THE PENNSYLVANIA STATE UNIVERSITY

The Graduate School

Department of Statistics

CONTRIBUTIONS TO ADAPTIVE WEB SAMPLING DESIGNS

A Thesis in

Statistics

by

Hong Xu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2007

The thesis of Hong Xu was reviewed and approved* by the following:

Steve K. Thompson
Professor of Statistics
Thesis Coadvisor

James L. Rosenberger
Professor of Statistics
Thesis Coadvisor, Chair of Committee

Vernon M. Chinchilli
Distinguished Professor of Health Evaluation Sciences

Runze Li
Associate Professor of Statistics

Mosuk Chow
Associate Professor of Statistics

BRUCE LINDSAY
Professor of Statistics, Department Head

*Signatures are on file in the Graduate School.

Abstract

CONTRIBUTIONS TO ADAPTIVE WEB SAMPLING DESIGNS

Investigation of the characteristics and estimation of quantities of hidden and hard-to-access population are of interest to scientists. Such populations are difficult to target because of their elusive nature or other prohibitive characteristics. So crafting designs of a representative sample and creating estimation methods to obtain efficient information from the sampled data are core challenges for people who investigate sampling theories and applications. Thompson (2006a) proposed an adaptive web sampling (AWS) scheme which takes into account the social networks between subjects to get a more efficient sample, and the procedures are more flexible than adaptive sampling. Three papers which contribute to AWS are included in this thesis. They focus on new resampling mythologies to improve the inferential estimation, produce designs with practical restrictions to minimize the cost and maximize the sampling simultaneously and model based estimation for non-responses. Simulated and real data sets are used to demonstrate implementation.

Contents

List of Tables

List of Figures

Acknowledgements

I am most grateful and indebted to my thesis advisor, Professor Steven K. Thompson, for the large doses of guidance, patience, and encouragement he has given me during my time here at Penn State. I am also grateful and indebted to my Co-advisor Professor James L. Rosenberger, for inspiration and enlightening discussions on a new topic of the cost issue. I am also thankful to my committee members Runze Li, Mosuk Chow, and Vernon M. Chinchilli. I thank my fellow students, Mike Kwanisai for insightful commen- tary and kind cooperation. My husband Zhe Zhang, daughter Emily Zhang, my family and relatives showed a great deal of patience and understanding during my studies that seemed to last forever. They always picked me up when the going was tough.

Hong Xu

The Pennsylvania State University

May, 2007

Chapter 1

Introduction

1.1 Background

The design of a representative sample from a rare, hidden and hard-to-access population and estimation methods to obtain efficient information from the sampled data are core challenges for people who investigate sampling theories and applications. The goal of this thesis is to develop more efficient resampling strategies to make more reliable population estimates, and to propose new practical sampling methods in the face of cost and nonresponses during the sampling procedures.

A rare population is defined as a small subset of the whole population, which could be one hundredth percentage, one thousandth percentage or even less (Kalton & Anderson (1986)). Such populations include very high/low income households; racial, ethnic, or religious groups; persons with specific illnesses; homeless persons; illegal drug users; individuals interested in continuing education; homosexual men in a metropolitan area; female smokers with high school education or less and HIV/AIDS risk people e.c.. Because of the relative rarity and elusive nature of these populations, conventional sampling designs such as simple random sampling are inefficient for producing data on the individuals of interest.

Methods of sampling rare populations have been reviewed by Sudman & Kalton (1986), Kish (1991), Kalton (1991), Kalton (1993). After that, adaptive sampling was introduced by Thompson & Seber (1996), and followed by adaptive clustered

sampling (Thompson (1990a)), stratified adaptive clustered sampling (Thompson (1990b)). Most of this work is applied to spatial science. Link-tracing sampling can also be called one type of adaptive sampling, which can be implemented both to spatial and social science. Potterat et al. (1993) used link tracing to study data from an HIV high risk population in Corolado Springs. Thompson (2006a) and Thompson (2006b) used the same data for his new adaptive web sampling designs, target and random walk designs. This thesis extends and advances the work done by Steven K. Thompson.

1.2 Structure of the thesis

This thesis is organized as follows. Chapter 2 gives a brief overview of social networks and their characteristics. Commonly used sampling methods, adaptive sampling (Thompson & Seber (1996)) and adaptive web sampling (Thompson (2006a)). Sampling inference are also discussed. The next three chapters are potentially papers which contribute to adaptive sampling and adaptive web sampling. Chapter 3 presents new resampling strategies for inference based on adaptive web sampling designs. A Markov Chain Monte Carlo (MCMC) procedure is the driving tool for the procedure in this chapter. Chapter 4 presents a class of adaptive web sampling designs under the cost constraints in terms of time, money and risk etc. One type of adaptive web sampling and cost model were described and implemented in this chapter. A simulated data set of banking wire transactions is used to illustrate the procedures discussed here. A data set from a high-risk population from a Colorado Springs study is also used as an application of the methods in this chapter. In chapter 5, a new model-based approach was proposed that is an extension of the model by Chow & Thompson (1998). This model accommodates non-random non-responses and shows how maximum likelihood estimates could be obtained. Chapter 6 sum-

marizes the results of this thesis and presents conclusions for the study. A discussion of the limitations of this study and suggestions for further research concludes Chapter 6.

Chapter 2

Literature Review

2.1 Sampling

Sampling is the process of selecting units (e.g., people, organizations, plants, animals) from a population of interest so that by only studying the sample we may estimate some characteristics of a population. For example, to estimate the proportion and characteristics of HIV/AIDS positive people in the USA, it would be too expensive to interview all Americans and ask who is infected or not. We could get just as valid information with a smaller sample and the estimates from it. So long a sample is as representative as possible of the whole population we are investigating. Even with the perfect questionnaire (if such a thing exists), our sampled data will only be useful if the respondents are typical of the population as a whole. For this reason, implementation of different sampling methods according to the properties or characteristics of the population is very important. Thus, to estimate the prevalence of a rare disease, the sample might consist of a number of medical institutions, each of which has records of patients treated. Or in a study of transmission of disease, a sampling of injection drug users is obtained by following social links from one member of the population to another (Thompson (2002)).

2.2 Sampling designs for targeting rare populations

In the sampling literature, there are many conventional sampling methods in which the selection procedure does not depend in any way on observations made during the survey. These designs include simple random sampling, stratified sampling, cluster and systematic sampling, and multistage sampling etc. In conventional sampling, the design is based entirely on a *prior* information, and is fixed before the study begins. Under such designs, researchers make decisions about the sample size before the sampling procedure begins. The sampling frame is usually available or can be easily obtained for such designs. And it is not difficult to obtain data on the units of interest when conventional sampling designs are used. However, conventional sampling methods are not appropriate for sampling the hidden and hard-to-access population such as HIV/AIDS persons, rare and endangered species. Because of the relative rarity and elusive nature of these populations, samples obtained by conventional methods tend to contain a very few number of elements from the population of interest. For example, the original impetus for the National Health and Social Life Survey (Laumann et al.1994), a US national probability sample survey of sexual behavior was in large part concern regarding the AIDS epidemic. The survey broke new ground in using a probability based sampling design and estimation methods for the study of human sexual behaviors. Funding constraints due to political controversies limited the sample size to 3432 people, which was considerably less than the originally proposed sample size. When the data were collected it was found that the people in the sample who reported having tested positive for the HIV virus numbered only six. For such a rare group, even a very large conventional sample would be unlikely to be adequate. Another weakness of conventional sampling for a hidden population is that, the selection probability is equal in conventional sampling, but for some populations, the individual selection probability may not be equal because

some individuals may have a higher inclusion probability than others. For example, in a survey to estimate rare diseases such as HIV or hepatitis C infections, a simple random sample of medical centers is selected and the records for the patients treated in the medical centers are obtained. However, some patients may be treated at more than one medical center, so these patients have a higher possibility of being included in the study than others.

There have also been a number of approaches to estimating the size of hidden populations. Adaptive clustered sampling (Thompson (1990a)) and adaptive stratified clustered sampling (Thompson (1990b))are used for geographic clustered distributed rare populations. Other adaptive sampling methods such as network sampling (*multiplicity sampling*, Birnaum & Sirken (1965)), link tracing sampling (Potterat et al. (1993)), respondent driven sampling (Heckathorn (1997), Heckathorn (2002)) and so on can be used for sampling populations with social network structure.

Link Tracing Sampling

The idea was first introduced by Coleman (1958). Every subject is interviewed during the study and asked questions on their sociometric relations and more subjects are included by following links. Potterat et al. (1993) and McCoy & Inciardi (1993) used such designs in their study of heterosexual transmission and cocaine use and associated sexual behaviors respectively. In link-tracing designs, investigators use links between people to find other people to include in the sample. Any sociometric relation of interest can define a link between two individuals in the population. The sampling design adapts based on observations made during the survey; for example, drug users may be asked to refer other drug users to the researcher. This is necessary because they could not know what social connections to follow or whom to include in the sample before the investigation. The key difference between

the conventional sampling and adaptive sampling is that in adaptive sampling, we take into account the information obtained during the sampling procedure. Despite having the advantage of conveniently increasing the sample size, it is frequently necessary to use link-tracing designs because it is sometimes the easiest practical way to identify members of rare and hard-to-reach populations (Spreen (1992); Steven K. Thompson (1 November,2002)).

In the social sciences, link-tracing designs provide the only practical way to observe and study social networks. The statistical literature on design and estimation with link-tracing designs includes procedures variously termed snowball sampling, random walks, and network sampling.

Snowball Sampling

Snowball sampling has considerable theoretical appeal. Goodman (1961) coined the term snowball sampling. Initial respondents are asked to identify the other people which are related to them to include in the sample and so on for a desired number of waves. The process stops after a certain number of waves or when there are no more newly mentioned subjects. Snowball sampling may be defined to include all or only a fixed number of subjects with whom they share a relationship. Snowball designs were also developed in the graph setting with a variety of initial probability sampling designs and any number of links and waves by Frank (1977), Frank (1978), and Frank (1979). Frank & Snijders (1994) discussed methods for estimating the size of a hidden population using snowball sampling.

Network Sampling

Much of the early work on network sampling was undertaken by Birnaum & Sirken (1965) and their associates at the National Center for Health Statistics. The links generally are symmetric, and new links added do not depend on the observed

information. One advantage of network sampling is that the inclusion probability for each unit in the sample data is known and easy to calculated. On the other hand, since it is necessary to ask additional screening questions and to spend resources locating identified members of the rare population, network sampling costs slightly more than standard procedures. In most of cases, network sampling costs more than is compensated by the reduction in sampling variance. Another shortcoming of such sampling designs is the requirement for accurate reporting about all persons in the network. Network sampling can be used for both networks in social science (Spreen. & Zwaagstra (1994)) or spatial science (Birnbaum & Sirken (1998)).

Random Walk Sampling

Klovdahl (1989) used the term "random walk" design to describe the situation where each subject is asked to name people with whom they have a social relationship. After that, one subject is randomly picked from the names at each stage of sampling. This sampling method is a modification of snowball sampling. Thompson (2006b) contributed uniform and target walk designs, which could be implemented more efficiently to network populations with isolated components. A random walk procedure applied at each step produced a design with the desired stationary probabilities. (Henzingera et al. (2000); Lawrence & Giles (1998)) used random walk procedures to investigating internet searching.

Respondent-Driven Sampling

Heckathorn (1997) first described "respondent-driven sampling". The respondent-driven sampling is based on an adaptive sampling design where the selection procedure is affected by the realized network in the population (Thompson & Seber (1996); Thompson & Frank (2000)). Salganik & Heckathorn (2004) discussed a population proportion estimation method based on such sampling designs.

2.3 Adaptive Web Sampling(AWS)

All the sampling designs described above are based on following links completely. At any stage of the procedure, new units selected depend on the values of the variables of interest associated with the units previously included, so the samples may contain a fair number of subjects of interest. But if the population is composed of more than one subset of linked components (Figure 2.1 and 2.2), link tracing may not be efficient to obtain sufficient subjects of interest. For example, in Figure 2.1, we will never reach cluster 2 if the sampling started from one unit in cluster 1. A new type of design introduced by Thompson (2006a) gains over link tracing type of sampling designs named Adaptive Web Sampling (AWS). AWS is more flexible in controlling how far/deep the sampling procedure could go, how the sample could be spread out, how large the sample size could be etc. It can be applied to any graph model with network structure. A network of spatially-based application of the designs, hidden human populations at risk for HIV/AIDS in Colorado Springs data, and a wintering waterfowl survey are evaluated in this paper. The work in Chapter 3 and Chapter 4 are based on the basic idea of such designs. Chapter 3 focuses on improvement of population proportion estimation under such designs; In chapter 4, we implement a cost model through such designs and evaluated it for different parameter values.

AWS is defined as follows: At any point in the sampling, the next unit or next set of units is with high probability selected from a distribution that depends on the values of variables of interest in an active set of units already selected. With low probability the next unit is selected from a distribution that does not depend on those values of variables of interest. The active set may consist of all the units selected so far, or the most recently selected units, or other possibilities such as the last two steps units or sequences of units (Thompson (2006a)). For example, in the study of injection drug users in relation to the spread of the HIV/AIDS prevalence, we first

picked one person, asked if he/she is HIV positive, but we do not include a new person by totally following the link from person one. Instead, with high probability say 90%, we selected a new person by following link, and with low probability the new person is randomly selected from the population. If person one did not report anyone with whom he share the injection, then person two will be randomly selected from the population. There are lots of variations on the general idea of AWS such as Random and Targeted Walk Sampling Designs also described by Thompson (2006b).

Figure 2.3 is a simple example of AWS design. The top left is a population of size 20. Red circle represent subjects which have characteristics we are interested in. Otherwise subjects are shown in yellow circles. The other plots illustrate three samples based on AWS. More subjects of interest are likely to be included in each sample. Figure 2.4 is an illustration of how AWS proceeds. The top left one is a small population of size 5. Units which are linked with each other are more likely to be selected than isolated units. An initial sample is shown in the top right plot, which is composed of node 1. At the second step, node 2 and node 3 could be included with high probability; node 4 and node 5 could be selected with lower probabilities. The bottom left showed node 2 is selected at the second step. Though node 3 has a higher potential selection probability than then others, it is possible that node 4 and node 5 could be selected in next wave. As shown in bottom right plot, node 5 is selected.

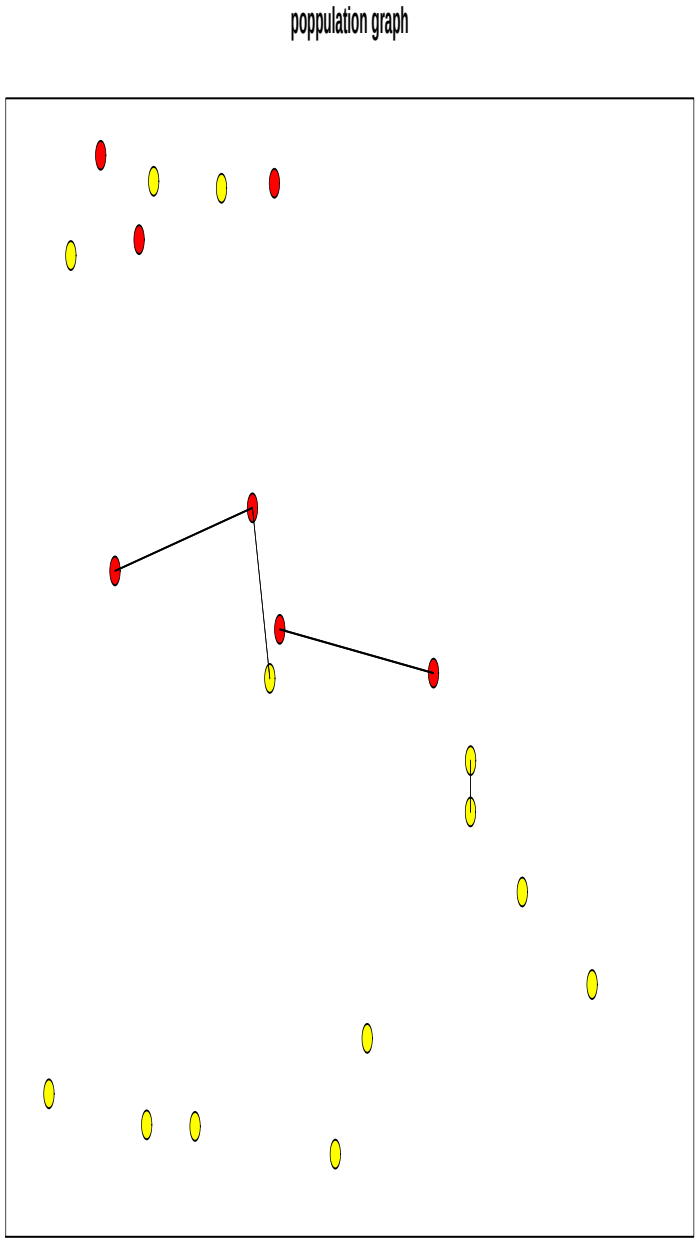


Figure 2.1. population size is 20; red dot: HIV positive people; yellow dot: HIV negative people; black line indicates link between two people.

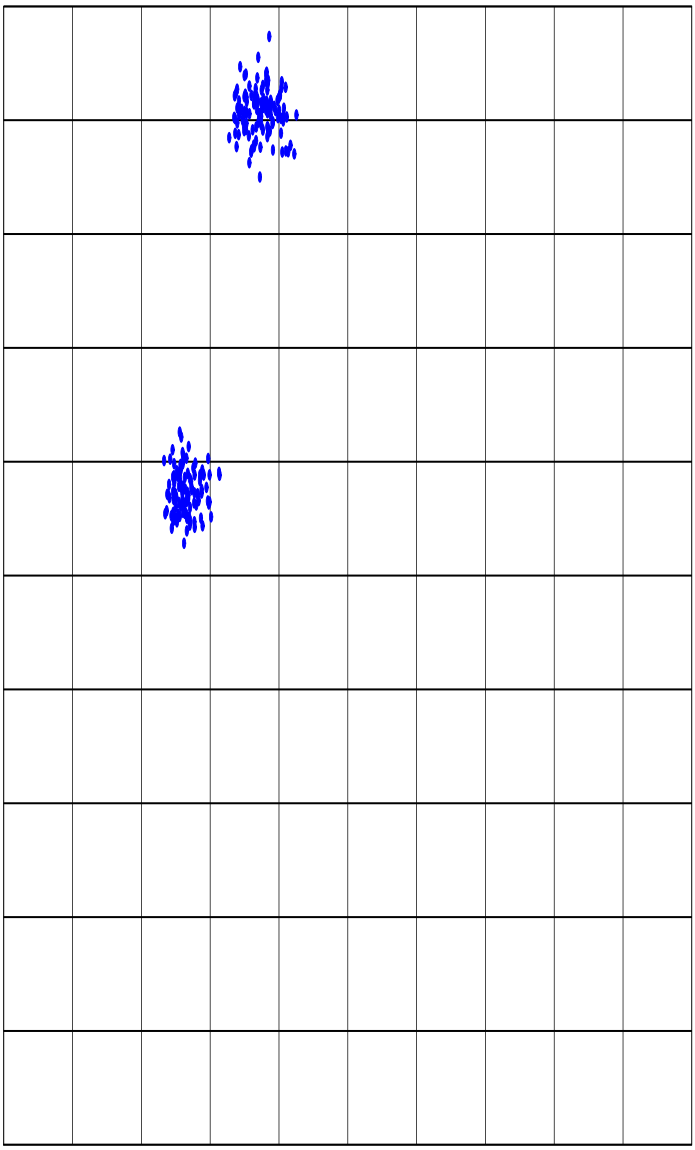


Figure 2.2. Spatial and population graph: blue dot: population in spatial environment

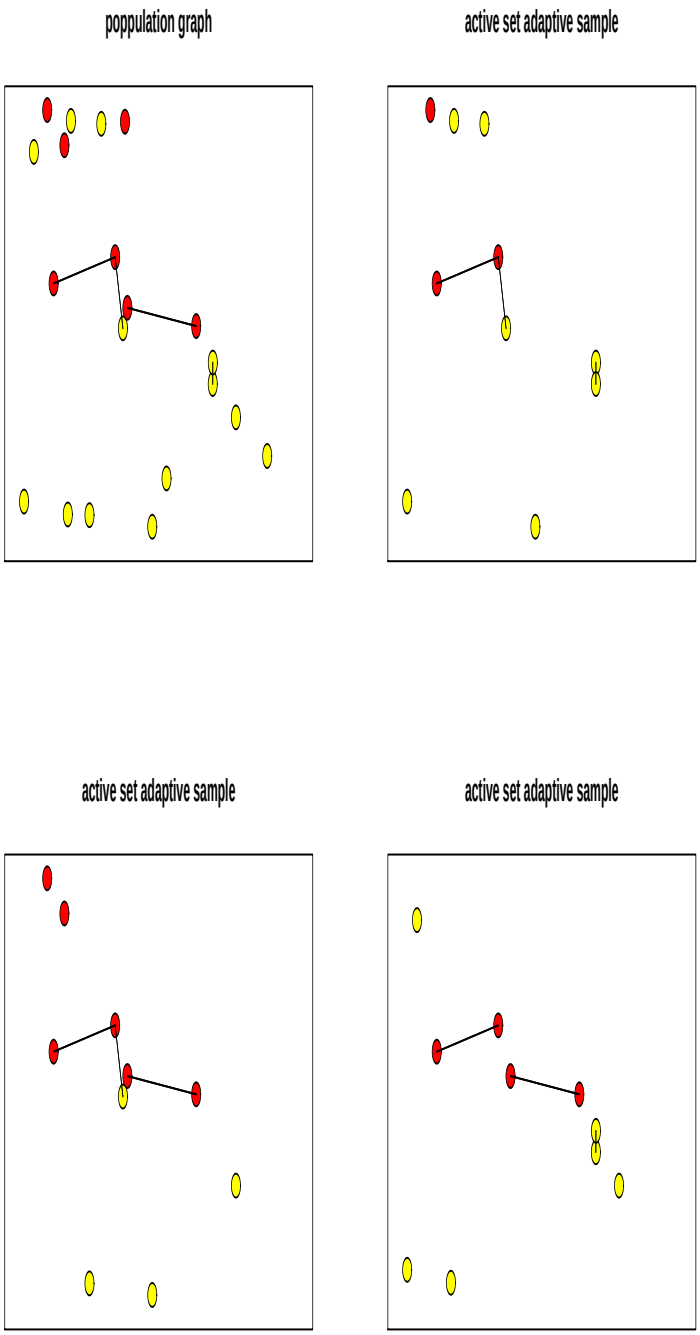


Figure 2.3. Top left: Population of size 20. Red circles represent units which have characters of interest. Other wise, units are shown in yellow circles. Other three plots are illustration of three samples based on AWS design.

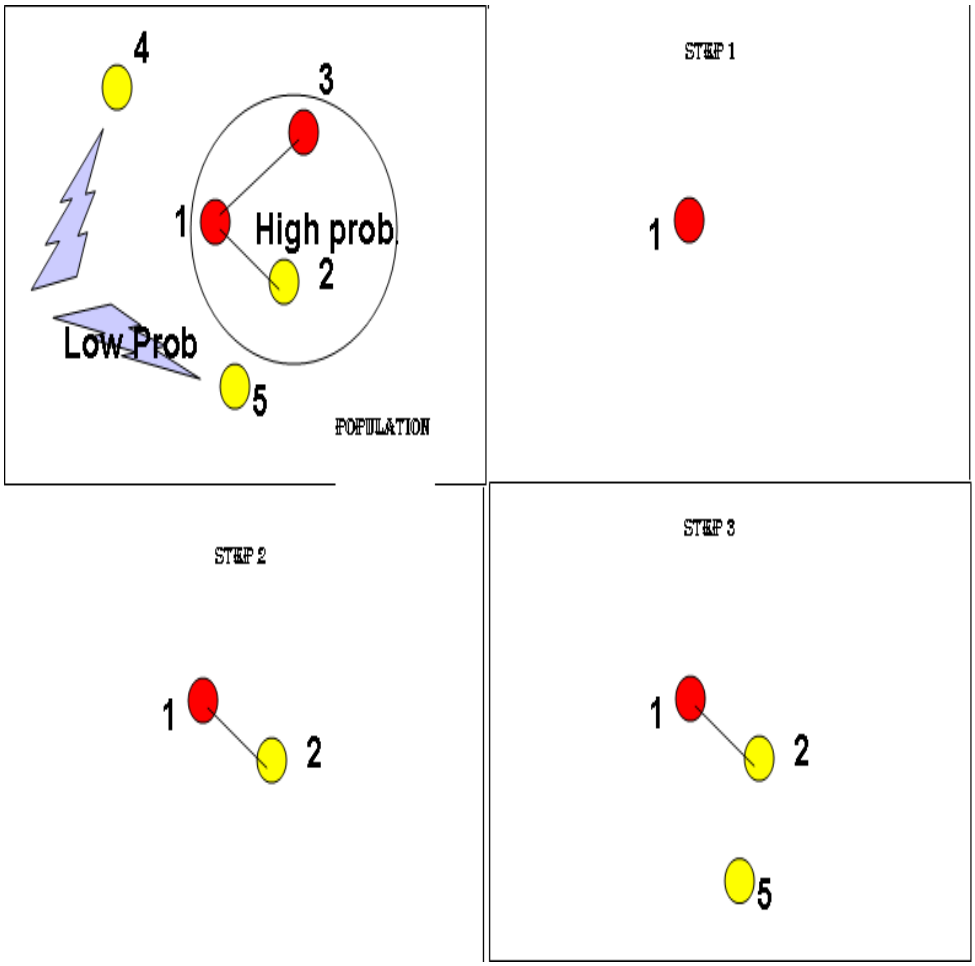


Figure 2.4. Top left: Population of size 5. Red circles represent units which have characters of interest. Other wise, units are shown in yellow circles. The other three plots are illustrations of how AWS could proceed at three steps.

2.4 Graph Model

Populations with network structure are often modeled as graphs with nodes of the graph representing individuals and the edges or arcs of the graph representing social links, relationships,or transactions. The population graph itself can be viewed as either a fixed structure or as a realization of the stochastic graph model. In the social environment, the links (arcs or edges) among people (nodes) can refer to friendship, marriage, sexual partners, or drug sharers. In a spatial environment, plots are

defined according to their geographic distribution. Two plots link to each other if they are in the same neighborhood. In the graph model, U is a population with N nodes:

$$U = \{1, 2, \cdots, N\}$$

and

$$y = (y_1, y_2, \cdots, y_N)$$

Where y_i indicates the variable we are interested, for example HIV positive or not, or dollar amount spent on heroin per month. In the spatial situation, y_i can refer to the number of species in plot i. The $N \times N$ matrix A indicating relationship between nodes. $A_{ij} = 1$ means there is a link from node i to node j, which means node i reported that he shared the drug-injection or needle with node j. The diagonal elements A_{ii} are set to zero.

$$A_{ij} = \begin{cases} 1 & \textit{link exists from i to j} \\ 0 & \textit{otherwise} \end{cases}$$

In a social environment, the relationship between people can be directional. For example, unit i reported to share the same injection with unit j, but unit j did not report such a relationship with i, so $A_{ij} = 1$ but $A_{ji} = 0$. If both of them reported sharing the same needle with each other, then $A_{ij} = A_{ji} = 1$, and if that is true for all i and j, then the matrix A is symmetric. The graph model is unidirectional. For example, the Y values for the population in Figure 2.4 is :

$$y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 0, y_5 = 0$$

and link matrix A is:

$$\left(\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

2.5 Sampling Inference

The objective of the sampling design is to infer the population characteristics based on the obtained sample data. These inferential outcomes include estimation, prediction, confidence interval, and test of hypothese. There are three categories of sampling inference methods: design-based, model-based and mixture of design and model based approaches.

An advantage of the design-based method is that little or no assumptions need to be made about the population characteristics. Both population and the population linkage-structure are viewed as fixed unknown values. Units included in the sample only depend on the sampling procedure, and the selection probability based on the designs are central to the inference. Another advantage is that regardless of whether it is a conventional or adaptive sampling procedure, the design-unbiased strategy will be unbiased no matter what the population itself is like. But the main disadvantage of design-based inference is that it is hard to compute the selection probability when the design becomes complicated. Most design-based estimators require knowledge of the inclusion probability (π_i) of each subject in the sample. The inclusion probability is the probability that subject i is included in the sample. It is hard to compute, especially without knowledge of the entire population network

structure. Sometimes approximations are used to estimate π_i. Heckathorn (2002) used the argument that in one-wave designs the inclusion probability for a node i is proportional to its degree d_i, i.e. $\pi_i \propto d_i$, to derive an estimator in a respondent-driven sampling study. Some estimators require knowledge of the population size, which in some practical cases, especially in hidden populations, is unknown.

Thompson (2006a) proposed three new methods for estimating the population proportion. Except for the first one which is only based on the initial sample, the other two not only consider the information in the initial sample, but also selection probabilities at each wave. Details are shown in Chapter 3, where three resampling strategies are proposed to improved those estimators.

Mathematical models for modeling graphs model have also been proposed and can be applied to a wide range of selection procedures. Under this method, the population is not fixed, but is a realization in the stochastic graph model. Based on the population characteristics, a joint distribution is assumed. Model-based methods also assume that the population model depends on parameters with a prior distribution. These models tend to be complex and solving them is not easy (Thompson & Frank (2000)). Newman et al. (2001) noted the difficulty of providing a realistic model for a social network.

There is, therefore, a need to develop sampling methods to make use of the advantages of design and model based methods. Researchers tend to find a way to balance the advantage of each by imposing design unbiasedness while seeking low mean square error under an assumed model. Partial knowledge is assumed about the probability distribution of the population parameters. For example, the distribution may be assumed to be of known parametric form but with unknown parameters. Or there is no parametric form, but the covariance matrix is proportional to a known matrix (Thompson & Seber (1996)).

2.6 The issues on sampling rare populations

The sampling design methods presented above are all implemented in the graph model. Sampling in graphs and inference from graph samples to the larger graph of interest is one of the core challenges in graph theory. To date, the graph theory has been insufficiently addressed in relation to its importance. The basic problem is that in the various literatures, inference about graphs is made as if the data represent a realization of the entire graph. Instead, the data are usually a sample, which in many cases is selected in such a way that it is not representative of the graph as a whole. An example is studies of the Internet, in which values for the average degree or the degree distribution are published. Typically, the data consist of a sample of around fifty thousand web pages (or, in the case of the physical Internet, a large number of routers) obtained by doing a 'web crawl', 'random walk', or other procedure in which links are followed from sites in the sample to add additional sites. Because such a selection procedure gives higher probability to sites with higher degree, the published estimates are substantially biased. The same bias is prevalent in studies of social networks, such as studies of hidden subpopulations at risk for HIV/AIDS in which social links are (necessarily) used in selecting the sample.

Designs for sampling in graphs, and inference methods based on design- and model-based approaches, is a dynamic new area of graph theory. The topic has fundamental theoretical importance, practical importance to almost every application involving networks, and challenging computational issues. It has been largely missed in the way graph theoretic issues have been conceptualized, and has come to the notice of researchers mainly when they discover something is wrong with their estimates. Also cost is always a issue for sampling in graphs in terms of time, money and risk, as in other situations where the researcher strives to obtain the most information about the population for the tests cost.

Chapter 3

Resampling designs for improved design-based inference

3.1 Introduction

Sampling in graphs is one of the core challenges of graph theory that has been insufficiently addressed so far in relation to its importance. The basic problem is that in the various literatures inference about graphs is made as if the data represent a realization of the entire graph. Instead, the data are usually a sample, which in many cases is selected in such a way that it is not in fact representative of the graph as a whole. So designs and associated inference for sampling in graphs are fundamentally and theoretically important.

The work in this chapter is based on Active Web Sampling (AWS) designs and associated estimators(Thompson (2006a)). The aim is to try new methods of resampling from the conditional distribution and estimating the Rao-Blackwell estimator from that. Rao-Blackwell estimators are improved estimators for population proportions based on conditional selection probabilities, and they are difficult to calculate when the sample size is large. In this chapter, three resampling methods are proposed and implemented in order to seek a more efficient way to approximate the exact Rao-Blackwell estimators.

Many sampling strategies can be used to sample hidden and hard-to-access populations, such as HIV/AIDs peoples, rare and endangered species. One such design is Adaptive Sampling (Thompson & Seber (1996)) design. In such designs,

additional neighboring units or individuals are introduced into a sample when a predefined condition is satisfied. This term is mainly used for designs that are based on geographic information. The units in a neighborhood are defined to link with each other, see Figure 2.2. In the social environment, *link-tracing* designs are used to include people by following links between them. Any social relation of interest can define a link between individuals in the population. Both adaptive sampling and link-tracing sampling designs are accomplished entirely by following links and work well in some situations. But those sampling designs are lack of flexibilities to control sample size, sample coverage. Since for such sampling design, snowball sampling for instance, new units are included by following link completely, the sample and statistic inference based on it are very sensitive to the starting points.

AWS designs gain more efficiency over the adaptive and link-tracing designs in some situation regarding to control the depth and breadth of sample coverage. Sample size could be predefined before the sampling procedure, and the designs are also easily implemented. The additional units which are included in the sample are not necessary by following links totally. Instead, at any point in the sampling, the next unit or next set of units is with high probability selected from a distribution that depends on the values of variables of interest in an active set of units already selected. With low probability the next unit is selected from a distribution not depending on those values of variables of interest. The active set may consist of all the units selected so far, or the most recently selected units, or other possibilities include the last two steps units or sequences of units (Thompson (2006a)). A simple example in the spatial setting is a population with two components (Figure 2.2). If the sampling started from units only in one component, we will never reach the other component by only following links and end up with only units in component one. AWS designs allow a jump from cluster one to cluster two, thus we could have a better chance of inclusion

of the units at each step. The sampling procedure stops when sample size satisfied a predefined number.

The preliminary estimators for population proportion can depend on initial sample, or conditional selection probability at each step. Such estimators can be improved by using Rao-Blackwell method. Thompson(Thompson (2006b)) introduced an idea of restamping method, which is to construct a Marked chain in order to get enough samples from the permutation sample space. Three restamping methods based on this idea are described in this paper. And their efficiencies are also compared.

3.2 Sampling Settings

3.2.1 Sampling Designs

AWS designs are used in studying population with network structure, which are often modeled as graphs with nodes of the graph representing individuals and the edges or arcs of the graph representing social links, relationships, or transactions. In the social environment, the links (arcs or edges) among people (nodes) can refer to friendship, marriage, sexual partners, or drug sharer. In a spatial environment, plots are defined according to the geographic distribution. Two plots link to each other if they are in a neighborhood.

For the design based method, the population is considered as a finite one with size N units and relations between the units. The population graph itself can be viewed as a fixed structure. Each unit is represented as a node and relations are represented as links. Population units are labeled as $U_1, U_2, \cdots, U_N$. Assume each of the N units has a value of interest y_i, which is a unknown constant associated with each population unit U_i. Randomness is introduced only through design itself. $A_{N\times N}$ is the link matrix among population units. $A_{ij} = 1$ if there is a link from unit i to unit

j, otherwise $A_{ij} = 0$. S_0 is an initial sample of n_0 elements with selection probability P_{S_0}. $S_1, S_2, \cdots$ are the new units selected at first step, second step and so on. S_i may include one unit or more units. If one new unit is added at each step, then at step k, selection probability for this new unit S_{k+1} is $q_{S_{k+1}|A_k}$ under AWS design until the sample size increases to n, which is predefined. A_k is the active set at step k, it could be recently selected units or the whole/part of units selected so far. And the next unit is included in the sample by following links from the Ak with probability d, and randomly selected from the unselected units with probability $1 - d$. So the next unit inclusion probability $q_{S_{k+1}|A_k}$ in the sample depending on the current active set A_k. If A_k is the units selected so far, and the units are selected without replacement, then the selection probability for next unit j can be written as:

$$q(j|A_k) = \begin{cases} d \times \frac{W_{A_k j}}{W_{A_k +}} + (1-d) \times \frac{1}{N - n_{A_k}} \\ \textit{link exists out from } A_k \\ \frac{1}{N - n_{A_k}} \\ \textit{no link exists out from } A_k \end{cases} \tag{3.1}$$

Where, W_{A_k+} could be associated with some variables which describe interested character outside from A_k.

is number of links from current active set A_k to unit j, w_{A_k+} is the total number of links outside from A_k, n_{A_k} is the number of units selected so far. The final sample is $S = \{S_0, S_1, \cdots, S_K\}$, where $K = n - n_0$ is the total number of steps. The probability of sample in the selection order is:

$$P(S) = P_0(S_0) q(S_1|A_0) q(S_2|A_1) \cdots q(S_K|A_{K-1}) \tag{3.2}$$

3.3 Design-Based Estimators

The design-based estimation(Thompson (2006a)) are described here. Those estimators are improved by using the Rao-Blackwell method. The design-based unbiased estimations of population mean and variance are based on the initial sample and conditional selection probability.

3.3.1 Est1

Based on Initial Sample values Suppose the initial sample only consists of one unit of selection probability π_0 with associated value y_0, then the unbiased estimator of population mean is:$\hat{\mu}_{01} = (1/N)y_0/\pi_0$. If the initial sample has more than one unit, $\hat{\mu}_{01}$ could be the unbiased estimator based on the initial sample design, such as H-T estimator and $\hat{\mu_{01}} = (1/N)\sum_{i\in S_0} y_i/\pi_i$. For an initial random sample,

$$\hat{\mu}_{01} = \overline{y_0} \tag{3.3}$$

3.3.2 EST2

Based on Conditional selection Probability: This estimator can be thought as a composite estimator of initial sample and nodes value with the selection probability step by step. The first part is the unbiased estimator of population total $\sum y_i$: $\hat{\tau_{S0}} = (N/n)\sum_{i\in S_0} y_i/\pi_i$. If the initial sample is random sample without replacement $\hat{\tau_{S0}} = (N/n)\sum_{S0} y_i = N\overline{y}_0$. The second part is the conditional selection probability after initial sample. Suppose at k-step with current active set A_{k-1}, the term is: $z_k = \sum_{j\in A_{k-1}} y_j + y_k/q_{A_{k-1}k}$. $q_{A_{k-1}k}$ is the selection probability of next unit at step k with y_k value. An unbiased estimator for the population mean:

$$\hat{\mu_{02}} = \frac{1}{Nn}\{n_0\hat{\tau_{S0}} + \sum z_i\} \tag{3.4}$$

3.3.3 EST3

Based on Generalized estimators: This estimator is the ratio of $N\hat{\mu}_{02}$ and $\hat{N}$. $\hat{N}$ is the estimator of population size N, which is also a composite estimator of two parts: estimator $\hat{N}_0$ of the population size N based on the initial sample and the estimator of population size at each step $\hat{N}_k$. $\hat{N}_0 = Ny_1, \hat{N}_i = \#Sk+1/P(S_ki)$. The ratio of two conditional probability-based estimators:

$$\hat{\mu}_{03} = \frac{N\hat{\mu}_{02}}{\hat{N}} \tag{3.5}$$

Where,

$$\hat{N} = 1/n\{n_0\hat{N}_0 + \sum \hat{N}_i\}$$

3.3.4 EST4

Based on mean of ratio estimator

$$\hat{\mu}_{04} = 1/n \sum \frac{z_i}{\hat{N}_i} \tag{3.6}$$

3.3.5 Example

An illustration will be given to show how the design works. Suppose Figure 3.1 is the population of size $N = 5$. Node 1 is linked with node 2 and node 3. Node 1 and node 3 are the ones with characteristics which we are interested in (red circles), and the associated value are $y_1 = 1$ and $y_3 = 1$. Node 2, node 4 and node 5 (yellow circles) have associated value $y_2 = y_4 = y_5 = 0$. The active set is the units selected, and new units are included by following link with probability $d = 0.9$. Unit selection probability is calculated by following(3.1). The total sample size is predefined to be 3. In other word, the sampling procedure stops once the sample size increases to 3.

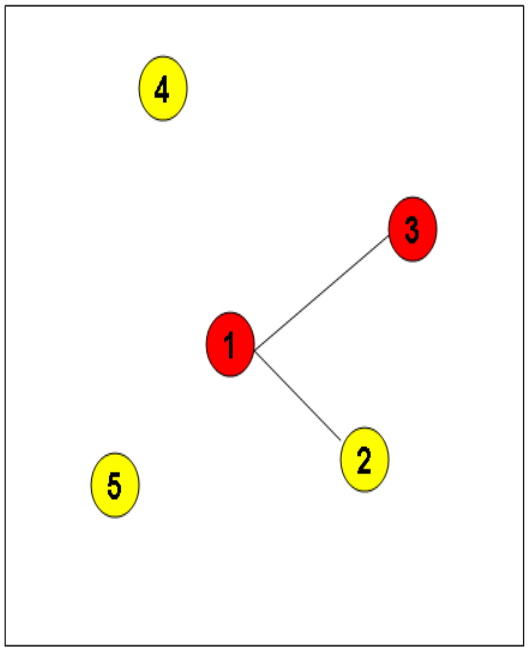


Figure 3.1. Realization of population of size 5. Red circles represent the units which have characteristics we are interested in. Yellow circles represent units which do not have the values we are interested in.

The sampling procedure:

Step-0: All the nodes in the population have equal selection probabilities.

$$P(S_0 = \{1\}) = P(S_0 = \{2\}) = P(S_0 = \{3\}) = P(S_0 = \{4\}) = P(S_0 = \{5\}) = \frac{1}{5}$$

Suppose one node is randomly selected to be the element in the initial sample S_0. In this example, we assume node 1 is selected at this step, so $S_0 = \{1\}$.

Step-1: The active set consists of all the nodes selected in the initial sample at step-0. A_i is used to denote the active set at step i. At step-1, $A_1 = \{1\}$ and the probability of selecting next unit j is:

if $j = 2$, then

$$P(S_1 = \{2\}|A_1 = \{1\}) = 0.9 * \frac{1}{2} + (1 - 0.9)\frac{1}{4} = 0.475$$

if $j = 3$, then

$$P(S_1 = \{3\}|A_1 = \{1\}) = 0.9 * \frac{1}{2} + (1 - 0.9)\frac{1}{4} = 0.475$$

if $j = 4$, then

$$P(S_1 = \{4\}|A_1 = \{1\}) = 0.9 * \frac{0}{2} + (1 - 0.9)\frac{1}{4} = 0.025$$

if $j = 5$, then

$$P(S_1 = \{5\}|A_1 = \{1\}) = 0.9 * \frac{0}{2} + (1 - 0.9)\frac{1}{4} = 0.025$$

Note that node 2 and node 3 have higher selection probability than node 4 and node 5. This is because node 2 and node 3 are related with node 1, and new nodes are included by following links with higher probability. So the nodes which are connected with the current active set must have higher selection probability than those which have no links to the current active set. Now suppose node 2 is selected, then $S_2 = \{2\}$ and our current sample with the nodes selection order described in subscript is: $S_c = \{1_{(1)}, 2_{(2)}\}$.

Step-2: We still use all the nodes selected so far to be the elements in the active set at this step. That is, $A_2 = \{1_{(1)}, 2_{(2)}\}$ with the order labeled in the subscript. The probability of selecting next node j is:

if $j = 3$, then

$$P(S_2 = \{3\}|A_2 = \{1_{(1)}, 2_{(2)}\}) = 0.9 * \frac{1}{1} + (1 - 0.9)\frac{1}{3} = 0.933$$

if $j = 4$, then

$$P(S_2 = \{4\}|A_2 = \{1_{(1)}, 2_{(2)}\}) = 0.9 * \frac{0}{1} + (1 - 0.9)\frac{1}{3} = 0.0333$$

if $j = 5$, then

$$P(S_2 = \{5\}|A_2 = \{1_{(1)}, 2_{(2)}\}) = 0.9 * \frac{0}{1} + (1 - 0.9)\frac{1}{3} = 0.0333$$

Node 3 has higher selection probability than node 4 and node 5 since it is linked with node 1. Node 4 and node 5 have the same selection probability since they do not have links with any other nodes in the population. Now suppose node 3 is included at this step, then $S_3 = \{3\}$ our current sample is $S_c = \{1_{(1)}, 2_{(2)}, 3_{(3)}\}$.

The sampling procedure stops at step-2 since the current sample have already included 3 nodes. Our final sample is $S = \{1_{(1)}, 2_{(2)}, 3_{(3)}\}$ with its selection order. According to (3.2), the sample selection probability is:

$$\begin{aligned} P(S = \{1_{(1)}, 2_{(2)}, 3_{(3)}\}) &= P(S_0) * P(S_1 = \{2\}|A_1) * P(S_2 = \{3\}|A_2) \\ &= 0.2 * 0.475 * 0.933 \\ &= 0.0886 \end{aligned} \tag{3.7}$$

3.4 Resampling Approaches

3.4.1 Rao-Blackwell Estimation

In sampling from the graph model, the original sampled data S consists of the sequence of labels of the units, in the order selected, together with the corresponded y value and the sample of paired units with associated relationship values which we are interested. The minimal sufficient statistics only consists of the labels of distinct units, together with the y value and relationship among paired units in sample ((Thompson

& Seber (1996))). Let S_r be the minimal sufficient statistics based on sampled data S, then

$$S_r = \{(i, y_i), (j, k), A_{jk}, i \in S^1, (j, k) \in S^2\}$$

Where (i, y_i) is node label and associated interested value. (j, k) is the label of paired nodes. A_{jk} is the network information of paired nodes j and k. S^1 is the sampled data with sequence of labels of units, but *without* selection Order. $S^2 = S^1 \times S^1$ is the sampling space for paired nodes.

Based on the Rao-Blackwell theorem, preliminary estimators can be improved by finding the conditional expectation of this estimator given the minimal sufficient statistics S_r (Rao 1945, Blackwell 1949). let μ_0 be the preliminary estimator, then the improved exact Rao-Blackwell estimator is:

$$\hat{\mu}_{rb} = E(\hat{\mu}_0|S_r) = \sum_{\underline{S}:r(\underline{S})=S_r} \hat{\mu}_0(\underline{S}) P(\underline{S}|S_r) \tag{3.8}$$

Where, the conditional distribution is:

$$P(\underline{S}|S) = P(\underline{S}) / \sum_{\underline{S}:r(\underline{S})=S} P(\underline{S}) \tag{3.9}$$

Note that $\underline{S}$ is an element in the permutation sampling space with reduced information S_r. For example, if S_r includes units $\{1, 2, 3, 4\}$ and the relationship information among them. The permutation sampling space should have $4! = 16$ elements, which are

$$\{\{1, 2, 3, 4\}, \{1, 2, 4, 3\}, \{1, 4, 3, 2\}, \{1, 3, 2, 4\}, \cdots\}$$

Suppose n is the number of units in S_r, the expectation of preliminary estimator given S_r is the expected of initial value over all $n!$ reordering of the sample data. When the sample size increase the calculation of this improved estimator is highly prohibited. The variance of this estimator also involves all reordering. Enumerating all the permutations and combinations of all the sequences given the sample space is really hard for large sample size. So new resampling approaches are really needed to avoid this high computation. The idea is to construct a Markov Chain in order to get enough sample from the permutation sample space. The resampling space covers all the reordering(permutation) of n units in the sampled data. To be different from the sampling procedure, we use

$$X_0, X_1, X_2, \cdots,$$

to denote the states of Markov Chain in the resampling procedure. The Markov Chain starts at X_0, which is the original sampled data S in the same order of the sample as it is actually selected. Each state represents one reordering of all the n units. The limiting probability distribution is the distribution of selecting the ordered sample, given the set of distinct sample units. So the mean of the preliminary estimates based on the ordered samples is the Rao-Blackwell estimate, and the average of the preliminary estimates over the chain approaches that.

Three resampling methods RS1, RS2 and RS3 are described in this chapter. Each of these methods is based on Markov Chains of accepted permutations. RS1 is proposed by Steven K Thompson(Thompson (2006b)) and the other two are proposed in this chapter. For RS1, the sample is used as the first permutation, and resampling is processed by using the designs to give a candidate permutation, comparing the candidate with the current one using Metropolis Hastings(Hastings (1970)), and so on. In that way, the candidate tends to have high probability, so a good chance of

acceptance, but is selected independently of previous selections. Permutation also start from the sampled data S in RS2 and RS3, but new candidate is selected depend on previous selections.

3.4.2 Independent Resampling (RS1) Procedure

IRS was described by Thompson (Thompson (2006a)). The approach is to independently generate a tentative permutation X_k at each step K. In fact, X_k is generated by applying the same sampling procedure, but with sample size n, to the data as if sample included the whole population, $N = n$. That is, to generate Y_k, we used the same design procedure, but instead of using the true population, the original sampled data in S_0 is used as our whole population. In this procedure, the proposal distribution is a conditional distribution given the minimal statistics based on S_0, which is also used as the first state X_0 in the Markov chain resampling procedure. Let $P_t(.|S_r)$ be the conditional selection distribution, S_r is the minimal statistics. The selection probability for the new permutation should be calculated the same way as (3.2) except the sampled data under the actual design is used as our whole population. At step k, the accepted permutation is:

$$X_k = \begin{cases} X_{k-1} & wp\ 1-\alpha_k \\ Y_k \sim P_t(.|S_r) & wp\ \alpha_k \end{cases} \tag{3.10}$$

where

$$\alpha_k = min\{1, \frac{p(Y_k)p_t(x_{k-1}|S_r)}{p(x_{k-1})p_t(Y_k|S_r))}\}$$

Since $P_t(.|S_r) \propto P_t(.)$, the accept rate at step k can be calculated:

$$\alpha = min\{1, \frac{p(Y_k)p_t(x_{k-1})}{p(x_{k-1})p_t(Y_k))}\}$$

3.4.3 Dependent Resampling Procedure I (RS2)

There are many other ways to do the resampling. In this paper, Two other methods were used. The first one is instead of independent sampling, Y_k is generated by randomly switching the order of two elements in the previous accepted permutation. Say, if $X_{k-1} = \{3, 1, 4, 5, 2\}$ is the accepted permutation at step $(k-1)$. $\{1, 2\}$ are the two randomly selected elements, then $Y_k = \{3, 2, 4, 5, 1\}$. Figure 3.2 is an illustration of RS1 procedure with four steps.

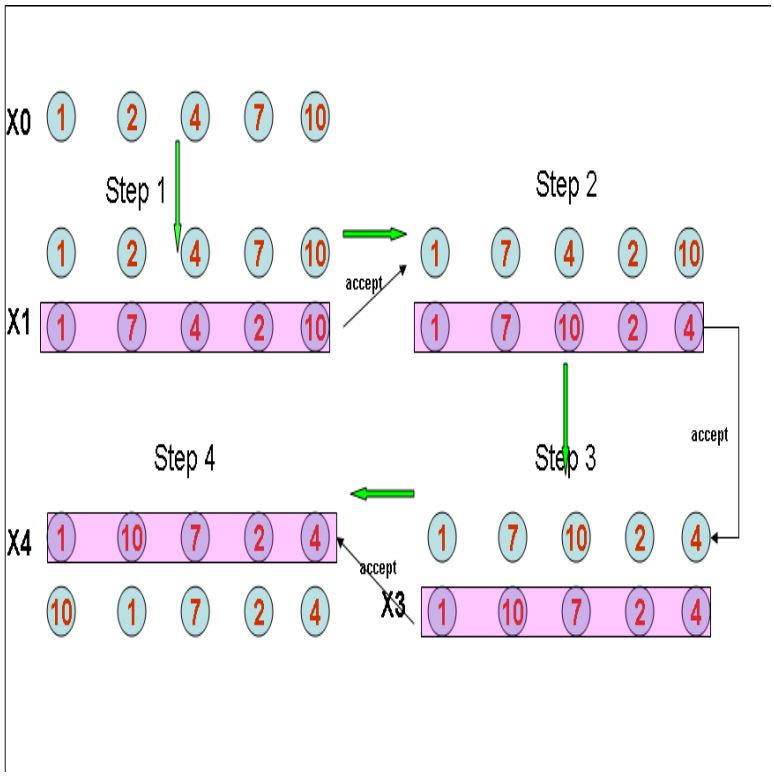


Figure 3.2. Four steps example of RS1 procedure. Light purple box indicates the accepted permutation at current step.

By doing this, the proposal distribution, given the present permutation, is uniform over all the permutations that can be obtained from that one by switching

two elements, and has zero probability for all other permutations.

$$P_{rs1}(Y=y)=\begin{cases} 1/\binom{n}{2} & \text{permutations based on switching the order of two units} \\ 0 & \text{other permutations} \end{cases} \tag{3.11}$$

Let $Y_1, Y_2, \cdots, Y_n$ be the sequence which is generated from P_{rs1}. Based on the accept-reject rate algorithm, the accepted sequence X_k is :

$$X_k=\begin{cases} X_{k-1} & wp\ 1-\alpha_k \\ Y_k \sim P_{rs1} & wp\ \alpha_k \end{cases} \tag{3.12}$$

Since new candidate permutation is generated from a uniform distribution, the probability of moving in that direction and the probability of moving back would be the same, which would cancel out one part of the M-H ratio. The other part of the M-H ratio should have the selection probability under the actual design for the present and the candidate. Thus,

$$\alpha_k = min\{1, \frac{p(Y_k)}{p(X_{k-1})}\} \tag{3.13}$$

3.4.4 Dependent Resampling Procedure II (RS3)

This resampling procedure is inspired by how the actual design is processed. During the actual design, at any step, the next unit selected is with high probability from the units which are connected with the current active set. Following this idea, if the next permutation is selected in the Markov Chain which constructed in the resampling procedure should be with high probability in the order that the connected units are close to each other, and with low probability by switching the order

of those unconnected units, then the acceptance rate may be higher than the permutations generated by switching the order of two randomly selected units. For example, $\{U_1, U_2, ...U_{10}\}$ is the original sample. Links exist only between U_1 and U_2, U_4 and U_5. In the resampling procedure, with high probability (say 0.9) the connected pair is selected from $\{(U_1, U_2), (U_4, U_5)\}$. With low probability (say 0.1) non-connected paired units are selected. The next permutation is generated by switching the order of this two units in the current state. If at the first step $\{(U_1, U_2)\}$ is chosen and the new sequence is accepted in M-H algorithm, we have $\{U_2, U_1, U_3, U_4, U_5, ..., U_{10}\}$. At the second step, (U_4, U_5) is chosen and switched order, the new sequence is $\{U_2, U_1, U_3, U_5, U_4, ..., U_{10}\}$. The proposal distribution is still uniformly distributed, denoted as P_{rs2} in (3.14). One part of M-H ratio still could be canceled out, the other part of it should has the calculation as in (3.13).

$$P_{rs2}(Y = y) = \begin{cases} 0.9/\binom{n}{2} & \text{permutations generated by switching order of two connected units} \\ 0.1/\binom{n}{2} & \text{permutations generated by switching order of two unconnected units} \\ 0 & \text{other permutations} \end{cases} \tag{3.14}$$

The accepted sequence X_k is :

$$X_k = \begin{cases} X_{k-1} & wp\ 1 - \alpha_k \\ Y_k \sim P_{rs2} & wp\ \alpha_k \end{cases} \tag{3.15}$$

3.4.5 Importance Sampling

For comparison, an alternative estimation approach using these same chain data, using an "importance sampling" type of estimator(G.Casella & C.P.Robert (1996)) instead of the average of over the accepted permutation is also addressed. The importance sampling type estimator uses the whole chain of permutations considered,

but accepted and unaccepted candidates. In other words, all the estimations during the resampling procedures are included and weighted. Let $P(X_k)$ be the probability of choosing sample X_k at step k under true population, and P_tX_k is probability of choosing sample X_k under the stationary distribution $Pt(.)$. The estimations based on importance sampling are calculated as:

$$E(\mu_i) = \sum_{i=1,2,\cdots n_r} \frac{P_t(X_k)}{P(X_k)} \hat{\mu}(X_k) / \sum_{i=1,2,\cdots n_r} \frac{P_t(X_k)}{P(X_k)} \tag{3.16}$$

So there are two approaches are used to approximating the Rao Blackwell estimator for the three resampling procedures. One approach averages the values of the preliminary estimator over the accepted permutations of the sample in the Markov chain. The other approach uses a weighted average of the values of the preliminary estimator over the whole chain including also the permutations of the sample that were not accepted by the Metropolis Hastings step, namely IS1, IS2 and IS3 for each resampling procedure. The relative weights are based on the ratios of the actual selection probability of the permutation under the design divided by the conditional probability of the permutation under the resampling design given the previous permutation. The importance sampling estimator also is divided by the sum of these relative weights.

3.5 Simulated Data

In this part, the AWS sampling designs and resampling approaches described above are implemented to sample from of population of size 200 (Figure 3.3). The illustration of three samples each of size 10 drawn from the population are also shown. The initial sample size is 4. Our goal is to employ different resampling approaches in order to estimate the population mean node value, namely population proportion estimation. We want to compare the efficiency and find the better MCMC procedure.

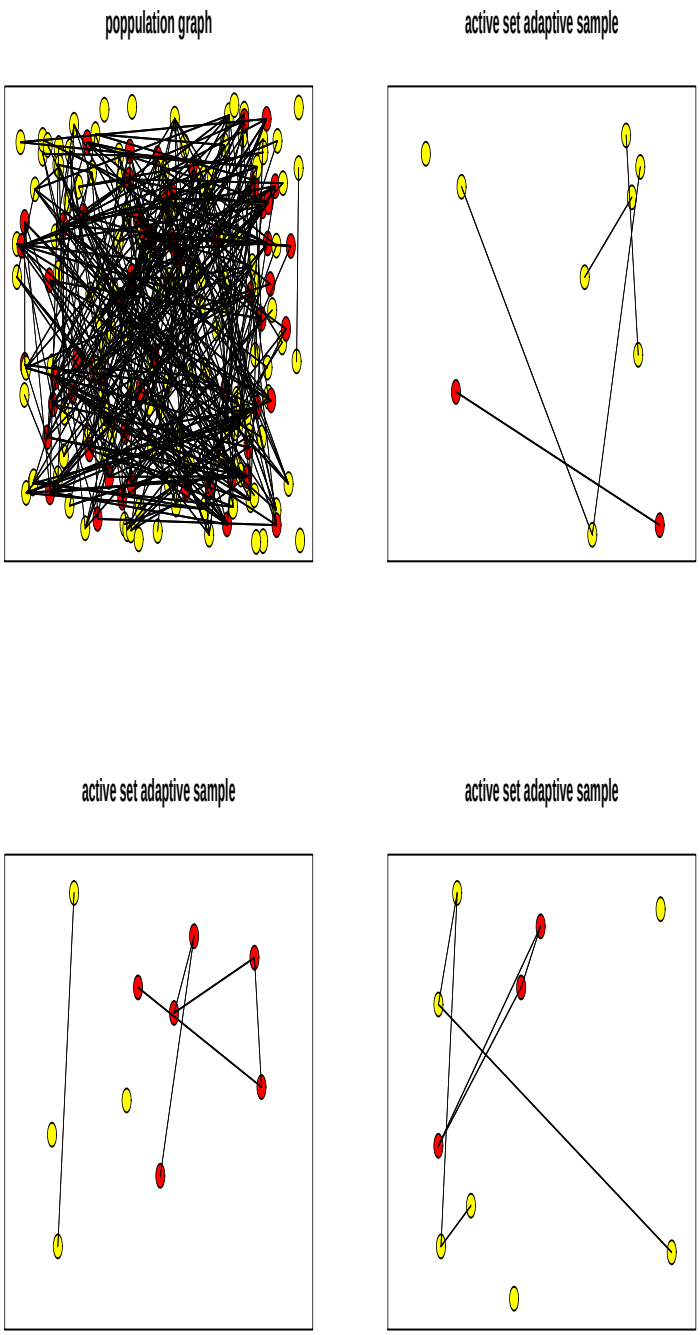
The efficiency of different resampling methods are compared based on $K = 2,000$ samples. The acceptance rate are also tracked. The true population proportion is 0.31.

3.5.1 Population mean node value estimation

The population mean are valued through preliminary estimators $\hat{\mu}_0$, the exact Rao-Blackwell estimators $\hat{\mu}_{rb}$, and three different MCMC resampling procedures which are used to approximate exact Rao-Blackwell estimators $\hat{\mu}_{rsi}, i = 1, 2, 3$. The main concern is the statistical efficiency in estimating the expectation of interest among these MCMC resampling procedures.

Most works used to investigate the MCMC efficiency are implemented by the assessment of burn in time, convergence and variance estimation (Gilks et al. (1996)). Methods are employed so far include between-within chains variation (Gelman & Rubin (1992)), Raftery & Lewis (1992a), Raftery & Lewis (1992b)) algorithm, Geweke algorithm(Geweke (1992)) etc. In physics literature, Goodman & Sokal (1989) defined the*integrated autocorrelation time* and *system relaxation time*. The concepts of them are also closely related to the convergence rate of the algorithm. Since at least some of our MCMC methods actually start in its stationary distribution, thus continues there, there is not a question of how long to reach the stationary resampling estimator. The question is how long it needs to run before the resampling estimator, which is a cumulative mean of preliminary estimators, has a small enough variance. So to compare the MCMC efficiency in our case, we consider the standard error of the cumulative means for chains at different length, and see which one has a smaller standard error among a set of fixed length chains. The standard error was based on the variance between chains, not the variance within chains.

Figure 3.3. Population 200 with Stochastic Block structure and 3 AWS samples of size 10.



The variation of the MCMC estimators is an addition of two parts (3.17). One is the average of conditional variance given the minimal sufficient statistics. The other is the variance of the conditional expectation given the minimal sufficient statistics S_r (3.17).

$$Var(\hat{\mu}_{rs}) = E(Var(\hat{\mu}_{rs}|S_r)) + Var(E(\hat{\mu}_{rs}|S_r)) \tag{3.17}$$

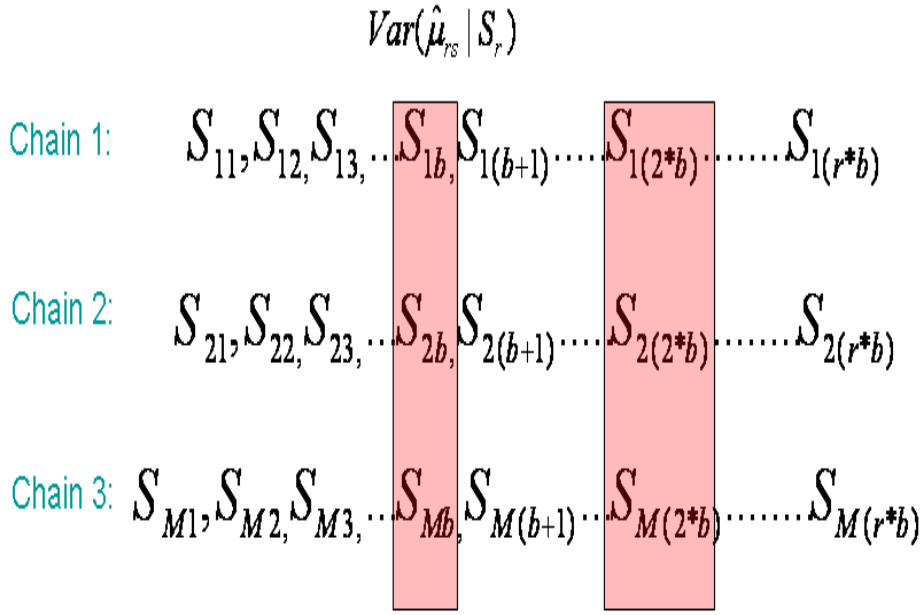
$$\begin{aligned} E(\hat{\mu}_{rs})|S_r) &= E(\frac{1}{b}\sum_{i=1}^{b}(\hat{\mu}_0(\underline{S})|S_r)) \\ &= \frac{1}{b}\sum_{i=1}^{b}(E(\hat{\mu}_0(\underline{S})|S_r)) \\ &= \hat{\mu}_{RB} \end{aligned} \tag{3.18}$$

Since for all preliminary estimator $\hat{\mu}_0$, (3.18) is satisfied, the variance of $E(\hat{\mu}_{rs})|S_r)$ is constant. It is enough to only consider $E(Var(\hat{\mu}_{rs}|d_S))$ to compare the variation of the MCMC estimators.

First, one set of M chains/samples each with length b are drawn from $P_t(.|S_r)$. Then sampled standard error are calculated at step b between the chains. Repeated the procedure, another set of M chains each with the same length are generated from the same conditional selection distribution, and standard errors are calculated the same way between chains. Such procedure is repeated until K sets of M chains are generated (Figure 3.4). $E(Var(\mu_{rs}|S_r))$ is approximated by the average of the K values of between chains standard error at each step. Let $\hat{\mu}_{ib}$ be the population proportion estimation at step b for the ith chain out of M chains, $\overline{\hat{\mu}_b}$ be the sample mean, and σ_k^2 be the variance of $\hat{\mu}_{ib}$, then

$$\hat{\sigma}_k^2 = var(\hat{\mu}_{ib}|S_r) = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\mu}_{ib} - \overline{\hat{\mu}_b})^2 \tag{3.19}$$

Figure 3.4. Expected variance given conditional selection probability calculation procedure.



The estimated value are presented in Table 3.1 which is based on M-H algorithm and Table (3.2) which is based on importance sampling algorithm. The standard error are calculated at iterations b equals to $b = 500, 1,000, 2,000, 4,000, 6,000, 8,000, 10,000$. Figure 3.5 and Figure 3.6 describe the changes of estimator variation with the increases of the number of interaction. RS1, RS2 and RS3 are three MCMC estimators. IS1, IS2 and IS3 are importance resampling estimators based on three resampling procedures. The decreases of $\hat{E}(Var(\hat{\mu}_{mc}|S))$ for independent resampling procedure decreases faster than both the depended resampling procedures based on importance sampling algorithm. For the MCMC algorithms, Standard error estimates based on the third resampling method (RS3) have larger values than estimates based on the

other two resampling methods. And the three resampling procedures to estimate population mean perform differently. For example, to estimate the mean based on initial sample, RS1 and RS2 have very close standard error which are smaller than the values based on RS3. For the estimators based on conditional selection probability(EST2), estimators based on RS2 has the smallest values than the others. The argument is that Independent resampling procedure performs best under importance sampling algorithm. And the procedure of randomly switching the order of two units is the least efficient methods since it has the highest variation. For accept-reject algorithm, the results are not consistent with the results under importance sampling algorithm.

Another way to evaluate how the resampling methods work is to compare the estimated expectation and variance of the estimators, and how the values are different from true population value which equals to 0.31 for the stochastic block structure population. Table 3.3 shows the results. All estimators are based on $K = 2,000$ samples. $\hat{\mu}_{rb}$ are the exact Rao-Blackwell estimators based on sample size 10. $\hat{\mu}_{RS}$ and $\hat{\mu}IS$ are based on average of $b = 10,000$ resampling values for MCMC and importance sampling methods. It is obvious that the $\hat{\mu}_{RB}$ is the best estimator since we select every one from permutation space with distribution $P(\underline{S}|S_r)$. It shows that the argument between the three resampling estimates thus requires more than 1,000 iterations. Figure 3.7 describes the distribution of empirical estimated values. The vertical line is true population mean node value. There is no distinguishable difference among estimates under three resampling methods since the density plot are almost identical. The Est2 may under estimate the population true value.

Figure 3.5. Average Cumulative Mean Standard deviation based on MCMC algorithm

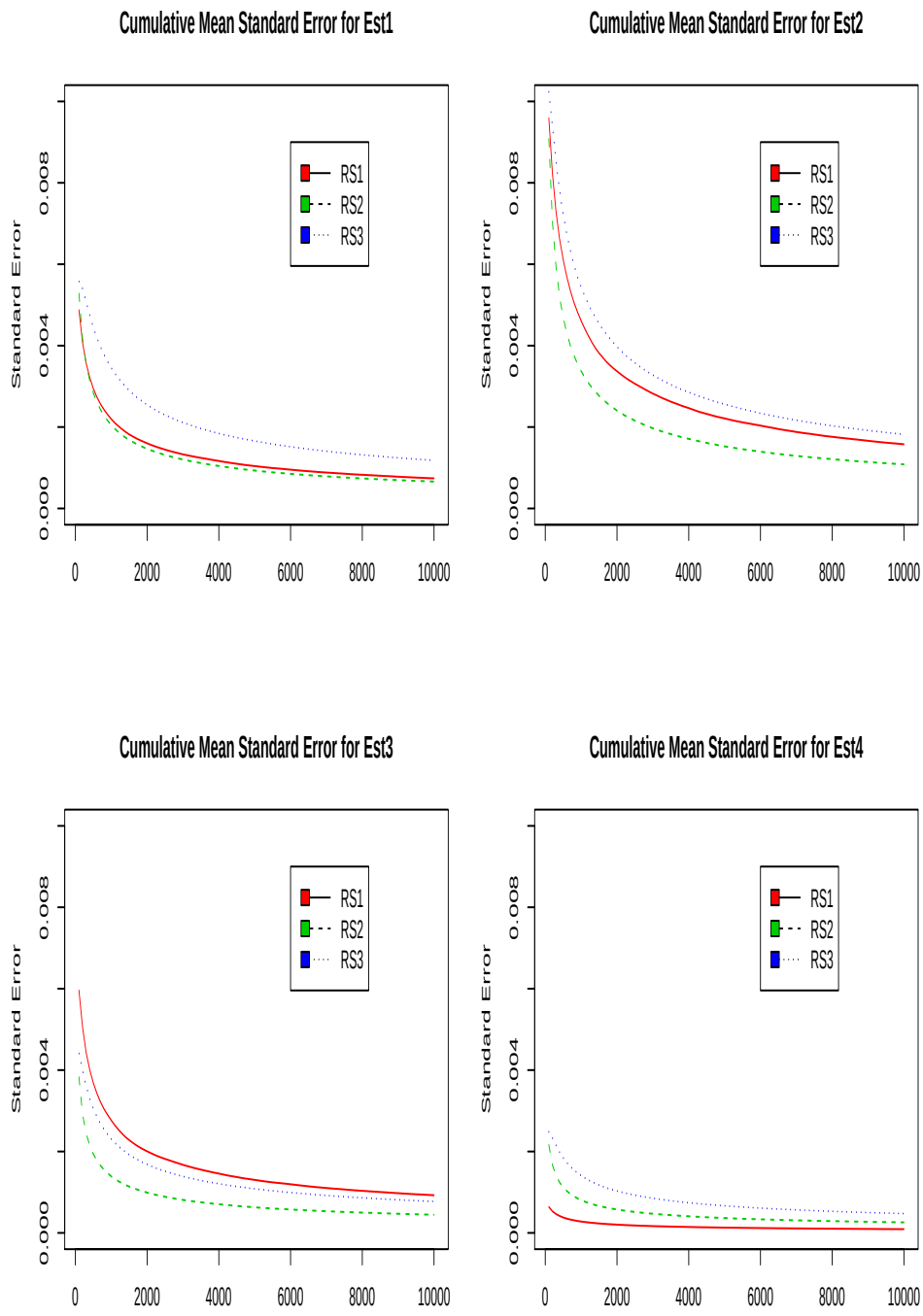


Table 3.1. Estimates of $E(var(\hat{\mu}_{rs}|S))$ based on M-H algorithm of RS1, RS2 and RS3

	b iterations						
Estimators	$b = 500$	$b = 1000$	$b = 2000$	$b = 4000$	$b = 6000$	$b = 8000$	$b = 10,000$
EST1							
$\hat{\mu}_{RS1}$	0.0153	0.0131	0.0111	0.0093	0.0084	0.0078	0.0074
$\hat{\mu}_{RS2}$	0.0140	0.0118	0.0099	0.0083	0.0075	0.0070	0.0066
$\hat{\mu}_{RS3}$	0.0243	0.0208	0.0176	0.0148	0.0134	0.0124	0.0118
EST2							
$\hat{\mu}_{RS1}$	0.0323	0.0278	0.0235	0.0199	0.0179	0.0166	0.0158
$\hat{\mu}_{RS2}$	0.0229	0.0194	0.0163	0.0136	0.0123	0.0114	0.0108
$\hat{\mu}_{RS3}$	0.0379	0.0323	0.0272	0.0228	0.0207	0.0192	0.0182
EST3							
$\hat{\mu}_{RS3}$	0.0192	0.0165	0.0139	0.0116	0.0105	0.0098	0.0093
$\hat{\mu}_{RS2}$	0.0094	0.0080	0.0067	0.0056	0.0051	0.0047	0.0045
$\hat{\mu}_{RS3}$	0.0161	0.0137	0.0115	0.0097	0.0088	0.0081	0.0077
EST4							
$\hat{\mu}_{RS1}$	0.0019	0.0016	0.0014	0.0012	0.0010	0.0010	0.0000
$\hat{\mu}_{RS2}$	0.0055	0.0046	0.0039	0.0033	0.0030	0.0027	0.0026
$\hat{\mu}_{RS3}$	0.0098	0.0084	0.0071	0.0060	0.0054	0.0050	0.0047

Figure 3.6. Average Cumulative Mean Standard deviation based on importance sampling method

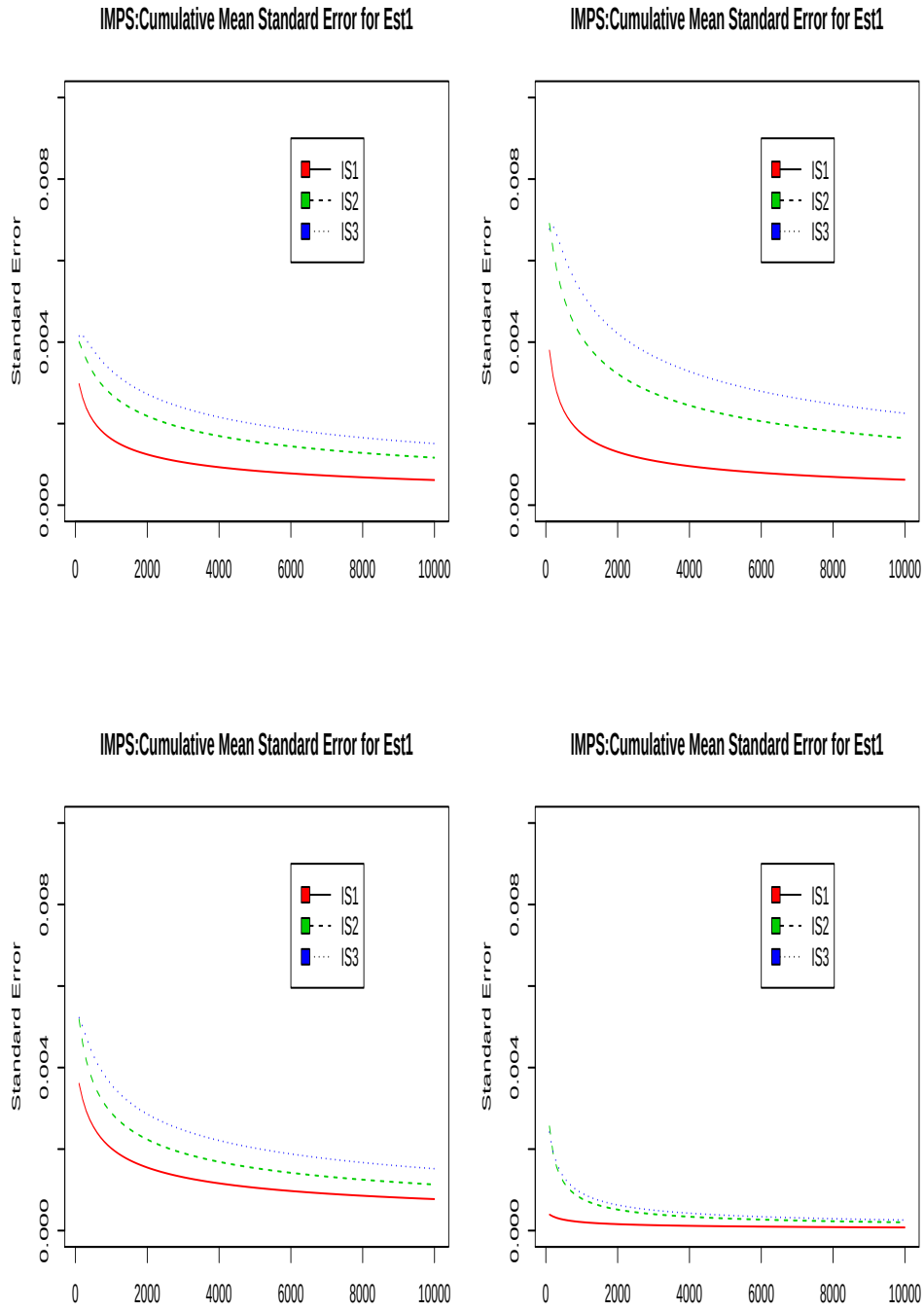


Table 3.2. Estimates of $E(var(\hat{\mu}_{rs}|S))$ based on importance sampling algorithm for IS1, IS2 and IS3 methods

	b iterations						
estimator	$b = 500$	$b = 1000$	$b = 2000$	$b = 4000$	$b = 6000$	$b = 8000$	$b = 10,000$
EST1							
$\hat{\mu}_{IS1}$	0.0120	0.0104	0.0089	0.0076	0.0069	0.0065	0.0062
$\hat{\mu}_{IS2}$	0.0211	0.0187	0.0163	0.0141	0.0130	0.0122	0.0116
$\hat{\mu}_{IS3}$	0.0264	0.0236	0.0208	0.0182	0.0168	0.0158	0.0151
EST2							
$\hat{\mu}_{IS1}$	0.0125	0.0107	0.0092	0.0078	0.0071	0.0066	0.0062
$\hat{\mu}_{IS2}$	0.0311	0.0271	0.0235	0.0202	0.0185	0.0172	0.0164
$\hat{\mu}_{IS3}$	0.0408	0.0361	0.0316	0.0274	0.0252	0.0236	0.0225
EST3							
$\hat{\mu}_{IS1}$	0.0149	0.0129	0.0111	0.0095	0.0087	0.0081	0.0077
$\hat{\mu}_{IS2}$	0.0215	0.0188	0.0162	0.0139	0.0127	0.0119	0.0113
$\hat{\mu}_{IS3}$	0.0276	0.0244	0.0213	0.0184	0.0170	0.0159	0.0152
EST4							
$\hat{\mu}_{IS1}$	0.0015	0.0013	0.0011	0.0010	9e-04	8e-04	8e-04
$\hat{\mu}_{IS2}$	0.0049	0.0040	0.0032	0.0026	0.0023	0.0022	0.002
$\hat{\mu}_{IS3}$	0.0060	0.0049	0.0040	0.0033	0.0030	0.0027	0.0026

Table 3.3. Estimated expectation and variance and mean squares error. Based on 2,000 samples and each with 1,000 re-samples. True population value os 0.31. μ_o is preliminary estimator, μ_{RB} is exact Rao-Blackwell estimator, $\mu_{RSi}, i = 1, 2, 3$ are resampling estimators. The sample size is 10 with initial sample size 4.

estimators		$E(\hat{\mu})$	$Var(\hat{\mu})$	$E(\hat{\mu} - 0.31)^2$
EST1	$\hat{\mu}_0$	0.3479	0.0215	0.0229
	$\hat{\mu}_{RB}$	0.3079	0.0224	0.0224
	$\hat{\mu}_{RS1}$	0.3375	0.0304	0.0311
	$\hat{\mu}_{RS2}$	0.3206	0.0310	0.0311
	$\hat{\mu}_{RS3}$	0.3278	0.0298	0.0301
EST2	$\hat{\mu}_O$	0.2934	0.0325	0.0327
	$\hat{\mu}_{RB}$	0.3124	0.0308	0.0308
	$\hat{\mu}_{RS1}$	0.3176	0.0347	0.0347
	$\hat{\mu}_{RS2}$	0.3130	0.0320	0.0320
	$\hat{\mu}_{RS3}$	0.3093	0.0312	0.0312
EST3	$\hat{\mu}_O$	0.3122	0.0326	0.0326
	$\hat{\mu}_{RB}$	0.3058	0.0314	0.0314
	$\hat{\mu}_{RS1}$	0.3103	0.0301	0.0301
	$\hat{\mu}_{RS2}$	0.3214	0.0293	0.0294
	$\hat{\mu}_{RS3}$	0.3097	0.0305	0.0305
EST4	$\hat{\mu}_O$	0.3016	0.0335	0.0335
	$\hat{\mu}_{RB}$	0.3158	0.0307	0.0307
	$\hat{\mu}_{RS1}$	0.3178	0.0368	0.0368
	$\hat{\mu}_{RS2}$	0.3269	0.0382	0.0384
	$\hat{\mu}_{RS3}$	0.3104	0.0403	0.0403

Figure 3.7. Samples based on AWS designs from population of size 20

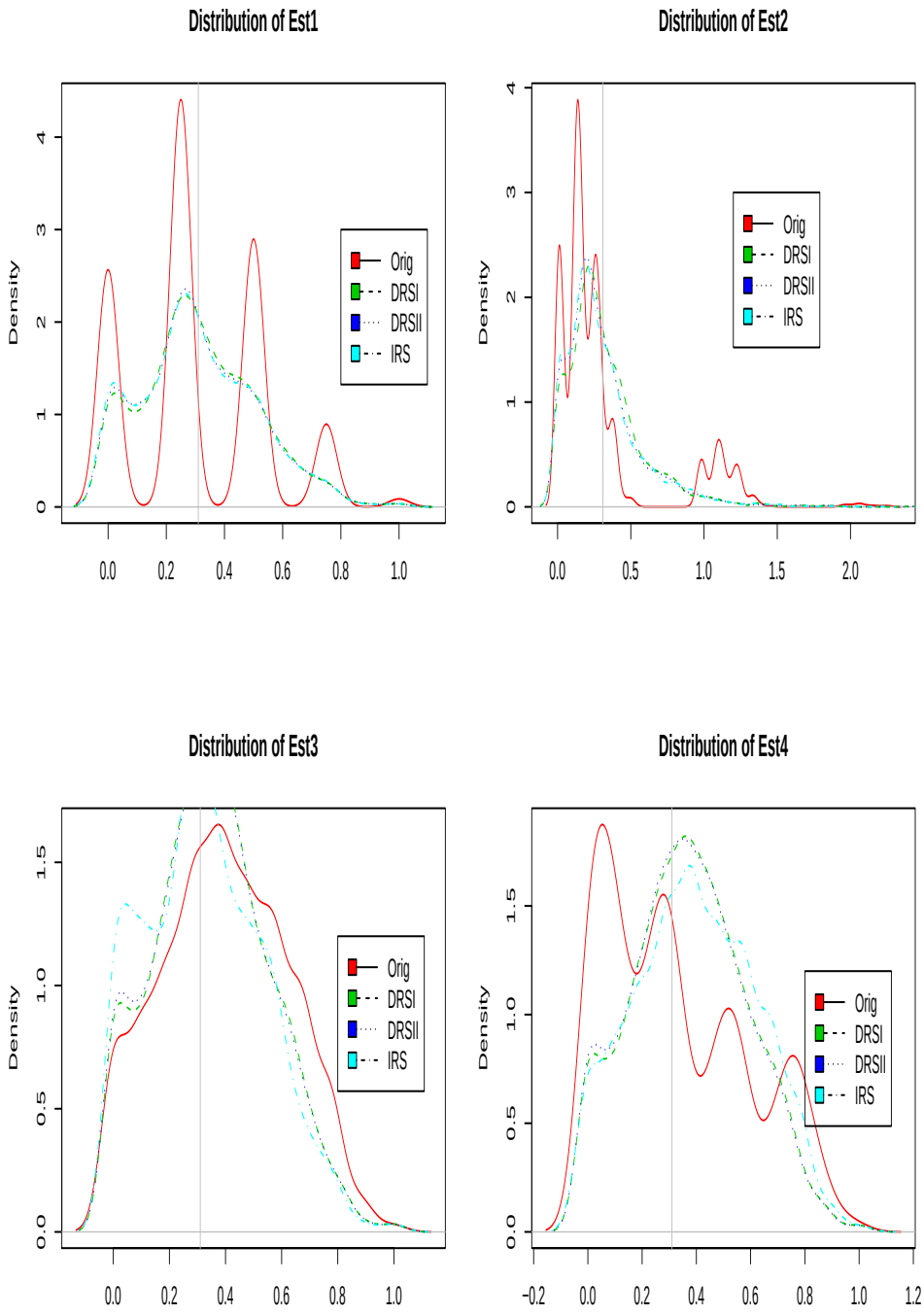


Table 3.4. Estimated expectation and variance and mean squares error. Based on 2,000 samples and each with 1,000 re-samples, importance resampling method. The sample size is 10 and initial sample size is 4.

estimators		$E(\hat{\mu})$	$Var(\hat{\mu})$	$E(\hat{\mu}-0.31)^2$
EST1	$\hat{\mu}_{IS1}$	0.3089	0.0097	0.0097
	$\hat{\mu}_{IS2}$	0.2967	0.0373	0.0375
	$\hat{\mu}_{IS3}$	0.3049	0.0536	0.0536
EST2	$\hat{\mu}_{IS1}$	0.3901	0.0179	0.0243
	$\hat{\mu}_{IS2}$	0.3078	0.0782	0.0782
	$\hat{\mu}_{IS3}$	0.3190	0.1469	0.1469
EST3	$\hat{\mu}_{IS1}$	0.4429	0.0129	0.0306
	$\hat{\mu}_{IS2}$	0.3085	0.0408	0.0408
	$\hat{\mu}_{IS3}$	0.3940	0.0506	0.0576
EST4	$\hat{\mu}_{IS1}$	0.8783	0.0160	0.3390
	$\hat{\mu}_{IS2}$	0.3205	0.0673	0.0674
	$\hat{\mu}_{IS3}$	0.3305	0.0686	0.0689

3.6 Discussion

We have seen that the output of the simulation schemes such as preliminary estimators can be improved by Rao-Blackwell estimators to reduce the variation. Although the computational implementation may seem involved, the exact Rao-Blackwell estimators can be easily programmed via recursive iterations when the sample size is no larger than 12. But the computation duty increases while the increases of the sample size, resampling procedures can be implemented instead of

Table 3.5. The acceptance rate for RS1, RS2 and RS3

rate(%)	RS1	RS2	RS3
	9.7	42.29	45.18

exact Rao-Blackwell calculation, either through M-H MCMC or importance sampling algorithms. The fact that our MCMC estimators starts from its stationary distribution, so continues from that, it is no problem for the estimators converge to the stationary distribution. The question is that how fast the variation change among a set of chains with fixed length. The result is the independent resampling procedure is the most efficient one under importance sampling algorithm. But for MCMC algorithm, the three resampling procedures perform differently. The acceptance rate for RS1 is lower than RS2 and RS3, since new permutations are generated independently.

Note that the above simulations are based on 200 population and the true population proportion is around 0.31. This proportion is relatively large compared to the definition of rare population, which is less than 0.05. In order to investigate the effect of the large population proportion, another simulation based on 0.025 true population proportion is implemented (3.6). The population size is also 200. The exact Rao-Blackwell estimators still have good computational accuracy. They performs best in terms of lowest mean square error. The estimators based on RS1 are more efficient than estimators based on RS2 and RS3 since they have lower bias and mean square error. There is not significant difference between the estimators based on RS2 and RS3.

Note that the results in this paper are based on simulations and no new methods have been developed to test which resampling procedure is better than RS1 which is proposed by Thompson(2006). Some improved methods for testing MCMC accuracy and variance comparison may exist, which should be pursuit in the future.

Table 3.6. Estimated expectation and variance and mean squares error. Based on 2,000 samples and each with 1,0000 re-samples. True population value os 0.025. $\mu_0 i, i = 1, 2, 3, 4$ are preliminary estimators. μ_{RB} is exact Rao-Blackwell estimators. $\mu_{RSi}, i = 1, 2, 3, 4$ are estimators based on three different resampling procedures. The total sample size is 10 and initial sample size is 4.

estimators	$E(\hat{\mu})$	$Var(\hat{\mu})$	$E(\hat{\mu} - 0.025)^2$
EST1			
$\hat{\mu}_{01}$	0.0260	0.0062	0.0062
$\hat{\mu}_{RB}$	0.0251	0.0030	0.0030
$\hat{\mu}_{RS1}$	0.0252	0.0034	0.0034
$\hat{\mu}_{RS2}$	0.0253	0.0034	0.0034
$\hat{\mu}_{RS3}$	0.0253	0.0034	0.0034
EST2			
$\hat{\mu}_{02}$	0.0260	0.0084	0.0084
$\hat{\mu}_{RB}$	0.0251	0.0052	0.0052
$\hat{\mu}_{RS1}$	0.0253	0.0054	0.0054
$\hat{\mu}_{RS2}$	0.0254	0.0055	0.0055
$\hat{\mu}_{RS3}$	0.0254	0.0055	0.0055
EST3			
$\hat{\mu}_{03}$	0.0240	0.0063	0.0063
$\hat{\mu}_{RB}$	0.0249	0.0029	0.0029
$\hat{\mu}_{RS1}$	0.0249	0.0030	0.0030
$\hat{\mu}_{RS2}$	0.0247	0.0035	0.0035
$\hat{\mu}_{RS3}$	0.0247	0.0031	0.0031
EST4			
$\hat{\mu}_{04}$	0.0241	0.0074	0.0074
$\hat{\mu}_{RB}$	0.0247	0.0023	0.0023
$\hat{\mu}_{RS1}$	0.0247	0.0024	0.0024
$\hat{\mu}_{RS2}$	0.0246	0.0027	0.0027
$\hat{\mu}_{RS3}$	0.0247	0.0027	0.0027

Chapter 4

Cost Optimization in Adaptive Web Sampling

4.1 Abstract

The main advantage of Adaptive Web Sampling (AWS) designs is that the sample size could be predefined and fixed prior to implementing the sampling strategy (Thompson (2006a)). However the randomness of the sample units still exists, and the probability of selection provides the basis for unbiased estimation. However the variation of the selected sample units has not been fully investigated to date. In this paper, we incorporate some practical restrictions, such as budget or time that determine which unit or set of units should be selected at each step and when the sampling process should stop. The objective is to minimize the cost and maximize the sampling procedure information by efficient design of the sampling options.

4.2 Introduction

Thompson & Seber (1996) introduced adaptive sampling designs mainly implemented for spatial environments. For spatially clustered populations such as animal or plant species and mineral or oil resources, adaptive cluster sampling and stratified adaptive cluster sampling are also introduced (Thompson (1990a),Thompson (1990b)) to improve the efficiency. Such designs can also be used to sample hidden human populations, the internet, and other populations with network structures. A weak point of such designs is the lack of control in how the initial sample is obtained

and sample coverage. Thompson (2006a) introduced a new and more flexible sampling design Adaptive Web Sampling(AWS) to overcome this shortcoming and the sample selection probability becomes step by step less dependent on the initial sample selection. These sampling designs have more flexibility and allow the possibility for designs under different active sets, different initial sample designs or different calculations for the selection probability at each step. However none of them are investigated under the constraints of budget, time or risk. In this chapter, we consider a type of AWS design which is implemented based on different design factors such as selection probabilities, initial sample size, and the number of units selected at each step. We are seeking to find an optimal sampling procedure based on the performance of population estimations, which depend on the different parameter values under the same cost model. The relative efficiency and mean squared error are measured by the comparison the variation of the estimation for population proportion.

We illustrate the application of one AWS methodology to the problem of estimation of the population proportion in two cases. One sample is a simulated wire transaction data set in Banking setting. The purpose is to identity the high risk fraud sub population of participants in the money transaction. The population of subjects for such a study are plentiful, but determination of the high risk subjects simply by observing money transaction everyday is very time consuming. Thus, it is important to obtain a set of representative subjects in the population in limited time. We will demonstrate how a AWS design with cost constraint in terms of time can assist achieving this goal. Another illustration is based on the Colorado Springs data, which is a study of the heterosexual transmission of HIV/AIDS in a "high-risk" population in Colorado Springs (Potterat et al. (1993), Rothenberg et al. (1995)). This empirical population was also used by Chow & Thompson (1998) as an example in a snowball sampling design in which all links were traced, with the exception of the last wave.

4.2.1 AWS designs

The AWS desing proceeds as follows (Thompson (2006a)): At any point during the sampling, the selection of the next unit or wave of units to include in the sample is with high probability d selected with a distribution that depend on the values of variables of interest in an "active set" of units already selected. With low probability $1 - d$, the next unit is selected from a distribution not depending on values of variables of interest. The active set could be the units selected so far, the most recently selected units or subset of the units selected.

Figure 4.1 is an illustration of how a sample is obtained through the application of AWS design. The population link matrix W_{ij} is a matrix of 1 and 0s. Where a 1 indicates a link between the units labeled in the rows and columns.

	1	2	3	4	5	6	7
1	0	0	1	1	1	0	0
2	0	0	0	1	1	0	0
3	1	0	0	1	0	1	0
4	1	1	1	0	1	0	1
5	1	1	0	1	0	1	0
6	0	0	1	0	1	0	0
7	0	0	0	1	0	0	0

The Nodes 1 and 2 are initially selected using any sampling designs from the entire population of nodes. Using the information obtained from interviewing node 1 and node 2, all the links connecting nodes 1 and 2 are observed. A red circle represents the nodes with characteristics we are interested in such as HIV positive, suspicious bank account etc.; blue circles represent HIV negative people. Node 6 represents one of the

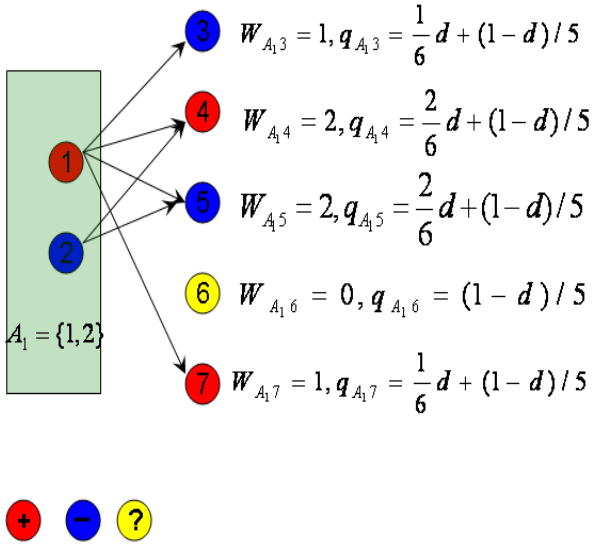


Figure 4.1. An illustration of adaptive web sampling. Nodes 1 and 2 are initially chosen at random. Weighted links are used to calculate transition probabilities from current active set to next selected units.

nodes that are not connected with nodes 1 and 2. Such nodes could not be observed so that their characteristics are unknown. We put a weighted link from active set A_1 to node 3 equals to 1, $W_{A_1 3} = 1$, since there is only one link out from A_1 to node 3; A weighted link from A_1 to node 4 equals to 2 since two links are out from A_1 to node 4. $W_{A_1 6} = 0$ since node 6 is isolated from A_1. New nodes are selected based on the observed weighted links. For the snowball type adaptive sampling, new units can only be selected from the nodes 3, 4, 5, and 7 since they are connected with the initial sample, thus we could not control the depth and width of such sampling procedures. The AWS design is more flexible in that new nodes can be selected not only from the connected ones but also can be selected from those isolated nodes. Suppose the nodes from the connected units have distribution f_c, and the nodes which are isolated from the current active set is distributed as f_u. At any point during the sampling procedure, new nodes are selected from a mixture distribution

Table 4.1. Transition probabilities from nodes 1 and 2 in Figure 4.1

$q_{A_1 3} = \frac{1}{6} * d + (1-d)/5$
$q_{A_1 4} = \frac{2}{6} * d + (1-d)/5$
$q_{A_1 5} = \frac{2}{6} * d + (1-d)/5$
$q_{A_1 6} = \frac{2}{6} * d + (1-d)/5$
$q_{A_1 7} = \frac{1}{6} * d + (1-d)/5$

$$f_m = df_c + (1-d)f_u \tag{4.1}$$

where, d, f_c, f_u could be chosen according to different sampling purpose. In this paper, we carried out a AWS design where d depends on a cost ratio. The cost ratio is defined as the cost of including connected units over the cost of including isolated units, denoted as r_c. f_c and f_u only depend on the observed links out from current active set. The transition probabilities for such designs are given by

$$q(j|A_k) = \begin{cases} d \times \frac{W_{A_k j}}{W_{A_k +}} + (1-d) \times \frac{1}{N - n_{A_k}} & \textit{link exists from } A_k \textit{ to } j \\ \frac{1}{N - n_{A_k}} & \textit{no link exists from } A_k \textit{ to } j \end{cases} \tag{4.2}$$

Where, $W_{A_k j}$ is the number of links from the current active set A_k to unit j, $W_{A_k +}$ is the total number of links from outside A_k, n_{A_k} is the number of units selected so far. In Figure 4.1, the transition probabilities from initial sample nodes 1 and nodes 2 to other nodes are:

4.2.2 Estimation

Based on Conditional selection Probability: Thompson (2006a) derived a population proportion estimator based on conditional selection probability. Such estimator can be thought as a composite estimator of the initial sample and the additional samples selected using the nodes value to determine the selection probability step by step. The first part is the unbiased estimator of the population total $\sum y_i$ based on the initial sample. And the second part is the estimators of $\sum y_i$ based on conditional selection probability at any step of the sampling procedure after initial sample. In his paper, the estimator is based on sampling with only *ONE* new unit included at each step. In this paper, EST2 is enlightened by the idea proposed by Thompson, but based on the sampling procedure which is taken by waves.

The population total $\sum y_i$ estimator based on the initial sample is:

$$\hat{\tau}_{S0} = (N/n) \sum_{i \in S_0} y_i/\pi_i$$

when units are selected with probability π. For a simple random initial sample without replacement which is implemented in this paper,

$$\hat{\tau}_{S0} = (N/n) \sum_{S0} y_i = N\overline{y}_0$$

At step K after the initial sample with active set A_k, Z_k is used to estimate the population total

$$Z_k = \sum_{j \in S_{ck}} y_j + \sum_{l \in S_k} y_l/q_{A_k l}$$

Here $q_{A_k l}$ is the selection probability of unit l in wave k with y_k value, $S_{ck} =$

$\{S_0, S_1, S_2, \cdots, \S_K\}$ is the units selected so far. An unbiased estimator for the population mean:

$$\hat{\mu}_1 = \frac{1}{Nn}\{n_0\hat{\tau}_{S0} + \sum Z_i\} \tag{4.3}$$

4.3 Cost Model

So far, there is a very sparse literature on costs for sampling hidden population due to less attention for cost efficiency modeling, less quantified decision making versus professional expertise or other reasons. For sampling rare populations, the costs due to uncertain difficulties of contacting sampling units, gaining cooperation of the sample units and the interviewer variation in efficiency are big issues. In some sampling situations, the cost of sampling, measured in terms of time, money or risk etc, may differ from wave to wave, and may also differ among units within each wave.

Dell & Clutter (1972) model cost in Ranked Set Sampling (RSS) and the sampling efficiency was evaluated through the ratio of population mean estimator between RSS and simple randomly sampling (SRS). In their cost model, c_s and c_q are the cost of stratification which involved sampling and ranking and the cost of qualification for one unit, respectively. RSS and SRS are compared in estimating the population mean through

$$RE(\mu_{RSS}, \mu_{SRS}) = \frac{c_q}{c_q + c_s} RP_{RSS} \tag{4.4}$$

where $RP_{RSS} = \frac{var(\mu_{SRS})}{var(\mu_{RSS})}$. Kaur et al. (1996) devised a more detailed cost model (KPST) that incorporates more cost categories such as set up costs, cost of ranking the units in a set, etc. W.Nahhas et al. (2002) extended their model by con-

sidering the cost of ranking which is assumed to be negligible in the KPST model. The motivation behind their work is that, in certain situations, sampling and ranking of units from the population are inexpensive while actual measurement of a unit is costly.

In this paper, our effort is motivated by the fact that sampling the rare population is expensive in terms of time and money under the convenience designs including SRS. At each step, new units are selected depending on the weighted links. For example in Figure 4.1, the weight matrix is the same as link matrix since we count the total number of connections out from A_k. In the wire transaction case, it could be number of total transactions, the total amount, the number of keyword hits, or other social connections between two accounts. For simplicity, we only consider the number of transaction records between business partners. Though we do not rank the units, the links with larger weights should have higher probability of being selected than the links with smaller weights. A cost model based on such sampling procedures could be derived as follows: Let C_1 be the cost for obtaining each unit which is connected with the current active set, and C_2 be the cost for jumping to an isolated one, respectively. Suppose n is the total AWS sample size, n_1 is the number of units included by following links, n_2 is the number of units selected randomly and $n_1 + n_2 = n$ C_0 is an overhead costs for AWS, then the total cost is:

$$C_{AWS} = C_0 + n_1 * C_1 + n_2 * C_2 \tag{4.5}$$

The following example will be used to illustrate the determination of parameter values in the AWS based on a pre-specified total cost value. The relative efficiency is compared to investigate the optimal parameter combination values. The parameters are listed in Table 4.2.

Table 4.2. Pre-specified parameter values for population proportion estimation for the example shown in Figure 4.2

$n0$	Initial sample size	{5,10,20}
ns	Number of units selected at each step	{2,4,8}
r_c	Cost ratio,$r_c = C_2/C_1$	{1,2,4,8}
d	Link tracing probability	20 values evenly across $(0, 1)$

4.4 Simulation

4.4.1 Simulation Setting

The population shown in Figure 4.2 is a simulated data set based on wire transaction in a banking center. Red circles represent suspicious accounts, and yellow circles represent normal accounts. Variable Y is used to describe the accounts status:

$$Y_i = \begin{cases} 1 & i \text{ is suspicious account} \\ 0 & i \text{ is normal account} \end{cases} \tag{4.6}$$

The black lines represent a wire transaction between two accounts. The accounts could be personal or company accounts. There are 6 suspicious accounts out of the set of 100 accounts. For simplicity, we assume that wire transactions always exits between paired accounts. In other words, there is a transaction from B to A once a wire transaction from A to B. Thus, the links matrix which describe the transaction among 100 accounts is symmetric, which is shown in Table 4.3.

There are millions of wire transaction records every week. How to capture the rare suspicious accounts from this huge data base efficiently is always an Anti Fraud issue for banks because of the time consuming nature of checking each account record. Here we assume C_1 units of time spent on checking an account which has a wire transaction with already checked accounts; C_2 units of time spent on checking an

account which is randomly picked from the unchecked account records. The AWS designs was implemented as follows. An initial sample of size n_0 was simply randomly sampled from the whole population. A fixed overhead cost C_0 for it is assumed. After that, n_s accounts were sampled with weight from those accounts which have wire transaction with any accounts in the initial sample. We call this is the first wave of the sampling procedure. And accounts obtained after the first wave are sample S_1. We add the cost for sampling and checking every account in the first wave. If the current cumulative cost C_w is less than a pre-specified total cost C_t, we use the accounts in S_1 to be the active set for next sampling wave untill $C_w >= C_t$. This means that when we do not have more time for investigating accounts, the sampling procedure should end. Suppose S is the set all the accounts we obtained by the time C_t is used up, we obtain the information of account status, transaction behavior among them, accounts selection order and all account selection probabilities. Based on this information, we can estimate the proportion of suspicious accounts in the whole population. To reduce the sampling bias, we repeated such procedure 2,000 times, and the sampling proportion and variance are averaged over the 2,000 samples. Combinations of the parameters specified in Table 4.2 are used in the simulation.

4.4.2 Simulation Result

Relative efficiency(RE) and mean square error(MSE) are used to compare how different link tracing probabilities effect the results in (4.7). Let RE_{n_0,n_0^*} be the RE between estimator based on initial sample n_0 and n_0^*, and which also depend on d and n_s,

$$RE_{n_0,n_0^*} = \frac{var(\hat{\mu}_{n_0})}{var(\hat{\mu}_{n_0^*})} \tag{4.7}$$

Figure 4.3 describes the RE for initial sample size of 10 and 5. Figure 4.4

describes the RE for initial sample size of 10 and 20. Both Figures show that RE decreases when link tracing probabilities increase. Thus we may have larger estimated variation when initial sample size increase. The number of units do not have a strong effect to the RE performance, as we can see since the three lines have similar behavior. This result can also be shown by the RE values. $RE_{5,10}$ is between 1.25 and 2.5 for different selection probability as seen in Figure4.4; $RE_{10,20}$ is larger than 1.5, as seen in Figure4.3. These results are reasonable since initial sample is randomly by disregarding their relationship. This leads to more dissimilarity and randomness which may produce larger variation. Higher cost ratio seems to lead to higher RE, but the performance is not significant. Figure 4.5, Figure 4.6 and Figure 4.7 depict the MSE at different values of the parameters. The value of MSE increases with the increase in selection probability. And the sharpest increases start at selection probability around 0.8. This result is most obvious for initial sample size 5. The same conclusion could be obtained from the RE figures. Note that MSE based on a cost ratio of 2 is right between MSE based on cost ratio of 1 and 4. Table 4.4.2 shows the the value of minimum mean square error and corresponding selection probabilities.

Figure 4.9 shows the estimated total sample size distribution. Figure 4.10 shows estimated sampling waves/depth distribution. Total estimated sample size is between 15 and 55 for initial sample size 5; Total estimated sample size is between 20 and 55 for initial sample size 10, and total estimated sample size is between 30 and 65 for initial sample size 20. The larger the initial sample size, the more a large sample size is obtained. The initial sample size seems to have no strong effect to how many units are sampled from population. But there are more sampling waves with larger initial sample size. This shows that, for larger initial sample sizes, more links are followed during the whole sampling procedure. But for different initial sample size, the sampling procedure stops around waves 8 and 9 for most of cases.

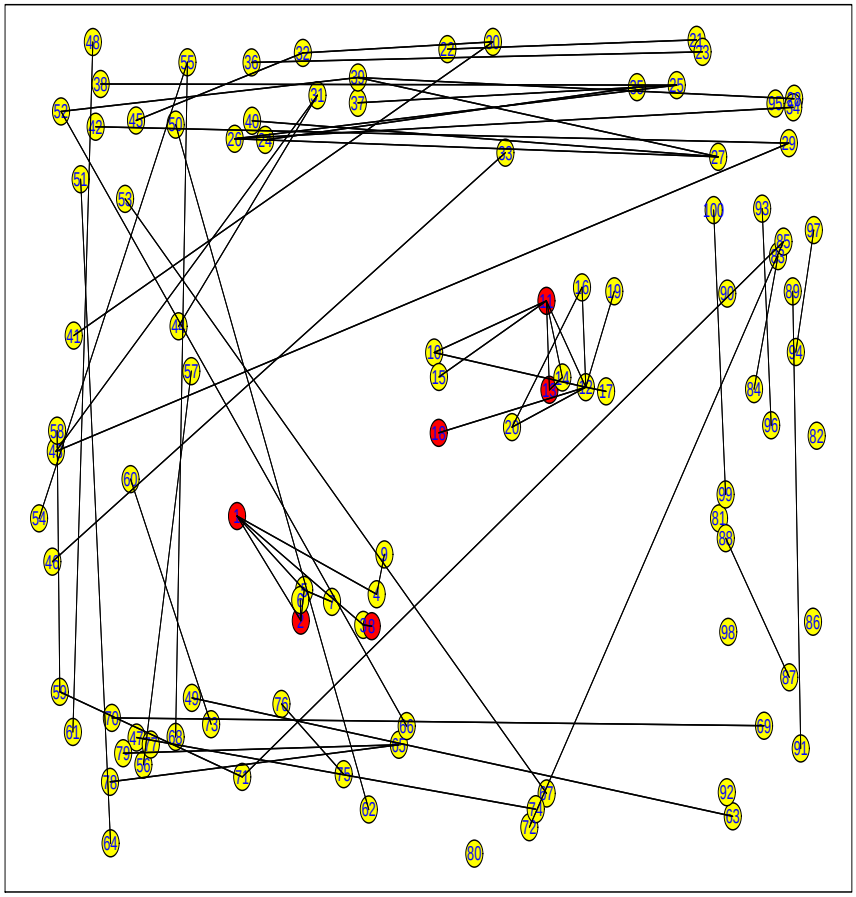


Figure 4.2. Simulated wire transaction record in Bank of America. Six suspicious Total number of accounts is 100.

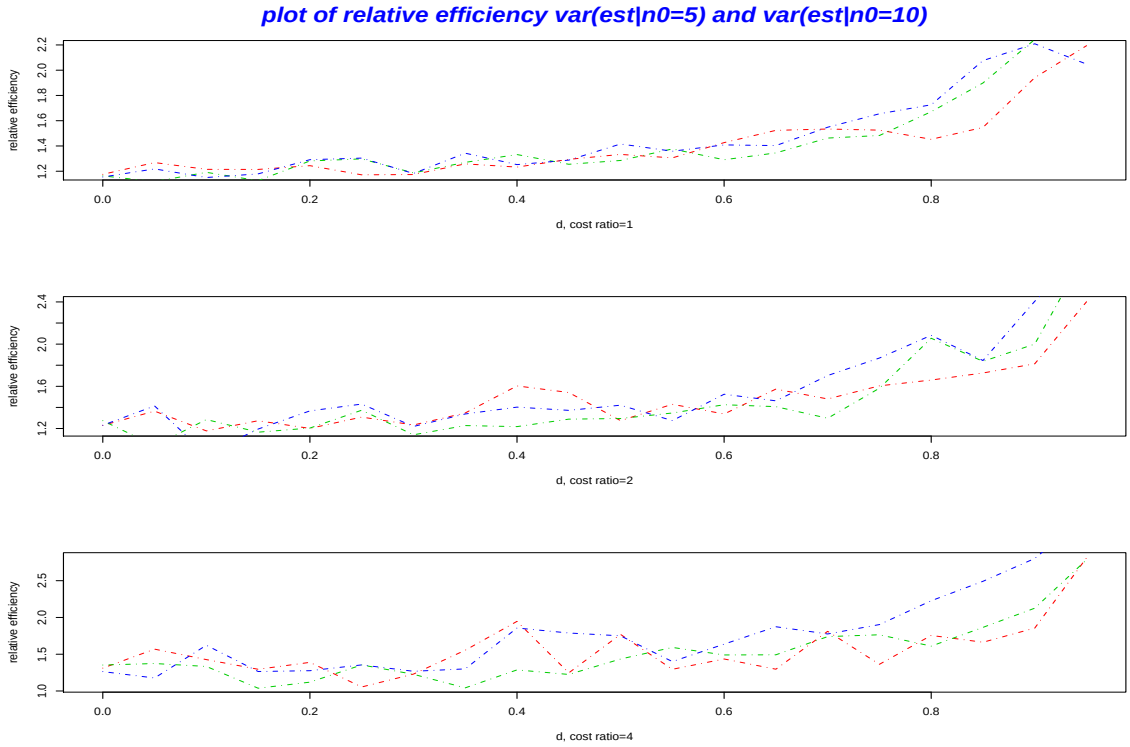


Figure 4.3. Relative efficiency between initial sample size 5 and 10, based on simulated population in Figure 4.2.

Table 4.3. Simulated Link Matrix for population in Figure 4.2

$$x[1,2] = x[2,1] = 1, x[2,6] = x[6,2] = 1, x[1,5] = x[5,1] = 1$$
$$x[1,3] = x[3,1] = 1, x[2,5] = x[5,2] = 1, x[5,7] = x[7,5] = 1$$
$$x[3,8] = x[8,3] = 1, x[1,4] = x[4,1] = 1, x[4,9] = x[9,4] = 1$$
$$x[13,14] = x[14,13] = 1, x[11,15] = x[15,11] = 1, x[11,10] = x[10,11] = 1$$
$$x[10,12] = x[12,10] = 1, x[12,16] = x[16,12] = 1, x[12,17] = x[17,12] = 1$$
$$x[12,18] = x[18,12] = 1, x[12,19] = x[19,12] = 1, x[12,20] = x[20,12] = 1$$
$$x[16,20] = x[20,16] = 1, x[21,22] = x[22,21] = 1, x[23,36] = x[36,23] = 1$$
$$x[24,35] = x[35,24] = 1, x[25,37] = x[37,25] = 1, x[25,38] = x[38,25] = 1$$
$$x[25,26] = x[26,25] = 1, x[26,27] = x[27,26] = 1, x[27,39] = x[39,27] = 1$$
$$x[27,40] = x[40,27] = 1, x[28,39] = x[39,28] = 1, x[29,42] = x[42,29] = 1$$
$$x[29,43] = x[43,29] = 1, x[30,41] = x[41,30] = 1, x[30,32] = x[32,30] = 1$$
$$x[31,43] = x[43,31] = 1, x[31,44] = x[44,31] = 1, x[32,45] = x[45,32] = 1$$
$$x[33,46] = x[46,33] = 1, x[26,34] = x[34,26] = 1, x[39,52] = x[52,39] = 1$$
$$x[47,74] = x[74,47] = 1, x[48,61] = x[61,48] = 1, x[49,63] = x[63,49] = 1$$
$$x[50,62] = x[62,50] = 1, x[51,64] = x[64,51] = 1, x[52,66] = x[66,52] = 1$$
$$x[53,67] = x[67,53] = 1, x[54,55] = x[55,54] = 1, x[55,68] = x[68,55] = 1$$
$$x[56,57] = x[58,59] = 1, x[59,71] = x[60,73] = 1, x[75,76] = x[65,78] = 1$$
$$x[65,79] = x[69,70] = 1, x[71,85] = x[72,85] = 1, x[98,85] = x[83,84] = 1$$
$$x[82,81] = x[82,80] = 1, x[86,73] = x[99,86] = 1, x[87,88] = x[89,91] = 1$$
$$x[95,92] = x[93,96] = 1, x[94,97] = x[99,100] = 1$$

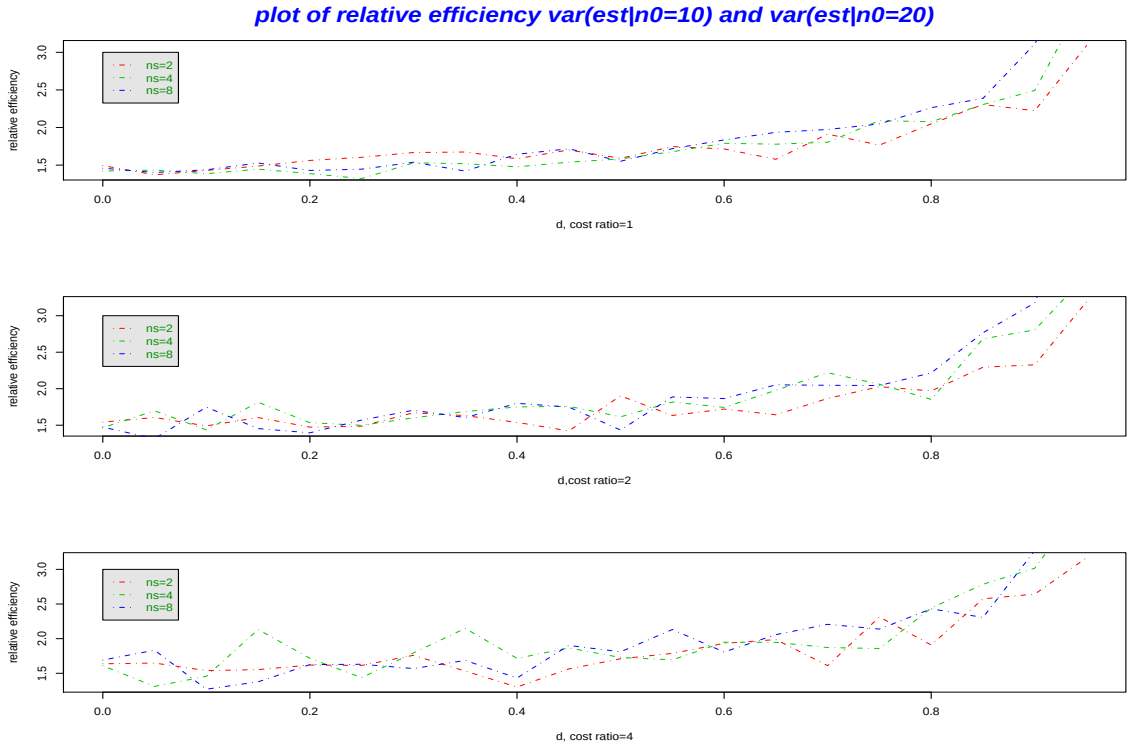


Figure 4.4. Relative efficiency between initial sample size 10 and 20, based on simulated population in Figure 4.2.

4.5 Colorado Spring Data

Colorado Spring data is from a study on the heterosexual transmission of HIV/AIDS in a "high-risk" population in Colorado Springs (Potterat et al. (1993); Rothenberg et al. (1995)). 8762 people are involved, out of which 595 were directly interviewed and 8538 refereed to as contacts by those interviewed. This study record people's relationship of sex, needle share, and drugs share. And also record everyone's HIV status, positive or negative. The population quantities interested in is pop-

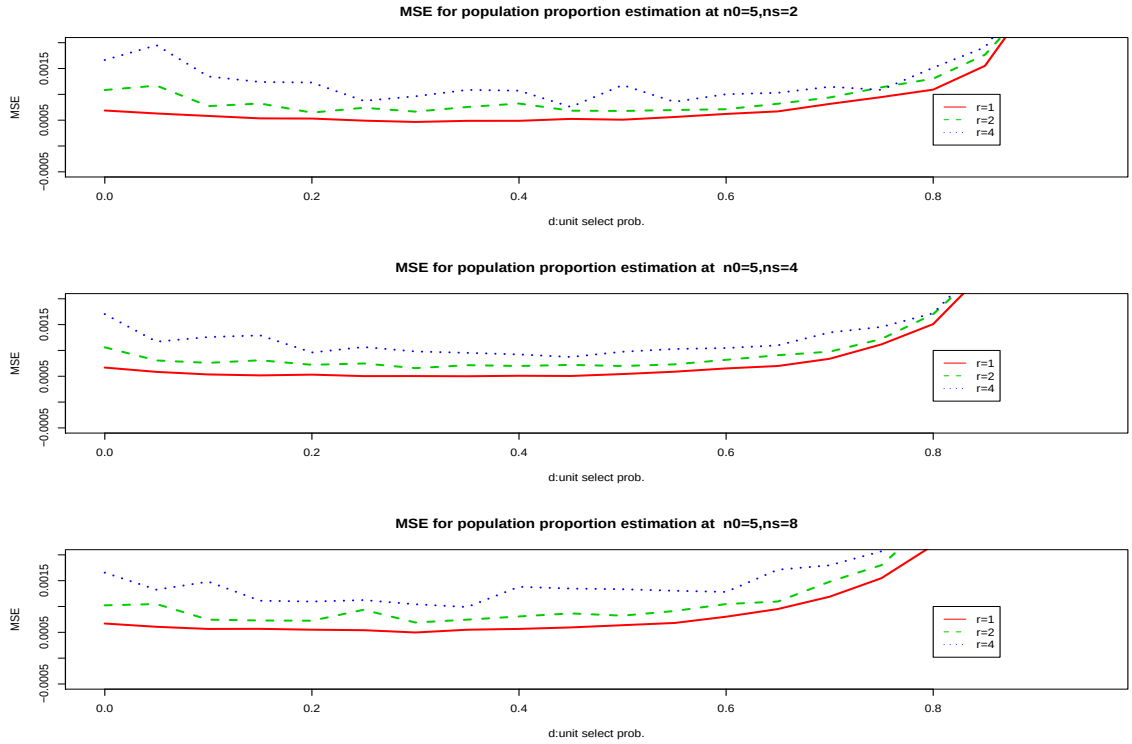


Figure 4.5. MSE of population proportion estimator, based on initial sample size 5. The true population proportion is 0.06 in Figure 4.2.

ulation proportion which is 0.5748.

Figure 4.11 describes the population structure. Red nodes represent injection drug users. Lines between paired nodes indicate drug-using relationships. Largest component contains 300 of the 595 individuals. Figure 4.12 describes the relative efficiency of $RE_{12,20}$. RE is larger than 1 means better performance and less variation for population proportion when initial sample is large. We got the similar results as from the simulated wire transaction data. The number of nodes selected at each wave do not have strong effect to the RE performance when link tracing probabilities less than 0.7. And the RE values become significantly different when link tracing probabilities larger than 0.7. The variation for initial sample size of 12 and the number of nodes included at each wave of 20 is significantly larger than the variation for

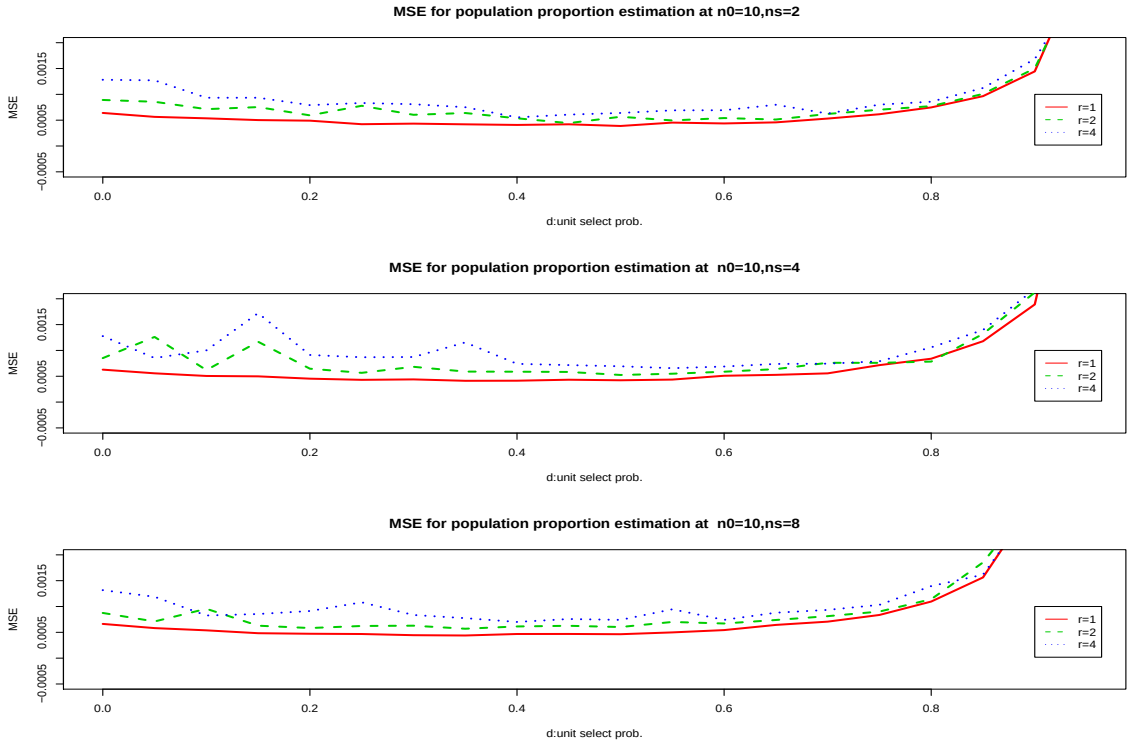


Figure 4.6. MSE of population proportion estimator, based on initial sample size 10. The true population proportion is 0.06 in Figure 4.2.

initial sample size 20 and the number of nodes included are 5 and 10. In Figure 4.12, the blue curve is significantly over the green and red curves when d is larger than 0.70. We got the similar results for estimated total sample size and sampling waves as results from simulated bank money transaction data.

Table 4.4. mse_{min}, the minimum value of MSE for population proportion estimation and corresponding selection probabilities d.

estimators	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
			$ns = 2$			
cost ratio	d	mse_{min}	d	mse_{min}	d	mse_{min}
$r = 1$	0.30	0.00047	0.50	0.00039	0.60	0.00027
$r = 2$	0.20	0.00065	0.45	0.00044	0.50	0.00030
$r = 4$	0.45	0.00076	0.40	0.00055	0.75	0.00035
			$ns = 4$			
cost ratio	d	mse_{min}	d	mse_{min}	d	mse_{min}
$r = 1$	0.35	0.00050	0.35	0.00041	0.55	0.00028
$r = 2$	0.30	0.00066	0.50	0.00052	0.55	0.00032
$r = 4$	0.45	0.00087	0.55	0.00066	0.60	0.00036
			$ns = 8$			
cost ratio	d	mse_{min}	d	mse_{min}	d	mse_min
$r = 1$	0.30	0.00050	0.35	0.00044	0.45	0.00030
$r = 2$	0.30	0.00069	0.35	0.00057	0.40	0.00036
$r = 4$	0.35	0.00099	0.40	0.00070	0.45	0.00040

Table 4.5. Pre-specified parameters value for population proportion estimation in Figure 4.2

$n0$	{12,20}
ns	{5,10,20}
$c2/c1$	{1,2,4}
d	25 values evenly across $(0, 1)$

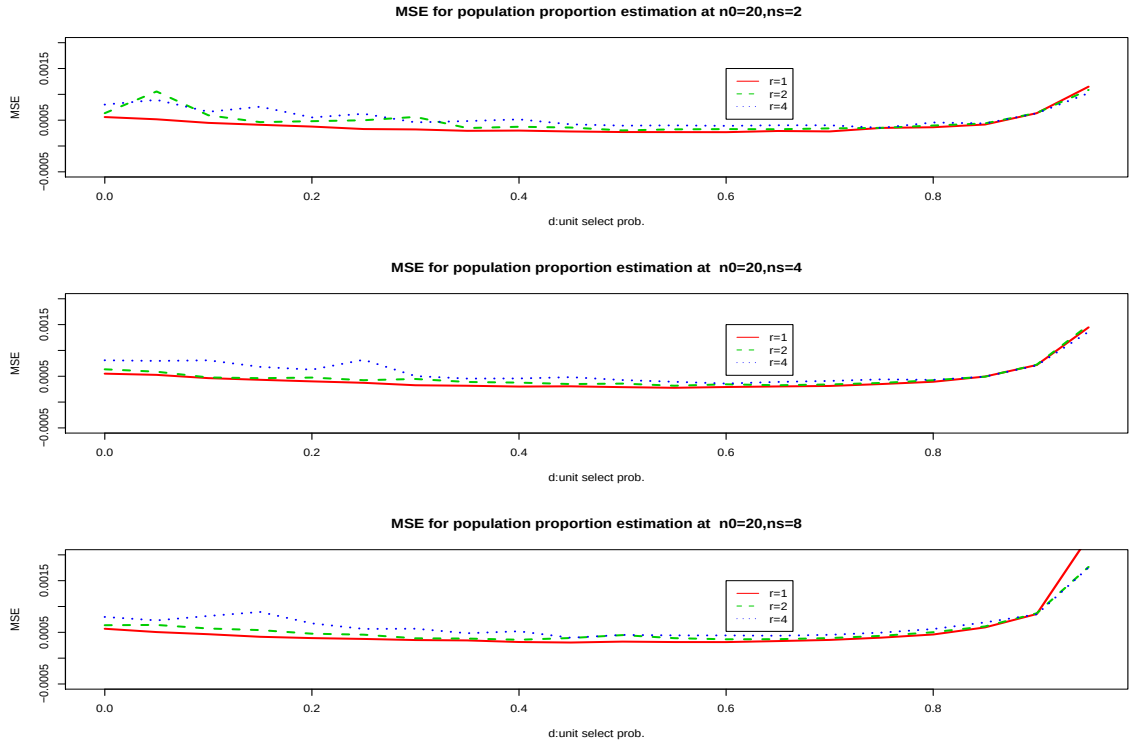


Figure 4.7. MSE of population proportion estimator, based on initial sample size 20. The true population proportion is 0.06 in Figure 4.2.

4.6 Discussion

The population proportion was estimated and optimized based on a type of AWS designs with cost model. Four parameters are in this sampling procedure: initial sample size, number of nodes included in each waves, cost ratio and link tracing probabilities. The relative efficiency between estimations with different initial sample size is evaluated. We summarize the results based on the results of the simulated wire transaction data and the empirical Colorado Spring data analysis as follows. Cost ratio is not sensitive to the estimation. This may because of the assumption for same fixed overhead cost for different initial sample. The number of nodes included is slightly sensitive to the estimation when the link tracing probability is larger than 0.70. Larger initial sample size lead to larger estimation variation.

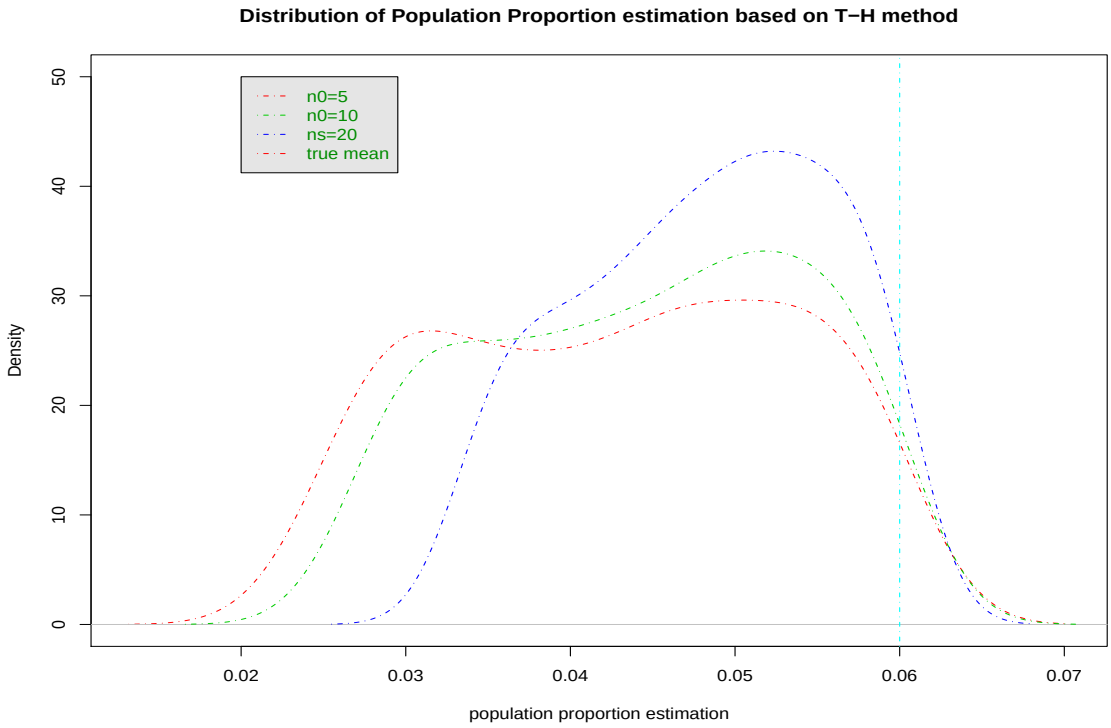


Figure 4.8. Population proportion distribution based selection probability with initial sample size 5, 10 and 20. The true population proportion is 0.06 in Figure 4.2

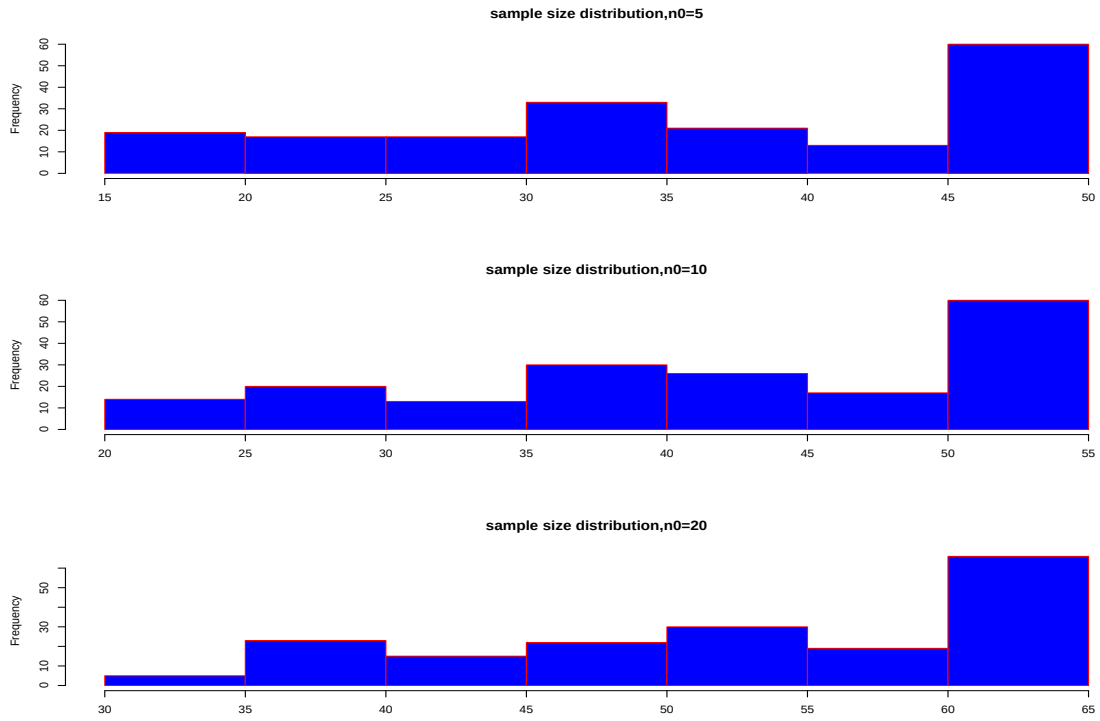


Figure 4.9. Distribution of estimated total sample size based on different initial sample size of 5, 10 and 20. Population is in Figure 4.2.

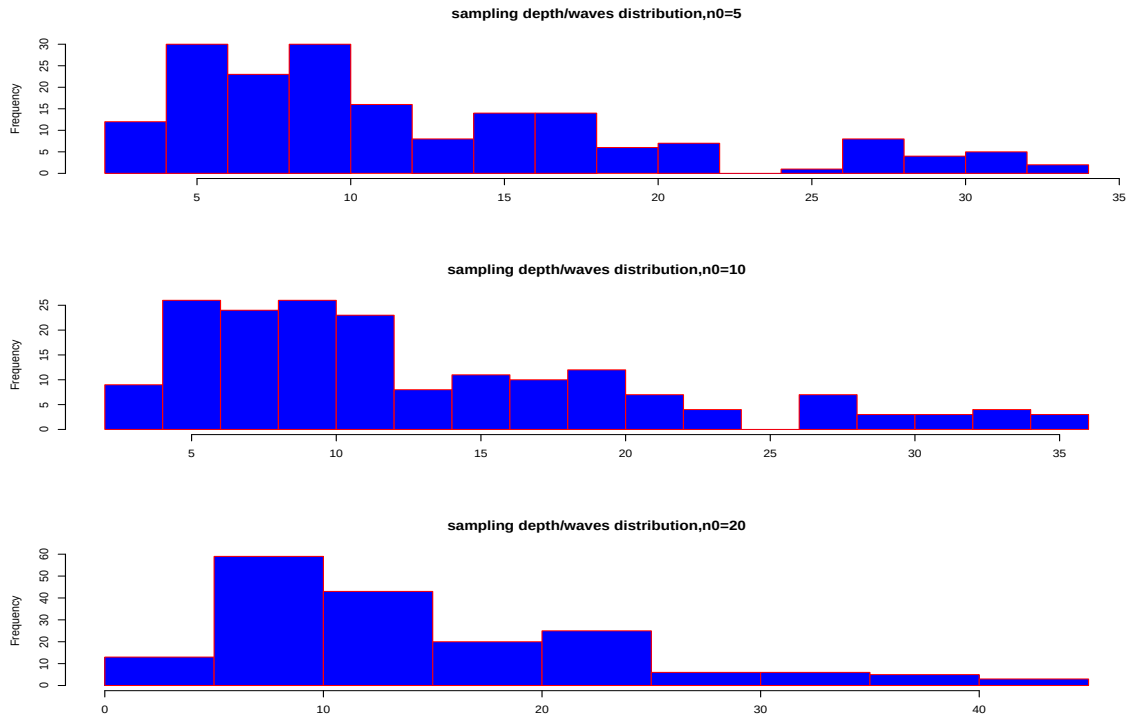


Figure 4.10. Distribution of estimated sampling waves/depth based on different initial sample size of 5, 10 and 20. Population is in Figure 4.2.

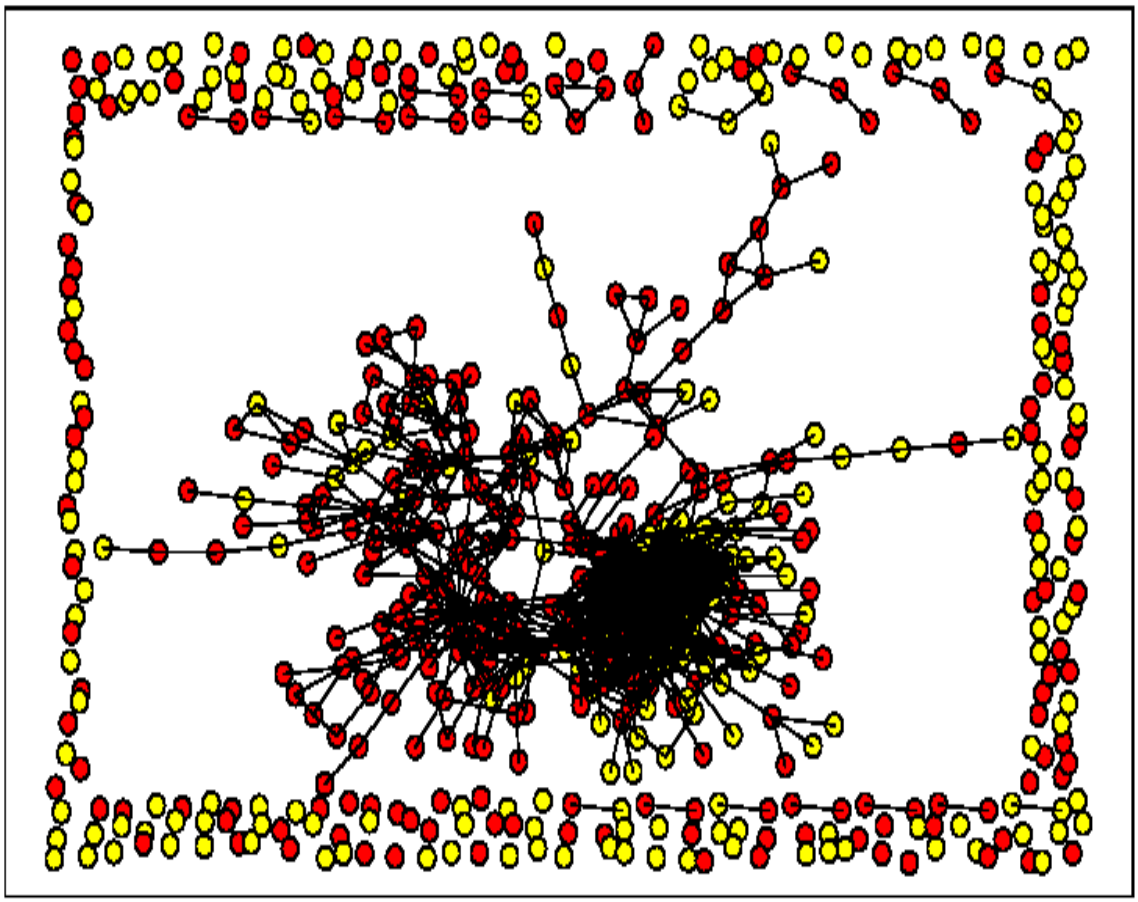


Figure 4.11. HIV/AIDS at-risk population. Dark node indicates injection drug use. Links indicate drug-using relationships. Largest component contains 300 of the 595 individuals.

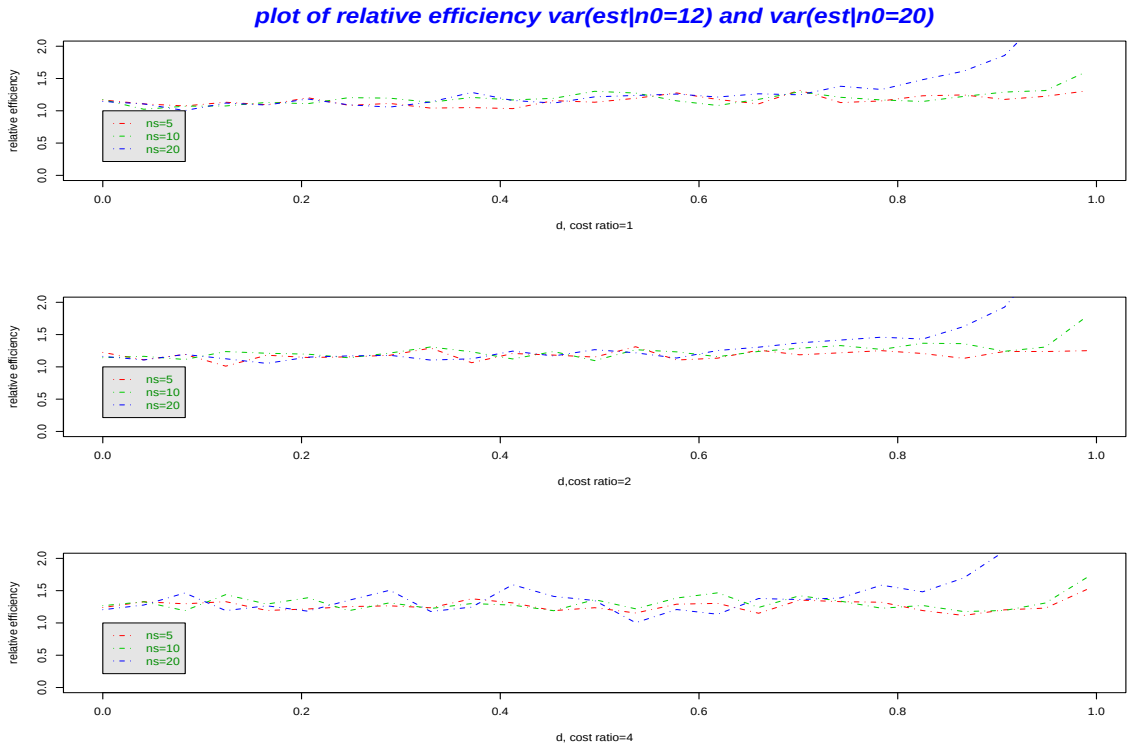


Figure 4.12. Relative efficiency between initial sample size 12 and 20, based on Colorado Spring data analysis.

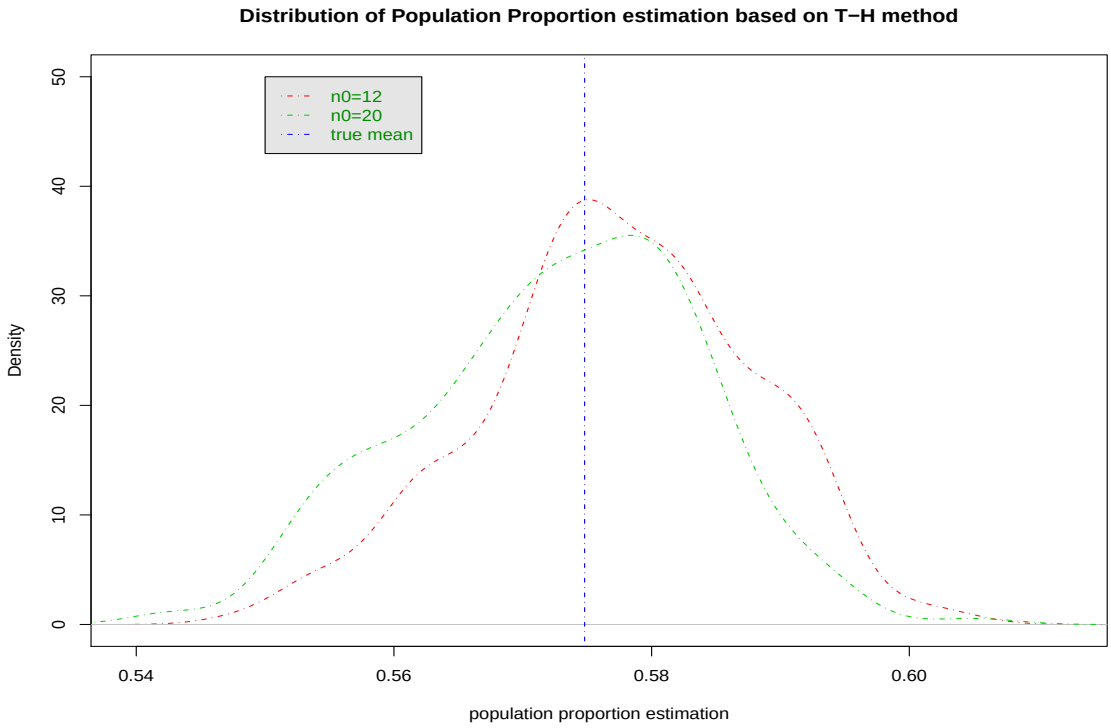


Figure 4.13. Population proportion distribution based on H-T estimation with initial sample size 12 and 20. The true population proportion is 0.5748.

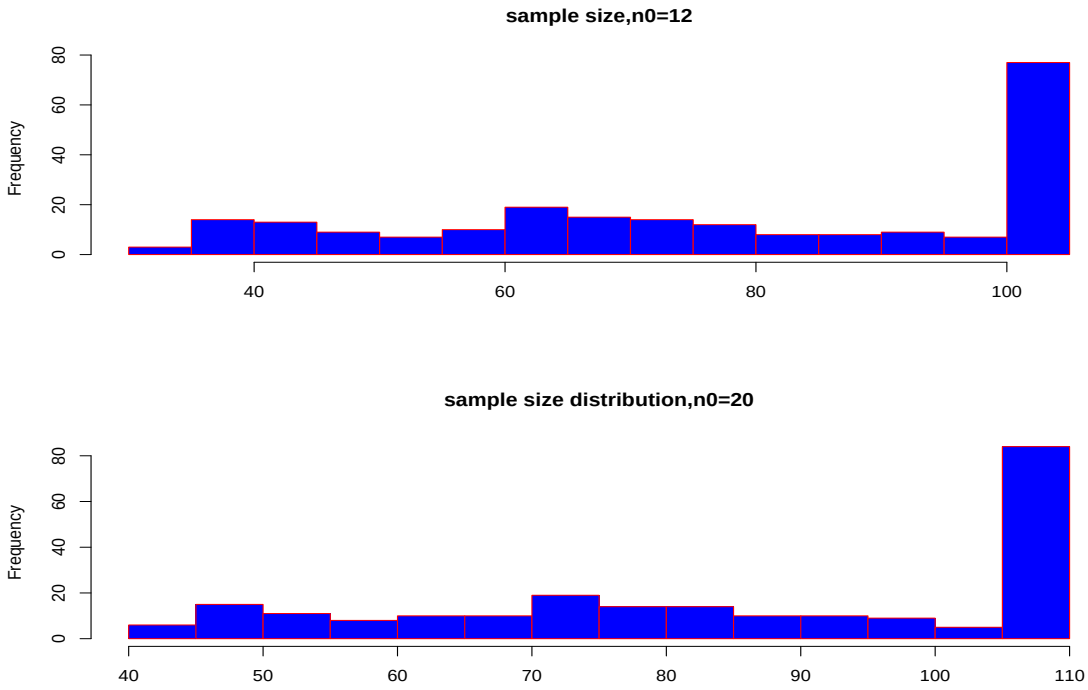


Figure 4.14. Distribution of estimated total sample size based on different initial sample size of 12 and 20. Population is in 4.2.

Chapter 5

Model Based Estimation for link-tracing designs with non-responses

5.1 Introduction

Link-tracing designs are mostly used for studying populations mostly on "sensitive" issues like in the study on illegal drug use, HIV/AIDS prevalence, prostitution activities and so forth. These studies mostly suffer from high non-response rates. In this paper we propose a model that is an extension of the model by Mosuk and Thompson (2002). We propose a model that accommodates non-random non-responses. We discuss how maximum likelihood estimates could be obtained. Both real and simulated data are used to for illustration. We assume that the population of interest has N subjects and there are different types of relations among the subjects. Each subject, u, has a value of interest Y_u. We assume that Y_u is a random variable that takes value 1 if subject u has a certain characteristic and 0 otherwise.

5.1.1 The Model

We assume a population of a known number of nodes N with each node taking two possible values 0 or 1. The proportion of 1's in the population is

$$z = \frac{\sum_{i=1}^{N} Y_i}{N} \tag{5.1}$$

where Y_i is the y-value of node i. If we define X_{uv} as the indicator variable that value 1 when there is a link between nodes u and v. Given the node u has a y-value i and

node v has a y-value j, we define the link probabilities

$$\lambda_{i+j} = P(X_{uv} = 1 | Y_u = i, Y_v = j) \tag{5.2}$$

where $i = 0, 1$ and $j = 0, 1$. So the population graph has link probabilities $\lambda = \{\lambda_{\mathbf{0}}, \lambda_{\mathbf{1}}, \lambda_{\mathbf{2}}\}$. Where λ_0 is the probability of a link between any pair of nodes with y-value 0, λ_1 is the probability of a link between a node with y-value 0 and a node with y-value 1. Similarly λ_2 is the probability of a link between a pair of nodes with y-value 1. Note that $\lambda_{i+j} = \lambda_{j+i}$.

In practice we don't know the λ's. The ratio in (5.1) is an estimator for $\theta_u = P(\text{a node has } y\text{-value } =1)$. Assuming a binary probit model implies that

$$\theta_u = Pr(Y_u = 1 | z_u, \beta) = \Phi(Z_u' \beta)$$

Where β is a $p \times 1$ vector of unknown parameters and Φ is the standardized normal distribution function. $Z_u' = (z_{u1}, z_{u2}, \cdots, z_{up})$ are vector of known covariate values.

In most link-tracing designs the value of interest Y_u is obtained from the sampled subject. Sensitive studies like the study of drug usage, HIV/AIDS and other Sexually Transmitted Diseases (STDs) and so forth suffer from high non-response rates. Which means parts of Y_u value in the sample are missing. Let r_u be the binary variable indicating whether y_u is observed ($r_u = 1$) or not ($r_u = 0$). If r_u is independent of the value of interest Y the analysis can be done by just excluding the non-responses in the estimation of parameters. This is *missing completely at random (MAR)*.

The model we consider here assumes that r_u may depend on y_u even conditional on covariates Z_u.

The aim of the work in this chapter is to develop a method for estimating the distribution of y_u

Let r_u be an indicator that takes value 1 when node u accepts to be interviewed and 0 otherwise. We define

$$\beta_i = P(r_u = 1|Y_u = i) \tag{5.3}$$

where $i = 0, 1$. The odds of a response from nodes u such that $y_u = i$ is $\frac{\beta_i}{1-\beta_i}$. We define odds ratio γ as

$$\gamma = \left(\frac{\beta_1}{1-\beta_1}\right) / \left(\frac{\beta_0}{1-\beta_0}\right) \tag{5.4}$$

When $\gamma = 1$ both nodes with y-value 1's and 0's have the same non-response rates. If $\gamma < 1$ nodes with y-value 1 tend to give more non-responses than nodes with y-value 0. The opposite is true when $\gamma > 1$. We note that γ take values in the range $(0, \infty)$.

The goal of this paper is to estimate θ when samples are drawn from populations with unknown γ. We also note that is certain studies it is of interest to study the pattern of non-responses. As a by-product, estimation and inference of γ helps to answer questions about the non-response behavior. In this paper we'll discuss how to obtain estimators for θ and γ using a snowball sample.

5.1.2 Applicability of the Model

In most link-tracing designs the value of interest Y is obtained from the sampled subject. Sensitive studies like the study of drug usage, HIV/AIDS and other Sexually Transmitted Diseases (STDs) and so forth suffer from high non-response rates. When non-responses are independent of the value of interest Y the analysis can be done by just excluding the non-responses in the estimation of parameters. This is *missing completely at random*.

The model we consider here assumes that the non-responses are related to the value of interest Y. We assume that each subject knows, or at least has an idea of his/her y-value. This information which the subject has affects his/her willingness to participate in the study. So β_i is the probability that a subject with y-value i will accept to be in the study. When a subjects rejects to be in the study, the investigator is not able to know the non-response's y-value, nor information about the existence of other links from the subject.

This model is more practical in dealing with non-responses which more often are believed to be related to the y-values.

5.2 Notation

Let S_0^c be the set of all nodes that are contacted and with and their links are followed. We define S_0^m as the set of nodes that have been contacted but refused to participate in the study. So nodes in S_0^m are contacted after following links from some nodes in S_0^c or if they have been selected in the initial sample. S_1 is the set of nodes in the last wave. All the nodes in S_1 have known y-values otherwise the nodes with belong to S_0^m.

The number of non-responses in the sample is denoted $\bar{r}(s)$ and $n_i(s) = n_i(S_0^c) + n_i(S_1)$ is the number of nodes with y-value i that are in the sample. $m_k(s)$ is the number of links of type k in the sample. We define $\bar{m}_k(s)$ as

$$\bar{m}_0(s) = n_0(S_0^c)(n_0(S_0^c) - 1)/2 + n_0(S_0^c)n_0(S_1) - m_0(s) \tag{5.5}$$

$$\bar{m}_1(s) = n_0(S_0^c)n_1(S_0^c) + n_0(S_0^c)n_1(S_1) + n_1(S_0^c)n_0(S_1) - m_1(s) \tag{5.6}$$

$$\bar{m}_2(s) = n_1(S_0^c)(n_1(S_0^c) - 1)/2 + n_1(S_0^c)n_1(S_1) - m_2(s) \tag{5.7}$$

Let z_i is the number of links from nodes $u \in S_0^c$ such that $y_u = i$ to the non-responses. Let $\bar{z}_i$ be the number of unobserved links from nodes $u \in S_0^c$ such that

$y_u = i$ to the non-responses. Since there are $n_i(S_0^c)$ nodes with $y_u = i$ in S_0^c and $\bar{r}(s)$ non-responses in S_0^m then

$$\bar{z}_i \quad = \quad n_i(S_0^c)\bar{r}(s) - z_i \tag{5.8}$$

Let $\bar{n}(s) = N - (n_0(s) + n_1(s) + \bar{r}(s))$ denote the number of nodes not in the sample. Also let R_i denote the number of responses, not observed, that could be obtained from the nodes not sampled i.e. nodes in $\bar{S}$.

5.3 Likelihood function with non-responses

For the complete data in which all node types and number of links of each type are known, the complete data likelihood is

$$L_c(\theta, \lambda; \mathbf{d}) \quad = \quad \prod_{i=0}^{1} \theta_i^{N_i} \prod_{i=0}^{1} \beta_i^{R_i^*} \prod_{k=0}^{2} \lambda_k^{M_k} \lambda_{k,0}^{C_k - M_k} \tag{5.9}$$

where N_i is the number of nodes with y-value i, R_i^* is the number of responses from nodes with y-value i, M_k is the number of links of type k and C_k is the total possible number of links of type k where $i, j = 0, 1$ and $k = 0, 1, 2$. Therefore $C_0 = N_0(N_0 - 1)/2$, $C_1 = N_0 N_1$ and $C_2 = N_1(N_1 - 1)/2$. Thompson and Frank (2000) showed that the observed data likelihood for snowball sampling in which all links are followed, except for the last wave, and there are no non-responses, i.e. $\beta_0 = \beta_1 = 1$ can be expressed as

$$L(\theta, \lambda; \mathbf{d}) \quad = \quad p(s|\mathbf{y_s}, \mathbf{x}) \prod_{i=0}^{1} \theta_i^{n_i(s)} \prod_{k=0}^{2} \lambda_k^{m_k(s)} \lambda_{k,0}^{\bar{m}_k(s)} \left[\sum_{j=0}^{1} \theta_j \prod_{i=0}^{1} \lambda_{i+j,0}^{n_i(s_0)} \right]^{n(\bar{s})} \tag{5.10}$$

where $n_i(s)$ is the number of nodes $u \in S$ such that $y_u = i$, $n_i(s_0)$ is the number of nodes $u \in S_0$ such that $y_u = i$, $n(\bar{s})$ is the number of nodes not sampled, $\lambda_{k,0} = 1 - \lambda_k$, $m_k(s)$ is the number of observed links of type k in the sample, $k = i + j$ for $\{i, j\} = \{0, 1\}$ and $\bar{m}_k(s)$ is the number of unobserved links of type k from the sample.

For a population that has non-responses the observed data likelihood is expressed as

where $n_i(s_0^c)$ is the number of nodes $u \in S_0^c$ such that $y_u = i$ and z_i is the number of links from node $u \in S_0^u$ such that $y_u = i$ to S_0^m.

The summand term with index $n(\bar{s})$ is the probability that there is no link between a node $u \in S_0^c$ such that $y_u = i$ and any of the $n(\bar{s})$ nodes in $\bar{S}$. The summand term with index z_i is the probability of a link from a node $u \in S_0^c$ such that $y_u = i$ to a node $v \in S_0^m$ that is a non-respondent i.e. $r_v = 0$. This probability can be derived as

$$\begin{aligned}
& P(\text{link to } v \in S_0^m, r_v = 0 | Y_u = i) \\
= \; & P(X_{uv} = 1, r_v = 0 | Y_u = i) \\
= \; & \sum_j P(X_{uv} = 1, r_v = 0, Y_v = j | Y_u = i) \\
= \; & \sum_j P(X_{uv} = 1, r_v = 0 | Y_u = i, Y_v = j) P(Y_v = j) \\
= \; & \sum_j \lambda_{i+j}(1 - \beta_j)\theta_j \\
= \; & \sum_j \theta_j(1 - \beta_j)\lambda_{i+j} \qquad (5.11)
\end{aligned}$$

Similarly the summand term with index $\bar{z}_i$ is the probability that there is no link from a node $u \in S_0^c$ such that $y_u = i$ to a node $v \in S_0^m$. Noting that $\theta_0 + \theta_1 = 1$ we can write

$$\begin{aligned}
& P(\text{no link to } v \in S_0^m, r_v = 0 | Y_u = i) \\
= \; & P(X_{uv} = 0, r_v = 0 | Y_u = i) \\
= \; & \sum_j P(X_{uv} = 0, r_v = 0 | Y_u = i, Y_v = j) P(Y_v = j) \\
= \; & \sum_j (1 - \lambda_{i+j})(1 - \beta_j)\theta_j \\
= \; & \sum_j \theta_j(1 - \beta_j)\lambda_{i+j,0} \qquad (5.12)
\end{aligned}$$

In this sampling design, the sample data is obtained in two stages. Firstly the investigator make contacts with the subjects by following relations. At this stage the investigator uniquely identifies the subject. In the second stage:- After being contacted the subject agrees or rejects to be in the study. When the subject agrees to be in the study, the subject's y-value is obtained and information on further links is obtained. When the subject rejects to be in the study, the investigator does not obtain the subject's y-value nor information on further links. The resultant sample hence comprises of *missing* values.

5.3.1 Predictive distribution of the unobserved quantities given the data

Let $\bar{n}_1(s)$ denote the unknown number of nodes in $\bar{S}$ with y-value 1 and R_i as the unknown number of responses that could be obtained from $\bar{S}$. From the sample data we know $\bar{r}(s)$, the number of non-respondents. We let $\bar{r}_i(s)$ be the unknown number of non-respondents in the sample with y-value i. Also z_i is the known number of links from nodes in the sample with y-value i to non-respondents. We define z_{ij} as the unknown number of links from respondents $u \in S_0^c$ such that $y_u = i$ to non-respondents $v \in S_0^m$ such that $y_v = j$ where $j = 0, 1$. We note that $z_i = \sum_j z_{ij}$. Let $\bar{M}_k$ be the unobserved number of links of type k and $\bar{C}_k$ the total possible links of type k from the nodes not sampled for $k = 0, 1, 2$. In fact $k = i + j$ so in the following sections we may write $i + j$ in place of k. So $\bar{M}_k$ are type k links in sets (S_0^m, S_0^m), (S_0^m, S_1), $(S_0^m, \bar{S})$, (S_1, S_1), $(S_1, \bar{S})$ and $(\bar{S}, \bar{S})$.

When $\bar{n}_i(s)$ and $\bar{r}_i(s)$ are known, the $\bar{C}_k$ values are

$$\begin{aligned} \bar{C}_0(\bar{n}_i(s), \bar{r}_i(s)) &= \binom{\bar{r}_0(s)}{2} + \bar{r}_0(s)(n_0(S_1) + \bar{n}_0(s)) + \\ &\quad \binom{n_0(S_1)}{2} + n_0(S_1)\bar{n}_0(s) + \binom{\bar{n}_0(s)}{2} \end{aligned} \tag{5.13}$$

$$\begin{aligned}\bar{C}_1(\bar{n}_i(s),\bar{r}_i(s)) &= \bar{r}_0(s)\bar{r}_1(s)+\bar{r}_0(s)[n_1(S_1)+\bar{n}_1(s)]+\\ &\quad \bar{r}_1(s)[n_0(S_1)+\bar{n}_0(s)]+n_0(S_1)n_1(S_1)+\\ &\quad n_0(S_1)\bar{n}_1(s)+n_1(S_1)\bar{n}_0(s)+\bar{n}_0(s)\bar{n}_1(s) \qquad (5.14)\end{aligned}$$

$$\begin{aligned}\bar{C}_2(\bar{n}_i(s),\bar{r}_i(s)) &= \binom{\bar{r}_1(s)}{2}+\bar{r}_1(s)(n_1(S_1)+\bar{n}_1(s))+\\ &\quad \binom{n_1(S_1)}{2}+n_1(S_1)\bar{n}_1(s)+\binom{\bar{n}_1(s)}{2} \qquad (5.15)\end{aligned}$$

The sufficient statistics are the unobserved quantities $\bar{n}_1(s)$, R_i, $\bar{r}_i(s)$ and z_{ij} for $i,j=0,1$. The terms $\bar{M}_{i+j}$ and $\bar{C}_{i+j}$ can be obtained from the sufficient statistics. Using the observed data, the joint predictive distribution for $\bar{n}_1(s)$, R_i, $\bar{r}_i(s)$ and z_{ij} is

$$\begin{aligned}L_p(.;\mathbf{d},\bar{n}_1(s),R_1,\bar{r}_i(s),z_{00},z_{11}) \propto\ & \binom{\bar{n}s}{\bar{n}_1(s)}\binom{\bar{n}_1(s)}{R_1}\binom{\bar{n}_0(s)}{R_0}\binom{\bar{r}(s)}{\bar{r}_1(s)}\binom{z_0^*}{z_{00}}\binom{z_1^*}{z_{11}}\\ & \theta_0^{n_0(s)+\bar{r}_0(s)+\bar{n}_0(s)}\theta_1^{n_1(s)+\bar{r}_1(s)+\bar{n}_1(s)}\\ & \beta_0^{n_0(s)+R_0}(1-\beta_0)^{\bar{r}_0(s)+\bar{n}_0(s)-R_0}\\ & \beta_1^{n_1(s)+R_1}(1-\beta_1)^{\bar{r}_1(s)+\bar{n}_1(s)-R_1}\\ & \lambda_0^{m_0(s)+z_{00}}\lambda_{0,0}^{\bar{m}_0(s)+n_0(S_0^c)\bar{n}_0(s)-z_{00}}\\ & \lambda_1^{m_1(s)+z_{10}+z_{01}}\lambda_{1,0}^{\bar{m}_1(s)+n_1(S_0^c)\bar{n}_0(s)+n_0(S_0^c)\bar{n}_1(s)-z_{10}-z_{01}}\\ & \lambda_2^{m_2(s)+z_{11}}\lambda_{2,0}^{\bar{m}_2(s)+n_1(S_0^c)\bar{n}_1(s)-z_{11}} \qquad (5.16)\end{aligned}$$

where $z_i^*=min\{z_i,n_i(S_0^c)\bar{r}_i(s)\}$ and $z_{ij}=z_i-z_{ii}$ for $i\neq j$. Considering (5.16) and factoring out a function of the unobserved number of 1's in $\bar{S}$,i.e. $\bar{n}_1(s)$ we get

$$\begin{aligned}p(\bar{n}_i(s)|\theta,\lambda,\beta,z_{ij},\bar{r}_1(s),R_1) &\sim \binom{\bar{n}s}{\bar{n}_1(s)}\prod_{i=0}^{1}\theta_i^{\bar{n}_i(s)}\prod_{i=0}^{1}\prod_{j=0}^{1}\lambda_{i+j,0}^{n_i(S_0^c)\bar{n}_j(s)}\\ &\sim \binom{\bar{n}s}{\bar{n}_1(s)}\prod_i\left\{\theta_i\prod_j\lambda_{i+j,0}^{n_j(S_0^c)}\right\}^{\bar{n}_i(s)} \qquad (5.17)\end{aligned}$$

We know that $\bar{n}_0(s)+\bar{n}_1(s)=\bar{n}(s)$ so (5.17) suggests that

$$\bar{n}_i(s)\sim B(\bar{n}(s),\Theta_i) \qquad (5.18)$$

where

$$\Theta_i = \frac{\Lambda_i}{\sum_i \Lambda_i} \tag{5.19}$$

and

$$\Lambda_i = \theta_i \prod_j \lambda_{i+j,0}^{n_j(S_0^c)} \tag{5.20}$$

Similarly using (5.16)

$$p(R_i|\theta,\lambda,\beta,z_{ij},\bar{r}_1(s),\bar{n}_1(s)) \quad \sim \quad B(\bar{n}_i(s),\beta_i) \tag{5.21}$$

$$p(\bar{r}_i(s)|\theta,\lambda,\beta,z_{ij},\bar{n}_1(s),R_1) \quad \sim \quad \binom{\bar{r}s}{\bar{r}_1(s)} \prod_{i=0}^{1} \{\theta_i(1-\beta_i)\}^{\bar{r}_i(s)} \tag{5.22}$$

Since $\bar{r}_0(s) + \bar{r}_1(s) = \bar{r}(s)$ so

$$\bar{r}_i(s)|\theta,\lambda,\beta,z_{ij},\bar{n}_1(s),R_1 \quad \sim \quad B(\bar{r}(s),\Phi_i) \tag{5.23}$$

where

$$\begin{aligned} \Phi_i &= P(Y_u = i|r_u = 0) \\ &= \frac{P(y_u = i, r_u = 0)}{P(r_u = 0)} \\ &= \frac{\theta_i(1-\beta_i)}{\sum_i \theta_i(1-\beta_i)} \end{aligned} \tag{5.24}$$

$$z_{00}|\theta,\lambda,\beta,\bar{n}_1(s),\bar{r}_1(s),R_1 \quad \sim \quad B(z_0^*,\lambda_0) \tag{5.25}$$

$$z_{11}|\theta,\lambda,\beta,\bar{n}_1(s),\bar{r}_1(s),R_1 \quad \sim \quad B(z_1^*,\lambda_2) \tag{5.26}$$

$$z_{01}|\theta,\lambda,\beta,\bar{n}_1(s),\bar{r}_1(s),R_1 \quad \sim \quad B(min\{(n_0(S_0^c)\bar{r}_1(s),z_0\},\lambda_1) \tag{5.27}$$

$$z_{10}|\theta,\lambda,\beta,\bar{n}_1(s),\bar{r}_1(s),R_1 \quad \sim \quad B(min\{(n_1(S_0^c)\bar{r}_0(s),z_1\},\lambda_1) \tag{5.28}$$

Given $\bar{r}_i(s)$ and $\bar{n}_i(s)$ we have $\bar{M}_{i+j}$ as the unobserved number of links of type $k = i + j$ in sets (S_0^m, S_0^m), (S_0^m, S_1), $(S_0^m, \bar{S})$, (S_1, S_1), $(S_1, \bar{S})$ and $(\bar{S}, \bar{S})$. Also

$$\bar{M}_{i+j} \quad \sim \quad B(\bar{C}_{i+j},\lambda_{i+j}) \tag{5.29}$$

where $\bar{C}_{i+j}$ are as given in (5.13), (5.14) and (5.15).

5.4 Estimation

In this section we discuss how estimation can be done using data obtained from a link tracing design with non-responses. We discuss how to get MLEs using the conditional predictive distributions derived in the previous section.

5.4.1 Maximum Likelihood Estimates

Getting maximum likelihood estimates involves taking the logarithm of equation (**??**) and then differentiating it w.r.t. the different parameters. Equating these derivatives to zero and simultaneously solving for the parameters would yield the maximum likelihood estimates. The maximum likelihood estimates are the values that maximize the likelihood function. Thompson and Frank (2000) this for their model and obtained a system of non-linear equations that can only be solved numerically. For our model, its even harder to obtain the system of non-linear equations so we suggest using recursive methods to find the maximum likelihood estimates.

MLE using EM Algorithm

The Expectation-Maximization (EM) algorithm is a method used to obtain MLEs mostly when the complete-data likelihood is much easier to work with than the observed-data likelihood. The EM algorithm by Dempster, *et.al.* (1977) maximizes the observed-data likelihood in two main steps - the E-step and the M-step. Let us suppose we want to find the MLE of a parameter (or vector of parameters) θ. In the E-step or expectation step we compute, for a fixed y_{obs} and parameter $\theta^{(t)}$ a function $Q(\theta|\theta^{(t)}) = E[l_c(\theta; Y)|y_{obs}, \theta^{(t)}]$ where l_c is the natural logarithm of the complete data likelihood given the observed data. In the M-step or maximization step, we optimize $Q(\theta|\theta^{(t)})$. Hence at the $(t+1)^{st}$ step the M-step finds $\theta^{(t+1)}$ that maximizes $Q(\theta|\theta^{(t)})$ to give $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$. The function Q does is non-decreasing and the

observed-data likelihood L_o is optimized at each iteration.

If the complete-data likelihood L_c is an exponential family then $l_c(\theta|Y)$ is a linear function of the sufficient statistics $T(Y)$. Where $T(Y) = (T_1(Y), T_2(Y), \ldots, T_s(Y))$ is an s-dimensional vector of complete-data sufficient statistics. For an exponential family the E-step of the EM algorithm involves replacing $T_j(Y)$ by $E[T_j(Y)|Y_{obs}, \theta^{(t)}]$ for $j = 1, 2, \ldots, s$.

For our model, the sufficient statistics are given N_i, R_i^*, M_{i+j} and C_{i+j}. From (**??**) the maximum likelihood estimates for θ_i, β_i and λ_{i+j} would be

$$\hat{\theta}_i = \frac{n_i(s) + \bar{r}_i(s) + \bar{n}_i(s)}{N} \tag{5.30}$$

$$\hat{\beta}_i = \frac{n_i(s) + R_i}{n_i(s) + \bar{r}_i(s) + \bar{n}_i(s)} \tag{5.31}$$

$$\hat{\lambda}_{i+j=k} = \frac{m_{i+j}(s) + \sum_{k=i+j} z_I + \bar{M}_{i+j}}{c_{i+j}(s) + \sum_{i+j=k} n_i(S_0^c)\bar{n}_j(s) + \sum_{k=i+j} z_i + \bar{C}_{i+j}} \tag{5.32}$$

At iteration t of the E-step, we therefore need to compute the expected value of the unobserved quantity given a fixed $\theta^{(t)}$ and the observed data y_{obs}. Using the predictive distributions in section 3 and the relation $E[X] = E[E(X|Y)]$ we obtain

$$E[\bar{n}_i(s)|y_{obs}, \theta^{(t)}, \beta^{(t)}, \lambda^{(t)}] = \bar{n}(s)\Theta_i^{(t)} \tag{5.33}$$

$$E[\bar{r}_i(s)|y_{obs}, \theta^{(t)}, \beta^{(t)}, \lambda^{(t)}] = \bar{r}(s)\Phi_i^{(t)} \tag{5.34}$$

$$E[R_i(s)|y_{obs}, \theta^{(t)}, \beta^{(t)}, \lambda^{(t)}] = \bar{n}(s)\beta_i^{(t)}\Theta_i^{(t)} \tag{5.35}$$

$$\begin{aligned} E[z_{ij}|y_{obs}, \theta^{(t)}, \beta^{(t)}, \lambda^{(t)}] &= E[z_i^*\Upsilon_{ij}^{(t)}|y_{obs}, \theta^{(t)}, \beta^{(t)}, \lambda^{(t)}] \\ &= min\{z_i, n_i(S_0^c)\bar{r}(s)\Phi_j^{(t)}\}\Upsilon_{ij}^{(t)} \end{aligned} \tag{5.36}$$

$$E[n_i(S_0^c)\bar{n}_j)|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}] = n_i(S_0^c)\bar{n}(s)\Theta_j^{(t)} \tag{5.37}$$

$$\begin{aligned} E[\bar{M}_{i+j}|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}] &= E[\bar{C}_{i+j}\lambda_{i+j}^{(t)}|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}] \\ &= \lambda_{i+j}^{(t)}D_{i+j}^{(t)} \end{aligned} \tag{5.38}$$

where $D_{i+j}^{(t)} = E[\bar{C}_{i+j}|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}]$ and $\bar{C}_{i+j}$ values are as given in (5.13), (5.14) and (5.15). After some calculations we get

$$\begin{aligned} D_0^{(t)} &= E[\bar{C}_0|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}] \\ &= \binom{\bar{r}(s)}{2}(\Phi_0^{(t)})^2 + n_0(S_1)\bar{r}(s)\Phi_0^{(t)} + \bar{n}(s)\bar{r}(s)\Phi_0^{(t)}\Theta_0^{(t)} \\ &+ \binom{n_0(S_1)}{2} + n_0(S_1)\bar{n}(s)\Theta_0^{(t)} + \binom{\bar{n}(s)}{2}(\Theta_0^{(t)})^2 \end{aligned} \tag{5.39}$$

$$\begin{aligned} D_1^{(t)} &= E[\bar{C}_1|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}] \\ &= 2\binom{\bar{r}(s)}{2}\Phi_0^{(t)}\Phi_1^{(t)} + n_1(S_1)\bar{r}(s)\Phi_0^{(t)} + \bar{n}(s)\bar{r}(s)\Phi_0^{(t)}\Theta_1^{(t)} \\ &+ n_0(S_1)\bar{r}(s)\Phi_1^{(t)} + \bar{n}(s)\bar{r}(s)\Phi_1^{(t)}\Theta_0^{(t)} + n_0(s_1)n_1(S_1) \\ &+ n_0(S_1)\bar{n}(s)\Theta_1^{(t)} + n_1(S_1)\bar{n}(s)\Theta_0^{(t)} + 2\binom{\bar{n}(s)}{2}\Theta_0^{(t)}\Theta_1^{(t)} \end{aligned} \tag{5.40}$$

$$\begin{aligned} D_2^{(t)} &= E[\bar{C}_1|y_{obs},\theta^{(t)},\beta^{(t)},\lambda^{(t)}] \\ &= \binom{\bar{r}(s)}{2}(\Phi_1^{(t)})^2 + n_1(S_1)\bar{r}(s)\Phi_1^{(t)} + \bar{n}(s)\bar{r}(s)\Phi_1^{(t)}\Theta_1^{(t)} \\ &+ \binom{n_1(S_1)}{2} + n_1(S_1)\bar{n}(s)\Theta_1^{(t)} + \binom{\bar{n}(s)}{2}(\Theta_1^{(t)})^2 \end{aligned} \tag{5.41}$$

For our model, the M-step of the EM algorithm involves just substituting the unknown quantities in equations (5.30), (5.31) and (5.32) by their expected values given the observed data and the current parameter estimates. If $\theta^{(t)}$, $\beta_i^{(t)}$ and $\lambda_{i+j}^{(t)}$ are the parameter estimates at the t^{th} step, then the $(t+1)^{st}$ step of the M-step computes

$$\hat{\theta}_i^{(t+1)} = \frac{n_i(s) + \bar{r}(s)\Phi_i^{(t)} + \bar{n}(s)\Theta_i^{(t)}}{N} \tag{5.42}$$

$$\hat{\beta}_i^{(t+1)} = \frac{n_i(s) + \bar{n}(s)\beta_i^{(t)}\Theta_i^{(t)}}{n_i(s) + \bar{r}(s)\Phi_i^{(t)} + \bar{n}(s)\Theta_i^{(t)}} \tag{5.43}$$

$$\hat{\lambda}_{i+j=k}^{(t+1)} = \frac{m_{i+j}(s) + \sum min\{z_i, n_i(S_0^c)\bar{r}(s)\Phi_j^{(t)}\}\Upsilon_{ij}^{(t)} + \lambda_{i+j}^{(t)}D_{i+j}^{(t)}}{c_{i+j}(s) + \sum_{i,j} n_i(S_0^c)\bar{n}(s)\Theta_j^{(t)} + \sum_{i,j} z_i + D_{i+j}^{(t)}} \tag{5.44}$$

where summation is over $\{i, j\} = \{0, 1\}$ such that $i + j = k$.

Illustrative steps for obtaining MLEs using the EM algorithm

0: Initialize $\theta_i^{(t)}$, $\beta_i^{(t)}$, $\lambda_{i+j}^{(t)}$

1: Compute $\Lambda_j^{(t)}$, $\Theta_i^{(t)}$, $\Phi_i^{(t)}$ and $\Upsilon_{ij}^{(t)}$, using (5.19), (5.19), (5.24) and (**??**) respectively

2: Compute $\theta_i^{(t+1)}$, $\beta_i^{(t+1)}$ and $\lambda_{i+j}^{(t+1)}$ using (5.42), (5.43) and (5.44)

4: Repeat steps 1 and 2 until either

(a) (i) $|\theta_i^{(t+1)} - \theta_i^{(t)}| < \varepsilon|\theta_i^{(t)}|$

(ii) $|\beta_i^{(t+1)} - \beta_i^{(t)}| < \varepsilon|\beta_i^{(t)}|$

(iii) $|\lambda_i^{(t+1)} - \lambda_i^{(t)}| < \varepsilon|\lambda_i^{(t)}|$

OR

(b) $|l_o(\theta_i^{(t+1)}, \beta_i^{(t+1)}, \lambda_{i+j}^{(t+1)}) - l_o(\theta_i^{(t)}, \beta_i^{(t)}, \lambda_{i+j}^{(t)})| < \varepsilon|l_o(\theta_i^{(t)}, \beta_i^{(t)}, \lambda_{i+j}^{(t)})|$

where ε is very small e.g. 10^{-5} and $l_o()$ is the natural log of the observed likelihood at the current parameter estimates.

At every iteration the EM algorithm maximize the observed-data likelihood. One of the drawbacks of the EM algorithm, just like any other optimization method, is that it not guaranteed to converge to the global maximum but to a local maximum. This happens especially when the function is not unimodal. Another drawback is that the EM may sometimes converge to the MLE at a painfully slow rate. Despite

these limitations, the EM algorithm has proved to be handy especially when analysis incomplete data.

Another method for finding MLE which is almost similar to the EM algorithm is Data Augmentation (DA). In DA method we first draw a value of the missing data from the conditional predictive distribution of Y_{mis}, $Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)})$. This is called the imputation (I-) step. The next step is the prediction (P-) step in which, conditioning on Y_{mis} we draw the new value of θ from its complete-data posterior, $\theta^{(t+1)} \sim P(\theta|Y_{mis}, Y_{obs})$. Repeating the I- and P-steps many times yields a stochastic sequence $\{(\theta^{(t)}, Y_{mis}^{(t)}\}$ whose stationary distribution is $P(\theta, Y_{mis}|Y_{obs})$. When t is large then $\theta^{(t)}$ is regarded as a random draw from $P(\theta|Y_{obs})$ and similarly Y_{mis} a random draw from $P(Y_{mis}|Y_{obs})$.

5.5 Testing for randomness

The model we have discussed assumes that the subjects' non-response are related to their y-values. When the non-responses are not a function of the y-values we have $\beta_0 = \beta_1 = \beta$ - a constant. When this is the case we say the y-values not observed due to non-responses are *missing at random*. In short, the missing pattern is not related to the y-value. In terms of the odds ratio γ defined in (5.4), when the missing pattern is at random then $\gamma = 1$.

Under missing at random $\beta_0 = \beta_1 = \beta$ and Θ in (5.19) does not depend on β. Thus from (5.18) $\bar{n}_i(s)$ is independent of β. Also from (5.24) we get $\Phi_i = \theta_i$. Thus $\bar{r}_i(s) \sim B(\bar{r}(s), \theta_i)$, independent of β. We note therefore that when missing-ness is at random then we do not need to know β for us to obtain MLEs for θ and λ_{i+j} expressed by (5.30) and (5.32) respectively. This implies that under randomness, knowledge of β does not affect the estimation of θ and λ_{i+j}. When $\beta = 1$ our model reduces to the model by Chow and Thompson (2002).

To test for the null hypothesis $H_{null} : \beta_0 = \beta_1 = \beta$ we will use a likelihood ratio test. Under the alternative hypothesis $H_{alt} : \beta_0 \neq \beta_1$ we can obtain the $l_0(\hat{\beta}_{alt})$ where $\hat{\beta}_{alt} = (\hat{\beta}_0, \hat{\beta}_1)$ are the MLEs for the full model. Similarly under H_{null} we obtain $\hat{\beta}_{null}$ as the MLE for β. The test statistic

$$T = 2l_o(\hat{\beta}_{alt}) - 2l_o(\hat{\beta}_{null}) \sim \chi^2_d \tag{5.45}$$

where d is the difference in the dimension of the parameters under H_{alt} and under H_{null}. In our case $d = 1$. Taking the logarithm of the observed data likelihood in (**??**) under H_{alt} and H_{null} and we obtain

$$\begin{aligned} l_o(\hat{\beta}_0, \hat{\beta}_1) &= \sum_i n_i(s) \ln[\hat{\beta}_i] + \sum_i z_i \ln[\sum_j \hat{\theta}_j(1-\hat{\beta}_j)\hat{\lambda}_{i+j}] \\ &+ \sum_i \bar{z}_i \ln[\sum_j \hat{\theta}_j(1-\hat{\beta}_j)\hat{\lambda}_{i+j,0}] + C \end{aligned} \tag{5.46}$$

$$\begin{aligned} l_o(\hat{\beta}) &= \sum_i n_i(s) \ln[\hat{\beta}] + \sum_i z_i \ln[(1-\hat{\beta}) \sum_j \hat{\theta}_j \hat{\lambda}_{i+j}] \\ &+ \sum_i \bar{z}_i \ln[(1-\hat{\beta}) \sum_j \hat{\theta}_j \hat{\lambda}_{i+j,0}] + C \end{aligned} \tag{5.47}$$

By (5.45) the test statistics is

$$\begin{aligned} T^* &= 2\sum_i n_i(s) \ln \frac{\hat{\beta}_i}{\hat{\beta}} + 2\sum_i z_i \ln \frac{\sum_j \hat{\theta}_j(1-\hat{\beta}_j)\hat{\lambda}_{i+j}}{(1-\hat{\beta})\sum_j \hat{\theta}_j \hat{\lambda}_{i+j}} + \\ & 2\sum_i \bar{z}_i \ln \frac{\sum_j \hat{\theta}_j(1-\hat{\beta}_j)\hat{\lambda}_{i+j,0}}{(1-\hat{\beta})\sum_j \hat{\theta}_j \hat{\lambda}_{i+j,0}} \sim \chi^2_1 \end{aligned} \tag{5.48}$$

where the ratios of $\ln()$ in (5.48) are defined and positive. For a 5% significance level, we reject the null hypothesis and say non-responses are not due to randomness if $T^* > \chi^2_{1,0.05} = 3.84$

5.6 Discussion

This paper presents an opening to doing estimation and hypothesis testing for network samples with non-responses but falls short in providing quantitative in-

ference on the estimates. Inference on the MLEs is hard to do using direct methods. This is because of the complex nature of the observed-data likelihood and hence the information matrix cannot be obtained analytically. One direction is using MCMC methods or Bayesian analysis. For the similar model by Chow and Thompson (2002) Bayesian estimation was used and HPD were obtained. The same approach, although harder here, can be used for this model.

Chapter 6

Conclusion and Future Work

Three papers with each concentrates on different applications or challenges of adaptive web sampling are described in this thesis. For the new resampling procedures, since the Markov chain we constructed starts from its stationary distribution, and from there convergency to its stationary distribution is not a problem, the problem is how fast each resampling procedure before it goes to a predefined smaller variance. In chapter 3, only between chains variation and mean square errors are used for comparison. Our results are based on carefully checked works and no new methods are developed to test the MCMC accuracy and which method is better. Also, some other technique for MCMC comparison and other better resampling procedures may exist, and those are valuable to be pursued. In chapter 4, one cost model based on a type of AWS is proposed with the implementation of simulated and real study data sets. More complected cost function would be justified from the practical aspects of the Bank data example. Regarding Chapter 5, simulation and real data study need to be shown. Bayesian approach may be more productive and easier to compute.

Bibliography

Birnaum, Z. & Sirken, M. (1965) Design of sample survey to estimate the prevvalence of rare disease: Three unbiased estimates. Vital and Health Statistics, Government Printing Office, Washington, DC .

Birnbaum, Z. & Sirken, M. (1998) Adaptive sampling in graphs. ASA Pro Srvy .

Chow, M. & Thompson, S. (1998) Estimation with link-tracing sampling designs - a bayesian approach. Proceedings of the Survey Research Section. Alexandria: American Statistical Association .

Coleman, J. (1958) Snowball sampling: Problems and techniques of chain referral sampling. Human Organization .

Dell, T. & Clutter, J. (1972) Ranked set sampling theory with order statistics background. Biometircs 28(545-553).

Dryver, A. (1999) . adaptive sampling designs and associated estimators. Phd Thesis, The Pennsylvania State University.

Erickson, B. (1978) Some problems of inference from chain data.

Erickson, B. (1979) Some problems of inference from chain data. Sociological Methodology .

Frank, O. (1977) Survey sampling in graphs. Journal of Statistical Planning and Inference pp. 235–264.

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

Frank, O. (1979) Estimation of population totals by use of snowball samples. In Perspectives on Social Network Research, Ed. P. Holland and S. Leinhardt, New York: Academic press pp. 379–347.

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

Frank Ove, S. T. (1994) Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics 10:53–67.

G.Casella & C.P.Robert (1996) Rao-blackwellization of sampling schemes. Biometrika 83(81-94).

Gelman, A. & Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7(457-511).

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Bayesian Statistics 4(169-194).

Gilks, W., S.Richardson & D.J.Spiegelhalter (1996) Markov Chain Monte Corlo in Practic(Interdisciplinary Statistics). Chapman & Hall., 175 Fifth Avenue, New York, NY, 10010,USA.

Goodman, J. & Sokal, A. D. (1989) Multigrid monte carlo method. conceptual foundations. Phy. Rev. D., 40(2035-2071).

Goodman, L. A. (1961) Snowball sampling. The Annals of Mathematical Statistics .

Hastings, W. K. (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57, Issue 1:97–109.

Heckathorn, D. D. (1997) Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems .

Heckathorn, D. D. (2002) Respondent-driven samplingii: Derivingvalid population estimates from chain-referral samples of hidden populations. Social Problems .

Henzingera, M. R., Heydonb, A., Mitzenmacherc, M. & Najork, M. (2000) On near-uniform url sampling. Computer Networks .

Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002) Latent space approaches to social network analysis. Journal of American Statistical Association 97(No.460, Theory and Method).

Kalton, G. (1991) Sampling considerations in research on hiv risk and illness. In Methodology Issues in AIDS Behavioral Research, D.G. Ostrow and R.C. kessler eds (53-74).

Kalton, G. (1993) Sampling rare and elusive populations. New York: Department of Economic and Social Information and Policy Analysis Statistics Division, United Nations .

Kalton, G. & Anderson, D. W. (1986) Sampling rare population. Journal of the Royal Statistical Society A 149.

Kaur, A., Patil, G., Shirk, S. & Taillie, C. (1996) Enviromental sampling with a concomitant variable: A comparison between ranked se sampling and stratified simple random sampling. Journal of Applied Statistics 23(231-255).

Kish, L. (1991) Toxonomy of elusivee populations. Journal of Official Statistics 7(339-347).

Klovdahl, A. (1989) Urban social networks: some methodological problems and possibilities. In: M. Kochen, Editor, The Small World, Ablex Publishing, Norwood, NJ.

Klovdahl, A., Potterat, J., Woodhouse, D., Muth, J., Muth, S. & Darrow, W. (1994) Social networks and infectious disease: the colorado springs study. Socical Science Medical .

Lawrence, S. & Giles, C. L. (1998) Searching the world wide web. Science .

Liu, J. S. (2001) Monte Carlo Strategies in Science Computing. Springer Verlag New York, Inc., 175 Fifth Avenue, New York, NY, 10010,USA.

McCoy, V. & Inciardi, J. A. (1993) Links women and aids: social determinants of sex-related activities. Women Health .

Newman, M. E. J., Strogatz, S. H. & Watts, D. J. (2001) Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E .

Potterat, J., Woodhouse, D., Rothenberg, R., Muth, S., Darrow, W., Muth, J. & Reynolds, J. (1993) Aids in colorado springs: Is there an epidemic? AIDS 7(1517-1521).

Raftery, A. E. & Lewis, S. M. (1992a) Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. Statistical Science 7(493-497.).

Raftery, A. E. & Lewis, S. M. (1992b) How many iterations in the gibbs sampler?. Bayesian Statistics 4,(ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith)(763-774,Oxford University Press.).

Rothenberg, R., Woodhouse, D. E., Potterat, J., S.Q., M., Darrow, W. & Klovdahl, A. (1995) Social networks in disease transmission: The colorado spring study. in

(eds)., r.h. needle, s.g. genser, and r.t. trotter ii, social networks, drug abuse, and hiv transmission. NIDA Research Monograph 151(3-19).

Salganik, M. & Heckathorn, D. (2002) Making unbiased estimates from hid- den populations using respondent-driven sampling. Working Paper,No. 128 Center for the Study of Ecoomy and Society, Department of Sociology, Cornell University(128).

Salganik, M. J. & Heckathorn, D. D. (2004) Sampling and estimation in hidden populations using respondent-driven sampling. SOCIOLOGICAL METHODOLOGY .

Spreen, M., C. M. (2000) Network sampling hard drug users. a structural analysis of the clients of aid agencies heerlen. Kwantitatieve Methoden .

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

Spreen., M. & Zwaagstra, R. (1994) Personal network sampling, outdegree analysis and multilevel analysis: introduction the network concept in studies of hodden population. International Sociology .

Steven K. Thompson, L. M. C. (1 November,2002) Adaptive sampling in research on risk related behaviors. Drug and Alcohol Dependence 68(Supplement 1):57–67.

Sudman, S. & Kalton, G. (1986) New developments in the sampling of special populations. Annual Review of Sociology 12(401-429).

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

Thompson, S. (1990b) Stratified adaptive cluster sampling. Biometrika 78:389–397.

Thompson, S. (1997) Thompson, s.k. (1997). adaptive sampling in behavioral surveys. in (eds., harrison, l. and hughes, a.), the validity of self-reported drug use: Improving the accuracy of survey estimates. NIDA Research Monograph 167, Rockville, MD: National Institute on Drug Abuse pp. 31–43.

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

Thompson, S. K. (1994) Factors influencing the efficiency of adaptive cluster sampling. Tech. rep., Department of Statisticse, Pennsylvania State University.

Thompson, S. K. (2002) Sampling. John Wiley and Sons, Inc., New York, 605 Third Avenue, New York, NY 10158-0012.

Thompson, S. K. (2006a) Adaptive web sampling. To appear,Biometrica .

Thompson, S. K. (2006b) Targeted random walk designs. Tech. rep.

Thompson, S. K. & Seber, G. A. F. (1996) Adaptive Sampling. John Wiley and Sons, Inc., New York, 605 Third Avenue, New York, NY 10158-0012.

W.Nahhas, R., Wolfe, D. A. & Chen., H. (2002) Ranked set sampling: Cost and optimal set size. Biometircs 58(964-971).

Yang, J. & Gupta., A. (2001) Incabs: A computer program for evaluating incabinet spectra. Proceedings of 16th International Conference on Structural Mechanics in Reactor Technology,Washington D.C.

Vita

Education

- *Ph.D.,Statistics, The Pennsylvania State University ,2007.*
- M.S., Statistics, The Pennsylvania State University, 2003
- *M.S.,Mathematic Education, The Tianjin Normal University, P.R.China, 1997.*
- B.A.,Mathematics, The Tianjin Normal University, P.R.China, 1994.

Professional Experience

- Quantitative Financial Analysis (08/06-present), Global Risk Team, Bank of America.
- Research Fellow (01/05-07/06), Statistics Department, The Pennsylvania State University.
- Statistician intern (06/04-01/05), Research and Development Department, Wyeth Pharmaceutical Company.
- Teaching Assistant(08/01-06/04), Statistics Department, The Pennsylvania State University