | 1.500 | 1.50000 | 1.50000 | 1.50000 |
| 2,0 | 1.000 | 1.000 | 1.000 | 1.00000 | 1.00000 | 1.00000 |
| 150,151;146,2 | 150.500 | 150.500 | 119.900 | 83.61111 | 83.61111 | 113.21111 |
| 150,146;151,2 | 148.250 | 112.250 | 119.900 | 156.61111 | 120.61111 | 113.21111 |
| 150,1;151,146,2 | 75.750 | 75.750 | 75.750 | 84.11111 | 84.11111 | 84.11111 |
| 150,0;151,146,2 | 75.250 | 75.250 | 75.250 | 83.61111 | 83.61111 | 83.61111 |
| 151,146;150,2 | 148.250 | 112.250 | 119.900 | 156.61111 | 120.61111 | 113.21111 |
| 151,1;150,146,2 | 75.750 | 75.750 | 75.750 | 84.11111 | 84.11111 | 84.11111 |
| 151,0;150,146,2 | 75.250 | 75.250 | 75.250 | 83.61111 | 83.61111 | 83.61111 |
| 146,1 | 73.500 | 73.500 | 73.500 | 73.50000 | 73.50000 | 73.50000 |
| 146,0 | 73.000 | 73.000 | 73.000 | 73.00000 | 73.00000 | 73.00000 |
| 1,0 | 0.500 | 0.500 | 0.500 | 0.50000 | 0.50000 | 0.50000 |
| MEAN | 75.000 | 75.000 | 75.000 | 75.00000 | 75.00000 | 75.00000 |
| BIAS | 0 | 0 | 0 | 0 | 0 | 0 |
| MSE | 2191.433 | 1845.833 | 1767.803 | 2021.98148 | 1676.38148 | 1603.36815 |

Using the information from table 3.1 it will be shown how to calculate the various estimators in table 3.2 for an initial sample of 151 and 146 and thus a final sample of 151,146;150,2.

$$\hat{\mu}_1 = \frac{1}{n}(w_3 + w_4) = \frac{1}{2}(150.5 + 146) = 148.25 \tag{3.55}$$

$$\bar{y}_e = \frac{1}{e_s}(y_1 + y_4) = \frac{1}{2}(2 + 146) = 74.0 \tag{3.56}$$

$$\hat{\mu}_{1+} = \frac{1}{n}(w_3^{'} + w_4^{'}) = \frac{1}{2}(150.5 + 74.0) = 112.25 \tag{3.57}$$

$$\begin{aligned} \hat{\mu}_{1RB} &= \frac{1}{5(2)}(\binom{2}{1}\binom{2}{1}[((1)74.0 + (1)150.5) + \\ &\quad \binom{2}{0}\binom{2}{2}((0)74.0 + (2)150.5)] \\ &= 119.90 \end{aligned} \tag{3.58}$$

$$\hat{\mu}_2 = \frac{1}{N}(y_2^*/\alpha_2 + y_3^*/\alpha_3) = \frac{1}{6}(301/(3/5) + 146/(1/3)) = 156.61111 \tag{3.59}$$

$$\hat{\mu}_{2+} = \frac{1}{N}(y_2^{'}/\alpha_2 + y_3^{'}/\alpha_3) = \frac{1}{6}(301/(3/5) + 74.0/(1/3)) = 120.61111 \tag{3.60}$$

$$\begin{aligned}\hat{\mu}_{2RB} &= \frac{1}{5(6)}(\binom{2}{1}\binom{2}{1}[((1)74.0/(1/3)+301/(3/5))+ \\ &\quad \binom{2}{0}\binom{2}{2}((0)74.0/(1/3)+301/(3/5))] \\ &= 113.21111\end{aligned} \tag{3.61}$$

Table 3.3: This table consists of all possible initial samples and there variances. In the first column numbers before the semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units.

| The Sample | $\widehat{var}(\hat{\mu}_1)$ | $\widehat{var}(\hat{\mu}_{1+})$ | $\widehat{var}(\hat{\mu}_{1RB})$ |
|---|---|---|---|
| 2,150;151,146 | 3675.3750000 | 543.3750000 | 200.6100000 |
| 2,151;150,146 | 3675.3750000 | 543.3750000 | 200.6100000 |
| 2,146 | 3456.0000000 | 3456.0000000 | 3456.0000000 |
| 2,1 | 0.1666667 | 0.1666667 | 0.1666667 |
| 2,0 | 0.6666667 | 0.6666667 | 0.6666667 |
| 150,151;146,2 | 0.0000000 | 0.0000000 | 200.6100000 |
| 150,146;151,2 | 3.3750000 | 543.3750000 | 200.6100000 |
| 150,1;151,146,2 | 3725.0416667 | 3725.0416667 | 3725.0416667 |
| 150,0;151,146,2 | 3775.0416667 | 3775.0416667 | 3775.0416667 |
| 151,146;150,2 | 3.3750000 | 543.3750000 | 200.6100000 |
| 151,1;150,146,2 | 3725.0416667 | 3725.0416667 | 3725.0416667 |
| 151,0;150,146,2 | 3775.0416667 | 3775.0416667 | 3775.0416667 |
| 146,1 | 3504.1666667 | 3504.1666667 | 3504.1666667 |
| 146,0 | 3552.6666667 | 3552.6666667 | 3552.6666667 |
| 1,0 | 0.1666667 | 0.1666667 | 0.1666667 |
| MEAN=MSE | 2191.4333333 | 1845.8333333 | 1767.8033333 |

Table 3.4: This table consists of all possible initial samples and there variances. In the first column numbers before the semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units.

| The Sample | $\widehat{var}(\hat{\mu}_2)$ | $\widehat{var}(\hat{\mu}_{2+})$ | $\widehat{var}(\hat{\mu}_{2RB})$ |
|---|---|---|---|
| 2,150;151,146 | 2713.3827160 | 183.3827160 | 486.9316049 |
| 2,151;150,146 | 2713.3827160 | 183.3827160 | 486.9316049 |
| 2,146 | 3456.0000000 | 3456.0000000 | 3456.0000000 |
| 2,1 | 0.1666667 | 0.1666667 | 0.1666667 |
| 2,0 | 0.6666667 | 0.6666667 | 0.6666667 |
| 150,151;146,2 | 2796.3271605 | 2796.3271605 | 486.9316049 |
| 150,146;151,2 | 245.3827160 | 183.3827160 | 486.9316049 |
| 150,1;151,146,2 | 2754.6882716 | 2754.6882716 | 2754.6882716 |
| 150,0;151,146,2 | 2796.3271605 | 2796.3271605 | 2796.3271605 |
| 151,146;150,2 | 245.3827160 | 183.3827160 | 486.9316049 |
| 151,1;150,146,2 | 2754.6882716 | 2754.6882716 | 2754.6882716 |
| 151,0;150,146,2 | 2796.3271605 | 2796.3271605 | 2796.3271605 |
| 146,1 | 3504.1666667 | 3504.1666667 | 3504.1666667 |
| 146,0 | 3552.6666667 | 3552.6666667 | 3552.6666667 |
| 1,0 | 0.1666667 | 0.1666667 | 0.1666667 |
| MEAN=MSE | 2021.98148 | 1676.3814814 | 1603.3681481 |

### 3.6 Simulations on Blue-Winged Teal Data

Simulations were performed on a real data set, Blue-Winged Teal Data (Smith *et al* 1995.) For each estimator 100,000 iterations were performed. A large number of iterations were used to most accurately estimate the true population variances of the estimators. Varying initial sample sizes were used. The same sample is used to calculate $\hat{\mu}$, $\hat{\mu}_+$, and $\hat{\mu}_{RB}$. Note, certain estimators were not calculated for when the condition is $y_i \geq 1$. The reason for this is that when the edge units are equal to zero the estimates are the same for several of the estimators. The formula used to estimate the variance is,

$$\widehat{var}(\hat{\mu}) = \frac{1}{100000 - 1} \sum_{i=1}^{100000} (\hat{\mu}_i - \bar{\mu}_h)^2 \tag{3.62}$$

Where $\hat{\mu}_i$ is the value for the relevant estimator for sample $i$ and

$$\bar{\mu}_h = \frac{1}{100000} \sum_{i=1}^{100000} \hat{\mu}_i \tag{3.63}$$

Table 3.5: blue-winged teal data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

Table 3.6: Results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_2)$ |
|---|---|---|---|---|
| 2 | 9.498920 | 234198.925180 | 224050.394832 | 220132.678734 |
| 5 | 19.466130 | 88334.507223 | 70431.757457 | 65833.976613 |
| 8 | 25.929880 | 51798.320718 | 33268.614156 | 29350.329462 |
| 10 | 29.049650 | 39605.444030 | 21381.643815 | 18118.724883 |
| 12 | 31.449190 | 31400.107040 | 14339.048320 | 11520.711107 |
| 15 | 34.223620 | 23170.593490 | 7928.127603 | 5725.659562 |
| 30 | 42.073920 | 6595.829140 | 782.311075 | 78.392251 |

Table 3.7: Results for the simulations on blue-winged teal data for $\hat{\mu}_{1s}$. Condition is $y_i \geq 20$

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1+})$ | $var(\hat{\mu}_{1RB})$ |
|---|---|---|---|---|
| 2 | 2.995570 | 873905.782164 | 873907.804405 | 868496.745805 |
| 5 | 7.260130 | 332174.993603 | 332173.691731 | 318413.034444 |
| 8 | 11.239110 | 192867.705128 | 192867.958692 | 178962.022980 |
| 10 | 13.761990 | 148384.829961 | 148385.521244 | 133216.518064 |
| 12 | 16.170850 | 117110.194664 | 117110.742053 | 102616.044688 |
| 15 | 19.632990 | 86156.798786 | 86156.666489 | 72376.536511 |
| 30 | 34.455060 | 24590.044883 | 24589.896351 | 14711.449196 |

Table 3.8: Results for the simulations on blue-winged teal data for $\hat{\mu}_{2s}$. Condition is $y_i \geq 20$

| $n$ | $E(\nu)$ | $var(\hat{\mu}_2)$ | $var(\hat{\mu}_{2+})$ | $var(\hat{\mu}_{2RB})$ |
|---|---|---|---|---|
| 2 | 2.995570 | 868445.385721 | 868447.431919 | 868446.283417 |
| 5 | 7.260130 | 317630.713901 | 317629.403346 | 317629.521898 |
| 8 | 11.239110 | 178101.521884 | 178102.081161 | 178101.731110 |
| 10 | 13.761990 | 132222.976874 | 132224.438263 | 132224.686988 |
| 12 | 16.170850 | 101512.176420 | 101512.753629 | 101513.019836 |
| 15 | 19.632990 | 71246.423619 | 71246.280207 | 71246.406654 |
| 30 | 34.455060 | 13862.738143 | 13862.456319 | 13862.413168 |

## 3.7 Comments

From the simulations it can be seen that for $\hat{\mu}_1$ a large reduction in the variance can be achieved by conditioning on the minimal sufficient statistic. For $\hat{\mu}_2$, there is no improvement when the condition is $y \geq 1$ and generally not much improvement for other criteria either. The new estimators and the fully Rao-Blackwellized $\hat{\mu}_2$ often offer little improvement. In order to illustrate why this is so and when the latter estimators are most appropriate consider the condition $y \geq 20$ for the blue-winged teal data. In that situation the edge units can range from 0 to 19 while the units that meet the condition may range from 20 to 13639. From this it is clear that the main contribution to reducing the variance comes from within networks meeting the condition, not the edge units. In situations where the variable of interest is not as directly correlated with the condition the new estimators and the fully rao-blackwellized $\hat{\mu}_2$ are estimators that have the potential to be considerably more efficient than their ordinary counterparts.

# Chapter 4

# Sampling Without Replacement of Clusters

## 4.1 Introduction

In the simplest form of adaptive cluster sampling an initial sample of $n$ units is selected by random sampling without replacement and, whenever the variable of interest for a unit in the sample satisfies a prespecified condition, neighboring units are added to the sample until no more units are found that meet the criterion (Thompson 1990). Even though the initial sample is selected without replacement, some units in the sample may be selected more than once because the initial sample may contain more than one unit in a given network of units satisfying the condition. A network and its associated edge units are called a cluster. An adaptive cluster sample can also be taken without replacement of networks (Salehi and Seber 1997), by selecting each unit of the initial sample at random from the population exclusive of networks already containing an initially selected unit. Even with this procedure, however, a unit which was previously added as an edge unit may be subsequently selected in the initial sample. In this chapter we describe a design in which clusters, rather than just networks, are selected without replacement.

## 4.2 Designs and Terminology

**Sampling without replacement of networks** (Salehi and Seber 1997). In this design, an initial unit is selected at random from the population and , if its $y$-value satisfies the

condition, its associated network and edge units are observed. The second initial unit is selected at random from the population exclusive of the units in the network already observed. In turn, each of the $n$ units is selected from the population exclusive of previously observed networks of units.

Although sampling is without replacement of networks, repeat observations may occur in the data. There are three ways a repeat observation may occur. First, an edge unit of more than one network may be observed more than once. The second way a repeat observation may occur is if a unit selected in the initial sample subsequently turns up as an edge unit when an adjacent network is selected. Finally, a unit observed as an edge unit may be selected in the sample.

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

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

## 4.3 Estimators

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

$$\hat{\mu}_{DR} = \frac{1}{Nn} \sum_{i=1}^{n} z_i, \tag{4.1}$$

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

$$z_i = \sum_{j=1}^{i-1} y_j + \frac{(1 - \sum_{j=1}^{i-1} p_j) y_i}{p_i} \tag{4.2}$$

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

As shown by Raj (1956), the conditional expectation of each $z_i$ given the initial $i - 1$ observations is the population total, so that, unconditionally, $z_i/N$ is an unbiased estimator of the population mean, for $i = 1, ..., N$. Thus their average, the Raj estimator, is an unbiased estimator of the population mean. Raj showed that this strategy would always have variance less than or equal to the variance of the Hansen-Hurwitz estimator used with sampling with replacement. Also shown by Raj is that the $z_i$ are uncorrelated, which is useful for an expression of the variance and developing a variance estimator. The following is a proof that $cov(z_i, z_j) = 0$. Let $j > i$ and let $x = ((y_1, p_1)..., (y_{j-1}, p_{j-1}))$

$$\begin{aligned} cov(z_i, z_j) &= E\left[cov(z_i, z_j \mid x)\right] + cov(E\left[z_i \mid x\right], E\left[z_j \mid x\right]) \\ &= E\left[cov(\text{a constant}, z_j)\right] + cov(z_i, \mu) \\ &= E\left[0\right] + 0 \end{aligned}$$

$$= \quad 0$$

(4.3)

With the adaptive designs of this paper, the conditional selection probabilities are not known for every sampled unit, and so a modification of the Raj estimator is needed. In addition, selection probability is in part determined by the size of a network whereas before each unit was given a selection probability.

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

For the design in which networks are selected without replacement, let $z_1^* = y_{1.}/p_1^*$, and, for $i = 2, 3, ..., n$,

$$z_i^* = \sum_{j=1}^{i-1} y_{j.} + \frac{(1 - \sum_{j=1}^{i-1} p_j^*) y_{i.}}{p_i^*}, \tag{4.4}$$

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

$$\hat{\mu}_{Net} = \frac{1}{Nn} \sum_{i=1}^{n} z_i^*. \tag{4.5}$$

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

$$var(\hat{\mu}_{Net}) = \frac{1}{N^2 n^2} \sum_{i=1}^{n} var(z_i^*) \tag{4.6}$$

An unbiased estimate of the variance is:

$$\widehat{var}(\hat{\mu}_{Net}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\frac{z_i^*}{N} - \hat{\mu}_{Net})^2 \tag{4.7}$$

When the design is selection without replacement of clusters, the relevant probabilities $p_i^*$ are unchanged, but the $z_i$ are $z_1 = y_{1.}/p_1^*$, and, for $i = 2, 3, ..., n$,

$$z_i = \sum_{j \in s_{i-1}} y_{j.} + \frac{(1 - \sum_{j \in s_{i-1}} p_j^*) y_{i.}}{p_i^*}, \tag{4.8}$$

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

$$\hat{\mu}_{Clust} = \frac{1}{Nn} \sum_{i=1}^{n} z_i, \tag{4.9}$$

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

$$var(\hat{\mu}_{Clust}) = \frac{1}{N^2 n^2} \sum_{i=1}^{n} var(z_i) \tag{4.10}$$

An unbiased estimate of the variance is:

$$\widehat{var}(\hat{\mu}_{Clust}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\frac{z_i}{N} - \hat{\mu}_{Clust})^2 \tag{4.11}$$

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

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

$$e_i = \begin{cases} 1 & \text{if } i \text{ is in the initial sample.} \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

$$d^* = \{(i, e_i, y_i) : i \in s\} \tag{4.13}$$

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

$$I(s_o^{'}, d^*) = \begin{cases} 1 & \text{if} \quad g(s_o^{'}) = d^* \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

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

$$\hat{\mu}_{Murthy} = \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*) \hat{\mu}(s_o) \Big/ \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*) \tag{4.15}$$

Unfortunately, for without replacement of clusters equation 4.15 cannot be reduced but for without replacement of networks it reduces to the estimator given in Salehi and Seber (1997):

$$\hat{\mu}_{MNet} = \frac{1}{N}\sum_{i=1}^{n}\frac{P(s|i)}{P(s)}y_{i.} \tag{4.16}$$

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

$$w_i = \frac{1}{m_i}\sum_{j\in\psi_i} y_j \tag{4.17}$$

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

$$var(\hat{\mu}_{MNet}) = \frac{1}{N^2}\sum_{i=1}^{K}\sum_{j<i}^{K} m_i m_j\big[1 - \sum_{s\ni i,j}\frac{P(s|i)P(s|j)}{P(s)}\big](w_i - w_j)^2 \tag{4.18}$$

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

$$\widehat{var}(\hat{\mu}_{MNet}) = \frac{1}{N^2P(s)^2}\big[\sum_{i=1}^{n}\sum_{j<i}^{n} m_i m_j (w_i - w_j)^2\{P(s)P(s|i,j) - P(s|i)P(s|j)\}\big] \tag{4.19}$$

Equation 4.15 for without replacement of clusters can also be viewed as

$$\hat{\mu}_{MClust} = E[\hat{\mu}_{Clust}|d^*] \tag{4.20}$$

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

$$\begin{aligned} var(\hat{\mu}_{MClust}) &= var(\hat{\mu}_{Clust}) - E[(\hat{\mu}_{Clust} - \hat{\mu}_{MClust})^2] \\ &= \frac{1}{N^2 n^2}\sum_{i=1}^{n} var(z_i) - \sum_{s_o \in \mathcal{S}} P(s_o)\big[\hat{\mu}_{Clust}(s_o) - \hat{\mu}_{MClust}(s_o)\big]^2 \end{aligned} \tag{4.21}$$

An unbiased estimate of the variance is:

$$\begin{aligned} \widetilde{var}(\hat{\mu}_{MClust}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{z_i}{N} - \hat{\mu}_{Clust})^2 \\ &\quad - \frac{1}{\sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*)} \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*)\big[\hat{\mu}_{Clust}(s_o) - \hat{\mu}_{MClust}\big]^2 \end{aligned} \tag{4.22}$$

A more efficient estimator of the variance is

$$\begin{aligned} \widehat{var}(\hat{\mu}_{MClust}) &= E[\widetilde{var}(\hat{\mu}_{MClust})|d^*] \\ &= \frac{1}{\sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*)} \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*) \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{z_i}{N} - \hat{\mu}_{Clust})^2 \\ &\quad - \frac{1}{\sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*)} \sum_{s_o \in \mathcal{S}} P(s_o) I(s_o^{'}, d^*)\big[\hat{\mu}_{Clust}(s_o) - \hat{\mu}_{MClust}\big]^2 \end{aligned} \tag{4.23}$$

The estimators $\hat{\mu}_{MClust}$ and $\hat{\mu}_{MNet}$ may be hard to compute for large $n$. The number of different permutations we have to compute is $n!$ for equation 4.15. The number of calculations can be reduced by noting that certain sets of permutations give the same

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

## 4.4 An Illustrative Example

The following example serves to illustrate the computation of the new estimators for given samples and shows, for a small population, the relative properties of the different types of estimators. The population is from Salehi and Seber (1997). The population consists of $N = 7$ units. The initial sample is a simple random sample of $n = 2$ units. Neighboring (adjacent) units are added whenever the condition $y_i \geq 5$ is satisfied. Table 1 lists basic quantities for every unit and network in the population. Tables 2 and 3 lists every possible sample and the value of each estimator for that sample with variances of the estimators given at the bottom of the table.

Table 4.1: The first line is the unit labels and the second line is their associated values. The population consists of the seven units. The following lines of the table are necessary components for calculating various estimators in adaptive cluster sampling, with n=2 and a condition of $y_i \geq 5$.

| unit $i$ | = | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $y_i$ | = | 5 | 130 | 1 | 0 | 2 | 7 | 120 |
| $y_{i.}$ | = | 135 | 135 | 1 | 0 | 2 | 127 | 127 |
| $y'_{i.}$ | = | 136 | 136 | 1 | 0 | 2 | 129 | 129 |
| $m_i$ | = | 2 | 2 | 1 | 1 | 1 | 2 | 2 |
| network #k | = | 1 | 1 | 2 | 3 | 4 | 5 | 5 |

Table 4.2: **An Example of without replacement of networks (Salehi and Seber 1997).** This table consists of all possible initial samples and a few possible associated estimates of $\mu$. In the first column numbers before the semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units.

| The Sample | p | $\nu$ | $\hat{\mu}_{Net}$ | $\hat{\mu}_{MNet}$ | $\hat{\mu}_{RBN}$ |
|---|---|---|---|---|---|
| 135,1 | 2/35 | 3 | 43.75 | 37.27 | 37.27 |
| 1,135 | 1/21 | 3 | 29.50 | 37.27 | 37.27 |
| 135,0;1 | 2/35 | 4 | 43.39 | 36.82 | 36.82 |
| 0,135;1 | 1/21 | 4 | 28.93 | 36.82 | 36.82 |
| 135,2;1 | 2/35 | 4 | 44.11 | 37.73 | 37.73 |
| 2,135;1 | 1/21 | 4 | 30.07 | 37.73 | 37.73 |
| 135,127;1,2 | 4/35 | 6 | 66.07 | 65.50 | 65.50 |
| 127,135;1,2 | 4/35 | 6 | 64.93 | 65.50 | 65.50 |
| 1,0 | 1/42 | 2 | 0.57 | 0.50 | 0.50 |
| 0,1 | 1/42 | 2 | 0.43 | 0.50 | 0.50 |
| 1,2 | 1/42 | 2 | 1.43 | 1.50 | 1.50 |
| 2,1 | 1/42 | 2 | 1.57 | 1.50 | 1.50 |
| 1,127;2 | 1/21 | 4 | 27.87 | 35.09 | 35.09 |
| 127,1;2 | 2/35 | 4 | 41.18 | 35.09 | 35.09 |
| 0,127;2 | 1/21 | 4 | 27.21 | 34.64 | 34.64 |
| 127,0;2 | 2/35 | 4 | 40.82 | 34.64 | 34.64 |
| 0,2 | 1/42 | 2 | 0.86 | 1.00 | 1.00 |
| 2,0 | 1/42 | 2 | 1.14 | 1.00 | 1.00 |
| 2,127 | 1/21 | 3 | 28.36 | 35.55 | 35.55 |
| 127,2 | 2/35 | 3 | 41.54 | 35.55 | 35.55 |
| Mean | | 3.96 | 37.86 | 37.86 | 37.86 |
| Variance | | | 401.23 | 371.34 | 371.34 |

Table 4.3: **An Example of without replacement of clusters.** This table consists of all possible initial samples and a few possible associated etimates of $\mu$. In the first column numbers before the semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units.

| The Sample | p | $\nu$ | $\hat{\mu}_{Clust}$ | $\hat{\mu}_{MClust}$ | $\hat{\mu}_{RBC}$ |
|---|---|---|---|---|---|
| 135,1 | 0 | — | — | — | — |
| 1,135 | 1/21 | 3 | 29.50 | 29.50 | 29.50 |
| 135,0;1 | 1/14 | 4 | 43.46 | 37.65 | 37.65 |
| 0,135;1 | 1/21 | 4 | 28.93 | 37.65 | 37.65 |
| 135,2;1 | 1/14 | 4 | 44.04 | 38.45 | 38.45 |
| 2,135;1 | 1/21 | 4 | 30.07 | 38.45 | 38.45 |
| 135,127;1,2 | 1/7 | 6 | 61.61 | 60.93 | 60.93 |
| 127,135;1,2 | 1/7 | 6 | 60.25 | 60.93 | 60.93 |
| 1,0 | 1/42 | 2 | 0.57 | 0.50 | 0.50 |
| 0,1 | 1/42 | 2 | 0.43 | 0.50 | 0.50 |
| 1,2 | 1/42 | 2 | 1.43 | 1.50 | 1.50 |
| 2,1 | 1/42 | 2 | 1.57 | 1.50 | 1.50 |
| 1,127;2 | 1/21 | 4 | 27.79 | 35.69 | 35.69 |
| 127,1;2 | 1/14 | 4 | 40.96 | 35.69 | 35.69 |
| 0,127;2 | 1/21 | 4 | 27.21 | 35.64 | 35.64 |
| 127,0;2 | 1/14 | 4 | 41.25 | 35.64 | 35.64 |
| 0,2 | 1/42 | 2 | 0.86 | 1.00 | 1.00 |
| 2,0 | 1/42 | 2 | 1.14 | 1.00 | 1.00 |
| 2,127 | 1/21 | 3 | 28.36 | 28.36 | 28.36 |
| 127,2 | 0 | — | — | — | — |
| Mean | | 4.19 | 37.86 | 37.86 | 37.86 |
| Variance | | | 377.32 | 354.99 | 354.99 |

## 4.5 Simulations on Blue-Winged Teal Data

Simulations were performed on the data in table 4.4. Estimates of the population mean were calculated 100,000 times for $n = 2$ and 20,000 times for $n = 5, 8, 10, 12$ for the estimation of variance and effective sample size in tables 4.5 and 4.6. To generate the estimates in table 4.6 a progam was written in C. The formula used to estimate the variance is,

$$\widehat{var}(\hat{\mu}) = \frac{1}{100000 - 1} \sum_{i=1}^{100000} (\hat{\mu}_i - \bar{\mu}_h)^2 \tag{4.24}$$

Where $\hat{\mu}_i$ is the value for the relevant estimator for sample $i$ and

$$\bar{\mu}_h = \frac{1}{100000} \sum_{i=1}^{100000} \hat{\mu}_i \tag{4.25}$$

Table 4.4: blue-winged teal data (Smith *et al.* 1995)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

Table 4.5: Simulation for blue-winged teal data. Without replacement of Networks (Salehi and Seber 1997).

| $n$ | $E(\nu)$ | $var(\hat{\mu}_{Net})$ | $var(\hat{\mu}_{MNet})$ |
|---|---|---|---|
| 2 | 9.6350 | 220909.11 | 218680.59 |
| 5 | 20.1873 | 67635.53 | 61622.01 |
| 8 | 27.0585 | 33021.83 | 25806.58 |
| 10 | 30.4778 | 22575.50 | 15270.38 |
| 12 | 33.1112 | 16194.88 | 8891.77 |

Table 4.6: Simulation for blue-winged teal data. Without Replacement of Clusters with exact computation of the estimators.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_{Clust})$ | $var(\hat{\mu}_{MClust})$ |
|---|---|---|---|
| 2 | 9.798 | 216169.50 | 211527.20 |
| 5 | 21.316 | 62452.13 | 54793.68 |
| 8 | 29.393 | 29426.98 | 21767.05 |
| 10 | 33.364 | 19646.04 | 12825.34 |
| 12 | 36.584 | 13953.65 | 8074.71 |

# Chapter 5

# Performance of Estimators for Sampling Without Replacement of Units in a Multivariate Setting

## 5.1 Introduction

Often when data is collected more than one variable is of interest. The same can be true when data is collected by adaptive cluster sampling. The condition may be set according to the primary variable of interest or an auxiliary variable but data on others variables of interest would be collected for the units observed. There has been a significant amount of research on how the Hansen-Hurwitz and Horvitz-Thompson type estimators perform. Generally it is found that the Horvitz-Thompson estimator performs better than the Hansen-Hurwitz estimator. Most if not all of this research was performed when a value of the variable of interest was chosen as the condition. For example, if the main variable of interest is frogs, the condition might be finding a single frog in the quadrant. Pretend now that the main purpose of the study was to collect data on frogs, but of also great concern was information on lizards and flies. Thus sampling might be done adaptively according to the spotting of a frog but data would be collected on all three of these species and each species would be of interest. In this case the Hansen-Hurwitz estimator might outperform the Horvitz-Thompson estimator when estimating the prevalence of a species other than frogs especially if the prevalence of the species is inversely related. This particular setting will be discussed in this chapter.

## 5.2 Estimators in Adaptive Cluster Sampling, Multivariate Setting

As in the typical finite population sampling situation, the population consists of $N$ units labeled $1, 2, .., N$ but in the multivariate setting each unit $i$ will have $p$ measurements associated with it. Thus there may be up to $p$ variables of interest and $\mathbf{y_i} = (y_{i1}, y_{i2}, ..., y_{ip})'$. Now we are concerned with a population matrix $N \times p$ instead of a population vector of size $N$, as in the univariate case. The population matrix will be considered fixed but unknown constants. The population mean for the $l^{th}$ measurement will be denoted,

$$\mu_l = \frac{1}{N} \sum_{i=1}^{N} y_{il} \tag{5.1}$$

This chapter will cover the estimation of the population mean for one of the measurements when the sample is taken adaptively according to another measurement.

### 5.2.1 Ordinary Estimators Redefined for the Multivariate Case

The notation for the univariate case and multivariate case for the ordinary estimators is very similar. The units selected in the initial sample are still denoted by $s_0$ and the units in the final sample by $s$. The set $s_0$ is the set of unit labels obtained in the initial sample and $s$ is the set of distinct unit labels in the final sample. Let $n$ denote the initial sample of size and $\nu$ the final sample size. Let $\psi_i$ denote the network which includes unit $i$ and $m_i$ the number of units in that network. The latter notation is the same and has the same value regardless of which measurement for unit $i$ we are presently concerned with. The only difference in multivariate case as opposed to the univariate case is that for a unit to meet the condition may depend on a measurement other than

the one presently interested in or a function of measurements. Let $w_{il}$ represent the average value of a unit in the network which contains unit $i$ for the $l^{th}$ measurement, that is

$$w_{il} = \frac{1}{m_i} \sum_{j \in \psi_i} y_{jl} \tag{5.2}$$

An unbiased estimator of the $l^{th}$ population mean is

$$\hat{\mu}_{1l} = \frac{1}{n} \sum_{i=1}^{n} w_{il} \tag{5.3}$$

The variance of $\hat{\mu}_{1l}$ for a sample taken without replacement is

$$var(\hat{\mu}_{1l}) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} (w_{il} - \mu_l)^2 \tag{5.4}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_{1l}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_{il} - \hat{\mu}_{1l})^2 \tag{5.5}$$

The covariance of $\hat{\mu}_{1l}$ and $\hat{\mu}_{1l'}$, $l \neq l'$, is (Thompson and Seber 1996),

$$cov(\hat{\mu}_{1l}, \hat{\mu}_{1l'}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{N} (w_{il} - \mu_l)(w_{il'} - \mu_{l'}) \tag{5.6}$$

and an unbiased estimator of the covariance is

$$\widehat{cov}(\hat{\mu}_{1l}, \hat{\mu}_{1l'}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_{il} - \hat{\mu}_l)(w_{il'} - \hat{\mu}_{l'}) \tag{5.7}$$

Now how to calculate $\hat{\mu}_{2l}$. Let $K$ still equal the number of distinct networks in the population, $\psi_k$ is the set of units in the $k^{th}$ network and $x_k$ denotes the number

of units that make up network $\psi_k$. Again the latter are fixed values regardless of the measurement of interest.

The sum of the $y_l-$ values in network $k$

$$y^*_{kl} = \sum_{i \in \psi_k} y_{il} \tag{5.8}$$

and the inclusion probability of network $k$

$$\alpha_k = 1 - \frac{\binom{N-x_k}{n}}{\binom{N}{n}} \tag{5.9}$$

$z_k$ is an indicator variable which equals one if any unit in the initial sample intersect the $k^{th}$ network.

$$z_k = \begin{cases} 1 & \text{if any unit of the } k^{th} \text{ network is in } s_0 \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

The estimator $\hat{\mu}_{2l}$ is

$$\hat{\mu}_{2l} = \frac{1}{N} \sum_{k=1}^{K} \frac{y^*_{kl} z_k}{\alpha_k} \tag{5.11}$$

The joint probability of two networks, $k$ and $h$ being intersected in the initial sample is

$$\alpha_{kh} = 1 - \frac{\{\binom{N-x_k}{n} + \binom{N-x_h}{n} - \binom{N-x_k-x_h}{n}\}}{\binom{N}{n}} \tag{5.12}$$

Also $\alpha_{kk} = \alpha_k$. The variance of $\hat{\mu}_{2l}$ is

$$var(\hat{\mu}_{2l}) = \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y^*_{kl} y^*_{hl} (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h} \tag{5.13}$$

An unbiased estimator of this variance is

$$\widehat{var}(\hat{\mu}_{2l}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y^*_{kl}y^*_{hl}z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \tag{5.14}$$

The covariance of $\hat{\mu}_{2l}$ and $\hat{\mu}_{2l'}$, $l \neq l^{'}$, is (Thompson and Seber 1996),

$$cov(\hat{\mu}_{2l},\hat{\mu}_{2l'}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y^*_{kl}y^*_{hl'}(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h} \tag{5.15}$$

and an unbiased estimator of the covariance is

$$\widehat{cov}(\hat{\mu}_{2l},\hat{\mu}_{2l'}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{h=1}^{K}\frac{y^*_{kl}y^*_{hl'}z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \tag{5.16}$$

### 5.2.2 Other Design Unbiased Estimators in the Multivariate Case

The other more efficient design unbiased estimators in the multivariate case are also calculated similar to the univariate. For the latter reason the notation will not be covered.

### 5.2.3 Univariate

Under standard conditions in adaptive cluster sampling for the the univariate case the Horvitz-Thompson based estimators tend to be more efficient than the Hansen-Hurwitz based estimators. Standard conditions being defined as larger networks tending to have larger $y$-values. In the situation where this is not the case the Hansen-Hurwitz based estimators can be more efficient (Salehi 1998)

### 5.2.4 Multivariate

Under typical settings in adaptive cluster sampling the Horvitz-Thompson based estimators tend to be more efficient than the Hansen-Hurwitz based estimators. The multivariate setting may not fall into the typical setting. The gray area is when estimation is done for one variable of interest but, units are adaptively added according to another variable of interest. In the latter case the large networks may not yield large $y-$values. Here the size of the network may be slightly correlated with or even negatively correlated with large $y-$values for the variable of interest. In such situations it may be better to use the Hansen-Hurwitz type estimators. The Horvitz-Thompson based estimators give large weights to large networks. In situations where the network size is not reflective of $y-$values for the variable of interest this may be is undesirable. On the other hand the Hansen-Hurwitz type estimators are based on the frequency networks are intersected in the initial sample. This is in part determined by the network size but the Hansen-Hurwitz type estimators are not as greatly influenced by network size.

## 5.3 Simulations

Table 5.1: blue-winged teal data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

Table 5.2: red-winged teal data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20 | 200 | 75 | 0 | 0 | 0 | 0 | 675 | 0 |
| 4000 | 13500 | 234 | 1335 | 4 | 0 | 0 | 35 | 0 | 55 |
| 0 | 0 | 0 | 0 | 97 | 0 | 4 | 0 | 1815 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1283 |

Table 5.3: blue-winged + red-winged teal data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20 | 203 | 80 | 0 | 0 | 0 | 0 | 675 | 0 |
| 4000 | 13500 | 234 | 1359 | 18 | 0 | 0 | 45 | 103 | 55 |
| 0 | 0 | 0 | 0 | 99 | 3 | 6 | 0 | 15454 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 122 |
| 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 1460 |

Table 5.4: Multivariate results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$ for blue-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_2)$ |
|---|---|---|---|---|
| 2 | 9.5694 | 237761.3 | 225036.3 | 220988.9 |
| 5 | 19.5159 | 88566.9 | 70545.0 | 65815.5 |
| 8 | 25.8871 | 51919.7 | 33198.4 | 29311.1 |
| 10 | 29.1585 | 39470.4 | 21221.3 | 18026.0 |
| 12 | 31.4184 | 31236.8 | 14567.2 | 11661.3 |
| 15 | 34.2465 | 23144.9 | 7925.3 | 5740.8 |
| 30 | 42.0664 | 6612.2 | 795.70 | 91.1 |

Table 5.5: Multivariate results for the simulations on red-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$ for blue-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_2)$ | $var(\hat{\mu}_{1+})$ | $var(\hat{\mu}_{2+})$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_{2RB})$ |
|---|---|---|---|---|---|---|---|
| 2 | 9.5694 | 1800569.5 | 1802226.4 | 1799969.7 | 1801627.6 | 1799496.6 | 1801583.2 |
| 5 | 19.5159 | 723445.2 | 726670.3 | 722014.3 | 725250.2 | 721884.2 | 725029.1 |
| 8 | 25.8871 | 400806.0 | 402948.0 | 399991.2 | 402142.9 | 399281.6 | 402081.6 |
| 10 | 29.1585 | 309877.9 | 311939.5 | 309685.2 | 311754.3 | 308924.8 | 311730.2 |
| 12 | 31.4184 | 244687.6 | 246407.6 | 244052 | 245772.1 | 243285.7 | 245762.5 |
| 15 | 34.2465 | 181680.2 | 183134.8 | 181200.6 | 182661.4 | 180532.0 | 182667.6 |
| 30 | 42.0664 | 51801.5 | 52398.5 | 51636.7 | 52235.4 | 51439.1 | 52227.4 |

Table 5.6: Multivariate results for the simulations on red-winged teal data for $\hat{\mu}_1$ and $\hat{\mu}_{2RB}$. Condition is $y_i \geq 1$ for blue-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{2RB})$ |
|---|---|---|---|
| 2 | 9.5694 | 1800569.5 | 1801583.2 |
| 5 | 19.5159 | 723445.2 | 725029.1 |
| 8 | 25.8871 | 400806.0 | 402081.6 |
| 10 | 29.1585 | 309877.9 | 311730.2 |
| 12 | 31.4184 | 244687.6 | 245762.5 |
| 15 | 34.2465 | 181680.2 | 182667.6 |
| 30 | 42.0664 | 51801.5 | 52227.4 |

Table 5.7: Multivariate results for the simulations on red-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$ for red-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_2)$ |
|---|---|---|---|---|
| 2 | 8.4328 | 359147.1 | 335576.7 | 329950.2 |
| 5 | 17.2317 | 137551.0 | 101811.2 | 94209.4 |
| 8 | 23.0647 | 78022.9 | 47151.3 | 40833.5 |
| 10 | 25.9387 | 60529.2 | 31559.0 | 25396.2 |
| 12 | 28.2725 | 47202.4 | 21873.4 | 16386.0 |
| 15 | 31.2130 | 35272.7 | 13149.5 | 8826.0 |
| 30 | 41.0636 | 9960.1 | 2288.6 | 1423.2 |

Table 5.8: Multivariate results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$ for red-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_2)$ | $var(\hat{\mu}_{1+})$ | $var(\hat{\mu}_{2+})$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_{2RB})$ |
|---|---|---|---|---|---|---|---|
| 2 | 8.4328 | 1773774.6 | 1773798.8 | 1773740.0 | 1773764.2 | 1773739.5 | 1773764.2 |
| 5 | 17.2317 | 679278.8 | 679304.8 | 679295.9 | 679321.9 | 679304.2 | 679338.6 |
| 8 | 23.0647 | 389922.0 | 389955.9 | 389937.8 | 389971.7 | 389936.9 | 389971.2 |
| 10 | 25.9387 | 302984.2 | 303015.3 | 302974.1 | 303005.2 | 302975.6 | 303003.7 |
| 12 | 28.2725 | 234830.2 | 234852.7 | 234816.6 | 234839.2 | 234814.9 | 234839.7 |
| 15 | 31.2130 | 174321.8 | 174339.7 | 174293.1 | 174311.1 | 174295.4 | 174315.6 |
| 30 | 41.0636 | 49500.4 | 49508.2 | 49484.0 | 49491.9 | 49483.5 | 49491.1 |

Table 5.9: Multivariate results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$ for either blue-wing or red-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_2)$ |
|---|---|---|---|---|
| 2 | 15.2240 | 177222.2 | 165114.7 | 159517.8 |
| 5 | 27.9991 | 66502.6 | 48269.4 | 43451.0 |
| 8 | 34.5132 | 38662.7 | 20229.9 | 17282.7 |
| 10 | 37.1890 | 29396.6 | 11659.1 | 9615.8 |
| 12 | 38.9337 | 23328.4 | 7256.2 | 5674.4 |
| 15 | 40.7538 | 17200.3 | 3436.7 | 2280.5 |
| 30 | 45.6354 | 4967.7 | 323.8 | 15.8 |

Table 5.10: Multivariate results for the simulations on red-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 1$ for either blue-wing or red-wing teal.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_2)$ |
|---|---|---|---|---|
| 2 | 15.2240 | 248686.1 | 222382.7 | 222664.5 |
| 5 | 27.9991 | 93642.0 | 58488.1 | 55606.1 |
| 8 | 34.5132 | 54501.5 | 24053.5 | 20611.4 |
| 10 | 37.1890 | 41642.4 | 13917.7 | 10313.1 |
| 12 | 38.9337 | 32343.9 | 8838.8 | 5368.8 |
| 15 | 40.7538 | 24035.2 | 4817.1 | 1820.3 |
| 30 | 45.6354 | 6822.5 | 849.1 | 8.8 |

Table 5.11: Multivariate results for the simulations on blue-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 100$ for blue-wing and red-wing teal combined.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_2)$ | $var(\hat{\mu}_{1+})$ | $var(\hat{\mu}_{2+})$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_{2RB})$ |
|---|---|---|---|---|---|---|---|
| 2 | 5.5644 | 570206.8 | 556527.1 | 570199.4 | 556519.5 | 558369.3 | 556519.2 |
| 5 | 12.6735 | 212359.6 | 195627.4 | 212360.0 | 195629.3 | 198097.7 | 195627.5 |
| 8 | 18.3111 | 124985.5 | 106469.2 | 124982.1 | 106464.8 | 107536.3 | 106465.0 |
| 10 | 21.6146 | 97401.5 | 76586.5 | 97400.4 | 76586.7 | 77783.4 | 76587.9 |
| 12 | 24.3579 | 75943.8 | 56974.2 | 75945.8 | 56974.1 | 58327.9 | 56974.0 |
| 15 | 28.0094 | 56811.9 | 37617.6 | 56811.8 | 37617.6 | 38791.6 | 37618.0 |
| 30 | 39.7095 | 16316.3 | 4932.7 | 16316.1 | 4932.5 | 5504.3 | 4932.5 |

Table 5.12: Multivariate results for the simulations on red-winged teal data for $\hat{\mu}_s$. Condition is $y_i \geq 100$ for blue-wing and red-wing teal combined.

| $n$ | $E(\nu)$ | $var(\hat{\mu}_1)$ | $var(\hat{\mu}_2)$ | $var(\hat{\mu}_{1+})$ | $var(\hat{\mu}_{2+})$ | $var(\hat{\mu}_{1RB})$ | $var(\hat{\mu}_{2RB})$ |
|---|---|---|---|---|---|---|---|
| 2 | 5.5644 | 650329.1 | 617786.0 | 650328.8 | 617785.7 | 618690.8 | 617802.4 |
| 5 | 12.6735 | 245840.6 | 202665.9 | 245833.2 | 202669.3 | 208190.7 | 202672.1 |
| 8 | 18.3111 | 142933.6 | 101153.1 | 142945.2 | 101156.6 | 105656.3 | 101158.1 |
| 10 | 21.6146 | 107958.4 | 67040.9 | 107945.8 | 67023.0 | 71569.1 | 67024.0 |
| 12 | 24.3579 | 83792.9 | 47171.6 | 83780.4 | 47160.3 | 51490.0 | 47157.7 |
| 15 | 28.0094 | 62635.5 | 27458.4 | 62617.9 | 27446.6 | 31588.7 | 27446.0 |
| 30 | 39.7095 | 17982.2 | 1224.1 | 17980.7 | 1221.4 | 3489.9 | 1221.3 |

## 5.4 Summary of the Simulation Results

Using the condition on one type of duck to estimate that type of duck the Horvitz-Thompson estimator was more efficient. When the prediction was on one duck but the condition was on both duck still the Horvitz-Thompson performed better. Only under the circumstances when the prediction was of one type of duck but the condition was on another the Hansen-Hurwitz estimator was more efficient. In conclusion, criterion which yield networks non-representative of the variable of interest the Hansen-Hurwitz estimator should be considered. For when the networks do not accurately reflect the variable of interest the fact that the Hansen-Hurwitz type estimators are more dependent on the particular initial sample than the Horvitz-Thompson type estimators is not bad.

# Chapter 6

# Adaptive Sampling in A Graph Setting

## 6.1 Introduction

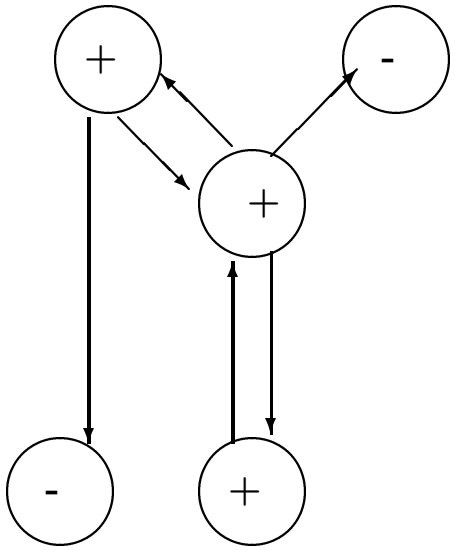
When analyzing social structures it is necessary to consider links among people, not only the attributes of the people. Thus it can be useful to use a link tracing design for collecting the sample. In addition, it can be very difficult to collect data on hidden populations and link tracing designs can also assist in this area (Frank and Snijders 1994). A human population's characteristics, such as social structure can be modeled as a stochastic graph. Attributes of individuals are represented by node values and relationships between individuals are represented by arc indicators. The focus of this chapter will cover methods for calculating the size a hidden population using a model based approach and maximum likelihood estimators. Section 6.4 covers the various models and estimators.

## 6.2 Adaptive Cluster Sampling

In a graph setting a neighborhood is defined as an arc between two nodes. A network is still defined as all units in the neighborhood of one another that meet the criterion. A network and its edge units are illustrated in figure 6.1. For the latter example the criterion will be HIV positive and an individual is in the neighborhood of another individual if there existed a sexual contact between the individuals. A matrix of

size $N \times N$, is used to represent the structure of the graph, and is denoted $A$. This type of matrix is often referred to as an adjacency matrix. An arc is an ordered pair of nodes representing the existence or absence of a directional relationship. If an arc exists from $i$ to $j$ then $a_{ij} = 1$ and $a_{ij} = 1$ for all $i = j$. $A$ is not considered a symmetric matrix here, thus $a_{ij} = 1$ does not imply $a_{ji} = 1$.

Fig. 6.1. The criterion to adaptively add units is there has to be an arc between nodes and the nodes have to be HIV (+). The network consists of the three HIV (+) people that have arcs between them. There two HIV (-) people which are also edges units of the network.



## 6.3 Design Unbiased Estimators

The standard estimators from the previous chapters can still be used to estimate the population mean for a variable of interest. As before if node (unit) $i$ is in the neighborhood of unit $j$ then $j$ must be in the neighborhood of unit $i$. An example would be if Mark is a friend of Joe then Joe must be a friend of Mark. Unfortunately, when dealing with people we cannot always assume that Joe and Mark will respond identically about their relationship. Also, it is often infeasible to survey the entire cluster of nodes. In some cases it can require too much time and money due to the size of the network and/or how dispersed the nodes are in the network. In addition, people may simply be unwilling to participate making it impossible to follow up all the links in the network to find the entire cluster. For these reasons it is important to explore model based approaches for estimating parameters of interest.

## 6.4 A Model Based Approach to Estimating Population Size

Often of interest is the size of a hidden population. A paper by Frank and Snijders (1994) illustrates a model based approach on how to calculate the size of a hidden population using snowball sampling. The following work will be based in part upon the model based approach in that paper. Snowball sampling can contain many waves of respondents and contacts. In snowball sampling the initial sample consists of all nodes-people-sampled in the initial sample. The first wave consists of all nodes mentioned by the initial sample but not in the initial sample. The second wave consists of all nodes mentioned by nodes in the first wave but in neither the initial sample nor the first

wave. This may continue until no new nodes are mentioned. The model based likelihood of Frank and Snijders (1994) method for calculating the size of the hidden population assumes that the initial sample of respondents was selected through a bernoulli sample but conditions on the initial sample size as if it was known a priori. In addition, it only utilizes information from the initial sample and the first wave of the snowball sample. In this section and the rest of the chapter the estimation of the size of the population of interest will be done without conditioning on the initial sample size in the latter manner. Also a likelihood model is proposed that can incorporate information from the second wave as well. Furthermore, a model is proposed where the population is considered to be made up of two sub groups with different arc probabilities in each group and between groups. The model will also assume a simple bernoulli sample is drawn from one group and thus the size of the second group will be estimated via the first and second wave information.

### 6.4.1 Estimation of Population Size using Snowball Sampling

We will assume that the initial sample of $n_0$ respondents was selected by a bernoulli sample. Each person is selected independently with probability $\alpha_0$. The initial sample size $n_0$ equals $|A_0|$ and is distributed binomial$(N, \alpha_0)$. $A_{ll} = A_l$ denotes the arcs from wave $l$ to wave $l$. $A_{lm}$ represents the arcs from wave $l$ to wave $m$. Wave zero is equivalent to the initial sample.

Let us assume the probability of the existence of an arc between two nodes $i$ and $j$ is identically independently distributed bernoulli with probability $\beta$. Thus $a_{ij} \quad \forall i \neq j$

is bernoulli with probability $\beta$. Let $a_{ij} = 1 \quad \forall i = j$, the node always has an arc with itself. The parameters of interest in this model are $\beta$, and $N$. The parameter of most importance is $N$, the size of the hidden population.

Let the number of times respondents in the initial sample are mentioned as contacts of the initial respondents be denoted $r_0$. Also, Let the total number of arcs mentioned by the initial respondents be denoted $t_0$.

$$r_0 = \sum_{i \in A_0} \sum_{j \in A_0} a_{ij} - n_0 \tag{6.1}$$

$$t_0 = \sum_{i \in A_0} \sum_{j \in A_0 \bigcup A_1} a_{ij} - n_0 \tag{6.2}$$

The distribution of $r_0$ is binomial $(n_0(n_0 - 1), \beta)$. Let $m_{0k}$ equal the number of nodes in the first wave that are mentioned $k$ times, where $k$ is defined from $0, ..., n_0$. Thus $m_{00}$ equals the number of nodes never mentioned not including the respondents in the initial sample. Let $m_{0s}$ denote the number of distinct contacts mentioned in the first wave. That is,

$$m0s = \sum_{k=1}^{n_0} m_{0k} \tag{6.3}$$

and

$$m_{00} = N - n_0 - m_{0s} \tag{6.4}$$

The distribution of $(m_{01}, ..., m_{0n_0})$ conditional on $n_0$ is multinomial $(N-n_0, p_{01}, ..., p_{0n_0})$ where,

$$p_{0k} = \binom{n_0}{k} \beta^k (1-\beta)^{n_0-k} \text{for } k = 0, ..., n_0 \tag{6.5}$$

The likelihood function conditioning on $n_0$ as if known a priori (Frank and Snijders 1994) is

$$\begin{aligned} L &= \binom{n_0(n_0-1)}{r_0} \beta^{r_0} (1-\beta)^{n_0(n_0-1)-r_0} \\ &\quad \times (N-n_0)! \prod_{k=0}^{n_0} \frac{p_{0k}^{m_{0k}}}{m_{0k}!} \end{aligned} \tag{6.6}$$

After some calculations the likelihood function can be shown to be equal to

$$\begin{aligned} L &= \frac{(N-n_0)!}{(N-n_0-m_{0s})!} \beta^{t_0} (1-\beta)^{n_0(N-1)-t_0} \\ &\quad \times \binom{n_0(n_0-1)}{r_0} \prod_{k=1}^{n_0} \frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \end{aligned} \tag{6.7}$$

Let:

$$a(x) = \binom{n_0(n_0-1)}{r_0} \prod_{k=1}^{n_0} \frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \tag{6.8}$$

and

$$b(x;\phi) = \frac{(N-n_0)!}{(N-n_0-m_{0s})!} \times \beta^{t_0}(1-\beta)^{(n_0(N-1)-t_0)} \tag{6.9}$$

Where $x = (m_{0s}, t_0)$ and $\phi = (\beta, N)$. From this it is clear by factorization theorem that $m_{0s}$ and $t_0$ are the sufficient statistics.

Solving for $\frac{dln(b_x(\beta,N))}{d\beta} = 0$ yields

$$\hat{\beta} = \frac{t_0}{n_0(N-1)} \tag{6.10}$$

The likelihood is then maximized for $N$ through brute force using a computer by increasing $N$ with a starting value of $n_0 + m_{0s}$ by 1 until the likelihood starts to decrease, that is until a maximum is achieved. The reason the starting value should be set at $n_0 + m_{0s}$ is because by the definition of those statistics that is the minimum number of people in the population of interest.

If in fact the sample is not a simple random sample and thus $n_0$ is a random variable then the researcher believes a likelihood function which considers $\alpha_0$ unknown and does not have the above restriction should be employed and is

$$\begin{aligned} L &= \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0}\binom{n_0(n_0-1)}{r_0}\beta^{r_0}(1-\beta)^{n_0(n_0-1)-r_0} \\ &\quad \times(N-n_0)!\prod_{k=0}^{n_0}\frac{p_k^{m_k}}{m_k!} \\ &= \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0}\frac{(N-n_0)!}{(N-n_0-m_0)!}\beta^{t_0}(1-\beta)^{n_0(N-1)-t_0} \\ &\quad \times\binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \end{aligned} \tag{6.11}$$

Let:

$$a(x) = \binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \tag{6.12}$$

and

$$b(x;\phi) = \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0}\frac{(N-n_0)!}{(N-n_0-m_{0s})!}\times\beta^{t_0}(1-\beta)^{(n_0(N-1)-t_0)} \tag{6.13}$$

Where $x = (n_0, m_{0s}, t_0)$ and $\phi = (\alpha_0, \beta, N)$. From this it is clear by factorization theorem that $n_0$, $m_{0s}$ and $t_0$ are the sufficient statistics.

Solving for $\frac{dln(b(x;\alpha_0,\beta,N))}{d\beta} = 0$ yields

$$\hat{\beta} = \frac{t_0}{n_0(N-1)} \tag{6.14}$$

and solving for $\frac{dln(b(x;\alpha_0,\beta,N))}{d\alpha_0} = 0$ yields

$$\hat{\alpha}_0 = \frac{n_0}{N} \tag{6.15}$$

Equation 6.11 can be maximized with respect to $N$ by substituting in equations 6.15 and 6.14 for $\alpha_0$ and $\beta$ respectively and substituting values of $N$ into equation 6.13 until a maximum is found.

### 6.4.2 An Extension Using the Second Wave

The previous work only used information that was obtained through the initial sample of respondents and thus didn't use information on the second wave. In this subsection we intend to estimate the population size using recruited respondents from the first wave as well as the initial respondents.

A respondent will be recruited with probability $\alpha_1$ from the contacts mentioned in the first wave meeting the condition. The condition will be a pre-specified minimum in-degree. Thus it is an adaptive snowball sample, in that the sampling design is in part

determined by the values observed in the sample.

The probability for recruiting a respondent from among the contacts meeting the condition, $\alpha_1$, will be considered an unknown constant. It is not always reasonable to assume that all the contacts mentioned meeting the condition will be recruited. For this reason it is assumed that there exists a probability $\alpha_1$ that a contact will be recruited, independent of a contacts potential responses.

The requirement for the in-degree can be anywhere from 1, just mentioned once to $n_0$ times, depending on the researcher. Since, a person mention contacts independently of who mentions them, this will not bias estimation and can be accounted for (Thompson and Frank 1999). Let $n_d$ represent the number of contacts in the first wave that meet the required in-degree, $d$ desired by the researcher and the number of recruited respondents will be denoted $n_1$. The data can be arranged such that more than one potential likelihood exists.

#### 6.4.2.1 The Likelihood

Let $r_1$ equal the number of times the recruited respondents (respondents from the first wave) are mentioned by other recruited respondents. Then $r_1 \mid n_1$ is binomial $(n_1(n_1 - 1), \beta)$ Let $t_1$ equal the total number of arcs mentioned by the $n_1$ recruited respondents. Let $m_{1k}$ equal the number of nodes that are not recruited respondents that are mentioned $k$ times, where $k$ is defined from $0, ..., n_1$. Then the distribution of $(m_{11}, ..., m_{1n_1})$ is multinomial $(N - n_1, p_{11}, ..., p_{1n_1})$ conditional on $n_1$.

$$p_{1k} = \binom{n_1}{k} \beta^k (1-\beta)^{n_1-k} \text{for } k = 0, ..., n_1 \tag{6.16}$$

The likelihood equation for including the recruited respondents information is

$$\begin{aligned} L &= P\left[n_0, r_0, (m_{01}, ..., m_{0n_0}), n_1, r_1, (m_{11}, ..., m_{1n_1})\right] \\ &= P\left[n_0\right] \times P\left[r_0, (m_{01}, ..., m_{0n_0} \mid n_0)\right] \times \\ &\quad P\left[n_1 \mid (m_{02}, ..., m_{0n_0}))\right] \times P\left[r_1, (m_{11}, ..., m_{1n_1}) \mid n_1\right] \end{aligned} \tag{6.17}$$

$$\begin{aligned} L &= \binom{N}{n_0} \alpha_0^{n_0} (1-\alpha_0)^{N-n_0} \times \\ &\quad \binom{n_0(n_0-1)}{r_0} \beta^{r_0} (1-\beta)^{n_0(n_0-1)-r_0} (N-n_0)! \prod_{k=0}^{n_0} \frac{p_{0k}^{m_{0k}}}{m_{0k}!} \times \\ &\quad \binom{n_d}{n_1} \alpha_1^{n_1} (1-\alpha_1)^{n_d-n_1} \times \\ &\quad \binom{n_1(n_1-1)}{r_1} \beta^{r_1} (1-\beta)^{(n_1(n_1-1)-r_1)} (N-n_1)! \prod_{k=0}^{n_1} \frac{p_{1k}^{m_{1k}}}{m_{1k}!} \end{aligned} \tag{6.18}$$

The likelihood function 6.18 can be manipulated into a more manageable form in a similar fashion as Frank and Snijders (1994) manipulated the likelihood equation 6.6. Letting $t_s = t_0 + t_1$ the likelihood can be written,

$$
\begin{aligned}
L &= \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\
&\quad \frac{(N-n_0)!}{(N-n_0-m_{0s})!}\beta^{t_0}(1-\beta)^{n_0(N-1)-t_0} \\
&\quad \times\binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \times \\
&\quad \binom{n_d}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_d-n_1} \times \\
&\quad \frac{(N-n_1)!}{(N-n_1-m_1)!}\beta^{t_1}(1-\beta)^{(n_1(N-1)-t_1)} \\
&\quad \times\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!} \\
&= \binom{n_d}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_d-n_1} \times \\
&\quad \binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!}\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!} \times \\
&\quad \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\
&\quad \frac{(N-n_0)!}{(N-n_0-m_{0s})!}\beta^{t_0}(1-\beta)^{(n_0(N-1)-t_0)} \times \\
&\quad \frac{(N-n_1)!}{(N-n_1-m_{1s})!}\beta^{t_1}(1-\beta)^{(n_1(N-1)-t_1)} \\
&= \binom{n_d}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_d-n_1} \times \\
&\quad \binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!}\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!} \times \\
&\quad \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\
&\quad \frac{(N-n_0)!}{(N-n_0-m_{0s})!}\frac{(N-n_1)!}{(N-n_1-m_{1s})!} \times
\end{aligned}
$$

$$
\begin{aligned}
& \beta^{(t_0+t_1)}(1-\beta)^{((n_0+n_1)(N-1)-(t_0+t_1))} \\
= \quad & \binom{n_d}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_m-n_1} \times \\
& \binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!}\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!} \times \\
& \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\
& \frac{(N-n_0)!}{(N-n_0-m_{0s})!}\frac{(N-n_1)!}{(N-n_1-m_{1s})!} \times \\
& \beta^{t_s}(1-\beta)^{((n_0+n_1)(N-1)-t_s)}
\end{aligned}
\tag{6.19}
$$

Solving for $\frac{dln(L(\alpha_0,\alpha_1,\beta,N))}{d\beta} = 0$ yields

$$\hat{\beta} = \frac{t_s}{(n_0+n_1)(N-1)} \tag{6.20}$$

and solving for $\frac{dln(L(\alpha_0,\alpha_1,\beta,N))}{d\alpha_0} = 0$ yields

$$\hat{\alpha}_0 = \frac{n_0}{N} \tag{6.21}$$

and solving for $\frac{dln(L(\alpha_0,\alpha_1,\beta,N))}{d\alpha_1} = 0$ yields

$$\hat{\alpha}_1 = \frac{n_1}{n_d} \tag{6.22}$$

Note that $\hat{\alpha}_1$ is a function of $n_1$ and $n_d$ and does not involve the main parameter of interest $N$. For this reason the likelihood is separated as follows:

Let:

$$\begin{aligned} L_1 &= \binom{n_d}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_d-n_1} \times \\ &\quad \binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!}\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!} \end{aligned} \tag{6.23}$$

and

$$\begin{aligned} L_2 &= \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\ &\quad \frac{(N-n_0)!}{(N-n_0-m_{0s})!}\frac{(N-n_1)!}{(N-n_1-m_{1s})!} \times \\ &\quad \beta^{t_s}(1-\beta)^{((n_0+n_1)(N-1)-t_s)} \end{aligned} \tag{6.24}$$

For a fixed $N$ the likelihood is maximized by $\beta = \frac{t_s}{(n_0+n_1)(N-1)}$ and $\alpha_0 = \frac{n_0}{N}$. Also, the function $L_1$ is constant for all values of $N$. Thus for any fixed $\beta$ and $\alpha_0$ the likelihood can be maximized by maximizing the function $L_2$ through a straight forward approach of incrementing $N$ by 1 until a maximum is reached.

### 6.4.3 Estimating Population Size When there are Two Subgroups With Different Arc Probabilities and Initial Sampling is Done Only Within One Group.

The former likelihood covered does not work very well when estimating population size when there are two subgroups with different arc probabilities. People of type 0 will be considered in the first subgroup and people of type 1 will be considered in the second subgroup. In this situation it is helpful to think of the population size as the sum of two population sizes. Let

$$N = N_0 + N_1 \tag{6.25}$$

For this subsection some of the notation is redefined accordingly. Let arcs in the subpopulation of size $N_0$ be independently identically distributed $\beta_0$ and arcs in the subpopulation of size $N_1$ be identically independently distributed $\beta_1$ and the arcs between the subpopulations in both directions be identically independently distributed $\beta_2$. Assume we start with an initial sample of size $n_0$. Let $n_{d0}$ equal the number of contacts from the initial sample of type 1 and let $t_{ds}$ equal the number arcs between the two subpopulations. The contacts of type 1 of the initial respondents will be recruited independently with probability $\alpha_1$. Let $r_0$ equal the number of times the initial respondents are mentioned by other initial respondents. Then $r_0 \mid n_0$ is binomial $(n_0(n_0-1), \beta_0)$ Let $t_0$ equal the total number of arcs mentioned by the $n_0$ initial respondents of type 0. Let $m_{0k}$ equal the number of nodes that are not initial respondents of type 0 that are mentioned

$k$ times, where $k$ is defined from $0, ..., n_0$. Then the distribution of $(m_{01}, ..., m_{0n_0})$ is multinomial $(N_0 - n_0, p_{01}, ..., p_{0n_0})$ conditional on $n_0$.

$$p_{0k} = \binom{n_0}{k} \beta_0^k (1 - \beta_0)^{n_0 - k} \text{for } k = 0, ..., n_0 \tag{6.26}$$

Let $r_1$ equal the number of times the $n_1$ recruited respondents are mentioned by other recruited respondents. Then $r_1 \mid n_1$ is binomial $(n_1(n_1 - 1), \beta_1)$ Let $t_1$ equal the total number of arcs mentioned by the $n_1$ recruited respondents of type 1. Let $m_{1k}$ equal the number of nodes that are not recruited respondents of type 1 that are mentioned $k$ times, where $k$ is defined from $0, ..., n_1$. Then the distribution of $(m_{11}, ..., m_{1n_1})$ is multinomial $(N_1 - n_1, p_{11}, ..., p_{1n_1})$ conditional on $n_1$.

$$p_{1k} = \binom{n_1}{k} \beta_1^k (1 - \beta_1)^{n_1 - k} \text{for } k = 0, ..., n_1 \tag{6.27}$$

$$\begin{aligned} L &= P\left[n_0, r_0, (m_{01}, ..., m_{0n_0}), n_{d0}, n_1, r_1, (m_{11}, ..., m_{1n_1}), t_{ds}\right] \\ &= P\left[n_0\right] \times P\left[r_0, (m_{01}, ..., m_{0n_0} \mid n_0)\right] \times \\ &\quad P\left[n_{d0} \mid n_0\right] \times P\left[n_1 \mid n_{d0}\right] \times \\ &\quad P\left[r_1, (m_{11}, ..., m_{1n_1}) \mid n_1\right] \times P\left[t_{ds} \mid (n_0, n_1)\right] \end{aligned} \tag{6.28}$$

$$
\begin{aligned}
L \;=\; & \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\
& \binom{n_0(n_0-1)}{r_0}\beta^{r_0}(1-\beta_0)^{n_0(n_0-1)-r_0}(N_0-n_0)!\prod_{k=0}^{n_0}\frac{p_{0k}^{m_{0k}}}{m_{0k}!} \times \\
& \binom{n_0N_1+n_1N_0}{t_{ds}}\beta_2^{t_{ds}}(1-\beta_2)^{n_0N_1+n_1N_0-t_{ds}} \times \\
& \binom{n_{d0}}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_{d0}-n_1} \times \\
& \binom{n_1(n_1-1)}{r_1}\beta_1^{r_1}(1-\beta_1)^{(n_1(n_1-1)-r_1)}(N_1-n_1)!\prod_{k=0}^{n_1}\frac{p_{1k}^{m_{1k}}}{m_{1k}!}
\end{aligned}
\tag{6.29}
$$

The likelihood function 6.29 can be simplified in a similar fashion as Frank and Snijders (1994) manipulated the likelihood equation 6.6.

$$
\begin{aligned}
L \;=\; & \binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0} \times \\
& \frac{(N_0-n_0)!}{(N_0-n_0-m_{0s})!}\beta^{t_0}(1-\beta)^{n_0(N_0-1)-t_0} \times \\
& \binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \times \\
& \binom{n_0N_1+n_1N_0}{t_{ds}}\beta_2^{t_{ds}}(1-\beta_2)^{n_0N_1+n_1N_0-t_{ds}} \times \\
& \binom{n_{d0}}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_{d0}-n_1} \times \\
& \frac{(N_1-n_1)!}{(N_1-n_1-m_1)!}\beta_1^{t_1}(1-\beta_1)^{(n_1(N_1-1)-t_1)}
\end{aligned}
$$

$$
\begin{aligned}
&\times\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!}\\
=\;&\binom{n_{d0}}{n_1}\alpha_1^{n_1}(1-\alpha_1)^{n_{d0}-n_1}\times\\
&\binom{n_0(n_0-1)}{r_0}\prod_{k=1}^{n_0}\frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!}\binom{n_1(n_1-1)}{r_1}\prod_{k=1}^{n_1}\frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!}\times\\
&\binom{N}{n_0}\alpha_0^{n_0}(1-\alpha_0)^{N-n_0}\times\\
&\binom{n_0N_1+n_1N_0}{t_{ds}}\beta_2^{t_{ds}}(1-\beta_2)^{n_0N_1+n_1N_0-t_{ds}}\times\\
&\frac{(N_0-n_0)!}{(N_0-n_0-m_{0s})!}\beta_0^{t_0}(1-\beta_0)^{(n_0(N_0-1)-t_0)}\times\\
&\frac{(N_1-n_1)!}{(N_1-n_1-m_{1s})!}\beta_1^{t_1}(1-\beta_1)^{(n_1(N_1-1)-t_1)}
\end{aligned}
\tag{6.30}
$$

Solving for $\frac{dln(L)}{d\alpha_0}=0$ yields

$$\hat{\alpha}_0=\frac{n_0}{N_0} \tag{6.31}$$

Solving for $\frac{dln(L)}{d\alpha_1}=0$ yields

$$\hat{\alpha}_1=\frac{n_1}{n_d} \tag{6.32}$$

Solving for $\frac{dlnL}{d\beta_0}=0$ yields

$$\hat{\beta}_0=\frac{t_0}{n_0(N_0-1)} \tag{6.33}$$

Solving for $\frac{dlnL}{d\beta_1}=0$ yields

$$\hat{\beta}_1 = \frac{t_1}{n_1(N_1 - 1)} \tag{6.34}$$

Solving for $\frac{dlnL}{d\beta_2} = 0$ yields

$$\hat{\beta}_2 = \frac{t_{ds}}{n_0 N_1 + n_1 N_0} \tag{6.35}$$

Let:

$$\begin{aligned} L_1 &= \binom{n_{d0}}{n_1} \alpha_1^{n_1} (1-\alpha_1)^{n_{d0}-n_1} \times \\ &\quad \binom{n_0(n_0-1)}{r_0} \prod_{k=1}^{n_0} \frac{\binom{n_0}{k}^{m_{0k}}}{m_{0k}!} \binom{n_1(n_1-1)}{r_1} \prod_{k=1}^{n_1} \frac{\binom{n_1}{k}^{m_{1k}}}{m_{1k}!} \end{aligned} \tag{6.36}$$

and

$$\begin{aligned} L_2 &= \binom{N}{n_0} \alpha_0^{n_0} (1-\alpha_0)^{N-n_0} \times \\ &\quad \binom{n_0 N_1 + n_1 N_0}{t_{ds}} \beta_2^{t_{ds}} (1-\beta_2)^{n_0 N_1 + n_1 N_0 - t_{ds}} \times \\ &\quad \frac{(N_0 - n_0)!}{(N_0 - n_0 - m_{0s})!} \beta_0^{t_0} (1-\beta_0)^{(n_0(N_0-1)-t_0)} \times \\ &\quad \frac{(N_1 - n_1)!}{(N_1 - n_1 - m_{1s})!} \beta_1^{t_1} (1-\beta_1)^{(n_1(N_1-1)-t_1)} \end{aligned}$$

(6.37)

In order to maximize the likelihood it is only necessary to maximize equation 6.37, $L_2$. Again, a closed form solution for $N$ can not be derived. A maximum can be obtained by incrementing $N_0$ by 1 while keeping $N_1$ fixed until a maximum is found with respect to $N_0$. Then incrementing $N_1$ by 1 while keeping $N_0$ fixed until a maximum is achieved now with respect to $N_1$. This process is done recursively until neither $N_0$ nor $N_1$ change.

### 6.4.4 Variance Estimation

From Frank and Snijders (1994) paper and from the researcher's own work it is evident that a closed form solution for expressing the variance and estimating the variance is not feasible. Frank and Snijders (1994) propose a modified jacknife estimator to estimate the variance. This estimator performs well in that paper and performs well in the simulations performed by the researcher.

The concept of the delete 1 jacknife estimator is to produce $n$ sample estimates of the parameter of interest, $\hat{\theta}_i$ for $i = 1, ..., n_0$ by omitting the $i^{th}$ respondent in the sample in order to obtain an estimate of the variance of $\hat{\theta}$ (Shao and Tu 1995).

$$var(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \bar{\theta})^2 \tag{6.38}$$

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i \tag{6.39}$$

The latter standard jacknifing estimator is an unbiased variance estimator under the following conditions:

$$\hat{\theta} \text{ is a function of iid random variables} \tag{6.40}$$

$$var(\hat{\theta}_i) = \frac{n}{n-1} var(\hat{\theta}) \tag{6.41}$$

and the correlation is,

$$\rho(\hat{\theta}_i, \hat{\theta}_j) = \frac{n-2}{n-1} \text{ for } i \neq j \tag{6.42}$$

#### 6.4.4.1 A Modified Jacknife Estimator

Unfortunately the latter conditions 6.40, 6.41 and 6.42 are not met for the models proposed for estimating $N$. The estimators for $N$ are a function of $n_0$ independently selected initial respondents. The estimators utilize not only the number of arcs mentioned but also such information as the number of distinct contacts, that is the number of arcs leading to new contacts. Thus the estimators of $N$ are more like a function of the relationship between the $n_0$ respondents than a function of $n_0$ i.i.d. random variables. For this reason a modified jacknife estimator for the variance was developed. The theory behind the following jacknife variance estimator and the estimator itself is developed in Frank and Snijders (1994)

The estimator $\hat{N}$ is analogous to the sample $\bar{z}$ of a $n \times n$ matrix of random variables. $z_{ij} \equiv 0 \quad \forall \quad i = j$ and are i.i.d. random variables with variance $\sigma^2$ otherwise. Let $\bar{z}$ equal the average of all the variables minus the diagonal.

$$\bar{z} = \frac{1}{n(n-1)} \sum_{i \neq j} z_{ij} \tag{6.43}$$

$$var(\bar{z}) = \frac{\sigma^2}{n(n-1)} \tag{6.44}$$

Let $\bar{z}_i$ denote the average of all the variables minus the $i^{th}$ row, column, and the diagonal. Then the variance of $\bar{z}_i$ is

$$\begin{aligned} var(\bar{z}_i) &= var(\frac{1}{(n-1)(n-2)} \sum_{j \neq k, j \neq i, k \neq i} z_{jk}) \\ &= \frac{\sigma^2}{(n-1)(n-2)} \\ &= \frac{n}{n-2} \frac{\sigma^2}{n(n-1)} \\ &= \frac{n}{n-2} var(\bar{z}) \end{aligned} \tag{6.45}$$

$$\begin{aligned} \rho(\bar{z}_i, \bar{z}_j) &= cov(\bar{z}_i, \bar{z}_j) \div (var(\bar{z}_i)^{1/2} var(\bar{z}_j)^{1/2}) \\ &= cov \Big( \frac{1}{(n-1)(n-2)} \sum_{h \neq k, h \neq i, k \neq i} z_{hk}, \\ &\quad \frac{1}{(n-1)(n-2)} \sum_{h \neq k, h \neq j, k \neq j} z_{hk} \Big) \div \\ &\quad (var(\bar{z}_i)^{1/2} var(\bar{z}_j)^{1/2}) \\ &= \frac{(n-3)(n-2)\sigma^2}{(n-2)^2(n-1)^2} \div \frac{\sigma^2}{(n-1)(n-2)} \end{aligned}$$

$$= \frac{n-3}{n-1} \text{ for } i \neq j \tag{6.46}$$

$$\bar{z}_{\cdot} = \frac{1}{n}\sum_{i=1}^{n} \bar{z}_i = \bar{z} \tag{6.47}$$

$$\begin{aligned}
E\left[\sum_{i=1}^{n}[\bar{z}_i - \bar{z}_{\cdot}]^2\right] &= E\left[\sum_{i=1}^{n}\left[\bar{z}_i^2 - 2\bar{z}_i\bar{z}_{\cdot} + \bar{z}_{\cdot}^2\right]^2\right] \\
&= E\left[\sum_{i=1}^{n}\bar{z}_i^2 - n\bar{z}_{\cdot}^2\right] \\
&= E\left[\sum_{i=1}^{n}\bar{z}_i^2 - n(Ez_{ij,i\neq j})^2\right] - nE\left[\bar{z}_{\cdot}^2 - (Ez_{ij,i\neq j})^2\right] \\
&= nvar(\bar{z}_i) - nvar(\bar{z}_{\cdot}) \\
&= n\left[var(\bar{z}_i) - var(\bar{z})\right] \\
&= n\left[var(\bar{z}_i) - \frac{n-2}{n}var(\bar{z}_i)\right] \\
&= n\left[\frac{2}{n}var(\bar{z}_i)\right] \\
&= 2var(\bar{z}_i)
\end{aligned} \tag{6.48}$$

Implying that an unbiased variance estimator is

$$\widehat{var}(\bar{z}) = \frac{n-2}{2n}\sum_{i=1}^{n}[\bar{z}_i - \bar{z}_{\cdot}]^2$$

(6.49)

Thus the following estimator for the variance of $\hat{N}$ is proposed (Frank and Snijders 1994).

$$\widehat{var}(\hat{N}) = \frac{n_0 - 2}{2n_0} \sum_{i=1}^{n_0} \left[\hat{N}_i - \hat{N}_\cdot\right]^2 \tag{6.50}$$

$\hat{N}_i$ is the estimator for $N$ calculated without the information from or concerning the $ith$ respondent in the initial sample.

$$\hat{N}_\cdot = \frac{1}{n_0} \sum_{i=1}^{n_0} \hat{N}_i \tag{6.51}$$

#### 6.4.4.2 For Estimating Variance When the Estimator Depends Solely on Adaptively Added Respondents

The previously covered modified jacknife estimator for variance works well when estimating $N$ and there is a single arc probability but not under the model described in section 6.4.3 where the population is actually broken down into two subgroups. The modified jacknife estimator still performs well for estimating the variance of $\hat{N}_0$ but not for the variance of $\hat{N}_1$. The jacknife estimator assumes that there is the deletion of only a single respondent at a time. This does not hold for both adaptive estimation designs. In the case where there is only one $\beta$ it does not seem to be a problem. This can be explained

by viewing the removal of a single initial respondent and then recruited respondents accordingly as merely the removal of a respondent and all the information that resulted from him, where some respondents contain more information than others. An important difference is that when estimating the model consisting of two distinct subgroups the recruited respondents bring additional information that is not merely more information on the same parameters but information pertaining to additional parameters that the initial respondent didn't have. Also the jacknife estimator for the variance of $\hat{N}_0$ and for the variance of $\hat{N}_1$ but the formulae would be:

$$var(\hat{N}_0) = \frac{n_0 - 2}{2n_0} \sum_{i=1}^{n_0} \left[\hat{N}_{0i} - \hat{N}_{0\cdot}\right]^2 \tag{6.52}$$

$$\hat{N}_{0\cdot} = \frac{1}{n_0} \sum_{i=1}^{n_0} \hat{N}_{0i} \tag{6.53}$$

Let $\hat{N}_{1i}$ represent the estimate of $N_1$ when a single initial respondent from the $n_0$ respondents is removed and the number of respondents from $n_1$ is reduced accordingly.

$$var(\hat{N}_1) = \frac{n_0 - 2}{2n_0} \sum_{i=1}^{n_0} \left[\hat{N}_{1i} - \hat{N}_{1\cdot}\right]^2 \tag{6.54}$$

$$\hat{N}_{1\cdot} = \frac{1}{n_0}\sum_{i=1}^{n_0}\hat{N}_{1i} \tag{6.55}$$

The weight $\frac{n_0-2}{2n_0}$ is for the removal of one respondent and when calculating $\hat{N}_1$ the number of respondents removed within this groups is random. From the formulae it is clear that the above estimator is not appropriate for estimating the variance of $\hat{N}_1$.

For the latter reasons a slightly modified version of Frank and Snijders (1994) is recommended for estimating the variance of $\hat{N}_0$ and $\hat{N}_1$ and they appear to work well. Let $\hat{N}_{0i}$ represent the estimate of $N_0$ when a single initial respondent from the $n_0$ respondents is removed and the number of respondents from $n_1$ and the information from the $n_1$ respondents is held constant including the estimate, $\hat{N}_1$.

$$var(\hat{N}_0) = \frac{n_0-2}{2n_0}\sum_{i=1}^{n_0}\left[\hat{N}_{0i}-\hat{N}_{0\cdot}\right]^2 \tag{6.56}$$

$$\hat{N}_{0\cdot} = \frac{1}{n_0}\sum_{i=1}^{n_0}\hat{N}_{0i} \tag{6.57}$$

Let $\hat{N}_{1i}$ represent the estimate of $N_1$ when a single recruited respondent from the $n_1$ respondents is removed and the number of respondents from $n_0$ and the information from the $n_0$ respondents is held constant including the estimate, $\hat{N}_0$.

$$var(\hat{N}_1) = \frac{n_1 - 2}{2n_1} \sum_{i=1}^{n_1} \left[\hat{N}_{1i} - \hat{N}_{1.}\right]^2 \tag{6.58}$$

$$\hat{N}_{1.} = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{N}_{1i} \tag{6.59}$$

#### 6.4.4.3 Suggested Minimum Sample Size

Very small sample sizes can lead to underestimation of the true variance. A sample of size two will yield an estimate of zero for the variance. It is recommended to use a sample size greater than or equal to the square root of the population size of interest (Frank Snijders 1994). The researcher's simulation results suggests that at least the square root of the population size should be used. In addition, the variability of the estimators greatly depend on the probability of an arc between nodes.

The recommended jacknife estimator of variance for the estimators of $N$ considers each estimator of $N$ to be like a function of the relationship between $n$ respondents. This is the premise for comparing the estimator to the average of a $n \times n$ matrix of i.i.d. random variable minus the diagonal. When there is a very small sample, $n$, the data does not behave in this manner. For this reason it is believed that the jacknife estimators of variance are consistent but not unbiased. Thus it is recommended to avoid very small samples when estimating variance. Sample sizes of approximately ten or more tend to

work well for estimating the variance, judging from the simulation results. Simulations in the three $\beta$ model where $n_1$ tended to be small, averaging approximately ten and thus many samples even smaller, the jacknife estimator generally underestimated the variance.

$$\begin{aligned} E(n_1) &= \sum_{n_0=0}^{N_0} \binom{N_0}{n_0} \alpha_0^{n_0} (1-\alpha_0)^{N_0-n_0} \times \\ & \quad N_1(1-(1-\beta_2)^{n_0}) \times \alpha_1 \end{aligned} \tag{6.60}$$

### 6.4.5 The Effect of Response Bias

One type of response bias which may occur is under reporting. People often may not mention everyone they know in the population of interest. Under reporting that is unrelated to the attributes of the individuals is not reported is not a major problem. This will bias the estimation of the probability of the existence of an arc, $\beta$, but not the estimation of $N$, our main concern. It will decrease the efficiency of the estimator. A reason for this can be seen easiest by considering the simplest case where there is a single $\beta$. Although we are estimating the size of the hidden population $N$ we are really only estimating $N$ minus the number of respondents and the number of people mentioned by the respondents. In theory the larger the $\beta$ the greater the proportion of the population of interest known and a larger amount of "recaptures" indicating this fact. Thus a

smaller variability of the estimator.

Another potential problem is false reporting or over reporting. This problem can be in part dealt with by partial verification of the respondents information. Thus, when using link tracing designs it may be possible to estimate the amount of over reporting.

Response error, in general, can be minimized through careful wording in the questionnaire and/or using a more restrictive condition for the existence of an arc. For example, changing a name generator from "Who do you know that does heroin?" to "Who have you done heroin with or given heroin to?".

When there is a belief that the population can actually be broken down into two subgroups and there is a different probability within the groups and between the groups then a model similar to or the one specifically mentioned in section 6.4.3 is recommended.

### 6.4.6 Examples

In this section will be several illustrations to give the reader a clear example how to calculate the statistics involved for the model with a single $\beta$ and a criterion of a minimum in-degree of 2 to be a potential recruited respondent.

Fig. 6.2. The first row is the initial sample with the arcs existing among the respondents in the initial sample.

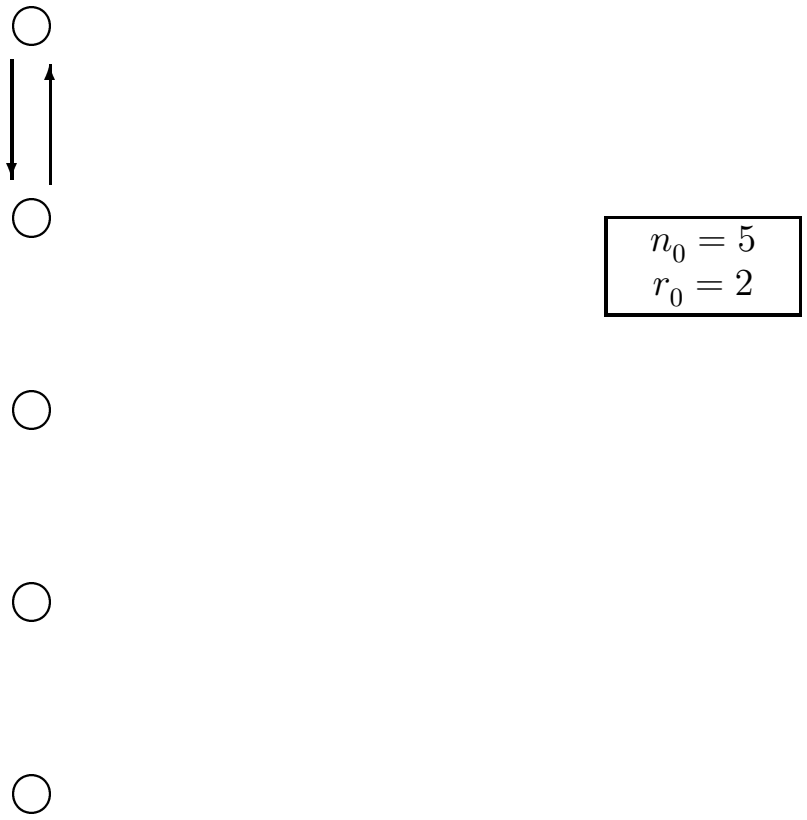
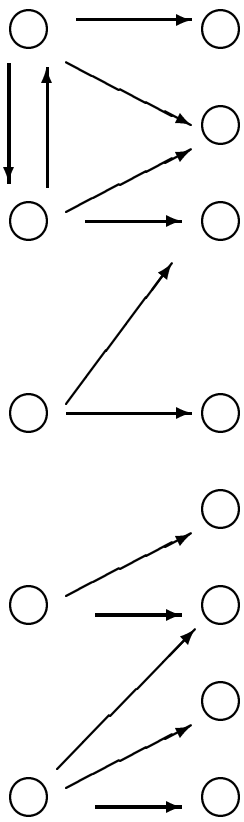
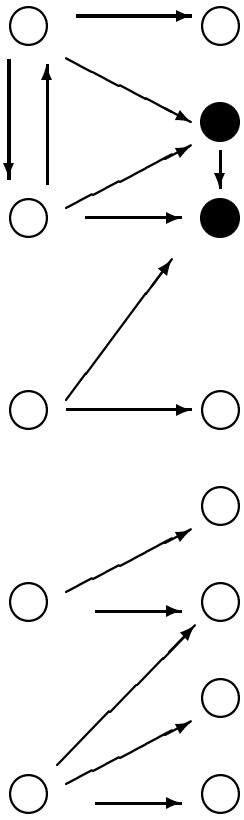


Fig. 6.3. The first row is the initial sample and the second row (first wave) are the contacts of the initial sample of which some will become respondents.



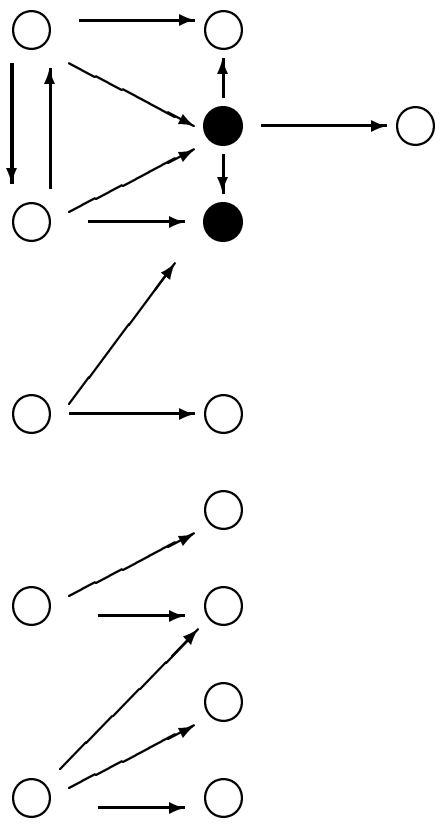
$$
\begin{aligned}
n_0 &= 5 \\
r_0 &= 2 \\
m_{01} &= 5 \\
m_{02} &= 3 \\
m_{03} &= 0 \\
m_{04} &= 0 \\
m_{05} &= 0 \\
m_{0s} &= 8 \\
t_0 &= 13
\end{aligned}
$$

Fig. 6.4. The single $\beta$ model with a criterion of the in-degree $\geq 2$ (Section 6.4.2). The first row is the initial sample and the second row (first wave) are the contacts of the initial sample of which some became respondents. The dark circles are the contacts meeting the criterion that were recruited and became respondents.



$$
\begin{aligned}
n_0 &= 5 \\
r_0 &= 2 \\
m_{01} &= 5 \\
m_{02} &= 3 \\
m_{03} &= 0 \\
m_{04} &= 0 \\
m_{05} &= 0 \\
m_{0s} &= 8 \\
t_0 &= 13 \\
n_1 &= 2 \\
r_1 &= 1
\end{aligned}
$$

Fig. 6.5. The single $\beta$ model with a criterion of the in-degree $\geq 2$ (Section 6.4.2). The first row is the initial sample and the second row (first wave) are the contacts of the initial sample of which some became respondents. The dark circles are respondents meeting the criterion that were also recruited. The third and final row consists of the contacts of the recruited respondents that were never seen before.



$$
\begin{aligned}
n_0 &= 5 \\
r_0 &= 2 \\
m_{01} &= 5 \\
m_{02} &= 3 \\
m_{03} &= 0 \\
m_{04} &= 0 \\
m_{05} &= 0 \\
m_{0s} &= 8 \\
t_0 &= 13 \\
n_1 &= 2 \\
r_1 &= 1 \\
m_{11} &= 2 \\
m_{12} &= 0 \\
m_{1s} &= 2 \\
t_1 &= 3 \\
t_s &= 16
\end{aligned}
$$

### 6.4.7 An Application

The Colorado Springs Network data set is such a situation where the data was collected through a link tracing design, more specifically, an adaptive link tracing procedure. In this study an initial sample was a convenience sample. High risk individuals were sought out at places such as std clinics, outreach programs, etc. If a person was mentioned by two different individuals in the initial sample the person was sought out as a potential respondent, the first wave. If a person was mentioned by someone in the initial and someone in the first wave or by two people in the first then they would be sought out as a potential respondent, the second wave. This study consisted of a total of 595 respondents of which 92 were recruited in this type of link tracing manner.

For the model for the Colorado Springs 1990 data set the desired in-degree is $\geq 2$. Thus the contacts in the first wave had to be mentioned twice in order to be considered a potential respondent. The selection process of a potential respondent is adaptive, since it is dependent upon observed in-degree of a node. Thus the major assumption is still that the observed in-degree (with the initial sample) of a person is independent of the out degree. Simply the recruited respondents should yield the same information as the other contacts. The difference this causes is the probability of a potential recruited respondent. A person has the probability of meeting the condition conditioning on $n_0$ equal to the probability of a node being mentioned at least twice. This is equal to $1 - \binom{n_0}{0}\beta^0(1-\beta)^{n_0-0} - \binom{n_0}{1}\beta^1(1-\beta)^{n_0-1} = 1-(1-\beta)^{n_0} - n_0\beta(1-\beta)^{n_0-1}$ Then the probability of a respondent in the first wave is $\alpha_1\left[(1-(1-\beta)^{n_0} - n_0\beta(1-\beta)^{n_0-1})\right]$.

In analyzing the Colorado Springs 1990 data set the population of interest will be considered intravenous drug users in Colorado Springs, a subset of the data. Respondents responding in the third year that were not obtained via link tracing and are an intranvenous drug user will be considered the initial respondents. Respondents recruited via link tracing in the actual study and are an intravenous drug user and named at least twice by the initial respondents will be used as the recruited respondents. It will be assumed that the initial sample of respondents was taken as a bernoulli sample from the population of interest. An estimate for the population size was made using Frank Snijders (1994) likelihood ignoring the adaptively added data and then an estimate was made using the adaptively added data by using the likelihood model in section 6.4.2. The number of initial respondents, $n_0$, was 187 and the number of distinct contacts mentioned by them, $m_{0s}$ was 389. The total number of arcs mentioned by the initial respondents, $t_0$ was 487. There were $n_1 = 12$ recruited respondents, and they named 41 contacts, $m_{1s}$ and also had a total of 41 arcs, $t_1$. Utilizing only the initial respondents yields an estimate, $\hat{N} = 1926$ and a jacknife estimate of variance of 66077.578284. Incorporating the recruited respondents yields an estimate, $\hat{N} = 1935$ and a jacknife estimate of variance of 63040.661157.

### 6.4.8 Simulations

Each simulation was run for 1000 iterations. The number of iterations will be denoted *iter*. Let $\hat{N}_*$ denote the estimator for the population of interest. The formula used to estimate the actual variance is,

$$\widehat{var}(\hat{N}_*) = \frac{1}{iter - 1} \sum_{i=1}^{iter} (\hat{N}_{*i} - \bar{N}_{*h})^2 \tag{6.61}$$

Where $\hat{N}_{*i}$ is the value for the relevant estimator for sample $i$ and

$$\bar{N}_{*h} = \frac{1}{iter} \sum_{i=1}^{iter} \hat{N}_{*i} \tag{6.62}$$

Table 6.1: $N = 100$. The condition is in-degree $\geq 2$ .75 probability for recruitment

| $\alpha$ | $\beta$ | $n_0$ | $n_1$ | $\hat{N}$ | jackvar | $\widehat{var}(\hat{N})$ |
|---|---|---|---|---|---|---|
| .10 | .07 | 10.02 | 0.00 | 105.06 | 876.69 | 880.47 |
| .10 | .07 | 10.02 | 10.37 | 104.03 | 572.89 | 465.20 |
| .10 | .03 | 10.04 | 0.00 | 116.92 | 4326.46 | 3659.52 |
| .10 | .03 | 10.04 | 2.44 | 105.67 | 3443.38 | 2898.39 |
| .15 | .03 | 15.00 | 0.00 | 109.36 | 2031.49 | 2140.59 |
| .15 | .03 | 15.00 | 4.69 | 103.61 | 1660.75 | 1122.22 |
| .20 | .03 | 20.04 | 0.00 | 102.36 | 506.22 | 357.62 |
| .20 | .03 | 20.04 | 7.09 | 101.27 | 386.50 | 292.16 |
| .20 | .05 | 19.94 | 0.00 | 101.54 | 151.00 | 134.94 |
| .20 | .05 | 19.94 | 15.68 | 101.59 | 115.78 | 98.12 |

Table 6.2: $N = 200$. The condition is in-degree $\geq 2$ .75 probability for recruitment

| $\alpha$ | $\beta$ | $n_0$ | $n_1$ | $\hat{N}$ | jackvar | $\widehat{var}(\hat{N})$ |
|---|---|---|---|---|---|---|
| .05 | .015 | 10.04 | 0.00 | 226.86 | 17819.12 | 18197.00 |
| .05 | .015 | 10.04 | 1.40 | 193.52 | 15533.29 | 14228.32 |
| .05 | .030 | 9.91 | 0.00 | 222.81 | 12527.91 | 9860.61 |
| .05 | .030 | 9.91 | 5.20 | 209.43 | 9302.13 | 5719.18 |
| .10 | .030 | 19.91 | 0.00 | 203.52 | 1124.88 | 897.92 |
| .10 | .030 | 19.91 | 16.38 | 203.17 | 911.60 | 586.95 |
| .10 | .015 | 19.97 | 0.00 | 218.87 | 9702.30 | 6185.16 |
| .10 | .015 | 19.97 | 4.92 | 206.32 | 7421.59 | 3981.26 |

Table 6.3: $N = 400$. Without recruiting any respondents. This table is used for comparison with models 6 and 7 of the next table.

| $\alpha$ | $\beta$ | $n_0$ | $n_1$ | $\hat{N}$ | jackvar | $\widehat{var}(\hat{N})$ |
|---|---|---|---|---|---|---|
| .05 | .03 | 19.79 | 0.00 | 403.54 | 2464.03 | 2186.16 |

The parameters used in each model for the next simulation are listed below in the order of appearance in the table.

Model 1: $\alpha_0 = .2 \quad \beta_0 = .07 \quad \beta_1 = .07 \quad \beta_2 = .0075 \quad N_0 = 100 \quad N_1 = 100$

Model 2: $\alpha_0 = .2 \quad \beta_0 = .03 \quad \beta_1 = .03 \quad \beta_2 = .0075 \quad N_0 = 100 \quad N_1 = 100$

Model 3: $\alpha_0 = .2 \quad \beta_0 = .03 \quad \beta_1 = .03 \quad \beta_2 = .0150 \quad N_0 = 100 \quad N_1 = 100$

Model 4: $\alpha_0 = .2 \quad \beta_0 = .07 \quad \beta_1 = .03 \quad \beta_2 = .0150 \quad N_0 = 100 \quad N_1 = 100$

Model 5: $\alpha_0 = .2 \quad \beta_0 = .07 \quad \beta_1 = .03 \quad \beta_2 = .0075 \quad N_0 = 100 \quad N_1 = 100$

Model 6: $\alpha_0 = .2 \quad \beta_0 = .03 \quad \beta_1 = .03 \quad \beta_2 = .0075 \quad N_0 = 400 \quad N_1 = 100$

Model 7: $\alpha_0 = .2 \quad \beta_0 = .03 \quad \beta_1 = .03 \quad \beta_2 = .0150 \quad N_0 = 400 \quad N_1 = 100$

Table 6.4: Initial sample is taken only from $N_0$. The condition is in-degree $\geq 1$ and in $N_1$ Again with $\alpha_1 = 0.75$ probability for recruitment.

| $n_0$ | $n_1$ | $\hat{N}_0$ | $\hat{N}_1$ | $jackvar(\hat{N}_0)$ | $\widehat{var}(\hat{N}_0)$ | $jackvar(\hat{N}_1)$ | $\widehat{var}(\hat{N}_1)$ |
|---|---|---|---|---|---|---|---|
| 20.00 | 10.54 | 100.67 | 102.53 | 58.12 | 50.44 | 617.96 | 904.85 |
| 19.95 | 10.45 | 102.78 | 110.94 | 388.88 | 362.57 | 3014.49 | 3183.04 |
| 20.06 | 19.54 | 102.17 | 100.50 | 364.66 | 300.22 | 560.54 | 400.36 |
| 19.77 | 19.42 | 100.59 | 103.28 | 59.41 | 51.72 | 613.26 | 513.71 |
| 20.10 | 10.56 | 100.69 | 106.39 | 55.09 | 42.83 | 2574.61 | 2678.65 |
| 20.16 | 10.42 | 405.49 | 108.31 | 2167.12 | 2160.27 | 3432.13 | 2378.00 |
| 19.92 | 19.40 | 403.19 | 101.68 | 2549.16 | 2728.70 | 620.37 | 457.35 |

### 6.4.9 Conclusions

Regardless of the model, a small increase in sample size can help tremendously, especially when arc probabilities are small.

For the model containing a single arc probability $\beta$ for all people in the population it is more advantageous to take a larger initial sample than to adaptively add respondents. In the case where the study was already been conducted and respondents were added adaptively such as in the Colorado Springs 1990 project, it can be very beneficial to incorporate the recruited respondents information.

The simulations show interesting and very important results in the case where the model consists of three different arc probabilities and the initial sample is taken within the a single subgroup, section 6.4.3. The results suggest that the variance of the estimators for $N_0$ and $N_1$ are similar to as if they were estimated separately by a single $\beta$ model and sampling was done within each group separately. This finding suggests that the design plays an important role in finding people of interest and except for sample size does not affect the estimation process under this model. Thus an adaptive link tracing design can be used to obtain reasonable estimates of the size of a very hard to reach hidden population.

## Chapter 7

# Future Research

There are still many areas of adaptive sampling that deserve attention. A general problem with all design unbiased estimators is that they are dependent upon the design being carried out properly. When the sampling is not carried out according to the design this can effect estimation of the parameters of interest greatly. Often, especially in adaptive sampling, it is not feasible to sample according to the design exactly. In addition, these sampling problems may be correlated with the parameters of interest by some aspect. For example, a researcher may not have enough funding to sample the entire network if it is too large and many design unbiased adaptive sampling estimators require the entire network to be observed.

For the latter reasons it is important to research model based estimators when adaptive design is used. As seen in chapter 6 model based estimators in conjunction with an adaptive design can be very useful. Many such estimators will require estimators for their variance. In conventional sampling, resampling methods are often used to obtain estimates of variance. These methods must be explored more thoroughly in the context of adaptive sampling. Often in adaptive sampling the number of units observed is dependent upon the $y-$values observed. Thus by removing one unit from the initial sample may lead to the removal of a random number of units from the final sample. This issue has not been researched for most if not all resampling techniques. The existence of

such techniques for adaptive sampling procedures would be very useful. In conclusion, model based estimators and resampling techniques for adaptive sampling designs deserve further research.

# Appendix A

# Proofs of Selected Formulae In Chapter 3

## A.1 Proof of Theorems 3.1 and 3.2

There are many similarities in the derivation of $\hat{\mu}_{1+}$ and $\hat{\mu}_{2+}$. For this reason we will use $\hat{\mu}$ to represent either $\hat{\mu}_1$ or $\hat{\mu}_2$ and similarly $\hat{\mu}_+$ in deriving the estimators. Both original estimators are improved by taking their expectation given the same sufficient statistic, $d^+$. $\hat{\mu}(s_0)$ is $\hat{\mu}$ as a function of the initial sample $s_0$. Let $S_0$ be a random variable that takes on values from the sample space $\mathcal{S}$ and $P(S_0 = s_0^{'}|D^+ = d^+)$ is the probability of that initial sample given $d^+$.

$$\begin{aligned} \hat{\mu}_+ &= E[\hat{\mu}|D^+ = d^+] \qquad \text{(A.1)} \\ &= \sum_{s_0^{'} \in \mathcal{S}_{'}} \hat{\mu}(s_0^{'}) P(S_0 = s_0^{'}|D^+ = d^+) \end{aligned}$$

$$\text{(A.2)}$$

The conditional probability of the initial sample given the sufficient statistic, $P(S_0 = s_0^{'}|D^+ = d^+)$, can be broken down into two parts. The first part is an indicator variable which will be one if the initial sample is compatible with $d^+$ or zero if it is not, equation

3.21, and the second is the number of initial samples compatible with $d^+$.

Then

$$
\begin{aligned}
P(S_0 = s_0^{'} | D^+ = d^+) &= \frac{P(S_0 = s_0^{'}, D^+ = d^+)}{P(D^+ = d^+)} \\
&= \frac{P(S_0 = s_0^{'}, D^+ = d^+)}{\sum_{s_0^{'} \in \mathcal{S}_{'}} P(S_0 = s_0^{'}, D^+ = d^+)} \\
&= \frac{P(S_0 = s_0^{'}) P(D^+ = d^+ | S_0 = s_0^{'})}{\sum_{s_0^{'} \in \mathcal{S}_{'}} P(S_0 = s_0^{'}) P(D^+ = d^+ | S_0 = s_0^{'})} \\
&= \frac{P(S_0 = s_0^{'}) I(s_0^{'}, d^+)}{\sum_{s_0^{'} \in \mathcal{S}_{'}} P(S_0 = s_0^{'}) I(s_0^{'}, d^+)} \\
&= \frac{\frac{1}{\binom{N}{n}} I(s_0^{'}, d^+)}{\sum_{s_0^{'} \in \mathcal{S}_{'}} \frac{1}{\binom{N}{n}} I(s_0^{'}, d^+)} \\
&= \frac{\frac{1}{\binom{N}{n}} I(s_0^{'}, d^+)}{\frac{1}{\binom{N}{n}} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+)} \\
&= \frac{I(s_0^{'}, d^+)}{\sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+)}
\end{aligned}
\tag{A.3}
$$

$P(S_0 = s_0^{'}|D^+ = d^+) = \frac{I(s_0^{'},d^+)}{\sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'},d^+)}$, which is simply one divided by the number of possible samples that yield the specified value $d^+$.

Let

$$L = \sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'}, d^+) \tag{A.4}$$

and let:

$$z_k^{'} = \begin{cases} 1 & \text{if} \quad k \in s \text{and} \quad e_k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

Recall that $e_s$ is the number of sample edge units. There are $\binom{e_s}{e_{s_0}}$ possible combinations for choosing the sample edge units in the initial simple random sample. Let $G = \binom{e_s}{e_{s_0}}$ Then $L$ equals the total number of combinations compatible with $d^+$. $L = G\prod_{k=1}^{K}\binom{x_k}{f_k z_k^{'}}$

$1 - \binom{N-1}{n}/\binom{N}{n} = \frac{n}{N}$ equals the probability of including any unit $i$ in $\hat{\mu}$ that doesn't satisfy the criteria. Thus each set of $e_{s_0}$ sample edge units given $d^+$ has an equal probability of being selected in the initial simple random sample.

Also, out of the $G$ combinations any single edge unit appears $\binom{e_s-1}{e_{s_0}-1}$ times.

And so out of the $L$ combinations any single edge unit appears $\binom{e_s-1}{e_{s_0}-1}\prod_{k=1}^{K}\binom{x_k}{f_k z_k^{'}}$

$= \binom{e_s - 1}{e_{s_0} - 1} \frac{L}{G}$

For the proof we will break up the estimators into two parts. The first part will be the without sample edge units and the second part will be the sample edge units.

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} w_i(1-e_i) + \frac{1}{n}\sum_{i=1}^{n} w_i e_i \\
&= \frac{1}{n}\sum_{i=1}^{n} w_i(1-e_i) + \frac{1}{n}\sum_{i=1}^{n} y_i e_i \\
&= \frac{1}{n}\sum_{i=1}^{n} w_i(1-e_i) + \frac{1}{n}\sum_{i \in s_0, e_i=1} y_i
\end{aligned} \tag{A.6}$$

$$\begin{aligned}
\hat{\mu}_2 &= \frac{1}{N}\sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k}(1-e_k^{'}) + \frac{1}{N}\sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k} e_k^{'} \\
&= \frac{1}{N}\sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k}(1-e_k^{'}) + \frac{1}{N}\sum_{k=1}^{K} \frac{y_k z_k}{n/N} e_k^{'} \\
&= \frac{1}{N}\sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k}(1-e_k^{'}) + \frac{1}{n}\sum_{k=1}^{K} y_k z_k e_k^{'} \\
&= \frac{1}{N}\sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k}(1-e_k^{'}) + \frac{1}{n}\sum_{i \in s_0, e_i=1} y_i
\end{aligned} \tag{A.7}$$

Let:

$$C_1 = \frac{1}{n}\sum_{i=1}^{n} w_i(1-e_i) \tag{A.8}$$

Let:

$$C_2 = \frac{1}{N}\sum_{k=1}^{K} \frac{y_k^* z_k}{\alpha_k}(1-e_k^{'}) \tag{A.9}$$

$C$ (sub 1 or 2) is the component of each estimator that does not contain the sample edge units. This component is constant for $I(s_0^{'}, d^+) = 1$. The second component containing the sample edge units, $\frac{1}{n}\sum_{i\in s_0, e_i=1} y_i$ is the same for both estimators.

$$\begin{aligned}
\hat{\mu}_+ &= \frac{1}{\sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'}, d^+)} \sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'}, d^+)\hat{\mu}(s_0^{'}) \\
&= \frac{1}{L}\sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'}, d^+)[C + \frac{1}{n}\sum_{i\in s_0^{'}, e_i=1} y_i] \\
&= C + \frac{1}{L}\sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'}, d^+)\frac{1}{n}\sum_{i\in s_0^{'}, e_i=1} y_i
\end{aligned}$$

$\sum_{s_0^{'}\in\mathcal{S}_{'}} I(s_0^{'}, d^+)\frac{1}{n}\sum_{i\in s_0^{'}, e_i=1} y_i$ equals the number of times each sample edge unit appears in all the possible combinations yielding $I(s_0^{'}, d^+) = 1$ times the sum of the distinct edge units.

$$= \frac{L}{G}\binom{e_s-1}{e_{s_0}-1}\sum_{i=1, e_i=1}^{e_s} y_i$$

$$= C + \frac{1}{L}\frac{L}{G}\binom{e_s-1}{e_{s_0}-1}\frac{1}{n}\sum_{i=1, e_i=1}^{e_s} y_i$$

$$
\begin{aligned}
&= C + \frac{\binom{e_s-1}{e_{s_0}-1}}{\binom{e_s}{e_{s_0}}} \frac{1}{n} \sum_{i=1,e_i=1}^{e_s} y_i \\
&= C + \frac{e_{s_0}}{e_s} \frac{1}{n} \sum_{i=1,e_i=1}^{e_s} y_i \\
&= C + \frac{e_{s_0}}{e_s} \frac{1}{n} e_s \bar{y}_e \\
&= C + \frac{1}{n} e_{s_0} \bar{y}_e \\
&= C + \frac{1}{n} \sum_{i \in s_0, e_i=1} \bar{y}_e
\end{aligned}
\tag{A.10}
$$

Since $y_k^{'} = w_i^{'} = \bar{y}_e$ for all $i, k \in s_0$ and $e_i, e_k^{'} = 1$ the rest follows.

$$
\hat{\mu}_{1+} = E[\hat{\mu}_1 | D^+] = \frac{1}{n} \sum_{i=1}^{n} w_i^{'} \tag{A.11}
$$

$$
\hat{\mu}_{2+} = E[\hat{\mu}_2 | D^+] = \frac{1}{N} \sum_{k=1}^{K} \frac{y_k^{'} z_k}{\alpha_k} \tag{A.12}
$$

## A.2 Proof of the Variances

$$
\begin{aligned}
var(\hat{\mu}_+) &= var(\hat{\mu}) - E[(\hat{\mu} - \hat{\mu}_+)^2] \\
&= var(\hat{\mu}) \\
&\quad - \sum_{d^+ \in \mathcal{D}^+} P(d^+) \frac{1}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+) (\hat{\mu}(s_0^{'}) - \hat{\mu}_+)^2 \\
&= var(\hat{\mu})
\end{aligned}
$$

$$
\begin{aligned}
& - \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+)(\hat{\mu}(s_0^{'}) - \hat{\mu}_+)^2 \\
= \quad & var(\hat{\mu}) \\
& - \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+) \times \\
& ((C + \frac{1}{n} \sum_{i \in s_0^{'}, e_i = 1} y_i) - (C + \frac{1}{n} e_{s_0} \bar{y}_e))^2 \\
= \quad & var(\hat{\mu}) \\
& - \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+)(\frac{1}{n} \sum_{i \in s_0^{'}, e_i = 1} y_i - \frac{1}{n} e_{s_0} \bar{y}_e)^2 \\
= \quad & var(\hat{\mu}) \\
& - \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+)(\frac{1}{n}( \sum_{i \in s_0^{'}, e_i = 1} y_i - e_{s_0} \bar{y}_e))^2 \\
= \quad & var(\hat{\mu}) \\
& - \frac{1}{n^2} \sum_{d^+ \in \mathcal{D}^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0^{'} \in \mathcal{S}_{'}} I(s_0^{'}, d^+)( \sum_{i \in s_0^{'}, e_i = 1} y_i - e_{s_0} \bar{y}_e)^2
\end{aligned}
\tag{A.13}
$$

# References

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

[2] Frank Ove and Snijders Tom (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. *Journal of Official Statistics* **10**, 53-67.

[3] Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc.* **B 17**, 269-278.

[4] Horvitz, D.G., and Thompson D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663-685.

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

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

[7] Raj, D. and Khamis, S. H. (1958). Some Remarks on Sampling With Replacement. *Annals of Mathematical Statistics* **29** 550-557.

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

[9] Salehi, M.M. (1998). Adaptive Cluster Sampling. Ph.D. Thesis, University of Auckland.

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

[11] Särndal Carl-Erik (1996). Efficient Estimators with Simple Variance in Unequal Probability Sampling. *Journal of the American Statistical Association* **91** 1289-1300.

[12] Shao, Jun and Tu, Dongsheng (1995). *The jackknife and bootstrap*. New York: Springer.

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

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

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

[16] 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.

# Vita

**Education:**
Ph.D. in Statistics (1999)
The Pennsylvania State University
B.A. in Mathematical Sciences/Statistics (1993)
Rice University

**Experience:**
Graduate Research Assistant (6/97-present)
-Funded by NIH Grant, Steven K. Thompson (P.I.)
Statistical Consultant (2/98-2/99)
-El Paso County Department of Health and Environment
Graduate Teaching Assistant (8/93-5/97)
Instructor for Introductory Statistics Course (Summers 1994 & 1995)

**Papers In Review & Technical Reports:**
"Improved Unbiased Estimators in Adaptive Cluster Sampling"
"Adaptive Cluster Sampling Without Replacement of Clusters"
"Improving Unbiased Estimators in Adaptive Cluster Sampling"
All papers were written by Arthur L. Dryver and Steven K. Thompson

**Computer Skills:**
Environments:
-UNIX, Windows, and Apple.
Statistical Software and Programming Languages:
-Basic, C, Fortran, LaTeX, LINDO, MATLAB, Minitab,
-PAJEK, PASCAL, SAS, SIMAN, S-Plus, and SPSS

**Awards and Presentations:**
NSF Travel Grant to present at University of Nebraska -Lincoln, 1997
Vollmar-Kleckner Travel Grant to present at JSM, 1998
Presented at Sunbelt XIX 1999