UCLA
UCLA Electronic Theses and Dissertations

**Title**
Population Size Estimation using Multiple Respondent-Driven Samples

**Permalink**
https://escholarship.org/uc/item/69s17218

**Author**
Kim, Brian

**Publication Date**
2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

# Population Size Estimation using Multiple Respondent-Driven Samples

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Brian Kim

2017

© Copyright by

Brian Kim

2017

ABSTRACT OF THE DISSERTATION

Population Size Estimation using Multiple Respondent-Driven Samples

by

Brian Kim

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2017

Professor Mark Stephen Handcock, Chair

Respondent-driven sampling (RDS) is commonly used to sample from hard-to-reach populations, such as female sex workers or people who inject drugs, since traditional methods are unable to efficiently survey members due to the typically highly stigmatized nature of the population. Estimating the size of these populations is often desirable so that organizations such as the Centers for Disease Control and Prevention (CDC) can provide the proper amount of aid. However, due to the nature of RDS, traditional methods of population size estimation do not provide good estimates. Therefore, in order to effectively utilize the data we can collect from these populations, we must develop new population size methods designed specifically for RDS data.

In this dissertation, I first explore some of the assumption in RDS using a data set of female sex workers in Kampala, Uganda. This is an exploratory look at the recruitment based on geographical location, with a focus on determining whether the RDS is able to actually reach all areas of the map instead of getting stuck in a certain region, as well as looking at the

effects of age on recruitment. Then, I introduce a new method of estimating population size that uses concepts from capture-recapture methods while modeling the RDS as a successive sampling process. This extends a current method of population size estimation based on RDS called Successive Sampling for Population Size Estimation (SS-PSE) to include more than one sample, incorporating the information from a capture-recapture design. I develop the Bayesian framework for the model including posterior sampling with a Markov chain Monte Carlo algorithm. Then, using simulation studies, I compare my method with various existing methods in the literature, as well as assessing the frequentist properties of my method.

The dissertation of Brian Kim is approved.

Yingnian Wu

Robert Erin Weiss

Frederic R Paik Schoenberg

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2017

*To my friends and family.*

TABLE OF CONTENTS

**1 Introduction** . . . . . **1**

1.1 Sampling Methods for Hard-To-Reach Populations . . . . . 2

1.2 Respondent-driven sampling . . . . . 4

**2 Exploration of RDS Assumptions: A Case Study of Female Sex Workers in Uganda** . . . . . **9**

2.1 Assumptions in RDS . . . . . 10

2.2 Introduction to Crane Data . . . . . 12

2.3 Approach for Assessment of RDS . . . . . 13

2.4 Assignment of Geographical Clusters . . . . . 14

2.5 Checking RDS Assumptions Using Permutation Tests . . . . . 16

2.5.1 Age . . . . . 16

2.5.2 Location . . . . . 18

2.5.3 Discussion . . . . . 19

2.6 Assessment of Markov Chain Assumptions . . . . . 20

2.7 Network Interactions Between Locations . . . . . 22

2.7.1 Modeling Location Network with Exponential Random Graph Models (ERGMs) . . . . . 24

2.8 Time to Recruitment and Distance . . . . . 26

2.9 Relational Event Modeling . . . . . 27

2.10 Discussion and Conclusions . . . . . 30

**3 Population Size Estimation Methods . . . . . 32**

3.1 Methods Based on Network Sampling . . . . . 32

3.1.1 RDS as Successive Sampling for Population Size Estimation . . . . . 33

3.1.2 Network Scale-Up . . . . . 33

3.2 Capture-Recapture and Multiple List Methods . . . . . 34

3.2.1 Ratio Estimators . . . . . 36

3.2.2 Chapman Estimator . . . . . 38

3.2.3 Bayesian Implementation of Hypergeometric Capture-Recapture Model 38

3.2.4 Log-Linear Models . . . . . 39

3.2.5 Rasch Model . . . . . 41

3.2.6 Grade of Membership Model . . . . . 42

3.2.7 Nonparametric Latent Class Model (NPLCM) . . . . . 44

3.3 Discussion of Current Methods . . . . . 45

**4 Improvements to Current Population Size Methods . . . . . 48**

4.1 Notation . . . . . 49

4.1.1 Illustrative Example for Notation . . . . . 51

4.2 Multivariate Wallenius' Non-central Hypergeometric (MWNCH) Model . . . 52
4.2.1 Extending Capture-Recapture . . . 52
4.2.2 Multivariate Wallenius' Noncentral Hypergeometric . . . 54
4.2.3 Multivariate Wallenius' Noncentral Hypergeometric as PPSWOR . . . 55
4.2.4 Likelihood . . . 56
4.2.5 Modeling the First List as PPSWOR . . . 58
4.2.6 Using Multivariate Wallenius' Noncentral Hypergeometric . . . 59
4.2.7 Computational Considerations . . . 60
4.3 Capture-Recapture SS-PSE . . . 61
4.3.1 Likelihood Formulation . . . 61
4.3.2 Modeling RDS as PPSWOR . . . 63
4.3.3 Bayesian Inference for unit size distribution . . . 65
4.3.4 Inference for population size . . . 67
4.3.5 Unit Size Distribution Model . . . 68
4.3.6 Prior specification . . . 68
4.3.7 Algorithmic Details . . . 69
4.4 Implementation . . . 71

**5 Analysis of Population Size Estimation Methods Using Simulation Studies 72**

5.1 Simulation of Population Networks . . . 73

5.2 Simulation of RDS . . . . . 75

5.3 Population Size Estimation Methods . . . . . 76

5.4 Initial Look at Performance of CR-SS-PSE . . . . . 78

5.5 Performance on Add Health Networks . . . . . 80

5.5.1 Faux Sycamore ($N = 715$) . . . . . 80

5.5.2 Faux Madrona ($N = 1000$) . . . . . 84

5.5.3 Add Health ($N = 2587$) . . . . . 88

5.6 Sensitivity to Population Size Prior . . . . . 88

5.7 Performance with Different Sample Fractions . . . . . 90

5.8 Performance with Mismatched Sample Sizes . . . . . 92

5.9 Using Simulated Networks from the National HIV Behavioral Surveillance Studies . . . . . 94

5.9.1 Simulations with $N = 10000$ . . . . . 97

5.9.2 Simulations with $N = 1000$ . . . . . 101

5.9.3 Simulations with $N = 5000$ . . . . . 104

5.10 Empirical Analysis of Frequentist Properties . . . . . 108

5.10.1 Faux Sycamore, Faux Madrona, and Add Health . . . . . 108

5.10.2 Calibrated Bayes with NHBS Networks . . . . . 110

5.11 Conclusions . . . . . 113

**6 Future Work and Conclusion . . . . . 118**

LIST OF FIGURES

1.1 A stylized diagram of what an RDS chain might look like. . . . . 5

1.2 Graphical representation of the recruitment tree for the respondent-driven sampling of PWID. The nodes are respondents, and the node color indicates cannabis usage. . . . . 6

2.1 A map showing the FSW locations along with their clusters. . . . . 15

2.2 The results of four permutation tests. We used 1000 permutations in each case. The vertical line represents the actual value. . . . . 17

2.3 Heatmap representations of the transition probability matrix. Red represents high proportions, while white represents low proportions. The rows represent the cluster of the recruiter and the columns represent the cluster of the recruit. . . . 21

2.4 Networks of recruitment after aggregating FSW into clusters based on location, with all waves and split into first 11 waves and last 14 waves. Each color represents a different geographical area of the map, split into four categories. For the network with all waves, the numbers shown are the size of the cluster. . . . . 23

3.1 Observed Data for a Simple Capture-Recapture . . . . . 35

3.2 Contingency table for a simple capture-recapture design . . . . . 36

5.1 Trace plots and histogram of the posterior distribution for $N$, $\lambda$, and $\nu$ using the Add Health data set ($N = 2587$) . . . . . 79

5.2 Simulation results from Faux Sycamore data set with all methods. The boxplot is of the posterior means. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI interval are also shown. The dashed line shows the true population size ($N = 715$). . . . . 81

5.3 A boxplot of posterior means using the six main methods with the Faux Sycamore network are shown. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 715$). 82

5.4 Simulation results from Faux Madrona data set. The boxplot is of the posterior means. The value underneath each boxplot represents the proportion of 95% SCI containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 1000$). . . 85

5.5 Boxplots of posterior means with six methods using the Add Health network. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 2587$). . . . . 87

5.6 Posterior Means with Low, Accurate, and High prior modes, along with a Flat prior. Posterior means are shown as a proportion of the true population size. For Add Health, the prior modes used were 1500, 2587, and 5000. For Faux Madrona, the modes were 700, 1000, 2000; for Faux Sycamore, 500, 715, and 1500. . . . . 89

5.7 A look at the estimation methods using different sample fractions with the Add Health network. Sample size is the size of each list ($n'$ and $n''$). . . . . 91

5.8 Posterior means from three estimation methods using different sample sizes for each list using the Add Health network. Only multiple list methods were included (no SS-PSE). . . . . 93

5.9 Posterior means from one city in the CDC PWID data set. The value underneath each boxplot represents the proportion of 95% SCIs containing the true population size. For each method, two points representing the median upper bound of the 95% SCIs and the median lower bound of the 95% SCIs are also shown. The dashed line shows the true population size ($N = 10000$). . . . . 96

5.10 Posterior means from the 40 simulated cities for $N = 10000$. The median and middle 95% of the posterior mean are shown with the point and line for each method. . . . . 98

5.11 Coverage results from the 40 simulated cities for $N = 10000$. Dashed lines show 90% and 95% coverage. . . . . 99

5.12 Posterior means from SS-PSE methods using uniform priors from one city in the CDC PWID data set. The value underneath each boxplot represents the proportion of 95% SCIs containing the true population size. For each method, two points representing the median upper bound of the 95% SCIs and the median lower bound of the 95% SCIs are also shown. The dashed line shows the true population size ($N = 10000$). . . . . 100

5.13 Posterior means from the 40 simulated cities with $N = 1000$. The median and middle 95% of the posterior mean are shown with the point and line for each method. . . . . 102

5.14 Coverage results from the 40 simulated cities for $N = 1000$. Dashed lines show 90% and 95% coverage. . . . . 103

5.15 Posterior means from the 40 simulated cities with $N = 5000$. The median and middle 95% of the posterior mean are shown with the point and line for each method. . . . . 105

5.16 Coverage results from the 40 simulated cities for $N = 5000$. Dashed lines show 90% and 95% coverage. . . . . 106

5.17 Posterior means for CR-SS-PSE for all three population sizes ($N = 10000, 1000, 5000$). The median and middle 95% of the posterior mean are shown. The dashed line shows the true population size. . . . . 107

5.18 The coverage of SCIs by width for Faux Madrona, Faux Sycamore, and Add Health networks. The dashed line indicates 95% coverage. . . . . 109

5.19 The coverage of SCIs by width for 40 Cities data for $N = 10000$. The dashed line indicates 95% coverage. . . . . 111

5.20 The median of the upper and lower bounds for the 95% and 100% SCI regions for each of the 40 cities for $N = 10000$. . . . . 112

5.21 The coverage of SCIs by width. The dashed line indicates 95% coverage. . . . . 114

5.22 The median of the upper and lower bounds for the 95% and 100% SCI regions. . 115

LIST OF TABLES

2.1 The 21 clusters and the number of FSW in each. . . . . 16

2.2 The centrality measures for each of the nodes in the social network analysis. . . 25

2.3 The coefficients from the ERGM with a location homophily term. . . . . 26

2.4 The results of the Negative Binomial regression. . . . . 27

2.5 The results of the Relational Event Model. . . . . 30

5.1 Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE ($Bias^2$/MSE) for each of the six methods with the Faux Sycamore network. . . . . 83

5.2 Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE ($Bias^2$/MSE) for each of the six methods with the Faux Madrona network. . . . . 84

5.3 Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE ($Bias^2$/MSE) for each of the six methods with the Add Health network. . . . . 86

5.4 Table of network characteristics for the CDC PWID networks. Number of Recruits percentage are among sample members who were given coupons. Some studies had a maximum of three coupons; the counts were set to 0 for the purposes of the last two rows. . . . . 95

5.5 Table of results for simulations in this chapter by whether they were **Accurate**, tended to **Over**estimate, or tended to **Under**estimate. Accurate results are shown in green while inaccurate (either over or underestimate) are shown in red. . . . . 117

# ACKNOWLEDGMENTS

If I were to try to name every person who has helped me reach this point, it would be an extreme disservice to those I would inevitably leave off. I can't even begin to name everyone that has helped me in one way or another during my time in graduate school — so much so that I won't even try. Of course, there are the obvious ones, the people who have had a direct impact in helping me write this very document and my closest friends who supported me throughout this endeavor. But there are also the people who may not have even known they were part of this journey of mine, from helping me relieve stress with games to giving me advice on the coffee-making that would fuel many late nights. And though they may not have had as much of an impact as others, I believe every single person deserves my gratitude in one form or another.

I admit this is not a strong suit of mine; my greatest fear is that I inadvertently forget someone in a long list of names. However, despite the conspicuous lack of names here, I do sincerely thank each and every person who has been a part of this stage of my life. Thank you to the professors, who taught me, guided me, and motivated me. Thank you to all my friends and family, who gave me endless support and happiness and joy. Thank you to the UCLA staff, in the Statistics department and outside, for providing such an amazing place of learning. Thank you even to people who don't fit into these neat categories, for everything they have done to help me reach this point.

Thank you all, for an unforgettable experience.

# VITA

| | |
|---|---|
| 2012 | B.A. Philosophy and B.A. Mathematics *cum laude*, Amherst College. |
| 2013–2016 | Teaching Assistant, Statistics Department, UCLA. |
| 2016 | TA of the Year, Statistics Department, UCLA. |
| 2014–2016 | Statistical Consultant, Institute for Digital Research and Education, UCLA. |
| 2015–2017 | Academic Mentor, Institute for Pure and Applied Mathematics, UCLA. |
| 2016–2017 | Collegium of University Teaching Fellows, Statistics Department, UCLA. |
| 2017 | Instructor: STAT 98T, Statistics Department, UCLA. |
| 2016–2017 | Dissertation Year Fellowship, Statistics Department, UCLA. |

# PUBLICATIONS

Kim, B., 2012. Using Pitch F/X Data to Evaluate Pitchers. Amherst College Undergraduate Honors Thesis, Mathematics.

# CHAPTER 1

# Introduction

There are some populations in which obtaining a simple random sample, or any sort of probability sample, can be impossible. This can be for a variety of reasons, including social stigma or unwillingness to self-identify. For example, in the population of people at high risk for HIV, such as men who have sex with other men (MSM), female sex workers (FSW) or people who inject drugs (PWID), the stigma of having HIV may discourage people in the population from participating in the study.

In many cases, finding the size of these hard-to-reach populations is of great interest. When deciding how much aid to send for HIV prevention, getting an accurate count of people at high risk for HIV is crucial for efficient allocation of resources (UNAIDS and Organization 2010). Since these populations are hard to reach, traditional methods of sampling, such as random telephone numbers or household surveys, are unfeasible. In addition, due to the stigma attached to many hidden populations, individuals may refuse to release information to protect theirs or others' privacy (Heckathorn 1997).

In this dissertation, we will propose a new method for estimating the size of these hard-to-reach populations. To do this, we consider the two main difficulties: finding a good way of sampling from a hidden population; and using the data collected to estimate the population

size while taking into account the nature of the non-probability sample that we use.

## 1.1 Sampling Methods for Hard-To-Reach Populations

Researchers have developed various methods to survey hidden populations. One method seeks to take advantage of natural patterns of the population of interest and aim to identify locations that may have a higher concentration. This method, called time-location sampling or venue-based sampling, involves selecting locations at which to sample at certain times. This allows the researcher to obtain samples and provides a sampling framework to calculate the inclusion probability of each unit in the population. For example, for a study of women at high risk for HIV, researchers may choose to survey people at locations with a higher concentration of women to try to find enough people who fit the selection criteria (Haley et al. 2014).

However, it is ineffective if there are many units that do not frequent the locations chosen. Time-location sampling relies on most of the members to aggregate at a relatively few places, something that may be unrealistic (Abdul-Quader et al. 2006). Researchers also must rely on accurately identifying the locations to do their sampling from, as well as obtain permission to recruit from the venues.

An alternative way of reaching these hidden populations involves exploiting their highly connected nature. For example, people who inject drugs are much more likely to know someone else who injects drugs. In addition, it is much more likely that a person in the population would even know that someone else is also in that population (e.g. it is more likely for a PWID to know who among people they know is also a PWID). Therefore,

researchers have developed various link-tracing sampling methods, in which the network links from sampled members of the population are traced out to unsampled members in the population in order to grow the sample (Spreen 1992, Handcock and Gile 2011, Gile and Handcock 2010).

One of these methods is snowball sampling, which starts by selecting a few initial contacts who provide information about others in the population. Researchers find these people and ask them to participate in the study. Those who do participate are then asked about others. The resulting pool of possible respondents is then tapped to obtain the next wave. This process is repeated until the overall sample is obtained (Coleman 1958, Goodman 1961).

This method provides a benefit over time-location sampling in that it allows a method of access to people who do not frequent public venues. Though the initial seeds may be biased by visibility in physical locations, subsequent samples from referrals can reach less visible individuals. A main downside to snowball sampling is that it requires participants to tell researchers about others in the population. Due to the stigma of being in the population, participants may be hesitant to identify their friends or acquaintances for inclusion into the study (Heckathorn 1997). In addition, even though the initial seeds are theoretically to be chosen at random from the population, this is usually impossible (which is why we would want to use this method of sampling in the first place). This can lead to samples that depend heavily on a convenience sample of inital seeds (Heckathorn 2011).

One method that has recently become very popular with researchers for studying hidden populations is respondent-driven sampling, or RDS.

## 1.2 Respondent-driven sampling

Respondent-driven sampling, which was introduced by Heckathorn (1997) as an alternative to traditional snowball sampling and time-location sampling techniques, employs a link-tracing design. RDS works in the following manner:

1. Start with a small initial sample, usually a convenience sample.
2. Give each respondent a few coupons to recruit others, with incentives for both the recruiter and the new recruit.
3. Include each recruit in the study and give them a limited number of coupons, typically approximately 3 (Malekinejad et al. 2008).
4. People who receive a coupon may choose to come in to join the study.
5. Each new participant is included in the study and given coupons to recruit others.
6. The process continues until a stopping condition is reached, such as a target sample size. If the chain stops before the stopping condition is reached (due to unfruitful referrals), new seeds may be chosen to start new chains.

Figure 1.1 shows a simple small RDS chain. Person 1 is the initial seed, who is included in the study and then given coupons to recruit others. Persons 2, 3, and 4 are all recruited into the study, and are in turn given coupons to recruit others. Person 2 does not recruit anyone into the study, and that branch ends. Person 3 recruits 5 and 6, while person 4 recruits 7. This typically continues on for longer than shown. Figure 1.2 shows real RDS chains with

multiple seeds of PWIDs from an RDS study was done in Nador, Morocco (Johnston et al. 2015).

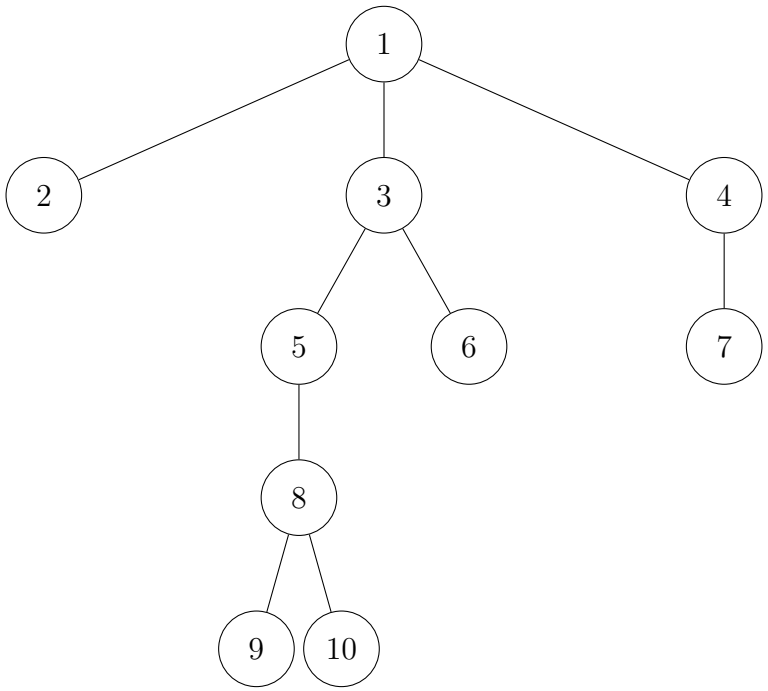



Figure 1.1: A stylized diagram of what an RDS chain might look like.

RDS has several benefits over more conventional methods. First, and most prominent, is that it enables researchers to survey a population for which a sampling frame does not exist. Since respondents know others in the population (a reasonable assumption for populations like MSM and FSW), it is much easier to ask them to find more people for the study rather than for researchers to try to find them. In addition, RDS offers a few benefits over other, more traditional chain-recruiting sampling techniques. First, RDS involves a dual incentive system, with incentives for both the recruiter for recruiting others as well as participants for joining the study. This means that recruitment into the study is encouraged even more than in other chain-recruiting techniques. In addition, the long chains generated by RDS result

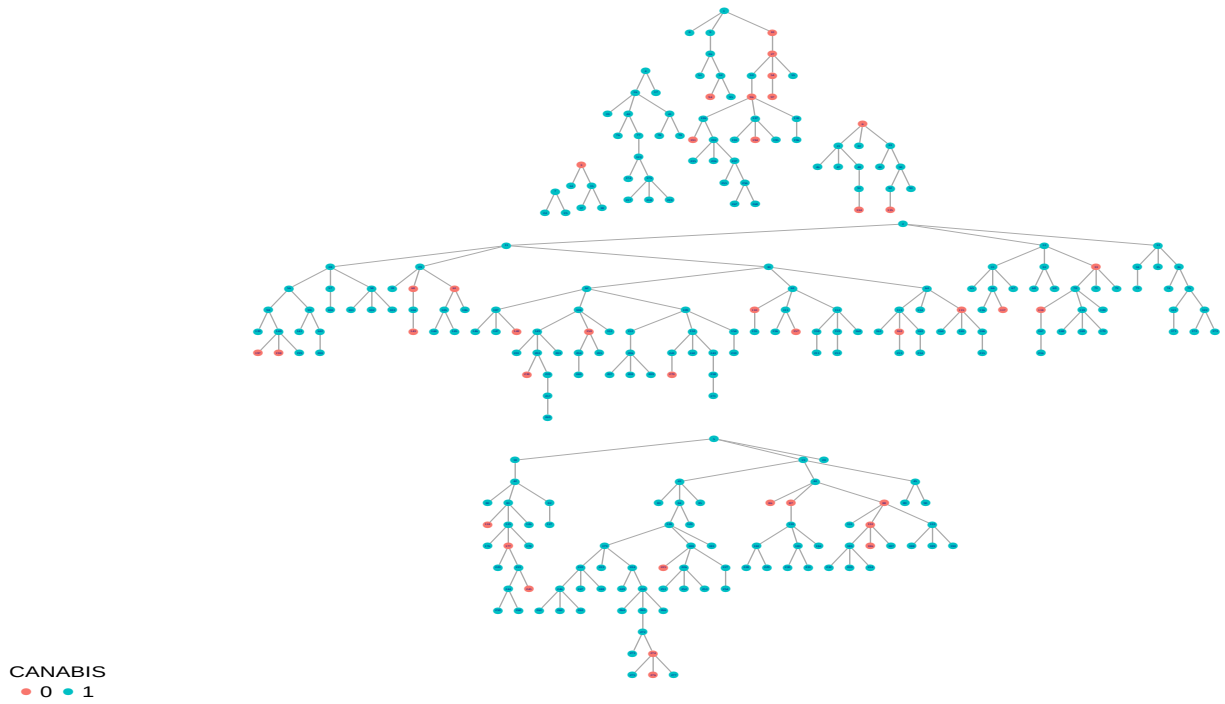



Figure 1.2: Graphical representation of the recruitment tree for the respondent-driven sampling of PWID. The nodes are respondents, and the node color indicates cannabis usage.

in samples that are not as prone to bias by the initial convenience sample, whereas other recruitment sampling methods may be (Heckathorn 1997, Gile 2011). In particular, RDS is able to reach the less visible members of the population that may be missed by time-location and snowball sampling (Kendall et al. 2008). Further, RDS does not require participants to give up any information about others. Instead, they are simply asked to recruit them into the study, giving agency to the possible recruits. This avoids the issue of asking respondents to reveal information about their friends or acquaintances (Heckathorn 1997).

Due to its many benefits, respondent-driven sampling has increasingly been the method of choice when surveying these hard-to-reach populations (UNAIDS and Organization 2010, Bengtsson et al. 2012, Platt et al. 2006). A review in 2008 found that there have been over 120 studies that have used RDS to sample most-at-risk populations for HIV (Malekinejad et al. 2008). Much of the previous work has been done using respondent-driven sampling to find proportions or other statistics (Volz and Heckathorn 2008, Gile 2011). Many studies comparing RDS with other methods have found that RDS is an effective and efficient method for sampling hard-to-reach populations (Abdul-Quader et al. 2006, Semaan 2010, Magnani et al. 2005, Platt et al. 2006).

However, the current literature on estimating the population size using only RDS data is quite limited, and has mostly focused on using a multiplier method (Paz-Bailey et al. 2011,Wattana et al. 2007). Some research has been done regarding using other forms of network sampling, such as incorporating general link-tracing design to capture-recapture (Vincent and Thompson 2016), but very few model-based methods have been developed specifically for RDS data (Handcock et al. 2014). Because RDS is so popular with researchers,

methods must be developed to estimate population size using RDS data.

In this dissertation, I will propose a new method for estimating the size of a hidden population using multiple respondent-driven samples. First, in Chapter 2, I will look at some of the assumptions behind RDS methods and assess how valid those assumptions are. In Chapter 3, I will review current methods in population size estimation, focusing on the methods that use network sampling or that use repeated sampling techniques. In Chapter 4, I will describe two new methods for population size estimation using RDS: the Multivariate Wallenius' Non-central Hypergeometric model and Capture-Recapture Successive Sampling - Population Size Estimation (CR-SS-PSE) model. Then, in Chapter 5, I will use simulation studies to compare my main model, CR-SS-PSE, with the methods currently used for population size estimation as described in Chapter 3. Finally, I will provide concluding remarks in Chapter 6.

# CHAPTER 2

# Exploration of RDS Assumptions: A Case Study of Female Sex Workers in Uganda

Even though Respondent-Driven Sampling is able to provide samples of hard-to-reach populations, there are concerns how it works in practice. There are a number of assumptions associated with RDS that may not be met in the real world. One is that respondents recruit randomly from their personal networks. This is an assumption that may provide a point of concern for RDS, since in reality, we might not expect this to be true. For example, if respondents only recruited from people their own age instead of truly randomly, then we would only obtain a sample within the population of that age group. These types of bottlenecks in the networks may lead to a sample from only a portion of the population.

Notably, though, we are interested in bottlenecking only if it affects the estimates. In other words, if we, for example, want to find what proportion of the respondents have HIV, then a bottleneck would only matter if the proportion is different on either side of the bottleneck. When estimating the size of the hidden population, however, we are concerned with bottlenecking because we might end up only estimating the size of one side of the bottleneck.

Alternatively, respondents may only recruit others who are close by, resulting in a geo-

graphical bias. Since it is more convenient to give coupons to people who are close, we might expect respondents to prefer recruiting from their immediate vicinity. In addition, we would expect that recruiting people further away takes more time, so the possible recuits may be less likely to come in (because, for example, it is further away from the testing center or they simply aren't willing to make the trip).

## 2.1 Assumptions in RDS

There are many assumptions in RDS and with methods that use RDS data. Gile (2011) touches on many of these for the Volz-Heckathorn estimator and the Successive Sampling estimator. Because the methods as described in Chapter 4 are based on a successive sampling approximation of RDS, we will mostly focus on those assumptions. The assumptions, taken from Gile (2011), are:

1. **Homophily weak enough:** If network homophily is too strong, then only a subset of the true population is being sampled. For example, if respondents only recruit from among people their own age, then the population becomes restricted to that age.

2. **Connected graph:** The populations of interest tend to be highly connected, which is why we use RDS. Lansky et al. (2012) tested this assumption by analyzing RDS in cities with multiple field sites and checking if cross-recruitment occurred. They found that in these cities, there was always cross-recruitment as long as the sites weren't separated in an artifical manner (one of the studies involved a field site that operated after a different one closed, meaning no cross-recruitments were possible).

3. **Sufficiently many sample waves:** The RDS process starts as a convenience sample, and it must be allowed to run for a sufficient number of waves to remove the biases of this convenience sample.

4. **All ties reciprocated:** Descriptions of personal networks are one-sided, since only one person is providing the information. Therefore, it is possible that there is incorrect specification of a relationship, since it might be considered a friendship by one but not by another. One way of alleviating some of this is by careful wording of the question, using phrases like, "How many people do you know that would give you a coupon?"

5. **Degree accurately reported:** Personal network size might be reported incorrectly for a number of reasons, including barrier effects and recall bias (Johnston et al. 2015, Killworth et al. 2006).

6. **Random referral:** Preferential recruitment is expected among members of the population, even if they are explicitly instructed to distribute coupons randomly. This could be because of a conscious choice, perhaps to allow friends to access the benefits of participating in the study, or simply due to an inherent propensity to favor certain people.

The main assumptions we will look at in this chapter are: homophily weak enough, sufficiently many sample waves, random referral.

## 2.2 Introduction to Crane Data

In 2012, the Centers for Disease Control and Prevention (CDC) conducted RDS surveys in Kampala, Uganda with a focus on collecting geographical information as part of what they called the Crane Survey. This was funded by the President's Emergency Plan For AIDS Relief (PEPFAR) and was a collaboration with Makerere University, CDC Uganda, and the Ministry of Health. The target populations for these surveys included female sex workers, men who have sex with men, and people who inject drugs, as well as others such as partner concurrency networks and people with disabilities.

As part of the survey, various medical tests and services were provided depending on the population, such as testing blood for various sexually-transmitted diseases and testing urine for drugs. In addition to disclosing their disease and drug status, respondents were asked a series of location-based questions. They were shown a blank map of the Kampala region and asked to put a pin at — in the case of female sex workers — the location they typically worked. In addition, when asked about their personal networks, respondents were asked about each region of the map separately. In each case, the respondents were assured of confidentiality and informed of the CDC's efforts to provide proper access to health care as the reason for obtaining the geographical information.

Though data was collected on multiple populations at high risk for HIV, we focus on just the population of female sex workers. There were 4 initial seeds, and respondents were given 3 coupons with which to recruit others. Data was collected on 1501 FSW, with 1334 providing geographical information. The missing locations were due to refusal to participate in the mapping exercise. Researchers determined that this refusal was at random and, in

particular, not associated with geographical area. Other characteristics of the participant were collected, such as age, and unique coupon linking was also performed to determine who the referral had come from.

## 2.3 Approach for Assessment of RDS

In this chapter, we offer an assessment of RDS using a data set that includes geographical location. We will do this by focusing a few main questions:

1. **How do respondents recruit in terms of geographic location?** That is, do they only recruit from those closer to themselves, or do they also recruit people who are far away? What is the "reach" of RDS? This is the main focus of using the geographical information that is unique to the way this RDS data set was collected.

2. **How does the recruitment process change over time and in later waves?** This is important for estimators that rely on stationarity.

3. **How do different variables such as age affect recruitment?** This is related to the assumption of weak enough homophily and random referral. Simply put, we want to know whether the RDS process is only recruiting from a small subset of the target population (e.g. only female sex workers of a certain age group rather than all female sex workers).

We start by assigning geographical clusters and grouping the respondents according to these clusters as described in Section 2.4. Then, in Section 2.5, we use permutation tests to look at whether recruitment is happening at random. In Section 2.6, we look at the Markov

chain assumptions by looking at the recruitment in clusters over time. In Section 2.7. we treat clusters as nodes of a network and do some network analysis on the resulting graph. In Section 2.8, we use a Negative Binomial regression to look at the effect of distance on time to recruitment. Lastly, in Section 2.9, we use a method called relational event modeling to look at how recruiting works between clusters.

## 2.4 Assignment of Geographical Clusters

The key innovation in this data was the inclusion of geographical locations in the RDS data. In particular, we want to know how recruiting happens across various locations in Kampala, and how different covariates might affect this. In much of our analysis, we decided to group FSW locations into clusters based on their location, reasoning that FSW who were close together were geographically similar. In other words, when a respondent marked a pin location, it was assumed that the actual area that they worked in was not simply at that single point.

We identified 21 clusters of FSW based on where the pins were most closely grouped together. The map and pin locations colored by cluster are shown in Figure 2.1. We chose these clusters by looking at the pin locations that were grouped together, as well as the description of the venues associated with the pin locations. We tried as much as possible to prevent large disparities in the sizes of the clusters, though some amount of variation was unavoidable due to the difference in popularity of venues.

Table 2.1 shows the number of FSW recruited from each cluster.

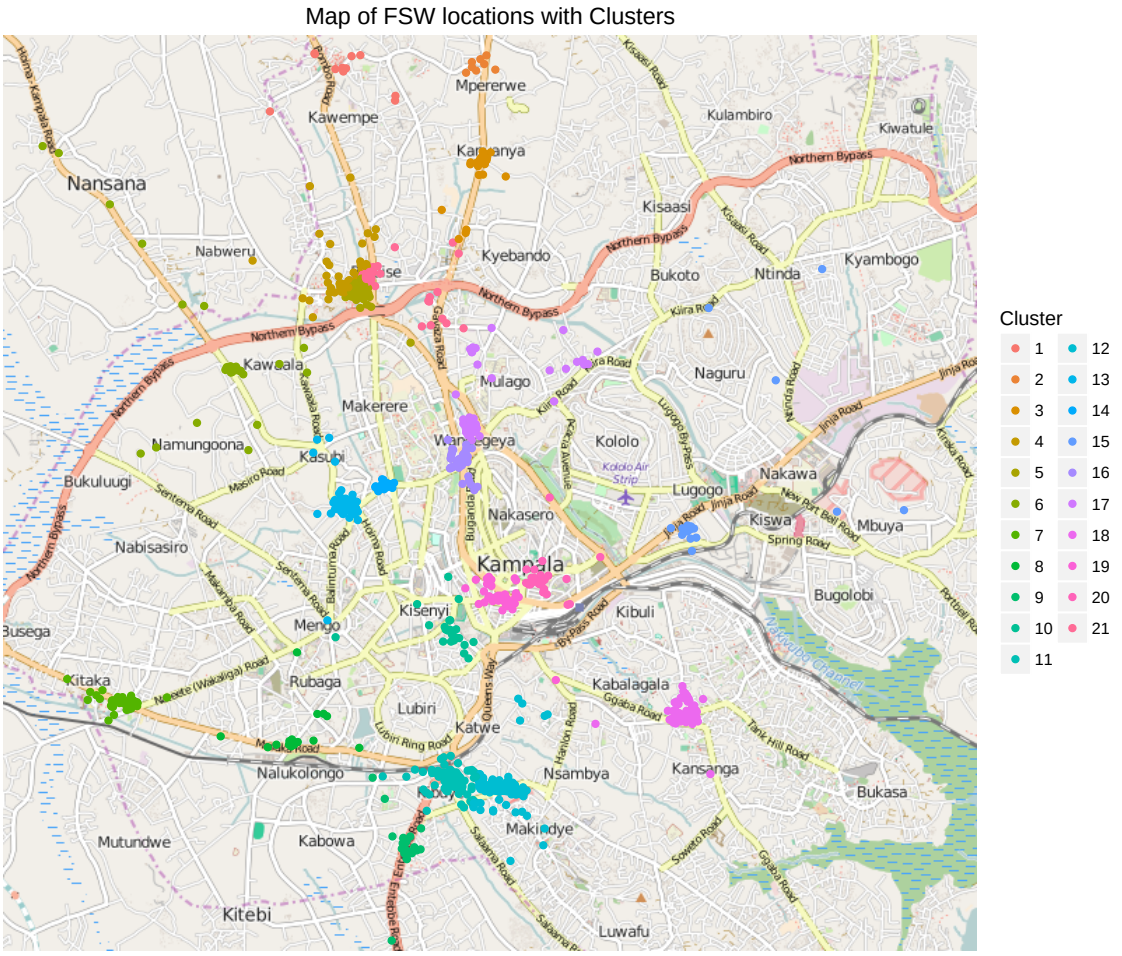



Figure 2.1: A map showing the FSW locations along with their clusters.

| **Cluster** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Number** | 19 | 13 | 28 | 118 | 150 | 31 | 38 | 18 | 32 | 23 | 127 |
| **Cluster** | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | |
| **Number** | 125 | 121 | 21 | 22 | 55 | 48 | 100 | 56 | 27 | 162 | |

Table 2.1: The 21 clusters and the number of FSW in each.

## 2.5 Checking RDS Assumptions Using Permutation Tests

One assumption of RDS is that respondents recruit randomly from within their personal networks. In order to check these, we performed permutation test to see if recruitment was related to other variables, such as age or location (Good 2000). In the permutation tests, we permuted the age or location for each person, making it random, then calculating a certain measure of interest. This was repeated 1000 times, which gave us a distribution of values that we would expect the measure to have if the age or location truly were random. This was then compared to the actual value of the measure from our RDS. A graphical representation of the results are shown in Figure 2.2.

### 2.5.1 Age

We first looked at whether FSW tended to recruit others who were of a similar age. To do this, we found the proportion of recruitments in which the recruiter and recruit were within 3 years of age — that is, out of all recruitments that happened, what proportion involved a difference in age between the recruiter and recruit of less than 3 years — then used a permutation test by making permuting the respondents' ages to see whether it was

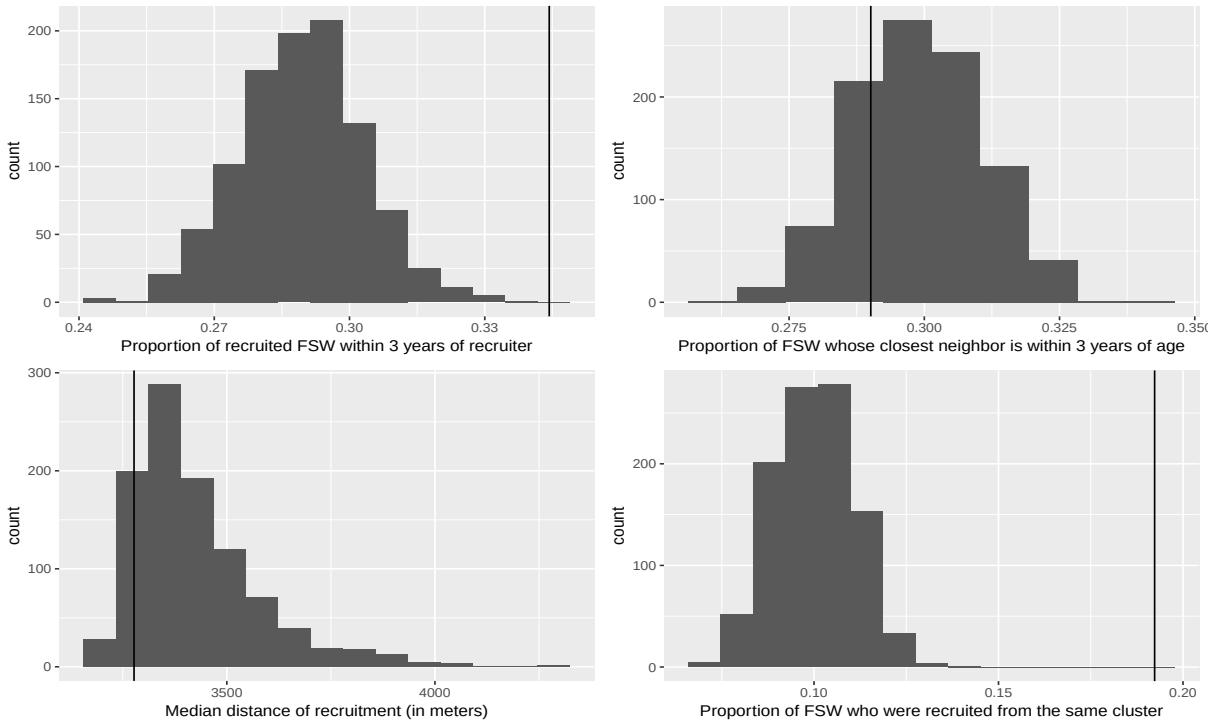



Figure 2.2: The results of four permutation tests. We used 1000 permutations in each case. The vertical line represents the actual value.

significantly greater than would be expected due to random chance. We found that FSW were recruited by someone within 3 years of age 34.4% of the time, and it was significantly greater than would be expected due to random chance. Figure 2.2 shows a histogram of the proportions from the 1000 permutations, as well as a line showing the observed proportion. We also used different age ranges and found that the results were similar (i.e. the proportion was significantly greater than would be expected due to random chance).

We wanted to be sure that the association between age and recruitment wasn't simply due to an effect of location. It is possible that certain areas tend to have younger or older FSW, and we found that FSW tend to recruit others of a similar age simply because they are recruiting whoever is close. In order to check this, we tested the proportion of FSW whose closest neighbor was within 3 years by permuting the ages once again. In our actual sample, 30.8% of FSW had a closest neighbor within 3 years of age, which wasn't significantly higher than random chance, leading us to believe that location isn't driving the association between age and recruitment.

### 2.5.2 Location

We wanted to test whether FSW were more likely to recruit other FSW who were closer in geographical distance to themselves rather than recruiting FSW who were further away. Notably, we did not use their distance along roads, but we believe that straight-line distance is close enough for our purposes. In our sample, on average, FSW recruited others who were 6200 meters away. However, since the distance between recruiter and recruit was highly skewed to the right (the median distance of recruitment was 3300), we decided to use the

median distance. In this permutation test, we permuted the locations of the new recruits, then computed the median distance of recruitment. So, each person recruited the same number of people as they actually recruited. We see in Figure 2.2 that the median distance to recruitment is not significantly different from random chance.

We also tested whether FSW tend to recruit within their own cluster. The results of both distance and cluster tests are shown in Figure 2.2. The median distance of recruitment wasn't significantly lower than might be expected due to random chance. However, the proportion of recruitments within the same cluster was 19.2%, which was significantly higher than we might expect from random chance.

### 2.5.3 Discussion

RDS assumes the participants recruit randomly within their personal networks. These permutations seem to show that this is not the case. In particular, we see that FSW prefer to recruit others of a similar age or in a close location. However, this does not mean that RDS is necessarily problematic. The RDS assumption is that respondents recruit randomly within their own networks — we expect the recruits to be dependent on the recruiter. If the effects of age and distance on recruitment are strong, then we might be worried that the seeds would play a large role in determining the nature of the sample. If the effects are weak, though, as long as the chains are run for a sufficiently long period of time, we should still be able to obtain samples that we can use to generalize to the whole population.

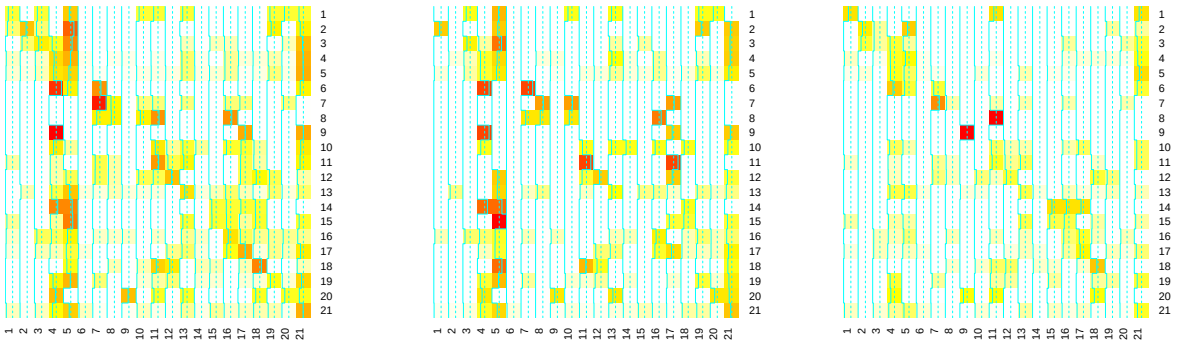
## 2.6 Assessment of Markov Chain Assumptions

In understanding the RDS process, we want to see how recruitment works across clusters. In other words, we want to know what the probabilities of moving from one cluster to another, and whether this changes over time. We first looked at all recruitments together, finding the probability of a transition based on the entire data set. To do this, we found all the recruitments from each cluster, then found the proportion of recruitments that went from that cluster to each of the other clusters. This gave us an estimated transition matrix for recruitment over the entire sample. We then split the data into the first 11 waves and last 14 waves, then compared the two halves to see how recruitment differed over time.

One main assumption of Markov Chain analysis is that the probability of a transition does not depend on any previous states and only depends on the current state (that is, the cluster that the recruiter is in). In addition, we also assume there is no time effect. That is, the probabilities of recruitment are the same throughout the entire sampling process.

We looked at Markov Chains for the transition probability between clusters, then looked at whether this changed over time. We separated the data set in half into the first 11 waves and the last 14 waves, choosing this cutoff point to keep the number of people recruited in each half approximately the same, then compared their respective transition probabilities. This let us see the difference in how recruiting happened between the first half and the second half of the RDS process. Figure 2.3 shows heatmaps of the Markov chain transition matrices.

It seems as though time doesn't have a huge effect on the way FSW recruit. The transition matrix for the first 11 waves is a bit different from that of the last 14, but there aren't any



(a) All Waves (b) First 11 Waves (c) Last 14 Waves

Figure 2.3: Heatmap representations of the transition probability matrix. Red represents high proportions, while white represents low proportions. The rows represent the cluster of the recruiter and the columns represent the cluster of the recruit.

glaring differences in areas being reached. For example, we see that FSW tend to recruit others in cluster 5 more often in the beginning of the RDS, and though this is not the case in later waves, people from are still being recruited.

In addition, the last 14 waves has fewer red spots, showing that in later waves, FSW aren't just recruiting from the same place over and over. In other words, there isn't a tendency to recruit from cluster 5 as there was in the beginning of the RDS. The few red spots that we see are due to a low number of recruitments happening from that particular cluster, resulting in an artificially high "100%."

## 2.7 Network Interactions Between Locations

In order to perform network analysis of geographical location, we treated each of geographical clusters as a single node and formed a directed tie between nodes if there was recruitment in that direction. For example, if someone from node A recruited someone from node B, then a tie was formed from node A to node B. Self ties were allowed, since someone from node A could recruitment someone else from node A. This gave us a directed network with a tie representing recruitment. Figure 2.4 shows a plot of the network with the directed ties.

We expect the nodes with more people to be more central, since they would have more chances to recruit others. In general, this seems to be the case, with larger nodes near the center of the graph containing more ties. However, there are also some clusters with around 100 people on the edges, with relatively few ties. These are blue nodes, representing nodes from the southern part of Kampala.

Figure 2.4 shows the two network plots from the first 11 waves and the last 14 waves. We see that they both have similar nodes at the center (red nodes representing the Central-Northern area of Kampala, with the exception of nodes 6 and 14) and at the edges (blue nodes representing the Central-Southern area of Kampala), and that they both exhibit similar levels of connectedness, with no isolated nodes and at least a few ties per node.

We calculated several measures of centrality (shown in Table 2.2) to see if we could detect bottlenecks and see if there were any interesting characteristics of this social network. Indegree and outdegree refer to the number of ties going out and going into the node, respectively. Betweenness is a measure of how frequently the node is in the shortest path between two other nodes, where a path is the steps on the graph to go from one node to another and

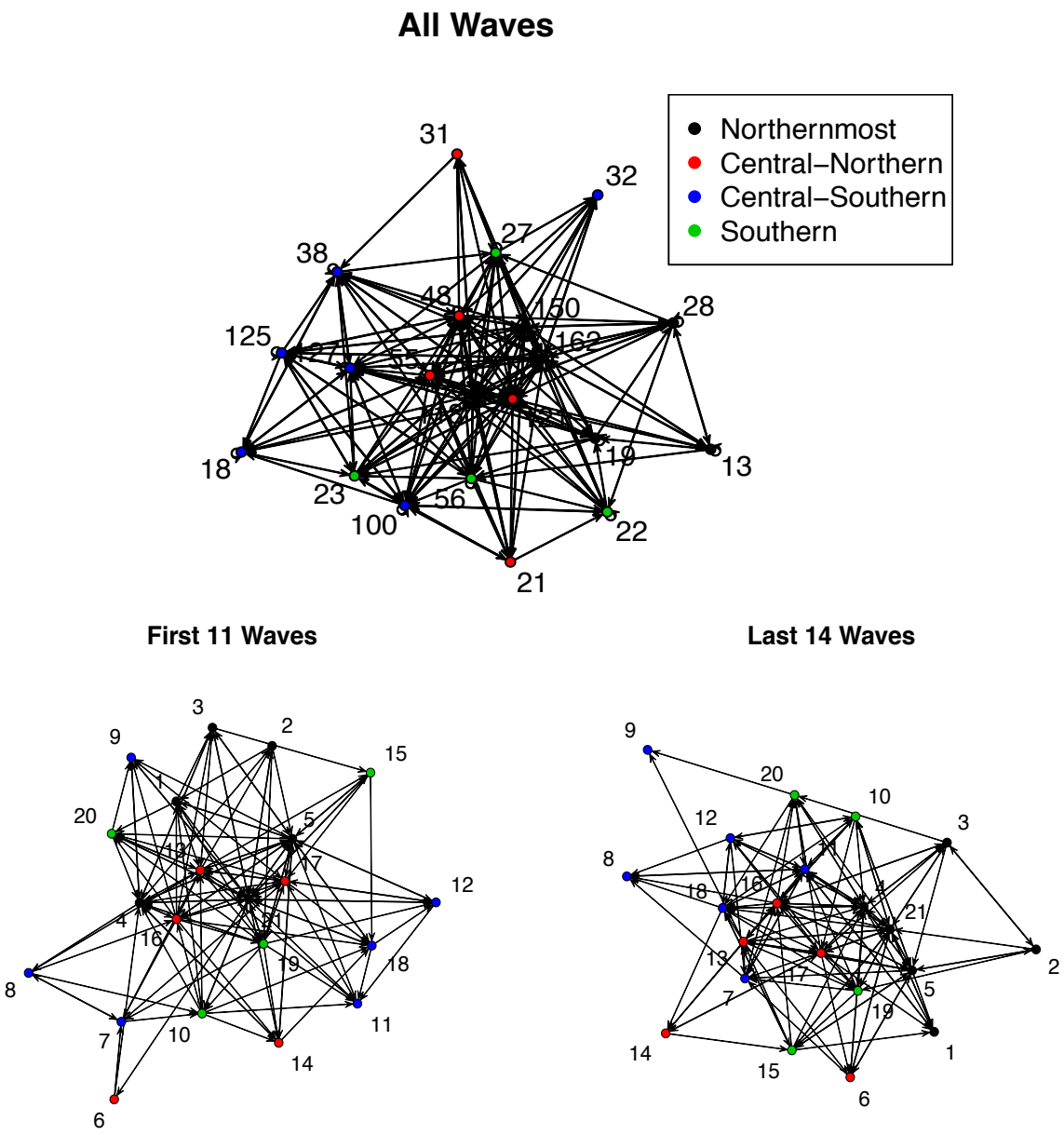



Figure 2.4: Networks of recruitment after aggregating FSW into clusters based on location, with all waves and split into first 11 waves and last 14 waves. Each color represents a different geographical area of the map, split into four categories. For the network with all waves, the numbers shown are the size of the cluster.

a shortest path is a path with the fewest steps from one node to another (Wasserman and Faust 1994). If we let $p_k$ be the $k^{th}$ point on the network, then the betweenness centrality of $p_k$, $c_B(p_k)$ is

$$c_B(p_k) = \sum_{i \neq j \neq k} \frac{g_{ij}(p_k)}{g_{ij}} \tag{2.1}$$

where $g_{ij}(p_k)$ is the number of shortest paths between $p_i$ and $p_j$ containing $p_k$ and $g_{ij}$ is the number of shortest paths between $p_i$ and $p_j$ (Freeman 1977).

The closeness of a node is a measure of shortest paths to each of the other nodes. It is given by

$$c_C(p_k) = \frac{N-1}{\sum_{i \neq k} d(p_i, p_k)}, \tag{2.2}$$

where $d(p_i, p_k)$ is the shortest distance between node $p_i$ and $p_k$, and $N$ is the number of nodes in the network (Freeman 1979). In this way, the closeness centrality measures the inverse of the average distance of a node to every other node in the network.

Notably, we see that none of the betweenness measures are 0, which means that each node is part of a shortest path between two other nodes. In addition, we don't see much variation in the closeness centrality, with the smallest being 0.51. This means that each node is fairly close to the others, with the most isolated node being, on average, two connections away from the other nodes.

### 2.7.1 Modeling Location Network with Exponential Random Graph Models (ERGMs)

We wanted to fit a model for the observed network statistics in order to better understand the properties of the recruitment network. To do this, we used exponential random graph

| **Cluster** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **Indegree** | 8 | 5 | 7 | 15 | 15 | 5 | 9 |
| **Outdegree** | 8 | 5 | 8 | 16 | 19 | 4 | 7 |
| **Betweenness** | 3.21 | 0.62 | 3.15 | 28.34 | 32.21 | 0.27 | 5.25 |
| **Closeness** | 0.62 | 0.56 | 0.62 | 0.83 | 0.95 | 0.56 | 0.61 |
| **Cluster** | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| **Indegree** | 7 | 6 | 9 | 13 | 7 | 12 | 9 |
| **Outdegree** | 4 | 3 | 10 | 9 | 10 | 17 | 6 |
| **Betweenness** | 1.19 | 0.17 | 7.54 | 10.28 | 2.70 | 21.24 | 1.15 |
| **Closeness** | 0.51 | 0.54 | 0.67 | 0.65 | 0.67 | 0.87 | 0.59 |
| **Cluster** | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| **Indegree** | 7 | 11 | 14 | 13 | 9 | 10 | 17 |
| **Outdegree** | 8 | 15 | 13 | 9 | 12 | 6 | 19 |
| **Betweenness** | 2.21 | 18.71 | 17.38 | 8.48 | 6.61 | 2.46 | 42.83 |
| **Closeness** | 0.62 | 0.80 | 0.74 | 0.65 | 0.71 | 0.59 | 0.95 |

Table 2.2: The centrality measures for each of the nodes in the social network analysis.

| | Estimate | Std. Error | P-value |
|---|---|---|---|
| Edges | -0.1589 | 0.1108 | 0.15226 |
| Homophily (Location) | 0.6466 | 0.2416 | 0.00773 |

Table 2.3: The coefficients from the ERGM with a location homophily term.

models (ERGMs) to determine the factors that were associated with the formation of a tie. We can think of ERGMs as assigning probabilities to all possible graphs with certain characteristics represented by network statistics, such as number of edges. ERGMs are part of the exponential family of models, and the coefficients that we obtain from the model can be interpreted similar to a logistic regression (Snijders et al. 2006).

We grouped the 21 nodes into four general areas. This allowed us to look at a measure of homophily to determine whether FSW from a certain cluster tended to recruit others from nearby clusters. We found that the homophily effect was significant, and that FSW tended to recruit more from nearby clusters. Table 2.3 shows the coefficients of the ERGM. Exponentiating the coefficient for the homophily term, we can see that the odds of recruiting from the same cluster was 1.9 times that of recruiting from a different cluster.

## 2.8 Time to Recruitment and Distance

We wanted to also look at how different variables affect the time to recruitment. That is, do new participants tend to come in sooner if they were recruited from areas relatively close to the recruiter?

For each FSW recruited into the study, we found their Euclidean distance to their re-

cruiter as well as the time it took for them to be recruited into the study. We calculated this time to recruitment by taking the number of days between when the recruiter was admitted into the sample and when the new participant was admitted into the sample. We then ran a negative binomial regression to see if distance had an effect on the number of days to recruitment. We found that that the effect of distance on time to recruitment was not significant. Table 2.4 shows the results of the negative binomial regression.

| | Estimate | Std. Error | z value | P-value |
|---|---|---|---|---|
| (Intercept) | 2.184 | 4.194e-02 | 52.074 | <.0001 |
| Distance of Recruitment | 8.104e-06 | 1.320e-05 | 0.614 | 0.539 |

Table 2.4: The results of the Negative Binomial regression.

## 2.9 Relational Event Modeling

Finally, we applied relational event models (REM) to try to model the time to recruitment and see what affects the recruitment process. REM is similar to a survival model, but instead of modeling a time to event, it models the time it takes for a certain relational event to happen (Butts 2008). A relational event is defined as a discrete action generated by an actor directed toward one or more targets. In this case, the relational event is a recruitment into the RDS of a person in a cluster by a person from another cluster. Actions are represented in the form $a = (i, j, k, t)$, where $i$ represents the sender of the action, $j$ represents the receiver of the action, $k$ represents the action type, and $t$ represents the time at which the action takes place. Let $s, r, c, \tau$ be the functions that return the sender, receiver, action type, and

time of an action, respectively (so $\tau(a)$ would return the time of action $a$). Given a set of time-ordered actions $a_1, a_2, \cdots$, let $A_t = \{a_i : \tau(a_i) \leq t\}$ be the set of all actions taken before or at time $t$. In addition, because we want to allow for past history to be able to influence what actions are possible (such as removing a possible receiver), we define the support set $\mathbb{A}(A_t)$ be the set of all sender/receiver/type combinations possible at time $t$ given the history thus far, $A_t$.

We start by describing the survival and hazard function. For an arbitrary random variable $X$ with density function $f$ and cumulative distribution function $F$, the survival function $S$ is defined as $S(x) = 1 - F(x)$. The hazard function $h$ is then defined as $h(x) = f(x)/S(x)$, the probability of a particular action occuring given that it has not occurred already.

The conditional likelihood for the $i$th event ($a_i$) occurring at time $\tau(a_i)$ is equal to the hazard for $a_i$ at time $\tau(a_i)$ multiplied by the survival functions for all possible events over the time interval from $\tau(a_i - 1)$ to $\tau(a_i)$. The joint likelihood for the entire event history at time $t$ is given by the product of all successive conditional likelihoods, along with a factor accounting for the time between the last event and $t$.

Let $M$ refer to the number of events in $A_t$ and $X_a$ refer to a covariate set associated with event $a$. Then, the likelihood for the relational event history is given by

$$p(A_t) = \left[ \prod_{i=1}^{M} \left[ h(\tau(a_i)|s(a_i), r(a_i), c(a_i), X_{a_i}, A_{\tau(a_i)}) \times \right.\right.$$
$$\left.\left. \prod_{a' \in \mathbb{A}(A_{\tau(a_i)})} S(\tau(a_i) - \tau(a_{i-1})|s(a'), r(a'), c(a'), X_{a'}, A_{\tau(a_{i-1})}) \right] \right]$$
$$\left. \times \prod_{a' \in \mathbb{A}(A_t)} S(t - \tau(a_M)|s(a'), r(a'), c(a'), X_{a'}, A_t) \right]. \tag{2.3}$$

This model was different from the ERGM and Markov Chain analysis we performed

earlier because it inherently includes a temporal element. For each of the previous analysis, we looked at the temporal nature of RDS by splitting the data up by waves. REM includes it as part of the model, which might give us deeper insights into how the process works.

One distinction in the way relational event models and Markov chain models work is that REM does not assume stationarity. That is, the Markov chain assumes that the probability of a transition from one cluster to another stays the same. REM, on the other hand, is able to incorporate temporal dependence, which we might deem necessary if we think the RDS behavior changes over time.

We performed the REM by treating each cluster as a single node rather than treating each person as a node. This was due to computational limitations associated with having so many nodes in the model. This does mean that the interpretation of the model should take this grouping into account.

We used covariates for the average age of each cluster and the size of each cluster. Since the `relevent` package for performing relational event modeling does not currently support self loops (that is, an action in which the sender and receiver are the same), we used two different methods. First, we tried it with taking out all self loops and running the model on the remaining events. This means we are essentially modeling the time to a recruitment outside of the cluster. Second, we tried creating a duplicate set of pseudo-nodes, with each node getting a matching pseudo-node and making self loops into ties to its pseudo-node pair. This gave us similar results to the first method, and so we provide just the results of the first.

Table 2.5 shows the coefficients along with their associated p-values. We see that sender

age, receiver age, and receiver size are all significant. The age coefficients are negative, meaning it was more likely for younger clusters to recruit and be recruited. The size coefficient is positive, which means bigger clusters are more likely to be recruited from.

We note that the model using only age and size is quite simple in its application and future work should be done to include more covariates.

| | Estimate | Std. Error | z value | P-value |
|---|---|---|---|---|
| Sender Age | -0.14887841 | 0.02055392 | -7.2433 | <.0001 |
| Receiver Age | -0.22462543 | 0.02079211 | -10.8034 | <.0001 |
| Receiver Size | 0.01394456 | 0.00075415 | 18.4905 | <.0001 |

Table 2.5: The results of the Relational Event Model.

## 2.10 Discussion and Conclusions

In our permutation tests and the ERGMs, we found that FSW tended to recruit nearby FSW. This gave us reason to believe that there might be bottlenecking or that recruitment might get stuck in certain areas. However, looking at the Markov Chains for recruitment and the network analysis, we were able to see that there weren't any regions getting excluded from the sampling process. Looking at the network plots, we were able to see no signs of bottlenecking, and while there were some nodes more central than others, these central nodes were connected to many others. This meant that the central nodes weren't acting as "gatekeepers" and did not inhibit the recruitment from reaching different parts of the city.

One limitation of the Relational Event Modeling was that we had to use clusters instead

of individuals as nodes. This meant that our covariates were aggregate measures rather than characteristics of the individual. This is important because we know that people make decisions based on their own characteristics. For example, a person would recruit preferentially from others of a similar age, not from a group of people who have an average age similar to their own group.

In general, it seems as though we are seeing preferential recruitment, as might be expected, by both age and distance. However, it does not seem as though this is affecting the ability of RDS to reach different areas of the city.

# CHAPTER 3

# Population Size Estimation Methods

Even though using RDS data for population size estimation has not been studied thoroughly, there is a rich literature on size estimation in general. The basis of these methods come from animal abundance. Researchers interested in finding the number of animals in a certain area developed methods to count them. These methods were applied to surveys of human populations, and extensions were developed to relax some of the assumptions.

In this chapter, we review some of the population size estimation methods used so far. First, in Section 3.1, we look at methods that use network information. Then, in Section 3.2, we look at the capture-recapture and multiple list methods first used in animal abundance and extended for other applications.

## 3.1 Methods Based on Network Sampling

Due to the highly connected nature of the populations of interest, there have been various methods that aim to use network properties of these populations to estimate their size. This can involve modeling the sampling process or simply using personal network information, such as how many people a respondent knows. The two we will focus on here are Successive Sampling - Population Size Estimation and Network Scale-up.

### 3.1.1 RDS as Successive Sampling for Population Size Estimation

Currently, one method, called Successive Sampling - Population Size Estimation (SS-PSE) (Handcock et al. 2014), estimates population size using a single RDS sample. While performing the RDS, information on each respondent's degree (i.e. how many people they know who are in the population) was collected, along with the order in which they were included in the study. The RDS process is treated as sampling with probability proportional to size without replacement (PPSWOR, Gile 2011).

Intuitively, if we are sampling with PPSWOR, we would expect the people with higher degrees to be sampled first and the people with lower degrees to be sampled later on. SS-PSE leverages this information to estimate the population size, modeling RDS with a successive sampling approximation, which accounts for the without-replacement nature of RDS and has been found to be effective in estimation of population means (Gile 2011). We discuss this process further in Chapter 4.

One limition of the method is that it only uses one sample. More conventional size estimation methods use multiple samples, so while the RDS process is actually modeled, there seems to be room for improvement.

### 3.1.2 Network Scale-Up

Another method looking to leverage information from networks is the network scale-up method (Bernard et al. 2010, Salganik et al. 2011). The network scale-up method uses information about respondents' personal networks and known population proportions to make size estimates. For example, suppose we want to know how many men who have sex

with other men (MSM) in a particular group of one million people. If a respondent knows 200 people (i.e. their personal network size is 200) and knows 2 people who are MSM, then we can conclude that 1% of the population are MSM. Known populations are used to estimate the size of each respondent's personal network — if a person reports knowing 2 people named "Joe" and there are 10,000 "Joes"s out of 1,000,000 people in the population, then that respondent's personal network size would be 200.

Unfortunately, while network scale-up seems quite reasonable for estimating population size in many cases, it is not promising for hard-to-reach populations. This method requires known characteristics of the population in order to scale up, and this is highly unlikely to be the case when the population of interest is hidden. In addition, one big assumption is that the unobserved and observed members have the same distribution of characteristics, which may not be true due to the hidden nature of the population.

## 3.2 Capture-Recapture and Multiple List Methods

A classic method of estimating the abundance of animals is using capture-recapture. A basic capture-recapture design as it applies to animal abundance can be described as follows:

1. Capture a certain number of animals.
2. Tag them, then release them back into the wild.
3. Perform a second capture (recapture).
4. Count how many of the recaptured animals are tagged.

5. Use the overlap to estimate the abundance.

The "overlap" should give us information about what proportion of population we are sampling with each capture. For example, suppose we capture 10 animals and recapture 10 animals. If the recapture contained 1 tagged animal, we woud estimate the abundance to be greater than if the recapture contained 9 tagged animals.

Figure 3.1: Observed Data for a Simple Capture-Recapture

| **Animal ID** | **List 1** | **List 2** |
|---|---|---|
| Animal 21 | 1 | 1 |
| Animal 12 | 1 | 1 |
| | ⋮ | ⋮ |
| Animal 2 | 1 | 0 |
| Animal 9 | 1 | 0 |
| | ⋮ | ⋮ |
| Animal 3 | 0 | 1 |
| Animal 17 | 0 | 1 |
| | ⋮ | ⋮ |

To get a better idea of what the data looks like, consider Figure 3.1. This is what the raw data might look like for a simple capture-recapture. A "1" means that the individual was caught in that list, while a "0" means the individual was not caught. We observe all combinations of "1"s and "0"s except the one of all "0"s.

We can also express the simple capture-recapture data in the form of a collapsed 2 by 2 table (if we have more than 2 lists, we would instead think of it as a J dimensional contingency table), as in Figure 3.2. Then, we see the data as capture histories as well as the frequencies of the capture histories. That is, we know the capture history of "1, 1", "1, 0", and "0, 1", while we do not observe the capture history "0, 0". This is a condensed version of the complete data, and multiple list methods typically approach the problem in this way, estimating the frequency of capture history of all "0"s.

| | | list 2 | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| list 1 | 1 | $a_{1,1}$ | $a_{1,0}$ | $a_{1,+}$ |
| | 0 | $a_{0,1}$ | $a_{0,0}$ | $a_{0,+}$ |
| | | $a_{+,1}$ | $a_{+,0}$ | $a_{+,+}$ |

Figure 3.2: Contingency table for a simple capture-recapture design

In the following sections, we review some of the methods for estimating the population size using multiple lists.

### 3.2.1 Ratio Estimators

Consider a basic capture-recapture design in which there are two lists: the initial capture and the subsequent recapture. We can view the lists as a two by two contingency table with the cells containing the number of people with each capture history. We know three of the four values in the table, but we do not know the value of $a_{0,0}$, the capture history of not

being caught in either list.

If we assume independence of the two lists and equal capture probabilities for each unit in the population, then the number of people caught in both lists should be $a_{1,1} = \frac{(a_{1,+})(a_{+,1})}{a_{+,+}}$, where $a_{1,+}$ is the number of people caught in the first list and $a_{+,1}$ is the number of people caught in the second list. Since $a_{+,+}$ is unknown and what we want to estimate, we can manipulate the equation to get what Paz-Bailey et al. (2011) and Berchenko and Frost (2011) refer to as the naive estimator,

$$\hat{N}_{\text{naive}} = \frac{(a_{1,0} + a_{1,1})(a_{0,1} + a_{1,1})}{a_{1,1}}, \tag{3.1}$$

where $a_{1,0}$ is the number of people who were captured in only the first sample and $a_{0,1}$ is the number of people who were captured in only the second sample. This is also called the Lincoln-Petersen estimator.

An estimator of the variance of $\hat{N}_{\text{naive}}$ is given by Paz-Bailey et al. (2011)

$$Var(\hat{N}_{\text{naive}}) = \frac{(a_{1,0} + a_{1,1})(a_{0,1} + a_{1,1})a_{1,0}a_{0,1}}{(a_{1,1})^3}. \tag{3.2}$$

This estimator assumes, among other things, that each individual has the same probability of being captured. However, because of the nature of RDS surveys, we know that this is not true. Paz-Bailey et al. (2011) attempted to account for this heterogeneity of capture probabilities by using an adjusted estimator, as described in Berchenko and Frost (2011),

$$\hat{N}_{\text{adj}} = \frac{\left(\sum_k I_k^{1,0} + \sum_k I_k^{1,1}\right)\left(\sum_k \frac{I_k^{0,1}}{\pi_k} + \sum_k \frac{I_k^{1,1}}{\pi_k}\right)}{\sum_k \frac{I_k^{1,1}}{\pi_k}}, \tag{3.3}$$

where $I_k^{i,j}$ is the indicator for whether individual k was sampled and $\pi_k$ is the sampling probability of the individual. In applications, these sampling probabilites must be estimated and are done according to personal network size (i.e. the number of people in the population that the respondent knows).

### 3.2.2 Chapman Estimator

Chapman (1951) proposed an alternative to the Lincoln-Petersen estimator, noting that it could be biased for small sample sizes ($n \ll N$). For example, consider the case in which there is no overlap between the two lists, $a_{1,1} = 0$. Then, the Lincoln-Petersen estimator becomes $\hat{N}_{naive} = \infty$, which is clearly undesirable. His estimator is given by

$$\hat{N}_{Chapman} = \frac{(a_{1,0} + a_{1,1} + 1)(a_{0,1} + a_{1,1} + 1)}{a_{1,1} + 1} - 1. \tag{3.4}$$

The Chapman estimator gives a finite value in the extreme case of no overlap because we are no longer dividing by 0 if $a_{1,1} = 0$. Generally, the same variance estimate is used for both the Lincoln-Petersen estimator and the Chapman estimator.

### 3.2.3 Bayesian Implementation of Hypergeometric Capture-Recapture Model

We can also use a Bayesian framework to develop a model based on the Hypergeometric distribution for capture-recapture data. Consider again the data as presented in Table 3.2. Let $A_{+,+}, A_{1,0}, A_{0,1}, A_{1,1}$ be the variables for the population and observed data such that their realizations are $a_{+,+}, a_{1,0}, a_{0,1}, a_{1,1}$, respectively. Using the Hypergeometric distribution,

$$L(A_{+,+} = a_{+,+} | A_{1,0} = a_{1,0}, A_{0,1} = a_{0,1}, A_{1,1} = a_{1,1})$$

$$\propto p(A_{1,1} = a_{1,1} | A_{+,+} = a_{+,+}, A_{1,0} = a_{1,0}, A_{0,1} = a_{0,1})$$
$$= \frac{\binom{a_{1,0}+a_{1,1}}{a_{1,1}}\binom{a_{+,+}-(a_{1,0}+a_{1,1})}{a_{0,1}}}{\binom{a_{+,+}}{a_{1,1}+a_{0,1}}}. \tag{3.5}$$

The posterior for the population size would then be

$$p(A_{+,+} = a_{+,+} | A_{1,0} = a_{1,0}, A_{0,1} = a_{0,1}, A_{1,1} = a_{1,1})$$
$$\propto \pi(a_{+,+}) \cdot \frac{\binom{a_{1,0}+a_{1,1}}{a_{1,1}}\binom{a_{+,+}-(a_{1,0}+a_{1,1})}{a_{0,1}}}{\binom{a_{+,+}}{a_{1,1}+a_{0,1}}}, \tag{3.6}$$

where $\pi(a_{+,+})$ is the population size prior. This formulation of the Hypergeometric Capture-Recapture model allows us to assign a prior to the population size, which is useful for our applications in Chapter 5, where we will compare a new method to existing methods. We use the same priors as in the new method, and further discussion of population size priors is given in Section 4.3.

### 3.2.4 Log-Linear Models

Log-linear models are one of the basic ways in which multiple lists are used to estimate population size. We consider the data as a contingency table of capture histories. Then, we need to estimate the number of people in the cell with capture history $0, ..., 0$, the case of being caught in none of the lists.

We looked at several different types of log-linear models in our analysis. The first was a simple model that assumed equal capture probabilities for each unit and for each list. Let $p$ be the probability of any individual being captured, $J$ be the number of lists, and $\boldsymbol{\omega}$ the $J$-dimensional binary vector representing the capture history of an individual. Since we are assuming equal capture probabilities for each unit and list, we can express the probability

of having a capture history $\boldsymbol{\omega}$ as

$$P(\boldsymbol{\omega}) = (1-p)^{J-\sum \omega_j} p^{\sum \omega_j}, \tag{3.7}$$

where $\sum \omega_j$ is the number of times a unit is caught. The expected number of units with capture history $\boldsymbol{\omega}$ is given by

$$\mu_\omega = N(1-p)^{J-\sum \omega_j} p^{\sum \omega_j},$$

where $N$ is the population size. We can express this in the form of a log-linear model

$$\mu_\omega = \exp\left(\log(N(1-p)^J) + \sum w_j \log(\tfrac{p}{1-p})\right).$$

If we let $\gamma = \log(N(1-p)^J)$ and $\beta = \log(\frac{p}{1-p})$, we can see the form of the log-linear model $E(\boldsymbol{Y}) = \exp(X\boldsymbol{\beta})$, where $\mathbf{Y}$ is a $(2^J-1)$-dimensional vector of the observed frequencies, $X$ is the $(2^J-1)$ by 2 design matrix with a first column of all 1's and a second column of $\sum \omega_j$, and $\boldsymbol{\beta}$ is equal to $(\gamma, \beta)^t$. The estimate of the population size can be estimated with $\hat{N} = n + \hat{\gamma}$, since $\exp(\gamma) = \exp(\log(N(1-p)^J)) = N(1-p)^J = N * P(\omega_0) = \mu_0$, where $\omega_0$ is the capture history of never being captured and $\mu_0$ is the expected number of individuals that were never captured.

This is the specification of the $M_0$ model as described by Rivest and Baillargeon (2007), the most basic of the log-linear models. Similar to the naive estimator in section 3.2.1, this model assumes independence of all lists and equal capture probabilities. Extensions of the model allow for heterogeneity in capture probabilities (h), difference in lists (t), and behavioral effects (b).

Consider the model $M_t$, the model in which we let the lists have different capture probabilities. In $M_0$, we had one coefficient for the number of times a unit was captured, $\sum \omega_j$. In $M_t$, we have a separate coefficient for each list, so our model looks like

$$\log \mu_\omega = \gamma + \sum(\omega_j \beta_j).$$

In this way, we have a model that accounts for the difference in lists. Further, we can adjust these models in order to take into account the heterogeneity to give us the $M_{th}$ models. There are several ways this can be done. Darroch et al. (1993) adds a column of $(\sum \omega_j)^2/2$ to the design matrix while assuming the logits of the individual capture probabilities to be distributed according to a mixed normal distribution. For a Poisson2 model as described by Rivest and Baillargeon (2007), a column of $2^{\sum \omega_j} - 1$ is added to the design matrix and the logit of individual capture probabilities is assumed to be mixed Poisson.

### 3.2.5 Rasch Model

The Rasch model (Fienberg et al. 1999) was developed for psychometric tests but can be applied to multiple list problems as well. In the Rasch model, the log odds of getting an answer correct is modeled with effects by respondent and by item difficulty. That is,

$$\text{logit}(p_{ij}) = \theta_i - \beta_j, \tag{3.8}$$

where $p_{ij}$ is the probability that the $i^{th}$ person gets the $j^{th}$ question correct, $\theta_i$ is the effect of the $i^{th}$ respondent, and $\beta_j$ is the effect of the $j^{th}$ question.

This type of model can also be applied to multiple list data by thinking of the respondent

effect as the catchability and item effect as the list effect. The model specification would look something like:

$$\left.\begin{aligned} X_{ij} \sim \text{Bernoulli}(p_j|\theta_i), \quad \log\left\{\frac{P_j(\theta_i)}{1-P_j(\theta_i)}\right\} = \theta_i + \beta_j, \quad & i = 1,\ldots,N, \\ & j = 1,\ldots,J, \\ \theta_i \sim F_\Theta(\theta_i), \quad & i = 1,\ldots,N, \\ \beta_j \sim G_\beta(\beta_j), \quad & j = 1,\ldots,J. \end{aligned}\right\} \tag{3.9}$$

Here, $X_{ij} = 1$ if the $i^{th}$ individual was caught in the $j^{th}$ list, and 0 otherwise. The $\theta_i$ represent the catchability parameter, which allow for heterogeneity in capture probability, and the $\beta_j$ represent the list parameter, allowing for a difference in lists. This model is more explicitly explained in Fienberg et al. (1999).

### 3.2.6 Grade of Membership Model

The Grade of Membership (GoM) model proposed by Manrique-Vallier and Fienberg (2008) tries to account for heterogeneity within individuals in the population using latent classes. We assume that there are $K$ latent classes and allow for partial membership of each unit in these latent classes. We further assume that every unit with the same membership has the same capture probability. That is,

$$P(X_{ij} = 1|i^{th} \text{ individual in the } k^{th}\text{class}) = \lambda_{jk}, \tag{3.10}$$

where $X_{ij}$ is an indicator variable for whether the $i^{th}$ individual was caught in the $j^{th}$ list, and $\lambda_{jk}$ is the capture probability in list $j$ for an individual in class $k$. This is used to model the heterogeneity in capture probabilities. Further, Manrique-Vallier and Fienberg (2008)

assume that the lists are conditionally independent given the membership vectors, reasoning that the membership in the latent classes explains all of the dependence between the lists.

Manrique-Vallier and Fienberg (2008) give each unit a membership vector, $g_i = (g_{i1}, \cdots, g_{iK})$, where $g_{ik} > 0, \sum_{k=1}^{K} g_{ik} = 1$, representing how much a member of each class that unit is. Thus, keeping in mind that the $J$ lists are conditionally independent, we get

$$p(x_i|g_i) = \prod_{j=1}^{J} \sum_{k=1}^{K} g_{ijk} \lambda_{jk}^{x_{ij}} (1 - \lambda_{jk})^{x_{ij}}. \tag{3.11}$$

Manrique-Vallier and Fienberg (2008) then use an augmented data representation to get

$$p(x, z|\lambda, g) = \prod_{i=1}^{N} \prod_{j=1}^{J} \sum_{k=1}^{K} \left( g_{ijk} \lambda_{jk}^{x_{ij}} (1 - \lambda_{jk})^{x_{ij}} \right)^{z_{ijk}}, \tag{3.12}$$

where $z_i = (z_{i1}, \cdots, z_{iJ}) \in Z = \{1, \cdots, K\}^J$ and $z_{ijk} = I(z_{ij} = k)$. Manrique-Vallier and Fienberg (2008) uses this augmented data likelihood to derive the full joint posterior:

$$p(N, \alpha, \lambda, g|x, z) \propto \binom{N}{n} p(N, \alpha, \lambda, g) \prod_{i=1}^{N} \prod_{j=1}^{J} \sum_{k=1}^{K} \left( g_{ijk} \lambda_{jk}^{x_{ij}} (1 - \lambda_{jk})^{x_{ij}} \right)^{z_{ijk}} I_{\{n+1, n+2, \cdots\}}(N), \tag{3.13}$$

where $p(N, \alpha, \lambda, g)$ is the joint prior distribution of $(N, \alpha, \lambda, g)$ and $I_{\{n+1, n+2, \cdots\}}(N)$ limits possible population sizes to be greater than the sample size. A full discussion of the MCMC for the posterior estimation is given in Manrique-Vallier and Fienberg (2008).

Manrique-Vallier and Fienberg (2008) applied this method to various multiple list data sets, such as killings in Chungui and diabetes, and found the method to work well compared to other multiple list methods such as log-linear models and the Rasch model. However, they have not yet applied the method to RDS data sets, nor have they compared it to network sampling based models.

### 3.2.7 Nonparametric Latent Class Model (NPLCM)

Manrique-Vallier (2016) proposes an alternative to GoM called the Nonparametric Latent Class Model (NPLCM). Similar to GoM, NPLCM uses latent variables to account for the heterogeneity in capture probabilties. As in equation 3.10, the probability that individual $i$ is captured in list $j$ is based on the latent class that the individual is in. However, this still has the model selection problem of requiring a choice of the number of latent classes.

Manrique-Vallier (2016) follows Dunson and Xing (2009) and uses a Bayesian nonparametric extension to the basic latent class model. Dunson and Xing (2009) propose using an infinite number of latent classes and using a prior specification that induces sparsity so that most of the probability mass is concentrated on a small (finite) subset, which effectively acts as the model selection process. This gives us a Dirichlet process mixture of product-Bernoulli distributions. Manrique-Vallier (2016) describes the generative process.

$$
\begin{aligned}
x_j|z &\overset{indep}{\sim} \text{Bernoulli}(\lambda_{jk}), && j = 1, \cdots, J \qquad (3.14)\\
z &\sim \text{Discrete}(\{1, 2, \cdots\}, (\pi_1, \pi_2, \cdots)) \\
\lambda_{jk} &\overset{iid}{\sim} \text{Beta}(1, 1) && j = 1, \cdots, J \text{ and } k = 1, 2, \cdots \\
(\pi_1, \pi_2, \cdots) &\sim \text{SB}(\alpha) \\
\alpha &\sim \text{Gamma}(a, b)
\end{aligned}
$$

where $x_j$ is whether the unit was caught in the $j^{th}$ list, $z$ is the latent class, $\lambda_{jk}$ is the probability of capture in the $j^{th}$ list for a unit in the $k^{th}$ class, $\pi_k$ is the probability of being the $k^{th}$ class, and $\text{SB}(\alpha)$ is the stick-breaking process with parameter $\alpha > 0$. The

stick-breaking procedure is given by

$$\pi_k = V_k \prod_{i=1}^{k-1}(1 - V_i) \qquad\qquad k = 1, 2, \cdots, \tag{3.15}$$

where $V_k$ are independent random variables with distribution Beta(1, $\alpha$). In practice, a large value, denoted by $K^*$, is chosen as the maximum number of latent classes, so $V_1, \cdots, V_{K^*-1}$ are independent random variables with distribution Beta(1, $\alpha$) and $V_{K^*} = 1$.

Manrique-Vallier (2016) describes the joint model for the observable sample as

$$p(\boldsymbol{\chi}|\boldsymbol{\lambda}, \boldsymbol{\pi}, N) \propto \binom{N}{n} \left[\sum_{k=1}^{K^*} \pi_k \prod_{j=1}^{J}(1 - \lambda_{jk})\right]^{N-n} \times \prod_{i=1}^{n}\sum_{k=1}^{K^*} \lambda_{jk}^{x_{ij}}(1 - \lambda_{jk}^{x_{ij}})^{1-x_{ij}}, \tag{3.16}$$

where $\boldsymbol{\chi} = (\mathbf{x_1}, \cdots, \mathbf{x_n})$, the complete observed data with $\mathbf{x_i} = (x_{i1}, \cdots, x_{iJ}) \in \{0, 1\}^J$ representing the capture history of the $i^{th}$ unit, $\boldsymbol{\lambda} = (\lambda_{jk})$ with $\lambda_{jk} \in (0, 1)$, and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ with $\sum_{k=1}^{K} \pi_k = 1$.

One important benefit of the NPLCM over the GoM model is that it requires only sampling from common distributions, which makes implementation relatively quick and easy. In addition, it is highly scalable in the sample sizes and number of lists. In our applications in Chapter 5, we use an `R` implementation of the method provided in the `LCMCR` package (Manrique-Vallier 2016).

## 3.3 Discussion of Current Methods

In this chapter, we have discussed two general approaches to the problem of population size estimation. The first approach involves using the highly connected nature of the target population, leveraging network information to estimate the size, while the second approach

involves a capture-recapture and multiple list scheme, finding the estimate using overlap information instead.

Each of these approaches have their strengths and weaknesses. Network-based methods account for the nature of the data. SS-PSE in particular specifically models the RDS process by treating it as sampling with PPWSOR. However, it also only uses one sample, and thus is unable to take advantage of any overlap information that could be obtained from having multiple lists.

On the other hand, the most basic capture-recapture models make many assumptions about the sampling process, such as equal probabilities of capture in individuals, equal probabilities of capture in lsits, independence in lists, and independence in individuals. More advanced methods (as seen in models such as $M_{th}$ and variants, the Grade of Membership model, and NPLCM) seek to relax these assumptions, but are not based on network sampling methods, such as RDS. The adjusted estimator described by Berchenko and Frost (2011) accounts for degree from using RDS in a capture-recapture setting, but there is currently no model-based method for estimating the population size designed to use multiple RDS samples. Due to the complex nature of the RDS process, there is a need for methods that are designed with RDS in mind and can provide better estimates when using RDS data.

In addition, we must consider the trade-off between complexity and accuracy. Currently, simple ratio estimators are still used in the field due to the relative simplicity in calculating and interpreting the results (UNAIDS and Organization 2010). However, we have reason to believe these methods are too simple and may not give good esimates (Manrique-Vallier and Fienberg 2008).

Our goal in the next chapter is to combine the strengths of SS-PSE with that of the multiple list methods and develop a novel method to estimate the population size by modeling the RDS process while using overlap information.

# CHAPTER 4

# Improvements to Current Population Size Methods

One of the biggest benefits of using RDS data is the ability to collect multiple samples on a hidden population relatively easily. In a review of RDS studies related to HIV surveillance, Malekinejad et al. (2008) found that RDS studies took, on average, 9 weeks to complete. Because of this, we want to adapt capture-recapture methods to work with RDS data so that we can use as much information as possible to obtain the most accurate results possible. Currently, capture-recapture and other multiple list methods have not been tailored for RDS data. Paz-Bailey et al. (2011) has used RDS in the recapture stage, but there have been no published studies using respondent driven sampling for each stage of multiple captures in population size estimation. In addition, Paz-Bailey et al. (2011) only used an adjusted ratio estimator. We aim to take a model-based approach, which will not only give us better estimates, but also better measures of variance.

We aim to improve on current methods by utilizing the degrees of each respondent and the order in which they were sampled to develop a method specifically for multiple RDS lists. In this chapter, we will look at two novel methods for estimating the size of a hidden population in this way. In 4.2, we explore using an extension of the hypergeometric distribution to model the second list, building it up starting from the basic hypergeometric model as mentioned

in Section 3.2.3. In 4.3, we introduce Capture-Recapture SS-PSE, a novel model that treats both lists as independent samples with PPSWOR.

## 4.1 Notation

Here, we will develop some of the notation that will be used in the following sections. In both the Multivariate Wallenius' Non-central Hypergeometric model (Section 4.2) and the Capture-Recapture SS-PSE model (Section 4.3), we assume the existence of a population with an associated unit size. This can generally be anything about the individual that affects their catchability, but for our purposes in RDS surveys, the unit sizes will be the personal network size, or degree. For both methods, our observed data consists of two lists, or captures, in which the personal network size of each observation is recorded in order of observation. That is, we are tracking both the unit size (degree) as well as the sequential order of the when units were including into the study. We note that in practice, this is done in the order that respondents come into the research center to be included in the study, regardless of which wave they are a part of. In addition, the second list contains information about whether the unit was captured in the first list. Notably, in these methods, we do not assume that we have a matching between the first and second list for the units captured in both. Instead, we only know that the unit was part of the first sample. This is similar to the information collected in previous RDS studies that have tried to use capture-recapture methodologies, as the researchers would ask whether the respondent had previously received a unique keychain (Paz-Bailey et al. 2011).

Let the total population size be $N$, with each individual unit in the population indexed

$1, \cdots, N$. That is, the members of the population are given a fixed, arbitrary order. Let $U_1, \cdots, U_N$ be the random variables representing the unit sizes of these units. Our observed data consists of the unit sizes, in sampling order, of the two lists, as well as which units in the second list were also caught in the first list. For our purposes, we do not use recruiter information (that is, who the recruiter was for each new recruit). Let $n'$, $n''$ be the sample size of the first and second list, respectively, and let $n$ be the total number of unique people observed in the samples. Let $n_0$ be the number of units in the second list that were already observed in the first list. Note that $n \leq n' + n''$ since units may be caught in both lists, $n = n' + n'' - n_0$ in general, and $n = n' + n''$ only if there is no overlap between the two lists.

Let $G' = (G'_1, \cdots, G'_{n'})$ be the random indices of the sequentially sampled units in sampling order with realization $g' = (g'_1, \cdots, g'_{n'})$ in the first list, and let $G'' = (G''_1, \cdots, G''_{n''})$ be the random indices of the sequentially sampled units in sampling order with realization $g'' = (g''_1, \cdots, g''_{n''})$ in the second list. Let $g^{unobs} = \{g_1^{unobs}, \cdots, g_{N-n}^{unobs}\}$ be the indices of all unobserved units in their population order. Note that since there were $n$ unique units observed, there are $N - n$ unobserved units.

Let $U'_{obs} = \{U_{g'_1}, \cdots, U_{g'_{n'}}\}$ be the unit sizes of the first list in sampling order, and $U''_{obs} = \{U_{g''_1}, \cdots, U_{g''_{n''}}\}$ be the unit sizes of the second list in sampling order, with realizations $u'_{obs} = \{u_{g'_1}, \cdots, u_{g'_{n'}}\}$ and $u''_{obs} = \{u_{g''_1}, \cdots, u_{g''_{n''}}\}$, respectively. Let $Y''_{obs} = \{Y''_1, Y''_2, \cdots, Y''_{n''}\}$ with realizations $y''_{obs} = \{y''_1, y''_2, \cdots, y''_{n''}\}$ represent the recapture information. In other words, $y''_i = 1$ if the $i^{th}$ unit was observed in the first list (i.e. is a recapture as opposed to a completely new observation) and 0 otherwise. Let $U_{unobs} = \{U_{g_1^{unobs}}, \cdots, U_{g_{N-n}^{unobs}}\}$ be the random unit sizes of the unobserved units in their population order, and $u_{unobs} = \{u_{g_1^{unobs}}, \cdots, u_{g_{N-n}^{unobs}}\}$

their realizations. Let $U = (U_1, \cdots, U_N)$ and $u = (u_1, \cdots, u_N)$. Note that $U$ and $U_{unobs}$ are ordered according to their unknown, arbitrary, fixed population labels, while the elements of $U'_{obs}$ and $U''_{obs}$ are in their respective order of observation. The random vector $G'$ maps between the fixed population ordering and the first list, while $G''$ maps between the fixed population ordering and the second list.

### 4.1.1 Illustrative Example for Notation

Suppose we have a population of 10 units, represented by $\{A, B, C, D, E, F, G, H, I, J\}$. The population has an arbitrary ordering, which we will assign in alphabetical order. So, in this population, $A$ is the first unit, $B$ is the second unit, and so on. Further, the units in this population have unit sizes according to the following table:

| **Population** | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| **(Arbitrary) Population Order** ($i$) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Unit Size** ($u_i$) | 6 | 4 | 2 | 3 | 4 | 1 | 4 | 2 | 3 | 9 |

So, in other words, $u_1 = 6, u_2 = 4, u_3 = 2$, and so on.

Suppose we take two samples from this population, with both sample sizes equal to 3. So, $n' = n'' = 3$. Suppose the first sample taken, in order, was $\{A, F, C\}$, and the second sample taken, in order, was $\{F, B, I\}$. Then, $g' = \{1, 6, 3\}$ since the first, sixth, and third elements of the population were sampled in that order. Similarly, $g'' = \{6, 2, 9\}$.

Our observed data in this case are the unit sizes of these units. So, $u'_{obs} = \{6, 1, 2\}$, and $u''_{obs} = \{1, 4, 3\}$. Since only the first observation in the second list had been previously captured, $y''_{obs} = \{1, 0, 0\}$.

## 4.2 Multivariate Wallenius' Non-central Hypergeometric (MWNCH) Model

In this section, we will develop a novel model using the Wallenius' Non-central Hypergeometric distribution as a natural extension of the capture-recapture methods which use the hypergeometric distribution.

### 4.2.1 Extending Capture-Recapture

The idea behind modeling RDS is that we want a mechanism that can, as closely as possible, approximate the network sampling process so that more "popular" people are more likely to be sampled. However, we still also want to keep the framework of the capture-recapture models, using the information in the overlap to estimate the population size. To that end, we will build on the basic hypergeometric distribution by trying to account for the sampling process of RDS.

Intuitively, we know that the units sampled in the first list are likely the more popular ones, since they were the ones sampled first. Therefore, when we sample the second list, we expect, on average, that the probability of sampling the people in the first list to be higher than the probability of sampling those who weren't in the first list. Even though each individual unit may have a separate probability of being included, we will model this difference in probability by grouping the units sampled in the first list and giving them a weight. That is, the units who were included in list 1 have a weight $\omega_1$, while the unobserved units from list 1 have weight $\omega_2$. Let $\omega = \omega_1/\omega_2$.

We can use the Wallenius' noncentral hypergeometric distribution proposed by Wallenius (1963). This distribution is a generalization of the hypergeometric distribution by allowing the items being sampled to have different weights. We will describe this distribution as a process of picking balls from an urn with red and blue balls, with different colors corresponding to a different probability of being selected. If $m_1$ is the number of red balls, $m_2$ is the number of blue balls, $\omega_1$ is the weight of the red balls, $\omega_2$ is the weight of the blue balls, and $\omega = \omega_1/\omega_2$, then, according to Wallenius' noncentral hypergeometric distribution, the probability of picking $x$ red balls out of $n$ draws is given by

$$p(x|m_1, m_2, N, \omega) = \binom{m_1}{x}\binom{N-m_2}{n-x}\int_0^1 (1-t^{\omega/D})^x(1-t^{1/D})^{n-x}dt, \tag{4.1}$$

where $D = \omega(m_1 - x) + (N - m_1 - (m_2 - x))$. Then, our likelihood is given by

$$\begin{aligned} L(N|n_0, n', n'', \omega) &\propto P(n_0|n', n'', N, \omega) \cdot \pi(N) \\ &= \pi(N) \cdot \binom{n'}{n_0}\binom{N-n'}{n''-n_0}\int_0^1 (1-t^{\omega/D})^{n_0}(1-t^{1/D})^{n-n_0}dt, \end{aligned}$$

where $D = \omega(n' - n_0) + (N - n' - (n'' - n_0))$.

This gives us a model similar to that of the hypergeometric models discussed earlier. However, it is unclear how exactly we should determine these weights. In addition, we must use one weight to represent the capture probability of every single unit in each group. We know this to be untrue, as the degree of each unit within the captured and uncaptured groups can vary. Along those lines, just as in the capture-recapture methods, we only use information about the sample sizes and overlap. However, we want to use all available information, including the order in which the samples were taken and the degree (or size) of the units.

### 4.2.2 Multivariate Wallenius' Noncentral Hypergeometric

We can expand on the above method by using the multivariate form as described by Chesson (1976). In the previous section, we used the univariate Wallenius' noncentral hypergeometric distribution. In the univariate case, we had only two possibilities (that is, to use the example typically used, two colors of balls in the urn). However, we can extend this to include any number of colors of balls, each with a different weight, using the multivariate version of the distribution. This gives us the ability to include more fine-tuned differences in sampling weights instead of assigning the same weight to each unit that was observed (or unobserved).

Suppose that the units in the population have two types of capture probability: high capture probability and low capture probability. We can represent this using two different weights, $\omega_1$ and $\omega_2$. These units were then either captured in the first list or not captured in the first list. This gives us four distinct groups: high weight and captured in first list, low weight and captured in first list, high weight and not captured in first list, and low weight and not captured in first list. So, in other words, we have $\boldsymbol{\omega} = \{\omega_1^+, \omega_2^+, \omega_1^-, \omega_2^-\}$, where the superscript represents whether the unit was observed in the first list or not and the subscript represents the categories of unit size. In addition, we have two categories of sizes, with $\omega_1^+ = \omega_1^-$ and $\omega_2^+ = \omega_2^-$. We will let $\omega_1 = \omega_1^+ = \omega_1^-$ and $\omega_2 = \omega_2^+ = \omega_2^-$.

Then, we can use the multivariate Wallenius' noncentral hypergeometric distribution. Let $\mathbf{N} = \{N_1^+, N_2^+, N_1^-, N_2^-\}$ be the total numbers of people in each of the four groups described above. That is, $N_1^+$ is the number of people with sampling weight $\omega_1$ who were observed in the first list, $N_2^+$ is the number of people with weight $\omega_2$ who were observed in the first list, $N_1^-$ is the number of people with weight $\omega_1$ who were not observed in the

first list, and $N_2^-$ is the number of people with weight $\omega_2$ who were not observed in the first list. So, $N_1^+ + N_2^+ + N_1^- + N_2^- = N$ and $N_1^+ + N_2^+ = n'$. Let $\mathbf{n} = \{n_1^+, n_2^+, n_1^-, n_2^-\}$ be the number in each group observed in the second list. So, $n_1^+ + n_2^+ + n_1^- + n_2^- = n''$. Let $\boldsymbol{\omega} = \{\omega_1, \omega_2, \omega_1, \omega_2\}$ Then, following Fog (2008),

$$P(\mathbf{n}|\mathbf{N}) = \left(\prod_{i=1}^{2} \binom{N_i^+}{n_i^+}\binom{N_i^-}{n_i^-}\right) \int_0^1 \prod_{i=1}^{2} (1 - t^{\omega_i/D})^{n_i^+} (1 - t^{\omega_i/D})^{n_i^-} \, dt, \tag{4.2}$$

where $D = \boldsymbol{\omega} \cdot (\mathbf{N} - \mathbf{n}) = \sum_{i=1}^{2} \omega_i (N_i^+ - n_i^+) + \sum_{i=1}^{2} \omega_i (N_i^- - n_i^-)$.

If we want to include more than just "small" and "large" as possible unit sizes, we can increase the number of unit size categories. That is, to generalize to any number of unit size categories, let $c$ be the number of unit size categories we want. Since there are two lists, this means that there will be $2c$ combinations of "weight" and "observed in first list." Let $\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_c, \omega_1, \cdots, \omega_c\}$ and $\mathbf{N} = \{N_1^+, \cdots, N_c^+, N_1^-, \cdots, N_c^-\}$. Let $\mathbf{n} = \{n_1^+, \cdots, n_c^+, n_1^-, \cdots, n_c^-\}$. Then,

$$P(\mathbf{n}|\mathbf{N}) = \left(\prod_{i=1}^{c} \binom{N_i^+}{n_i^+}\binom{N_i^-}{n_i^-}\right) \int_0^1 \prod_{i=1}^{c} (1 - t^{\omega_i/D})^{n_i^+} (1 - t^{\omega_i/D})^{n_i^-} \, dt, \tag{4.3}$$

where $D = \boldsymbol{\omega} \cdot (\mathbf{N} - \mathbf{n}) = \sum_{i=1}^{c} \omega_i (N_i^+ - n_i^+) + \sum_{i=1}^{c} \omega_i (N_i^- - n_i^-)$.

### 4.2.3 Multivariate Wallenius' Noncentral Hypergeometric as PPSWOR

If we take the models in the previous section even further, instead of choosing to group similar unit sizes together, we can instead give each unit size its own category. In other words, each unit would be given a weight equal to its degree. This means that the multivariate Wallenius' Noncentral Hypergeometric distribution is modeling exactly a PPSWOR process.

Why do we want to model it as a PPSWOR process? Due to the difficulty in dealing

with the complex RDS process, there has been considerable work in using different ways of approximating it. The traditional method is to conceptualize it as a Markov Chain process with a stationary distribution reached after a long "burn-in" period (Heckathorn 1997, Volz and Heckathorn 2008). Alternatively, a successive sampling approximation for RDS has also been proposed (Gile 2008, Gile 2011, Gile and Handcock 2015). The successive sampling scheme, also known as sampling with probability proportional to size without replacement, has seen use in survey sampling and ecology (Andreatta and Kaufman 1986, Bickel et al. 1992, Nair and Wang 1989). Gile (2008) and Gile (2011) provide justification for using a successive sampling approximation to the RDS process, arguing that this can reduce finite population bias compared to methods based on a with-replacement assumption, such as the Salganik-Heckathorn estimator (Salganik and Heckathorn 2004) and the Volz-Heckathorn estimator (Volz and Heckathorn 2008).

### 4.2.4 Likelihood

Let c be the maximum unit size. This can be larger than the largest observed unit size due to possibly larger unobserved unit sizes. Let $\mathbf{N} = \{N_1^+, \cdots, N_c^+, N_1^-, \cdots, N_c^-\}$, where $N_i^+$ is the number of people with sampling weight $\omega_i$ who were observed in the first list and $N_i^-$ is the number of people with sampling weight $\omega_i$ who were caught in the second list, but not in the first.. We set the weights to be proportional to size, so $\omega_1^+ = \omega_1^- = 1$, $\omega_2^+ = \omega_2^- = 2$, and so on. Therefore, $N_1^+ + \cdots + N_c^+ + N_1^- + \cdots + N_c^- = N$ and $N_1^+ + ... + N_c^+ = n'$. In other words, $\mathbf{N}$ represents the population tabulated by unit size and whether it was caught in the first list. Let $\mathbf{n} = \{n_1^+, \cdots, n_c^+, n_1^-, \cdots, n_c^-\}$ be the number in each group observed in

the second list. In other words, just as $\mathbf{N}$ is the tabulation of the population by unit size and capture status in the first list, $\mathbf{n}$ is the tabulation of the second list by unit size and capture status in the first list. Note that $n_i^+ \leq N_i^+$ and $n_i^- \leq N_i^-$ for all $1 \leq i \leq c$, and that $n_1^+ + \cdots + n_c^+ + n_1^- + ... + n_c^- = n''$. We will assume that the unit sizes are independent and identically distributed by some probability mass function $f(\cdot|\eta)$.

For the sake of simplicity, we will use $\mathbf{U_{obs}} = \{U'_{obs}, U''_{obs}, Y''_{obs}\}$, with realizations $\mathbf{u_{obs}} = \{u'_{obs}, u''_{obs}, y''_{obs}\}$, to represent all observed data, including both lists as well as information about their overlap.

Therefore,

$$L(N, \eta|\mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto p(\mathbf{U_{obs}} = \mathbf{u_{obs}}|N, \eta)$$

$$= p(U'_{obs} = u'_{obs}|N, \eta) \cdot p(U''_{obs} = u''_{obs}, Y''_{obs} = y''_{obs}|U'_{obs} = u'_{obs}, N, \eta) \tag{4.4}$$

$$= p(U'_{obs} = u'_{obs}|N, \eta) \cdot p(\mathbf{n}|U'_{obs} = u'_{obs}, N, \eta) p(U''_{obs} = u''_{obs}, Y''_{obs} = y''_{obs}|U'_{obs} = u'_{obs}, \mathbf{n}, N, \eta). \tag{4.5}$$

Note that we can get from (4.4) to (4.5) because $\mathbf{n}$ is just the tabulated version of the information in $U'_{obs}, U''_{obs}, V''_{obs}$. We can sum over all possible $u, g', g''$ to get

$$L(N, \eta|\mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto p(\mathbf{U_{obs}} = \mathbf{u_{obs}}|N, \eta)$$

$$= p(U'_{obs} = u'_{obs}|N, \eta) \cdot p(U''_{obs} = u''_{obs}, Y''_{obs} = y''_{obs}|U'_{obs} = u'_{obs}, N, \eta)$$

$$= p(U'_{obs} = u'_{obs}|N, \eta) \cdot p(U''_{obs} = u''_{obs}, Y''_{obs} = y''_{obs}|U'_{obs} = u'_{obs}, \mathbf{n}, N, \eta) p(\mathbf{n}|U'_{obs} = u'_{obs}, N, \eta)$$

$$= \sum_u \sum_{g', g''} \Big[ p(U'_{obs} = u'_{obs}|G' = g', U = u, N, \eta) \cdot p(U''_{obs}, V''_{obs}|G'' = g'', U = u, U'_{obs}, \mathbf{n}, N, \eta)$$

$$p(G' = g'|U = u) p(G'' = g''|\mathbf{n}, U = u) p(\mathbf{n}|U = u, U'_{obs} = u'_{obs}, N, \eta) p(U = u|\eta) \Big] \tag{4.6}$$

$$= \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} p(G' = (1 \cdots n')|U = v) p(G'' = g^*|\mathbf{n}, U = v) p(\mathbf{n}|\mathbf{N}) \prod_{j=1}^{N} f(u_j|\eta), \tag{4.7}$$

where $\mathcal{U}$ is the set of equivalence classes of unit sizes consistent with our observed data $\mathbf{u_{obs}}$ given that the observed units are indexed from 1 to $n$ in order of observation; and $g^*$ is such that $g_i^* = k$ if the $i^{th}$ element of the second list was $k^{th}$ unit in the first list, and the units that were sampled for the first time in the second list take the values $(n'+1, n'+2, \cdots, n)$ in order. In other words, we assign the indices $(1, \cdots, n)$ to all observed units in the order they were first observed. Since the likelihood is the same for different values of $g'$ and $g''$ as long as they are consistent with the observd data, we can do this process of assigning a particular set of indices and account for the number of possible sequences with a factor outside the sum. Note that $p(\mathbf{n}|U = u, U'_{obs} = u'_{obs}, N, \eta) = p(\mathbf{n}|\mathbf{N})$ since the only additional information $U = u, U'_{obs} = u'_{obs}$ provides over $\mathbf{N}$ is the order in which the first list was sampled, which is not relevant for the noncentral hypergeometric distribution.

### 4.2.5 Modeling the First List as PPSWOR

We will treat the first list in the same way as in SS-PSE (Handcock et al. 2014), treating the RDS process as PPSWOR. According to the successive sampling procedure, we sample the first unit with probability

$$p(G'_1 = k) = \frac{u_k}{\sum_{j=1}^{N} u_k}, \qquad k = 1, \cdots, N, \tag{4.8}$$

and subsequent units $(2, \cdots, n')$ from the remaining units (that is, without replacement) with probability

$$p(G'_i = k|G'_1 = g'_1, \cdots, G'_{i-1} = g'_{i-1}) = \frac{u_k}{\sum_{k \notin \{g'_1, \cdots, g'_{i-1}\}} u_k}, \qquad k \notin \{g'_1, \cdots, g'_{i-1}\}. \tag{4.9}$$

Then, the probability of an observed sequence $g$ given a population of unit sizes is given by

$$p(G' = g|U = u) = \prod_{k=1}^{n} \frac{u_{g'_k}}{r'_k}, \tag{4.10}$$

where

$$r'_k = \sum_{i=1}^{N} u_i - \sum_{j=1}^{k-1} u_{g'_j} \qquad k = 1, \cdots, n.$$

So, the likelihood is

$$\begin{aligned} L(N, \eta|\mathbf{U_{obs}} = \mathbf{u_{obs}}) &\propto p(\mathbf{U_{obs}} = \mathbf{u_{obs}}|N, \eta) \\ &= \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \left[ \left( \prod_{k=1}^{n} \frac{u_{g_k}}{r_k} \right) p(G'' = g^*|\mathbf{n}, U = v) p(\mathbf{n}|\mathbf{N}) \prod_{j=1}^{N} f(u_j|\eta) \right]. \end{aligned} \tag{4.11}$$

### 4.2.6 Using Multivariate Wallenius' Noncentral Hypergeometric

The Multivariate Wallenius' Non-central Hypergeometric distribution describes a process in which we are selecting, without replacement, from a population of unequal sampling probabilties. Thus, if we use the multivariate Wallenius' noncentral hypergeomtric distribution with sampling weight equal to unit size, then we are modeling the sampling process as PPSWOR.

So, we get

$$p(\mathbf{n}|U = u, U'_{obs}, N, \eta) = p(\mathbf{n}|\mathbf{N}) = \left( \prod_{i=1}^{c} \binom{N_i^+}{n_i^+} \binom{N_i^-}{n_i^-} \right) \int_0^1 \prod_{i=1}^{c} (1 - t^{\omega_i^j / D})^{n_i^+} (1 - t^{\omega_i^j / D})^{n_i^-} dt, \tag{4.12}$$

where $D = \boldsymbol{\omega} \cdot (\mathbf{N} - \mathbf{n}) = \sum_{i=1}^{c} \omega_i^+ (N_i^+ - n_i^+) + \sum_{i=1}^{c} \omega_i^- (N_i^- - n_i^-)$. This gives us

$$L(N, \eta|U'_{obs}, U''_{obs}, V''_{obs}) \propto p(U'_{obs}, U''_{obs}, V''_{obs}|N, \eta) \tag{4.13}$$

$$= \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \left( p(G' = (1 \cdots n') | U = v) p(G'' = g^* | \mathbf{n}, U = v) p(\mathbf{n}|\mathbf{N}) \prod_{j=1}^{N} f(u_j|\eta) \right), \quad (4.14)$$

where $p(\mathbf{n}|\mathbf{N})$ is given by (4.12).

### 4.2.7 Computational Considerations

Consider the likelihood in (4.14). For simplicity, we will denote the expression inside the summation as

$$Q(v, \eta) = \left( \prod_{k=1}^{n} \frac{u_{g_k}}{r_k} \right) p(G'' = g^* | \mathbf{n}, U = v) p(\mathbf{n}|\mathbf{N}) \prod_{j=1}^{N} f(u_j|\eta).$$

In practice, it is difficult to find the sum $\sum_{v \in \mathcal{U}} Q(v, \eta)$ due to the size of $\mathcal{U}$. In order to compute it, we must sum over all possible unit size combinations for the unobserved units. Handcock et al. (2014) is able to circumvent this issue by re-expressing the sum as a product of (much more managable) sums. However, due to the nature of the Multivariate Wallenius' Non-central Hypergeometric distribution, we are unable use the same method.

It is possible to compute the likelihood for select cases in which $n$ is extremely close to the true value of $N$ (and thus, the number of unobserved units needed to be summed over is relatively low), but the nature of this sum makes it extremely difficult to implement for any realistic situations. We found that the nature of the non-central hypergeometric distribution makes this quite intractable, and though the theory seems sound, we were unable to obtain applicable results from this method.

Because of this, we adjust our method, modeling the second list using a variant of the successive sampling procedure instead of using the Wallenius' Non-central Hypergeometric distribution. This novel method, which we call Capture-Recapture SS-PSE, is described in

the following section.

## 4.3 Capture-Recapture SS-PSE

The difficulty in using the multivariate Wallenius' Non-Central Hypergeometric distribution to model the second list was in computing the distribution over all possible unobserved units. In this section, we will propose a different model that is similarly based on the idea of using two independent lists, both of which we model as sampling with probability proportional to size without replacement.

### 4.3.1 Likelihood Formulation

We start by describing the likelihood. Recall that $n'$ and $n''$ refer to the sample size of the first list and second list, respectively, and that $n_0$ refers to the size of the overlap while $n$ refers to the overall unique sample size (so that $n_1 + n_2 - n_0 = n$).

As before, we have an ordered sampling design, with $G' = (G'_1, \cdots, G'_{n'})$ representing the random indices of the sequentially sampled units with realization $g' = (g'_1, \cdots, g'_{n'})$ in the first list, and $G'' = (G''_1, \cdots, G''_{n''})$ the random indices of the sequentially sampled units with realization $g'' = (g''_1, \cdots, g''_{n''})$ in the second list. $U'_{obs} = \{U_{g'_1}, \cdots, U_{g'_{n'}}\}$ is the unit sizes, in order of the first list, and $U''_{obs} = \{U_{g''_1}, \cdots, U_{g''_{n''}}\}$ the ordered unit sizes in order of the second list, with realizations $u'_{obs} = \{u_{g'_1}, \cdots, u_{g'_{n'}}\}$ and $u''_{obs} = \{u_{g''_1}, \cdots, u_{g''_{n''}}\}$, respectively. $Y''_{obs} = \{Y''_{g''_1}, \cdots, Y''_{g''_{n''}}\}$ with realizations $y''_{obs} = \{y''_{g''_1}, \cdots, y''_{g''_{n''}}\}$ represent the recapture information. In other words, $y''_{g_i} = 1$ if unit $g_i$ was observed in the first list and 0 otherwise. $U = (U_1, \cdots, U_N)$ and $u = (u_1, \cdots, u_N)$ are the unit sizes of the population.

As before, we will assume that the unit sizes are independent and identically distributed by some probability mass function $f(\cdot|\eta)$.

For the sake of simplicity, we will use $\mathbf{U_{obs}} = \{U'_{obs}, U''_{obs}, Y''_{obs}\}$, with realizations $\mathbf{u_{obs}} = \{u'_{obs}, u''_{obs}, y''_{obs}\}$, to represent all observed data, including both lists as well as information about their overlap.

$$
\begin{aligned}
&L(N, \eta|\mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto p(\mathbf{U_{obs}} = \mathbf{u_{obs}}|N, \eta) \\
&= p(U'_{obs} = u'_{obs}|N, \eta) \cdot p(U''_{obs} = u''_{obs}, Y''_{obs} = y''_{obs}|U'_{obs} = u'_{obs}, N, \eta) \\
&= p(U'_{obs} = u'_{obs}|N, \eta) \cdot p(U''_{obs} = u''_{obs}|Y''_{obs} = y'_{obs}, U'_{obs} = u'_{obs}, N, \eta) \cdot \\
&\quad p(Y''_{obs} = y'_{obs}|U'_{obs} = u'_{obs}, N, \eta) \\
&= \sum_u \Bigg[ \Bigg( \sum_{g'} p(U'_{obs} = u'_{obs}|U = u, G' = g', \eta) p(G' = g'|U = u, \eta) \Bigg) \cdot \\
&\qquad \Bigg( \sum_{g''} p(U''_{obs} = u''_{obs}|U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = U, G'' = g'', \eta) \cdot \\
&\qquad\quad p(G'' = g''|Y''_{obs} = y''_{obs}, U'_{obs} = u'_{obs}, U = u) p(Y''_{obs} = y''_{obs}|U'_{obs} = u'_{obs}, U = U, \eta) \Bigg) \cdot \\
&\qquad p(U = u|\eta) \Bigg] \\
&= \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} p(G' = (1 \cdots n')|U = v) p(G'' = g^*|U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = v, \eta) \cdot \\
&\qquad\qquad p(Y''_{obs} = y''_{obs}|U'_{obs}, U = U, \eta) \prod_{j=1}^{N} f(u_j|\eta).
\end{aligned}
\tag{4.15}
$$

Here, $\mathcal{U}$ is the set of equivalence classes of unit sizes possible for the $N$ units given that the observed data was $\mathbf{u_{obs}}$. Intuitively, the elements of $\mathcal{U}$ include all possible unit sizes for each of the $N$ units, except $n$ of them are constrained to be the observed unit sizes. Since the likelihood is equivalent for all values of $g'$ and $g''$ as long as they are have the same unit sizes, we assign the labels sequentially starting from 1 and incrementing up when we sample

a previously unobserved unit, then multiply by the number of permutations outside the sum. So, in the first list, we have $g' = \{1, \cdots, n'\}$ and in the second list, we have $g'' = g^*$, where the values of $g^*$ takes on the original label from the first list if it was already observed in the first list, and the next available sequential value if it was not observed in the first list. In other words, the newly-observed units in $g^*$ are in order from $n_1 + 1$ to $n$ (recall that $n$ refers to the combined sample size, or the number of unique units sampled in the two lists), while the previously-observed units retain their original labeling. Since we are choosing $n'$ indices from $N$ possible in the first list and $n' - n_0$ indices from $N - n'$ possible in the second list, the multiplicative factor is

$$\frac{N!}{(N-n')!} \cdot \frac{(N-n')!}{(N-n'-(n''-n_0))!} = \frac{N!}{(N-n)!}. \tag{4.16}$$

### 4.3.2 Modeling RDS as PPSWOR

We treat the RDS samples as a successive sampling procedure. The first list is the same as in SS-PSE (Handcock et al. (2014)) and in the first list for the MWNCH model as detailed in Section 4.2.5, which gives us

$$L(N, \eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \prod_{k=1}^{n_1} \frac{u_{g'_k}}{r'_k} p(G'' = g^* | U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = v, \eta) \cdot$$
$$p(Y''_{obs} = y''_{obs} | U'_{obs}, U = U, \eta) \prod_{j=1}^{N} f(u_j | \eta), \tag{4.17}$$

where

$$r'_k = \sum_{i=1}^{n'} u_{g'_i} - \sum_{j=1}^{k-1} u_{g'_j}. \tag{4.18}$$

For the second list, we split it up into two parts: whether the units in second list were in the first list or not, and the order in which they were captured given the information about

whether they were in the first list. We start with the latter.

Given we know whether the unit was in the first list or not, we can treat the sampling process as PPSWOR out of the two groups: captured in first list and not captured in first list. Let $g_k^+, k \in \{1, \cdots, n_1\}$ refer the indices of units caught in the first list and $g_k^-, k \in \{n_1+1, \cdots, N-n_1\}$ refer to the indices of units not caught in the first list. Then, we get

$$p(G'' = g^* | U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = v, \eta) = \prod_{k=1}^{n_2-n_0} \frac{u_{g_k^+}}{r_k^+} \prod_{k=1}^{n_0} \frac{u_{g_k^-}}{r_k^-}, \tag{4.19}$$

where

$$r_k^+ = \sum_{i=1}^{n''} u_{g_i''} - \sum_{j=1}^{k-1} u_{g_j^+} \tag{4.20}$$

and

$$r_k^- = \sum_{i=1}^{n''} u_{g_i''} - \sum_{j=1}^{k-1} u_{g_j^-}. \tag{4.21}$$

So, our likelihood becomes

$$L(N, \eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \prod_{k=1}^{n_1} \frac{u_{g'_k}}{r'_k} \prod_{k=1}^{n_2-n_0} \frac{u_{g_k^+}}{r_k^+} \prod_{k=1}^{n_0} \frac{u_{g_k^-}}{r_k^-} \cdot$$
$$p(Y''_{obs} = y''_{obs} | U'_{obs}, U = U, \eta) \prod_{j=1}^{N} f(u_j | \eta). \tag{4.22}$$

Recall the $Y''_{obs}$ represents a vector of indicator variables for whether the units in the second list were captured in the first list. Again, we treat the process as PPSWOR, so

$$p(Y''_{obs} = y''_{obs} | U'_{obs} = u'_{obs}, U = U, \eta) = \prod_{k=1}^{n_2} \frac{1}{r''_k} \prod_{k=1}^{n_2-n_0} r_k^+ \prod_{k=1}^{n_0} r_k^- \tag{4.23}$$

where

$$r''_k = \sum_{i=1}^{n''} u_{g_i''} - \sum_{j=1}^{k-1} u_{g_j''}. \tag{4.24}$$

After cancelling terms, we get

$$L(N,\eta|\mathbf{U_{obs}}=\mathbf{u_{obs}}) \propto \frac{N!}{(N-n)!}\sum_{v\in\mathcal{U}}\prod_{k=1}^{n_1}\frac{u_{g'_k}}{r'_k}\prod_{k=1}^{n_2-n_0}u_{g_k^+}\prod_{k=1}^{n_0}u_{g_k^-}\prod_{k=1}^{n_2}\frac{1}{r''_k}\prod_{j=1}^{N}f(u_j|\eta). \tag{4.25}$$

### 4.3.3 Bayesian Inference for unit size distribution

Handcock et al. (2014) starts by developing inference for the unit size distribution conditional on known $N$. We do the same for our model. The posterior is given by

$$p(\eta|\mathbf{U_{obs}}=\mathbf{u_{obs}}) \propto \pi(\eta)\cdot L[\eta|\mathbf{U_{obs}}=\mathbf{u_{obs}}], \tag{4.26}$$

where $\pi(\eta)$ is the prior for the unit size distribution parameter.

West (1996) and Handcock et al. (2014) note that the likelihood is difficult to deal with and use

$$p(U=u|G'=g',G''=g'',\eta) = \frac{N!}{(N-n)!}\prod_{k=1}^{n'}\frac{u_{g'_k}}{r'_k}\prod_{h=1}^{n''}\frac{u_{g''_h}}{r''_h}\prod_{j=1}^{N}f(u_j|\eta). \tag{4.27}$$

So, from (4.27),

$$p(U_{unobs}=u_{unobs}|\eta,\mathbf{U_{obs}}=\mathbf{u_{obs}}) \propto \prod_{k=1}^{n_1}\frac{1}{r'_k}\prod_{h=1}^{n_2}\frac{1}{r''_h}\prod_{j=n+1}^{N}f(u_j|\eta). \tag{4.28}$$

West (1996) and Handcock et al. (2014) note that the $r'_k$ and $r''_k$ terms are difficult to deal with and use a method involving augmenting the data. We adapt the method to include multiple lists.

Following West (1996), for $k\in\{1,\cdots,n'\}$, $h\in\{1,\cdots,n''\}$, let $\psi'_k$ and $\psi''_h$ have the exponential distribution with rate parameter $r'_k$ and $r''_h$, respectively. We can see that

$$\begin{aligned}\int_0^\infty r'_k e^{-r'_k\psi'_k}d\psi'_k &= 1\\ \implies \int_0^\infty e^{-r'_k\psi'_k}d\psi'_k &= \frac{1}{r'_k}\end{aligned} \tag{4.29}$$

and

$$\int_0^\infty r''_h e^{-r''_h \psi''_h} d\psi''_h = 1$$
$$\implies \int_0^\infty e^{-r''_h \psi''_h} d\psi''_h = \frac{1}{r''_h}. \tag{4.30}$$

In other words,

$$p(\psi'_k = \psi' | \eta, U_{unobs} = u_{unobs}, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r'_k \exp(-r'_k \psi') \tag{4.31}$$

and

$$p(\psi''_h = \psi'' | \eta, U_{unobs} = u_{unobs}, \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r''_h \exp(-r''_h \psi''). \tag{4.32}$$

We can then augment the data with $\Psi' = (\psi'_1, \cdots, \psi'_{n_1})$ and $\Psi'' = (\psi''_1, \cdots, \psi''_{n_2})$, where the components of $\Psi'$ and $\Psi''$ are all conditionally independent of one another. Let $\Psi = (\Psi', \Psi'')$. Then,

$$\begin{aligned}
&p(U_{unobs} = u_{unobs}, \Psi | \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&= p(\Psi' = \psi' | \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) p(\Psi'' = \psi'' | \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \cdot p(U_{unobs} = u_{unobs} | \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&\propto \prod_{j=1}^{n'} e^{-r'_j \psi'_j} \prod_{j=1}^{n''} e^{-r''_j \psi''_j} \prod_{j=n+1}^{N} f(u_j | \eta).
\end{aligned} \tag{4.33}$$

Using (4.18), (4.20), and (4.33),

$$\begin{aligned}
&p(U_{unobs} = u_{unobs} | \Psi', \Psi'', \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&\propto \prod_{j=1}^{n'} e^{-r''_j \psi'_j} \prod_{k=1}^{n''} e^{-r''_j \psi''_j} \prod_{j=n+1}^{N} f(u_j | \eta) \\
&\propto \prod_{i=1}^{n'} e^{-\psi'_i \sum_{j=n'+1}^{N} u_{g'_j}} \prod_{i=1}^{n'} e^{-\psi'_i \sum_{j=i}^{n'} u_{g'_j}} \prod_{i=1}^{n''} e^{-\psi''_i \sum_{j=n''+1}^{N} u_{g''_j}} \prod_{i=1}^{n''} e^{-\psi''_i \sum_{j=i}^{n''} u_{g''_j}} \prod_{j=n+1}^{N} f(u_j | \eta) \\
&\propto \prod_{j=n+1}^{N} \exp\left( - u_j \sum_{i=1}^{n'} \psi'_i \right) \exp\left( - u_j \sum_{i=1}^{n''} \psi''_i \right) f(u_{g_j} | \eta)
\end{aligned}$$

$$= \prod_{j=n+1}^{N} \exp\left( - u_j \left( \sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i'' \right) \right) f(u_{g_j}|\eta). \tag{4.34}$$

We see that the unobserved units are conditionally independent from the unnormalized PMF $\exp(-u_j(\sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i''))f(u_{g_j}|\eta)$. In addition, they are independent of all observed information. Thus, we can get draws from the augmented posterior,

$$p(\eta, U_{unobs} = u_{unobs}, \Psi|\mathbf{U_{obs}} = \mathbf{u_{obs}}), \tag{4.35}$$

using a three component Gibbs sampler.

### 4.3.4 Inference for population size

In the previous section, we assumed known $N$. However, when estimating the population size, we do not know $N$ and want to estimate it. To do this, we adjust the method in the previous section to treat $N$ as a parameter.

Following Handcock et al. (2014), we derive the conditional for N.

$$\begin{aligned}
&p(N|\eta, \Psi', \Psi'', \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&\propto \pi(N) p(\mathbf{U_{obs}} = \mathbf{u_{obs}}|N, \eta, \Psi', \Psi'') \\
&= \frac{N!}{(N-n)!}\pi(N) \sum_{v \in \mathcal{U}} \left[ \prod_{j=n+1}^{N} \exp\left( - u_j \left( \sum_{i=1}^{n_1} \psi_i' + \sum_{i=1}^{n_2} \psi_i'' \right) \right) f(u_{g_j}|\eta) \right] \\
&= \frac{N!}{(N-n)!}\pi(N) \prod_{j=n+1}^{N} \left[ \sum_{v_j=1}^{\infty} \exp\left( - v_j \left( \sum_{i=1}^{n_1} \psi_i' + \sum_{i=1}^{n_2} \psi_i'' \right) \right) f(j|\eta) \right] \\
&= \frac{N!}{(N-n)!}\pi(N) \left[ \gamma(\sum_{i=1}^{n_1} \psi_i' + \sum_{i=1}^{n_2} \psi_i'', \eta) \right]^{N-n}, \text{where } \gamma(\alpha, \eta) = \sum_{j=1}^{\infty} e^{-\alpha j} f(j|\eta).
\end{aligned} \tag{4.36}$$

We can use (4.36) to obtain samples from the joint augmented posterior,

$$p(N, \eta, U_{unobs} = u_{unobs}\Psi|\mathbf{U_{obs}} = \mathbf{u_{obs}}), \tag{4.37}$$

which we can then use to obtain the marginal posterior distribution of $N$ and $\eta$.

The full details of the MCMC algorithm are given in Section 4.3.7.

### 4.3.5 Unit Size Distribution Model

For our purposes, we need a super-population model for the unit sizes. In our case, the unit sizes are the personal network sizes, or degrees. There has been a considerable amount of work done on modeling the degree distribution of a network. Handcock et al. (2014) notes that certain long-tailed distributions such as the Poisson-log-normal and the Waring and Yule distributions, which allow for power-law over-dispersion (Handcock and Jones (2006)), are not able to represent the under-dispersion in degree counts, suggesting the Conway-Maxwell-Poisson distribution (Shmueli et al. (2005)) as an alternative.

For the applications in this paper, we chose to use the Conway-Maxwell-Poisson distribution as it offers greater flexibility over similar distributions such as the Poisson while using only one additional parameter.

### 4.3.6 Prior specification

We can parametrize the Conway-Maxwell-Poisson distribution in terms of its mean and standard deviation. We then put priors on the log mean and variance parameters, using the Normal distribution for the prior log mean, $\mu$, given the prior standard deviation, $\sigma$, and scaled Inverse Chi-squared for the variance, $\sigma^2$, so

$$\log(\mu)|\sigma \sim N(\mu_0, \sigma/df_{mean}) \tag{4.38}$$

and

$$\sigma^2 \sim \text{Inv}\chi^2(\sigma_0^2; df_{sigma}). \tag{4.39}$$

In our applications, we use diffuse priors with $df_{mean} = 1$ and $df_{sigma} = 5$.

For the population size, Fienberg et al. (1999) suggests a prior of the form

$$\pi(N) = \frac{(N-l)!}{N!}, \qquad n < N < N_{max}, \tag{4.40}$$

where $N_{max}$ is a maximum value for the population size set to be very large. They use this prior with $l = 1$ in their applications. Handcock et al. (2014) uses instead a two parameter class of priors,

$$\pi(N) = \frac{\beta n (N-n)^{\beta-1}}{N^{\alpha+\beta}} \qquad \text{for } N > n, \alpha > 0, \beta > 0. \tag{4.41}$$

This was the prior used in Handcock et al. (2015) and can be thought of using knowledge about the sample fraction $(n/N)$ as a Beta$(\alpha, \beta)$ distribution. Handcock et al. (2014) notes that this class of priors was chosen after consultation with field researchers and based on budget and logistic considerations for their choice of sample size.

### 4.3.7 Algorithmic Details

Here, we describe in the detail the algorithm for drawing from the joint posterior. We follow Handcock et al. (2014) and West (1996).

1. Initialize $N$ at a point estimate and $U_{unobs}$ at a set of unit sizes.

2. Sample $\eta$ from

$$p(\eta | U_{unobs}, \mathbf{U_{obs}} = \mathbf{u_{obs}}.\Psi', \Psi'', N) = \pi(N) \cdot \prod_{j=1}^{N} f(u_j|\eta) \tag{4.42}$$

This is done using a Metropolis-Hastings algorithm. In our applications, we used the Conway-Maxwell-Poisson distribution as our unit size distribution, so we had two parameters $\eta = ((\log(\mu), \sigma^2)$. We used a Gaussian proposal for the log mean and an Inverse-$\chi^2$ for the variance.

3. Sample $\Psi', \Psi''$ from

$$p(\psi'_k = \psi' | \eta, U_{unobs} = u_{unobs}, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r'_k \exp(-r'_k \psi')$$

and

$$p(\psi''_k = \psi'' | \eta, U_{unobs} = u_{unobs}, \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r''_k \exp(-r''_k \psi'').$$

These are independent standard Exponential draws.

4. Sample $N$ from equation (4.36). In order to make computation easier, we set $N_{max}$ a maximum value for $N$. We compute (4.36) for each value between $n$ and $N_{max}$ and use this to sample a value between $n$ and $N_{max}$ directly.

5. Sample $U_{unobs}$ from equation (4.34). This is done using a rejection sampling method, similar to the one described in West (1996) and used in SS-PSE by Handcock et al. (2014).

   The rejection sampling process is

   (i) Draw $d$ from $f(\cdot|\eta)$ and, independently, $u \sim U(0, 1)$.

   (ii) If $\log(u) > -(\sum_{i=1}^{n'} \psi'_i + \sum_{i=1}^{n''} \psi''_i) \cdot d$, reject $d$ and return to (i). Otherwise, save $d$ and repeat until $N - n$ elements of $u_{unobs}$ have been sampled.

6. Repeat until convergence.

## 4.4 Implementation

Due to the intractible nature of the MWNCH model and the relative ease with which we can implement the CR-SS-PSE model, we have only implemented the CR-SS-PSE and use it for our applications. In the next chapter, we show simulation results comparing it to other currently used methods of estimating population size.

We have implemented the CR-SS-PSE model using much of the code in the `sspse` package (Handcock and Gile (2015)). Though we can theoretically use other distributions to model the unit size, we have currently only implemented the model using the Conway-Maxwell-Poisson distribution.

# CHAPTER 5

# Analysis of Population Size Estimation Methods Using Simulation Studies

We want to compare our method, Capture-Recapture Successive Sampling-Population Size Estimation (CR-SS-PSE), with other methods currently used. An assessment of how accurately our method approximates the truth typically involves a discussion of asyptotics properties. However, in the case of population size estimation, it is unclear what asymptotics properties would even be. Since the true population size is fixed and finite (and what we want to estimate), the concept of increasing $n$ indefinitely doesn't make sense. For this reason, we use simulation studies for our analysis.

We should note that there are two things being simulated. First, we simulate a population network. Though using real, complete network populations would be ideal, in general, complete network information about a certain population is extremely difficult to collect. Because of this, we used simulated populations while trying to keep the network characteristics are close to reality as possible. Then, we simulate the RDS process on these networks. In our applications, we assumed the RDS assumptions discussed in Chapter 2 were true. Though we do not necessarily think this is always the case, we believe that they are close enough to represent what really happens.

When we apply traditional capture-recapture models, we implicitly make assumptions about the RDS process. That is, we may be treating the RDS as a simple random sample as in the basic Lincoln-Petersen estimator or $M_0$ model, or a weighted sample as in the Rasch model. However, we have reason to believe that the RDS process is actually very different from this (Gile 2011). Therefore, in these simulations, we try to get as close to real respondent-driven samples as possible to see how each of these methods perform.

In the next two sections, we describe the network and RDS simulations. Then, in Section 5.3, we describe the methods to be used before showing the results of our simulations.

## 5.1 Simulation of Population Networks

In our simulations, we wanted to emulate real populations as much as possible. Therefore, we tried to use simulated networks generated with network characteristics as close to reality in the populations of interest as possible. In general, we do not expect people to form networks completely randomly. We want to incorporate some of the factors that affect network structure, as it can, in turn, greatly affect how the RDS process samples from the population. Therefore, we will look at several key aspects of network formation when simulating our populations.

Exponential random graph models (ERGMs) were used to generate each of the simulated population networks discussed in this chapter (Gile and Handcock 2010, Snijders et al. 2006, Spiller et al.). That is, $y$, an $N$ by $N$ binary matrix representation of a network (with $y_{ij} = 1$ if there is a tie between node $i$ and node $j$, and $y_{ij} = 0$ otherwise) is represented as a realization of the random variable $Y$ with distribution

$$P_\eta(Y = y|x) = \exp\{\eta \cdot g(y, x) - \kappa(\eta, x)\} \qquad y \in \mathcal{Y}, \tag{5.1}$$

where $x$ are covariates, $g(y, x)$ is a $p$-dimensional vector of network statistics, $\eta \in \mathbb{R}^p$ is a parameter vector, $\mathcal{Y}$ is the set of all possible undirected networks with $N$ nodes, and $\exp\{\kappa(\eta, x)\} = \sum_{u \in \mathcal{Y}} \exp\{\eta \cdot g(u, x)\}$ is the normalizing constant (Handcock et al. 2014).

The ERGM provides a way for us to put a probability on every possible network with a certain number of nodes, $N$, based on characteristics of the network we want to simulate. In the simulations discussed in this chapter, we use parameters for homophily and differential activity, as well as overall mean degree that acts as a baseline level of propensity to form ties. We use the ERGM to draw networks from the space of all possible networks with a probability described by (5.1).

For the first part of our simulation studies, we use two simulated network — called Faux Sycamore and Faux Madrona — as well as a real network from the Add Health data set. A description of the methods used to generate the Faux Sycamore and Faux Madrona networks, as well as a discussion of the structure of the Add Health network, is provided in Section 5.5.

Finally, we looked simulated networks based on RDS data collected on people who inject drugs (PWID) from two sets of 20 cities collected by the CDC (Centers for Disease Control and Prevention 2012, Centers for Disease Control and Prevention 2015). Even though the cities were repeated in these studies, we will treat them as essentially unique city-year pairs, since the time difference would result in a different population, and so we will refer to them as 40 separate cities. We will refer to these 40 cases as the CDC PWID studies. These

networks are described in more detail in Section 5.5.

## 5.2 Simulation of RDS

In addition to simulating the population network, we simulate the actual RDS process that is performed on these networks. In the simulations performed in this chapter, we will define a *trial* as consisting of collecting two respondent-driven samples and recording the order of observation, personal network size (unit size) as well as the overlap between the two samples. The latter is only collected in the second sample as the answer to the question, "Were you in the first sample?" In other words, we do not track unique matching between the two lists.

For the Add Health, Faux Madrona, and Faux Sycamore networks, we started with a random sample of 10 initial seeds, which is consistent with a review of over 120 RDS studies that found an average of 10 seeds used (Malekinejad et al. 2008), and the number used in a pilot study run by the CDC (Abdul-Quader et al. 2006). We used two coupons for each respondent. In the simulation, starting with the seeds, we sampled two recruits randomly from the nodes connected to each respondent. Then, we used the newest wave of recruits to repeat the process. If a respondent had only one available link to an unsampled person, that person only recruited one person. We stopped the RDS recruitment when we reached our target sample size. If the chain ran into a dead end, we were to keep going by selecting new seeds at random from the population. In practice, we did not enounter a case in which this was necessary.

The recruiting was assumed to have been done at random from each respondent's personal network. We note that the homophily in the network structure affects this random

recruitment. Even if a respondent randomly selects someone, they are more likely to choose someone similar to themselves because they know more people similar to themselves. As we have noted in Chapter 2, even though we see a homophily effect, we do not expect extreme cases, such as respondents recruiting no one in a different age group.

For the 40 simulated networks based on CDC PWID studies, we simulated networks using network characteristics for each of the cities, then performed one RDS trial for each the networks. The networks were generated for each of three different population sizes, $N = \{1000, 5000, 10000\}$. The RDS simulation that was performed on each of the networks was based on the real RDS that was taken in the real city. That is, the simulated RDS had the same number of starting seeds, number of coupons given out, number of seeds with or without the trait (HIV infection), and distribution of the number of recruits per respondent as the actual RDS run in that city.

## 5.3 Population Size Estimation Methods

We compare a variety of methods using the data collected from the RDS simulations. In particular, we want to compare the performance of our model — Capture-Recapture SS-PSE — to the performance of previous methods using network information (SS-PSE) and multiple list information (Hypergeometric Capture-Recapture, NPLCM).

The results of the Lincoln-Petersen estimator, the Chapman estimator, and the Hypergeometric model are all extremely similar in their point estimates (as would be expected due to their similarity), so we will show only one case with all three methods in Section 5.5.1 and choose to only show the Hypergeometric model for the rest. The same is true for the

log-linear models described in Section 3.2.4.

In the following sections, we show six different methods to estimate the population size: Capture-Recapture SS-PSE, the Hypergeometric Capture-Recapture (HCR) model, SS-PSE using a separate RDS with sample size $n'$, an Independent SS-PSE model using both lists, and Combined SS-PSE using the both samples. For the Independent SS-PSE model, we found the posterior distribution of $N$ using SS-PSE with first list, then used that exact distribution as the prior for $N$ in the SS-PSE model for the second list. For the Combined SS-PSE, we took the units in the first list in order, then the units in the second list who were not captured in the first list, in order. That is, we simply included all the unique units in the order in which they were recruited, throwing out any duplicates. We did this because we are treating the RDS process as a Successive Sampling process without replacement, so we treat the two lists as one big list without replacement by throwing out the duplicates. Finally, we used the Nonparametric Latent Class Model (NPLCM) described in Section 3.2.7. We found the posterior means as our point estimate, as well as the shortest contiguous 95% highest posterior density region as our interval estimate, which we will call our shortest credible interval (SCI).

Unit sizes were modeled using the Conway-Maxwell-Poisson distribution (hyper parameters $\mu_0 = 7$, $df_{mean} = 1$, $\sigma_0 = 3$, $df_{mean} = 5$) in both the CR-SS-PSE model and the SS-PSE model. For the population size prior, the same Beta distribution with $\alpha = 1$ and mode of twice the sample size $n$ was used for each method unless otherwise mentioned. The implementation of NPLCM currently only allowed for a population size prior of $\pi(N) \propto 1/N$, $n < N < N_{\max}$, where a maximum value of N, $N_{\max}$, is chosen, so we used that in all of

our applications (Manrique-Vallier 2016). A burn-in of 5000 and thinning interval of 10 was used for CR-SS-PSE and SS-PSE. NPLCM used a burn-in of 10000 and a thinning interval of 10.

In the following sections, we will detail the results of our simulation studies. First, in Section 5.4, we will use various methods to estimate the size of the population in the Add Health data set. This will serve as an initial look at the performance of the method before using 1000 trials in the later sections. In Section 5.5, we look at the performance of CR-SS-PSE compared to existing methods for population size estimation. Then, in Section 5.6, we look at the sensitivity of CR-SS-PSE to different priors for the population size. In Section 5.7, we look at the effects of using different sample fractions compared across different methods. In Section 5.8, we look at how the estimates are affected when the sample sizes in the two lists are imbalanced. In Section 5.9, we look at the CDC PWID data. Lastly, in Section 5.10, we take an empirical look at frequentist properties of CR-SS-PSE.

## 5.4 Initial Look at Performance of CR-SS-PSE

We first used CR-SS-PSE to estimate the size of the Add Health Network. The Add Health data set was constructed from a series of questionnaires given to students in school. The students were asked to nominate friends either in the same school or in a "sister" school, and a friendship network was generated based on these responses (Harris et al. 2009). We used a network of $N = 2587$ for our simulations. Though the networks generated were directed, we treated them as undirected by turning each tie into an undirected one.

The trace plots and posterior distributions for the population size $N$ and the unit size dis-

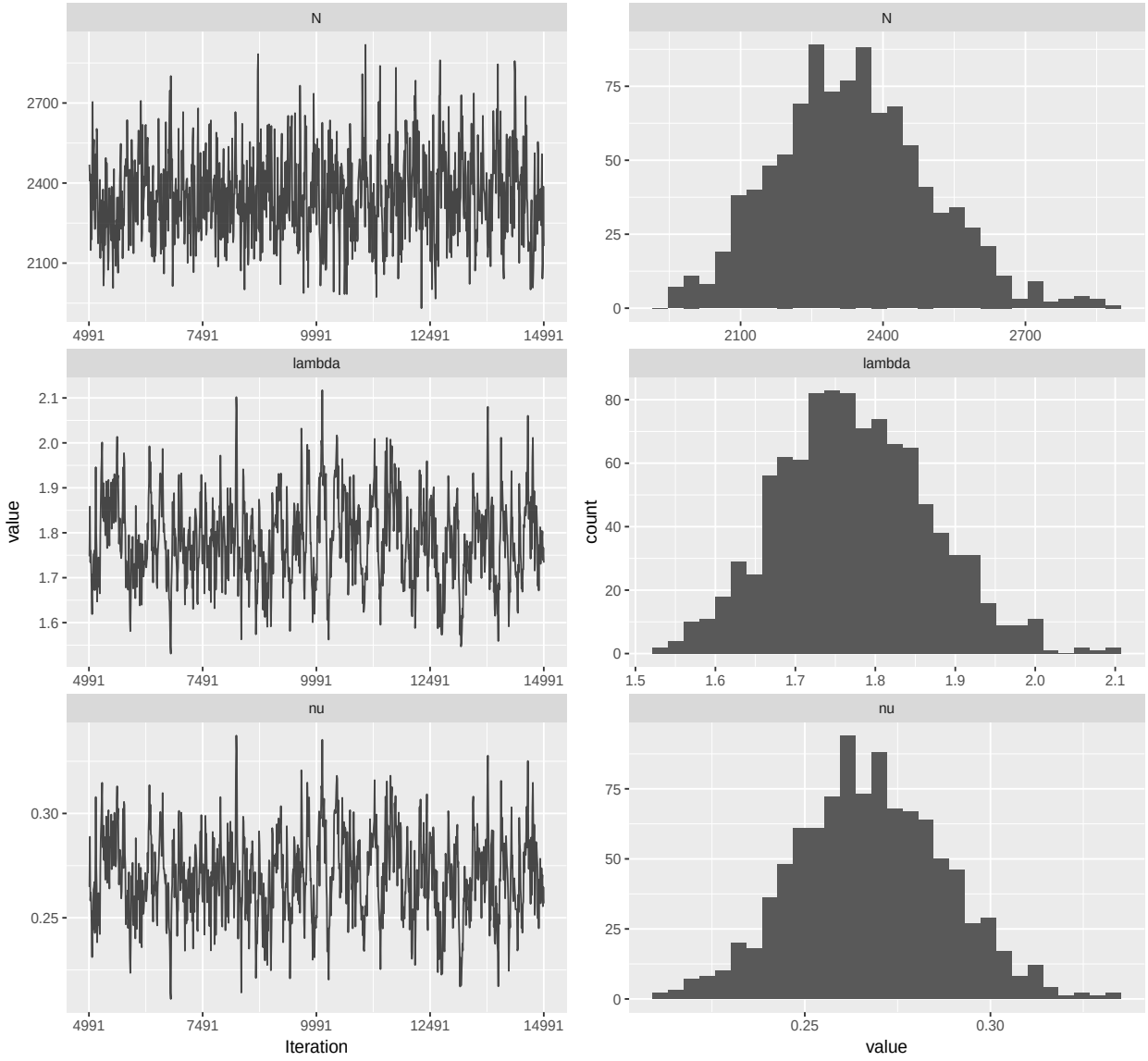



Figure 5.1: Trace plots and histogram of the posterior distribution for $N$, $\lambda$, and $\nu$ using the Add Health data set ($N = 2587$)

tribution parameters $\eta$ are shown in Figure 5.1. The MCMC seems to be well-behaved, with good convergence. The posterior distribution looks reasonable based on our true population size of $N = 2587$.

## 5.5 Performance on Add Health Networks

In this section, we look at three networks: the Faux Sycamore , Faux Madrona, and Add Health networks. The Faux Sycamore and Faux Madrona networks are simulated networks designed to have certain network characteristics similar to networks from the Add Health data set, while the Add Health network is a real friendship network of high school and middle school students. Each case is based on 1000 RDS trials (of two lists each) on the network, so for the simulated networks, there was only one simulation of the network with multiple simulations of the RDS procedure.

### 5.5.1 Faux Sycamore ($N = 715$)

The Faux Sycamore network data set was taken from the `RDS` package in `R` (Handcock et al. (2016)). This simulated data was used by Gile and Handcock (2010) in simulation studies to test RDS estimators and was created to include characteristics similar to the data from the CDC surveillance program (Abdul-Quader et al. 2006). In this network, individuals had a dichotomous characteristic of being infected or uninfected. The network was generated using homophily and differential activity. Homophily was such that two infected individuals were five times as likely to be linked with a tie than a mixed pair of individuals. Differential activity on their disease status was such that the mean degree of infected individuals was

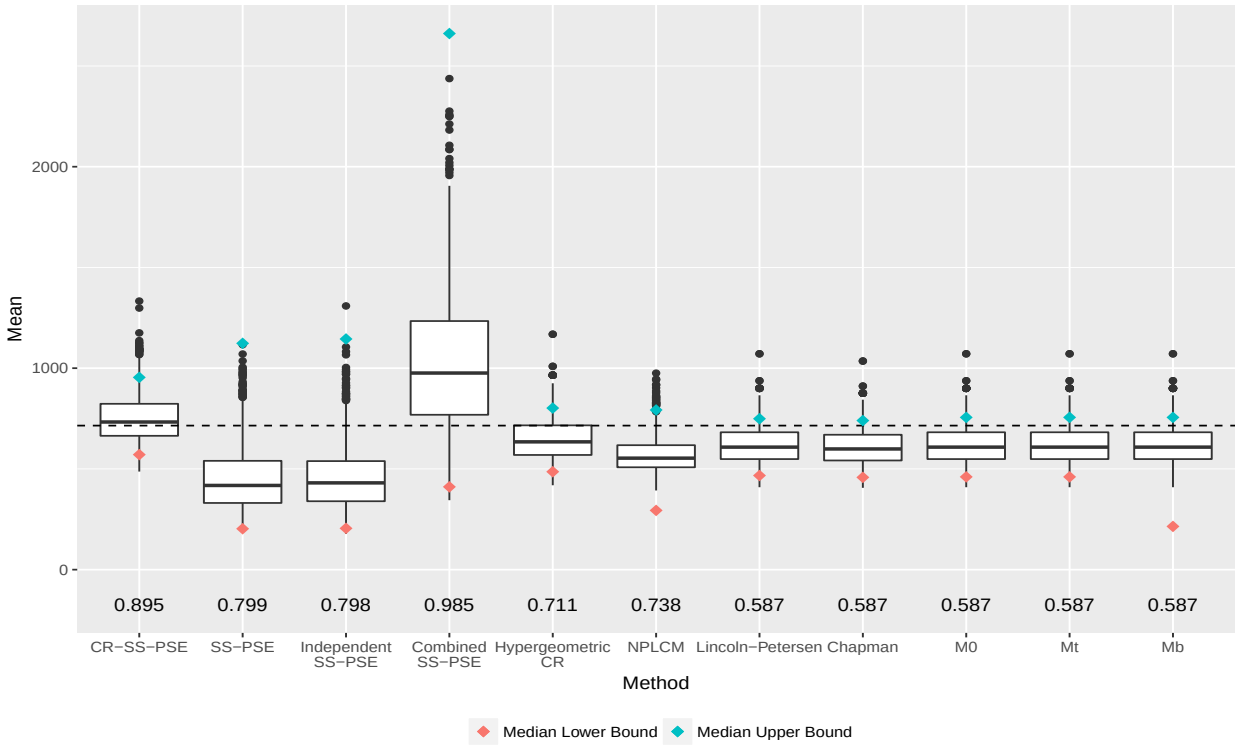



Figure 5.2: Simulation results from Faux Sycamore data set with all methods. The boxplot is of the posterior means. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI interval are also shown. The dashed line shows the true population size ($N = 715$).

twice that of uninfected individuals.

The sample size for each list was $n' = n'' = 150$. For this network, we used the six main methods described in Section 5.3 with 1000 simulations, as well as five other capture-recapture methods described in Chapter 3: Lincoln-Petersen estimator, Chapman estimator, M0, Mt, and Mb. Figure 5.2 shows the posterior means of all of these methods as well as coverage rates. We took the upper and lower bounds of the 95% SCIs and found the median of each, which are shown as blue and red diamonds for the median upper bound and median

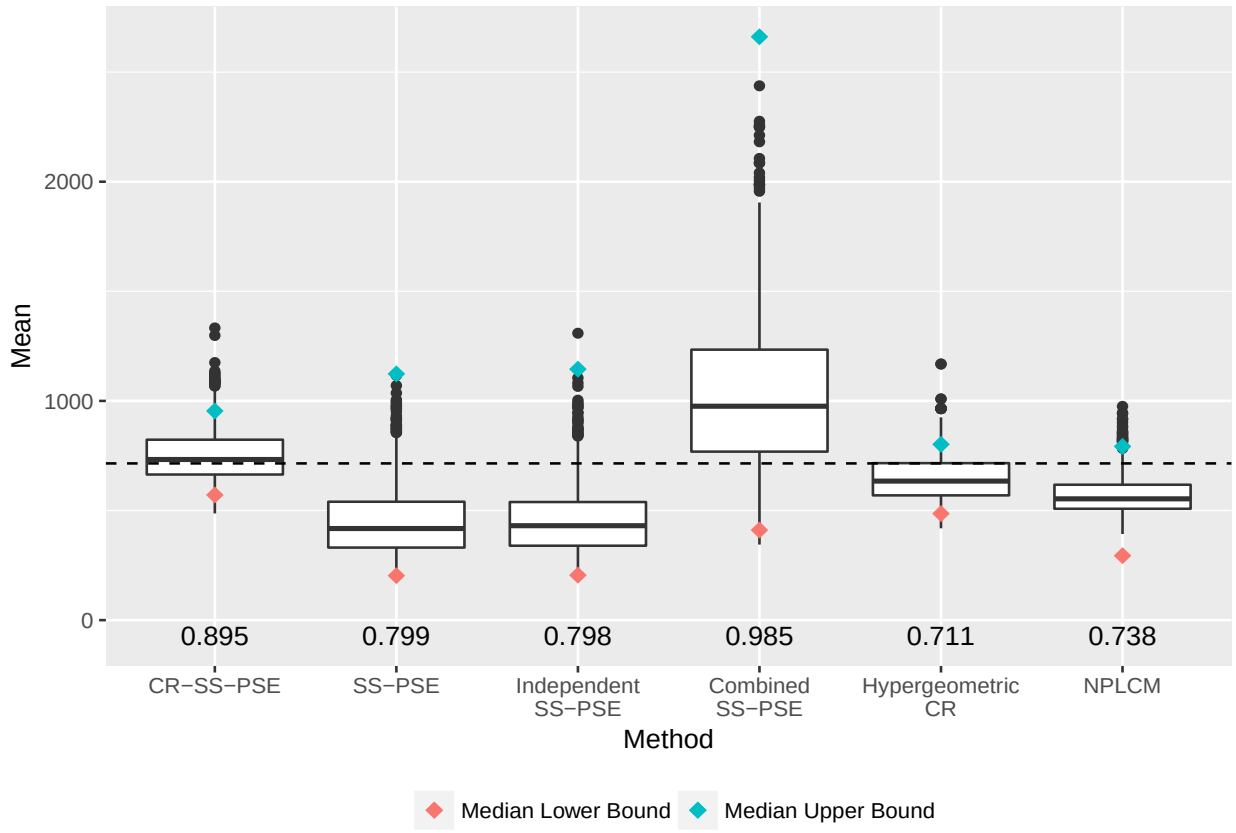



Figure 5.3: A boxplot of posterior means using the six main methods with the Faux Sycamore network are shown. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 715$).

| Method | MSE | Bias | Variance | Bias Proportion of MSE |
|---|---|---|---|---|
| CR-SS-PSE | $1.7 \times 10^4$ | 36 | $1.6 \times 10^4$ | 0.077 |
| SS-PSE | $9.6 \times 10^4$ | $-264$ | $2.7 \times 10^4$ | 0.722 |
| Independent SS-PSE | $9.3 \times 10^4$ | $-259$ | $2.6 \times 10^4$ | 0.719 |
| Combined SS-PSE | $2.2 \times 10^5$ | 307 | $1.2 \times 10^5$ | 0.434 |
| Hypergeometric CR | $1.7 \times 10^4$ | $-70$ | $1.2 \times 10^4$ | 0.287 |
| NPLCM | $3.0 \times 10^4$ | $-145$ | $8.6 \times 10^3$ | 0.708 |

Table 5.1: Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE ($\text{Bias}^2/\text{MSE}$) for each of the six methods with the Faux Sycamore network.

lower bound, respectively. Furthermore, the coverage rates are shown below each boxplot. Since the results of the last five methods are similar to that of the Hypergeometric Capture-Recapture, we will focus on the first six for the rest of our analysis in Figure 5.3 .

CR-SS-PSE seems to do the best job, with the posterior means centered at the true value. SS-PSE and Independent SS-PSE both underestimate the population size, while the Combined SS-PSE model overestimates the population size. The Hypergeometric Capture-Recapture model and NPLCM both also tended to underestimate the true population size. Table 5.1 gives some numerical summaries of the performance of the point estimates. CR-SS-PSE and Hypergeometric CR have the lowest mean squared error (MSE) values at $1.7 \times 10^4$, while CR-SS-PSE has the lowest bias. This generally supports our conclusions from Figure 5.3, in that CR-SS-PSE seems to have the best point estimate, with Hypergeometric CR a

| Method | MSE | Bias | Variance | Bias Proportion of MSE |
|---|---|---|---|---|
| CR-SS-PSE | $2.6 \times 10^6$ | 52 | $2.5 \times 10^4$ | 0.095 |
| SS-PSE | $1.3 \times 10^8$ | $-360$ | $7.2 \times 10^4$ | 0.642 |
| Independent SS-PSE | $1.2 \times 10^8$ | $-347$ | $7.8 \times 10^4$ | 0.608 |
| Combined SS-PSE | $3.6 \times 10^8$ | 599 | $3.8 \times 10^5$ | 0.487 |
| Hypergeometric CR | $7.2 \times 10^6$ | $-85$ | $2.0 \times 10^4$ | 0.266 |
| NPLCM | $3.3 \times 10^7$ | $-182$ | $1.8 \times 10^4$ | 0.646 |

Table 5.2: Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE ($\text{Bias}^2/\text{MSE}$) for each of the six methods with the Faux Madrona network.

close second.

### 5.5.2 Faux Madrona ($N = 1000$)

Similar to the Faux Sycamore network, Faux Madrona was simulated using ERGMs. Just as with the Faux Sycamore network, the network was generated using homophily and differential activity. As before, homophily was such that two infected individuals were five times as likely to be linked with a tie than a mixed pair of individuals, and differential activity on their disease status was such that the mean degree of infected individuals was twice that of uninfected individuals.

The sample size for both lists in this case was $n' = n'' = 200$. Figure 5.4 shows the posterior means for each method with 1000 simulations, as well as median bounds and

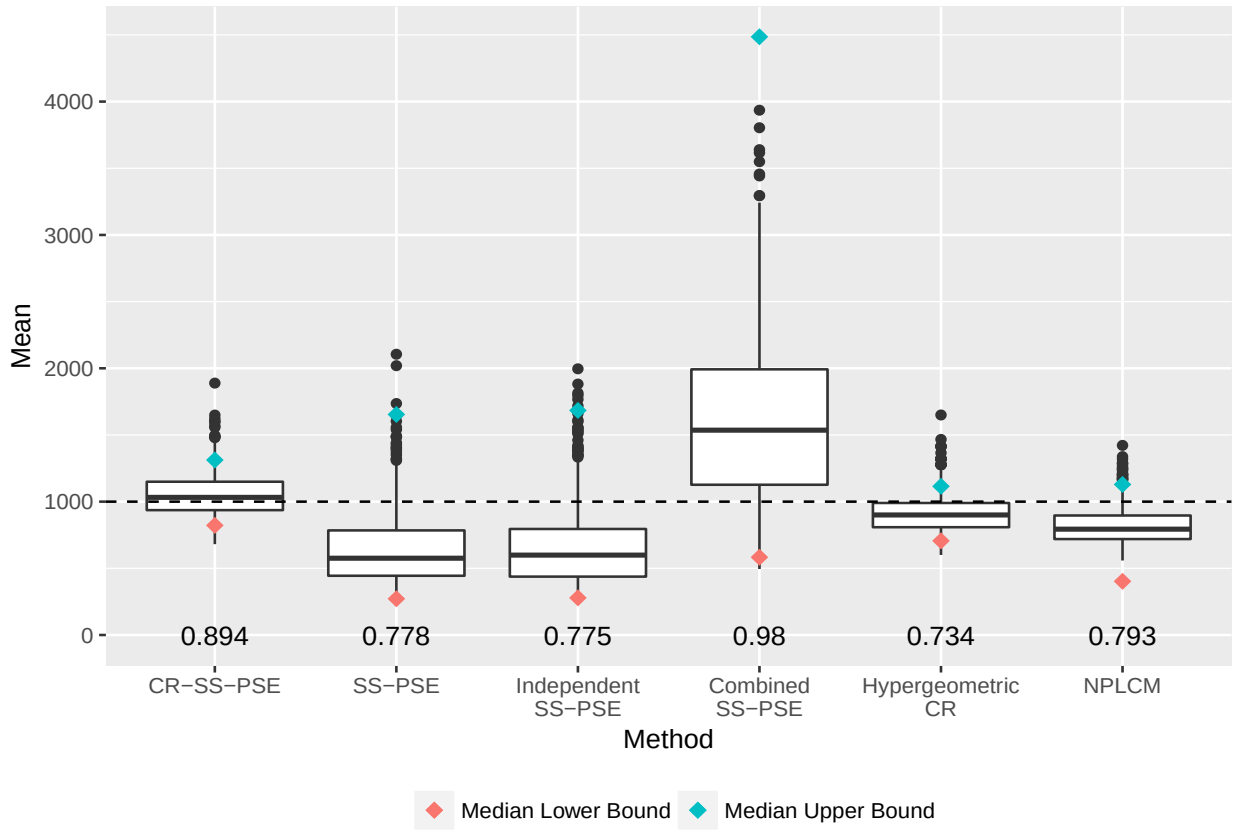



Figure 5.4: Simulation results from Faux Madrona data set. The boxplot is of the posterior means. The value underneath each boxplot represents the proportion of 95% SCI containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 1000$).

| Method | MSE | Bias | Variance | Bias Proportion of MSE |
|---|---|---|---|---|
| CR-SS-PSE | $1.7 \times 10^5$ | 80 | $1.7 \times 10^5$ | 0.037 |
| SS-PSE | $2.4 \times 10^6$ | $-1497$ | $2.0 \times 10^5$ | 0.919 |
| Independent SS-PSE | $2.5 \times 10^6$ | $-1523$ | $1.6 \times 10^5$ | 0.935 |
| Combined SS-PSE | $6.5 \times 10^6$ | 1373 | $4.7 \times 10^6$ | 0.288 |
| Hypergeometric CR | $2.9 \times 10^5$ | $-433$ | $9.9 \times 10^4$ | 0.656 |
| NPLCM | $5.1 \times 10^5$ | $-617$ | $1.3 \times 10^5$ | 0.749 |

Table 5.3: Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE ($\text{Bias}^2/\text{MSE}$) for each of the six methods with the Add Health network.

coverage rates. We see similarities to what we saw from the Faux Sycamore data set. CR-SS-PSE performs the best, with a median posterior mean very close to the true population size, and most of the other methods tend to underestimate while the Combined SS-PSE method overestimates. Again, we see a high coverage for the Combined SS-PSE despite relatively inaccurate point estimates, speaking to the widths of its interval estimates.

Table 5.2 shows the performance of the point estimates. CR-SS-PSE has the lowest MSE, followed by Hypergeometric CR and NPLCM. CR-SS-PSE also has the lowest bias. Combined SS-PSE has the highest variance of posterior means, which is supported by the wide boxplot in Figure 5.4.

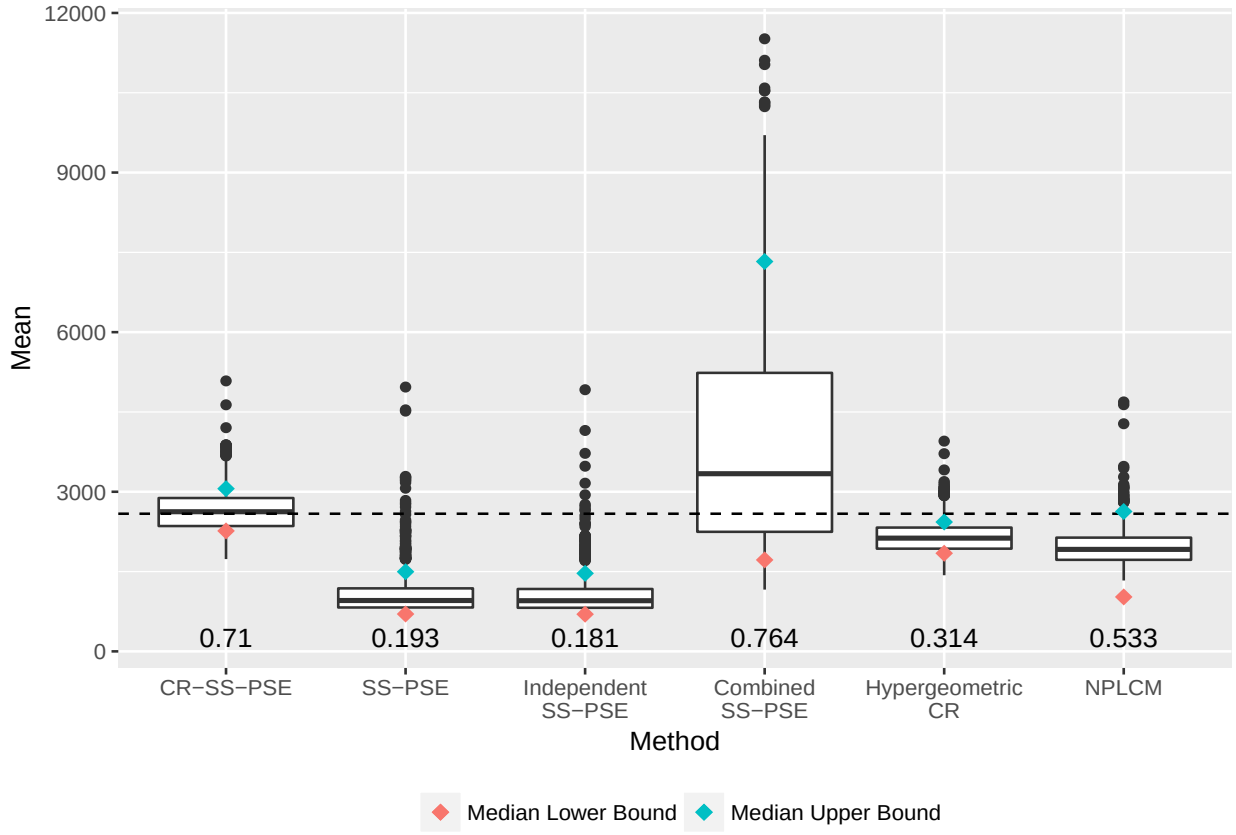



Figure 5.5: Boxplots of posterior means with six methods using the Add Health network. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 2587$).

### 5.5.3 Add Health ($N = 2587$)

In this section, we look at the same Add Health network as we did in Section 5.4. One aspect of this network that may affect the performance compared to the two previous networks is the possible existence of a bottleneck. Since this network consists of two schools, we might be concerned about a possible bottleneck due to a high level of homophily by school.

We again use 1000 simulated RDS trials, apply each of the six methods, and compare their posterior means and 95% SCIs as shown in Figure 5.5. We again see that CR-SS-PSE performs the best, with every single other method besides the Combined SS-SPE underestimating the true population size. The difference is even more drastic in this case, with every single point estimate in the SS-PSE and Independent SS-PSE methods coming under the true population size. Table 5.3 shows this difference even more clearly. CR-SS-PSE again has the lowest MSE, and the bias for the SS-PSE methods are all quite high.

In addition, note that the blue diamond for Hypergeometric CR is below the true population size. This means that the median upper bound for the 95% SCI is below the true population size, which in turn means that Hypergeometric CR has very small interval estimates. So, even though the point estimates seem to be getting close, the extremely narrow intervals result in low coverage rates, as can be seen here.

## 5.6 Sensitivity to Population Size Prior

Recall that in our applications, we are using a population size prior

$$\pi(N) = \frac{\beta n (N-n)^{\beta-1}}{N^{\alpha+\beta}} \qquad \text{for } N > n, \alpha > 0, \beta > 0. \tag{5.2}$$

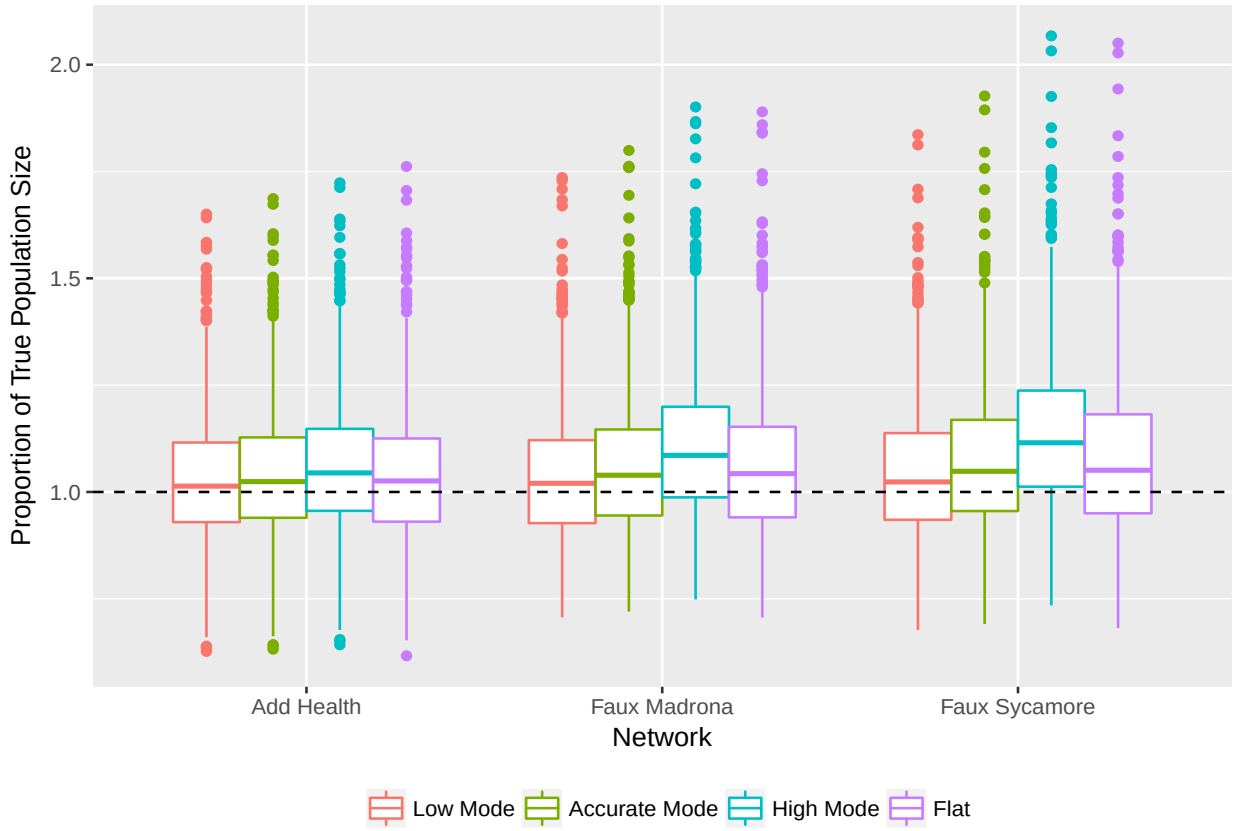



Figure 5.6: Posterior Means with Low, Accurate, and High prior modes, along with a Flat prior. Posterior means are shown as a proportion of the true population size. For Add Health, the prior modes used were 1500, 2587, and 5000. For Faux Madrona, the modes were 700, 1000, 2000; for Faux Sycamore, 500, 715, and 1500.

In the previous section, for each case, we used a population size prior with mode equal to twice the total sample size (mode $= 2n$) with $\alpha = 1$. In this section, we test the sensitivity of our model to the population size prior. To do this, we used the three networks (Faux Madrona, Faux Sycamore, and Add Health) and used CR-SS-PSE to get the posterior means with low, accurate, and high prior modes, as well as a flat prior with a maximum $N$. For Add Health, the prior modes used were 1500, 2587, and 5000. For Faux Madrona, the modes were 700, 1000, 2000; for Faux Sycamore, 500, 715, and 1500. For the flat priors, we used $N_{\max} = 20000$ for the Add Health network and $N_{\max} = 10000$ for the Faux Sycamore and Faux Madrona networks. We used the same sample sizes as in the previous sections (i.e. $n' = n'' = 150$ for Faux Sycamore, $n' = n'' = 200$ for Faux Madrona, and $n' = n'' = 500$ for Add Health). The posterior means using CR-SS-PSE with different priors are shown in Figure 5.6.

Though there are slight differences, all three networks seem to be very similar to each other. Notably, we see that though the priors do make a difference, it is usually very slight, and the estimates are generally good even with inaccurate priors.

## 5.7 Performance with Different Sample Fractions

We want to see how our method performs when we use different sample fractions. So, using the Add Health population network, we used a range of sample sizes and looked at the performance of our methods for each case using 1000 simulations each. The posterior means of the six population size estimation methods with different sample sizes taken from the Add Health network are shown in Figure 5.7.

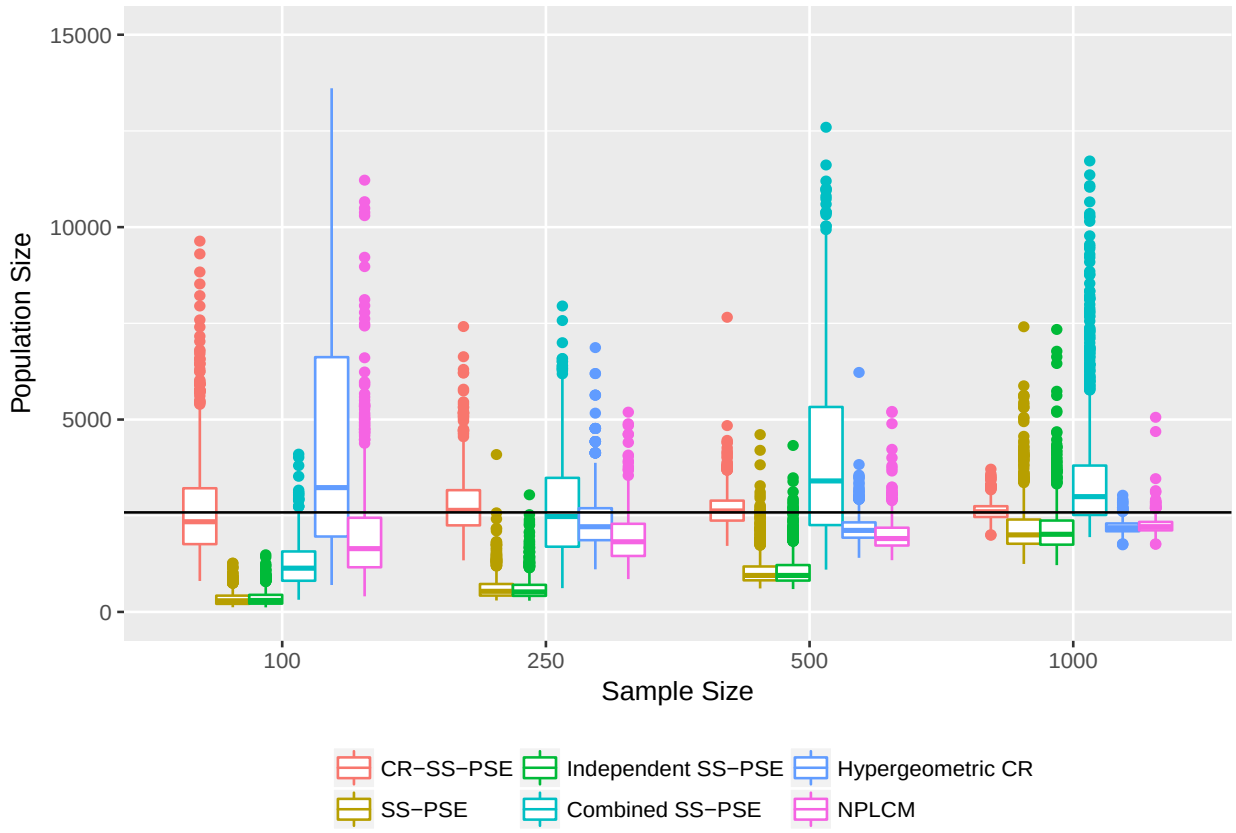



Figure 5.7: A look at the estimation methods using different sample fractions with the Add Health network. Sample size is the size of each list ($n'$ and $n''$).

As would be expected, the width of the distribution of posterior means tends to get smaller as the sample size goes up. For CR-SS-PSE, this is essentially the only thing that changes, as it is still centered at the true population size regardless of the sample size. The SS-PSE methods first underestimated, then were accurate, then overestimated as the sample size went up. The trend went the opposite way for the Hypergeometric Capture-Recapture method, as it underestimated for higher sample sizes.

## 5.8 Performance with Mismatched Sample Sizes

In the cases we have looked at thus far, we have used two lists with equal sample sizes (i.e. $n' = n''$). In reality, this may not be feasible. For example, one list may have to be cut short for various reasons, such as running out of funds. Alternatively, the second list may be able to recruit more people than expected, and the sample size may be increased.

Because of this, we want to assess the performance of our method when we have very different sample sizes, such as $n' < n''$. To do this, we took three different types of samples: one with $n' = 200, n'' = 800$, one with $n' = n'' = 500$, and one with $n' = 800, n'' = 200$. This essentially looks at how the methods performs with different balances of 1000 responses (with non-unique respondents) between the two lists. We used only CR-SS-PSE, Hypergeometric Capture-Recapture, and NPLCM for this, since SS-PSE does not use multiple lists. As before, we used 1000 trials for each of these cases.

The posterior means are shown in Figure 5.8. We can see that all of the same trends seem to hold for each method. There are very slight differences in each, but generally, the CR-SS-PSE estimates tend to be the best, with the other two tending to underestimate.

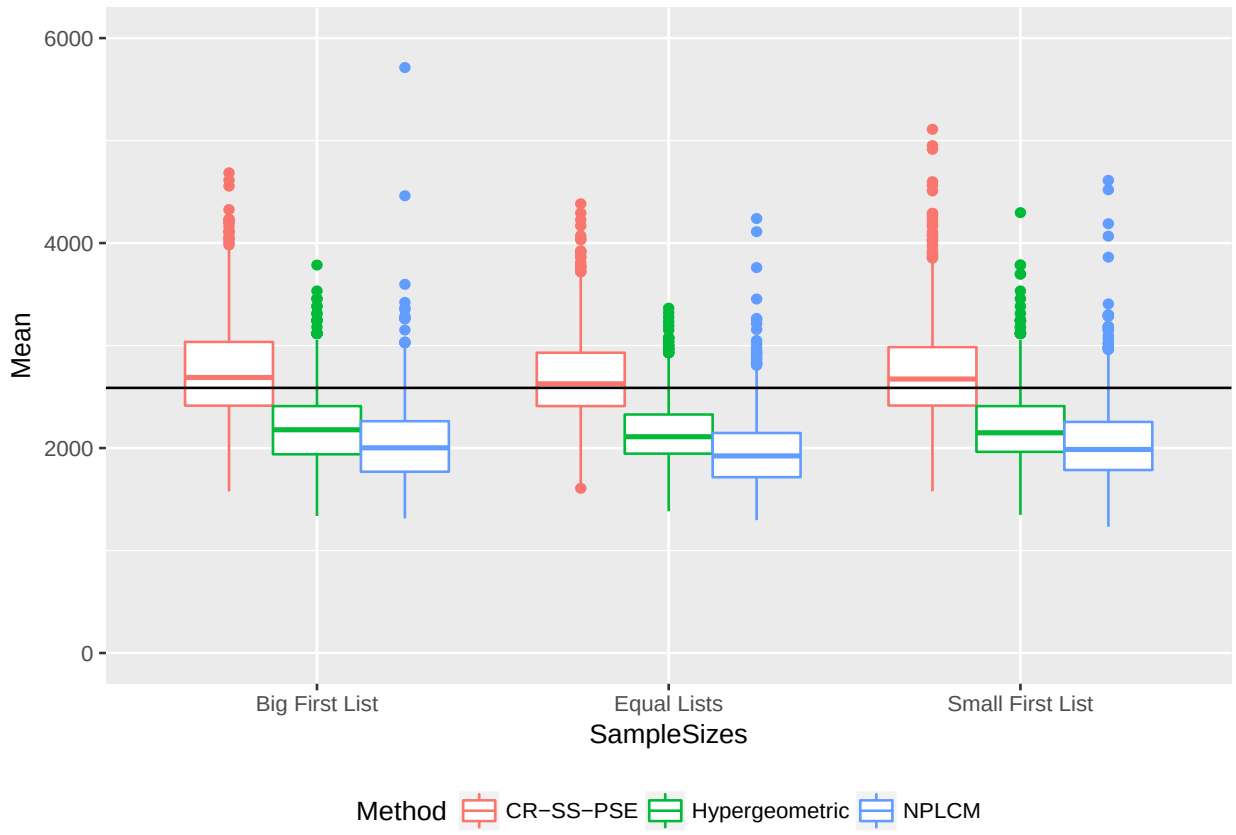



Figure 5.8: Posterior means from three estimation methods using different sample sizes for each list using the Add Health network. Only multiple list methods were included (no SS-PSE).

Importantly, it seems as though the results are very similar across different balances of the two lists.

## 5.9 Using Simulated Networks from the National HIV Behavioral Surveillance Studies

We want the simulated population networks to be as close to reality as possible. In order to try to emulate real networks, we used simulated networks that were based on RDS studies for people who inject drugs (PWID) by the CDC's National HIV Behavioral Surveillance system (NHBS) in 2009 and 2012 (Centers for Disease Control and Prevention 2012, Centers for Disease Control and Prevention 2015). There were 20 U.S. cities sampled in each year, for a total of 40 samples. To create the simulated networks, four network characteristics were measured: the prevalence of a binary trait, HIV infection, homophily for that trait, mean degree, and differential activity. Using the estimated mean values of these characteristics from the RDS studies, natural parameters were calculated for the ERGMs, which were then used to simulate many networks for each of the 40 cities: 1000 networks for the $N = 10000$ case and 250 each for the $N = 5000$ and $N = 1000$ cases. A summary of the network characteristics used is given in Table 5.4 (taken from Spiller et al.).

For each of these networks, one RDS trial was simulated using the same number of starting seeds, number of coupons given out, number of seeds with or without the trait, and the distribution of the number of recruits as was used in the original RDS study. Every population size estimation method in this section is the same as in previous sections in terms of methods and priors used.

| Characteristic | Mean | SD | Median | Min. | Max. |
|---|---|---|---|---|---|
| Prevalence | 0.104 | 0.0653 | 0.091 | 0.018 | 0.286 |
| Mean Degree | 10.64 | 5.096 | 9.88 | 4.45 | 35.39 |
| Homophily | 1.226 | 0.2281 | 1.19 | 0.91 | 1.99 |
| Differential Activity | 0.931 | 0.2098 | 0.92 | 0.53 | 1.44 |
| Sample Size | 519.1 | 108.85 | 539.5 | 206 | 700 |
| Number of Seeds | 8 | 3.31 | 8 | 3 | 16 |
| Number of Seeds With Trait | 1.1 | 1.18 | 1 | 0 | 5 |
| Number of Seeds Without Trait | 6.8 | 3.21 | 7 | 1 | 16 |
| Number of Seeds Missing Trait | 0.13 | 0.404 | 0 | 0 | 2 |
| % of Coupons Returned | 30.60% | 6.60% | 33.20% | 20.00% | 49.80% |
| Number Recruits = 0 | 33.90% | 7.01% | 35.50% | 21.40% | 48.00% |
| Number Recruits = 1 | 21.80% | 5.07% | 22.10% | 9.10% | 32.10% |
| Number Recruits = 2 | 17.70% | 3.54% | 18.20% | 10.00% | 25.10% |
| Number Recruits = 3 | 10.50% | 2.69% | 10.00% | 4.60% | 16.00% |
| Number Recruits = 4 | 1.70% | 2.00% | 0.67% | 0% | 7.70% |
| Number Recruits = 5 | 0.54% | 0.65% | 0.30% | 0% | 2.40% |

Table 5.4: Table of network characteristics for the CDC PWID networks. Number of Recruits percentage are among sample members who were given coupons. Some studies had a maximum of three coupons; the counts were set to 0 for the purposes of the last two rows.

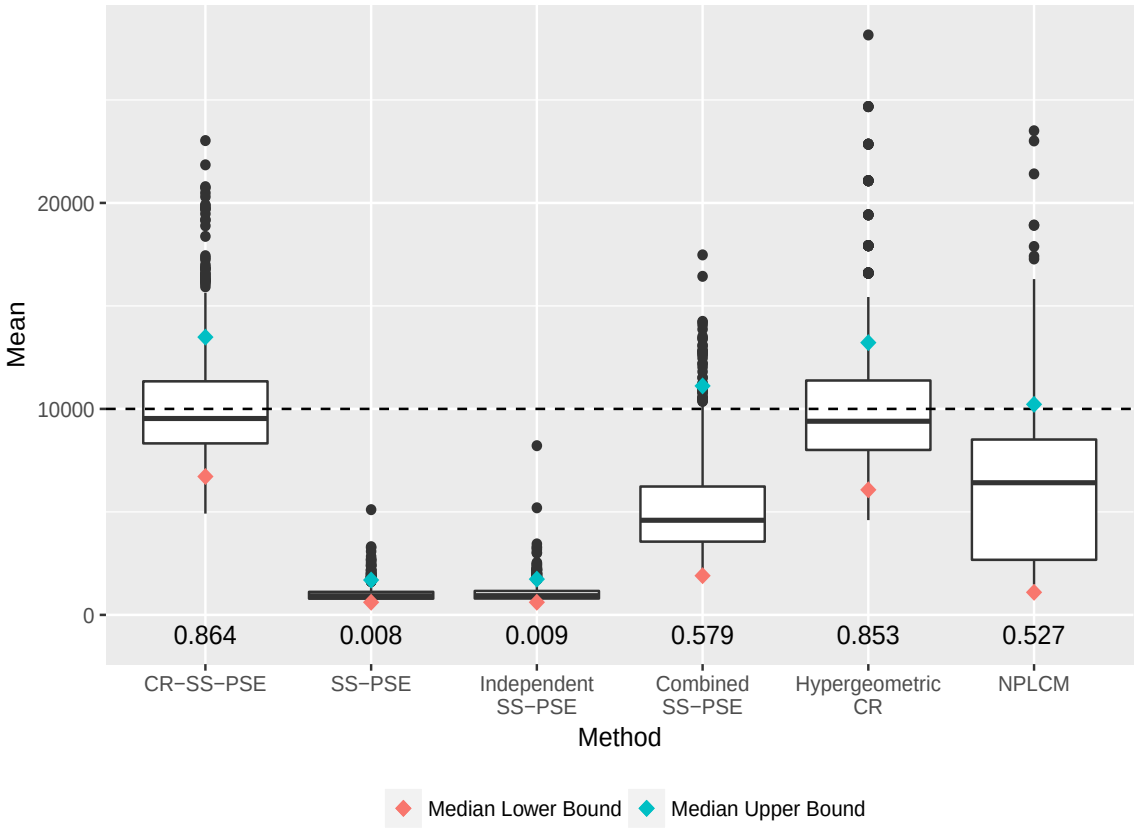



Figure 5.9: Posterior means from one city in the CDC PWID data set. The value underneath each boxplot represents the proportion of 95% SCIs containing the true population size. For each method, two points representing the median upper bound of the 95% SCIs and the median lower bound of the 95% SCIs are also shown. The dashed line shows the true population size ($N = 10000$).

We looked at the posterior means for one of the cities in Figure 5.9. We again see that CR-SS-PSE does a good job of estimating the population size, while the SS-PSE methods do not perform well. In this case, we also see that the Hypergeometric Capture-Recapture method also performs well and in fact very similarly to the CR-SS-PSE. In general, we see that most of the methods tend to underestimate the population size, with the SS-PSE and Independent SS-PSE models performing the worst.

### 5.9.1 Simulations with $N = 10000$

Figure 5.10 shows the posterior means from all 40 cities with a population size of $N = 10000$, with the point showing the median posterior mean and the line representing the middle 95% of the posterior means (that is, the 97.5 and 2.5 percentile). We see much of the same trends that we saw in the individual case. The CR-SS-PSE method seems to be doing very well, though the Hypergeometric Capture-Recapture model seems to be doing very similarly. In general, the centers of the distribution of posterior means for these two methods are at the true population size of 10000, but for a few cities, CR-SS-PSE does tend to underestimate slightly, while the Hypergeometric Capture-Recapture tends to overestimate slightly.

The SS-PSE methods do not perform well in any of the cases, consistently underestimating in all 40 cities. We suspect this may be because of the priors — since the sample fraction is so small, the SS-PSE methods may have a population size posterior distribution very similar to that of the population prior distribution. To analyze this further, we looked at just the SS-PSE methods for the same city as the example shown at the end of Section 5.9. WIth this city, we used a uniform prior with $N_{\max} = 40000$, representing that we have

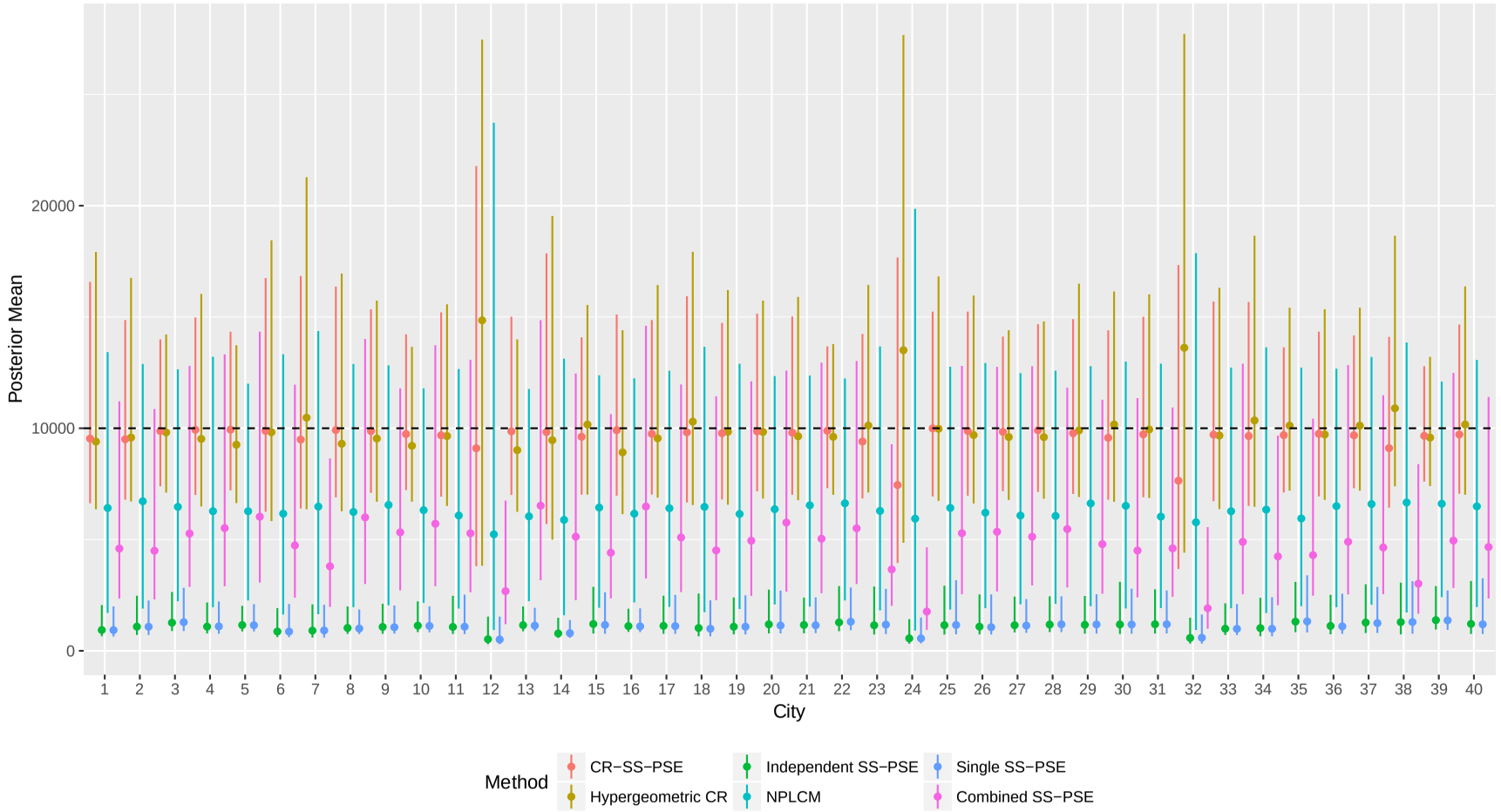



Figure 5.10: Posterior means from the 40 simulated cities for $N = 10000$. The median and middle 95% of the posterior mean are shown with the point and line for each method.

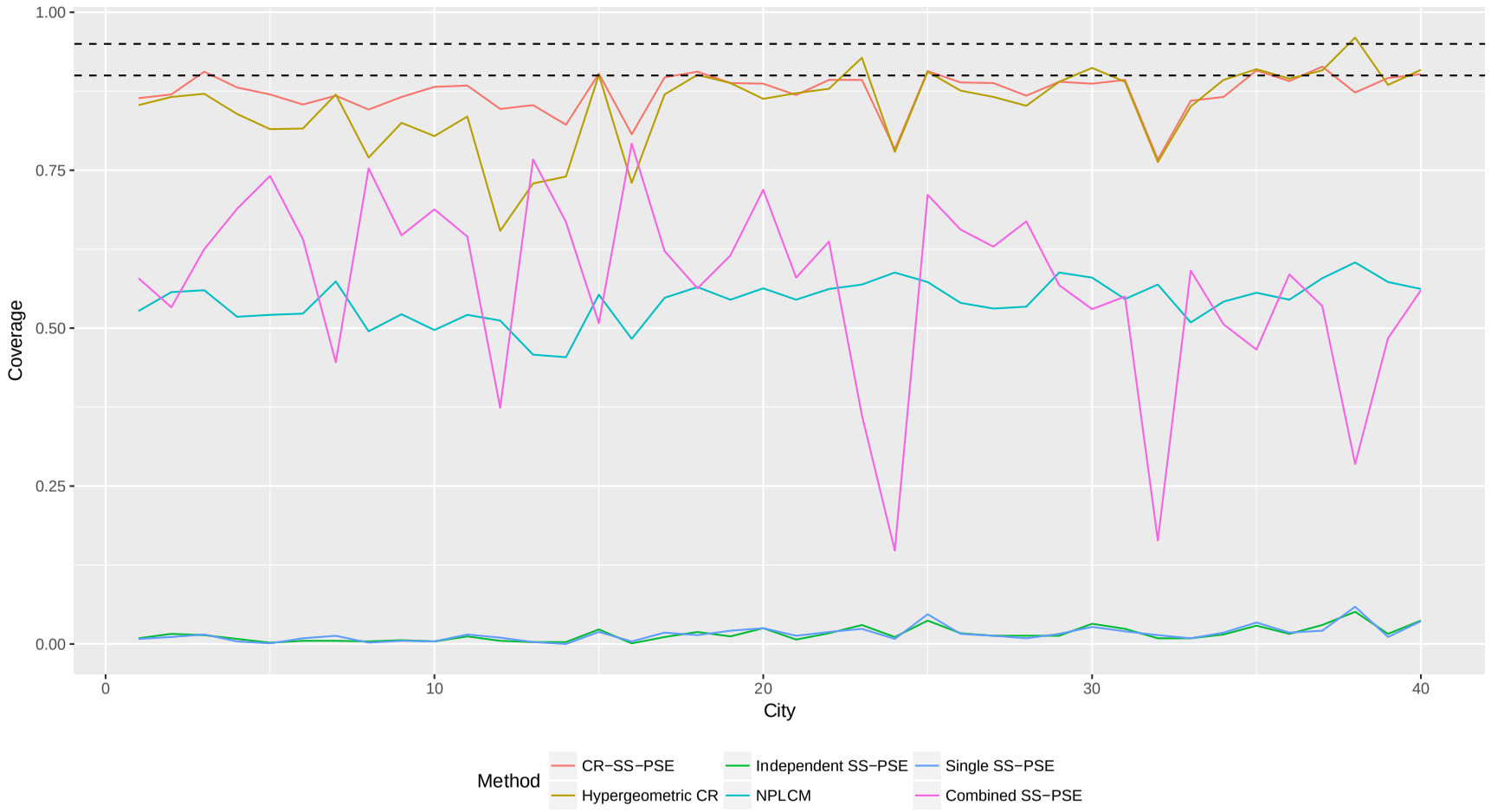



Figure 5.11: Coverage results from the 40 simulated cities for $N = 10000$. Dashed lines show 90% and 95% coverage.

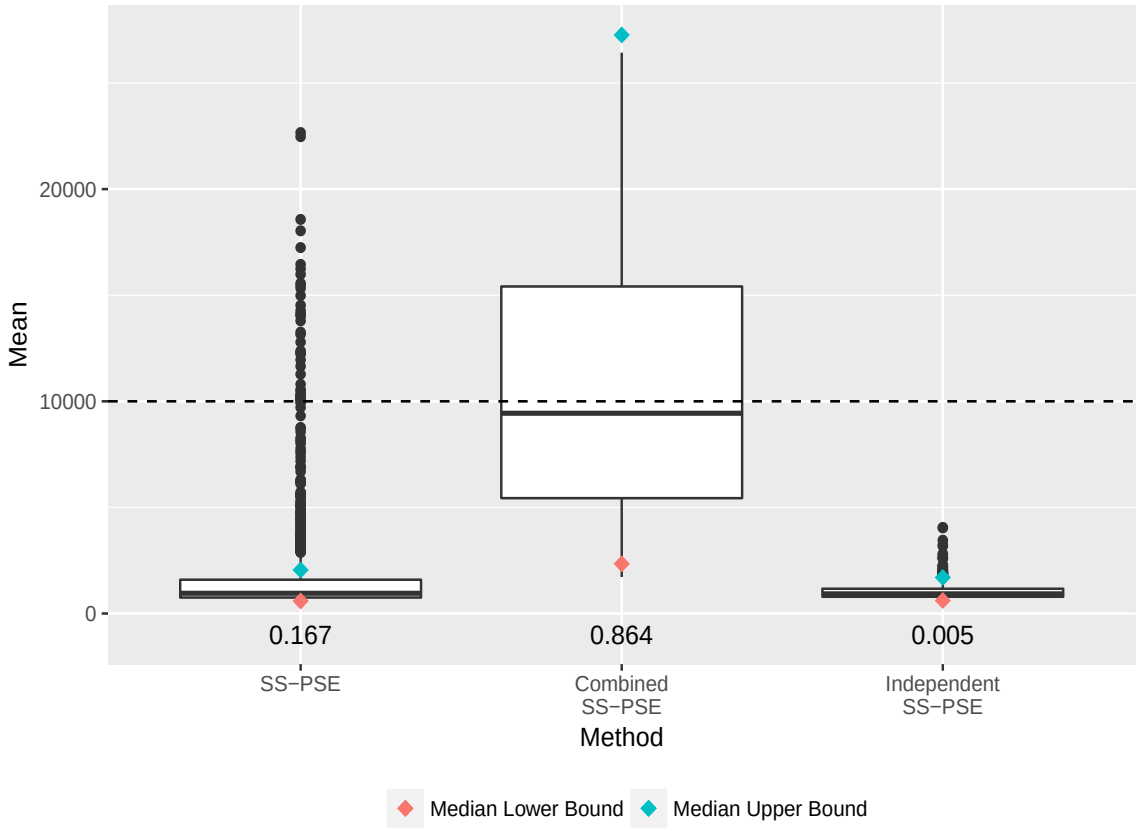



Figure 5.12: Posterior means from SS-PSE methods using uniform priors from one city in the CDC PWID data set. The value underneath each boxplot represents the proportion of 95% SCIs containing the true population size. For each method, two points representing the median upper bound of the 95% SCIs and the median lower bound of the 95% SCIs are also shown. The dashed line shows the true population size ($N = 10000$).

very little prior knowledge of the population size. The posterior means of this are shown in Figure 5.12. We see that the SS-PSE methods actually perform much better in this case, with Combined SS-PSE actually performing relatively well despite quite a high variance in posterior means. However, CR-SS-PSE still outperforms the SS-PSE methods, even with better priors.

Figure 5.11 shows the coverage for each of the 40 cities. Though none of the methods reach 95% coverage except for one case with the Hypergeometric Capture-Recapture method, we see that CR-SS-PSE and HCR are both very similar and relatively high, getting close to the 90% coverage level. As can be expected, the performance of the other methods by this metric falls off considerably.

### 5.9.2 Simulations with $N = 1000$

Figure 5.13 shows the results for $N = 1000$. We see that CR-SS-PSE again performs very well. The SS-PSE methods this time generally overestimated, with the centers of the posterior means tending to be around 1500, with a few exceptions. Hypergeometric Capture-Recapture and NPLCM both do better, though Hypergeometric Capture-Recapture does tend to slightly underestimate the population size.

Looking at the coverages in Figure 5.14, we can see that the Hypergeometric Capture-Recapture method is the worst, with generally very poor coverage even though the point estimates are actually fairly close to the true value. The SS-PSE methods tended to have fairly high coverage rates despite poor point estimates because they had very wide SCIs. CR-SS-PSE coverage rates tended to hover between 90% and 95%.

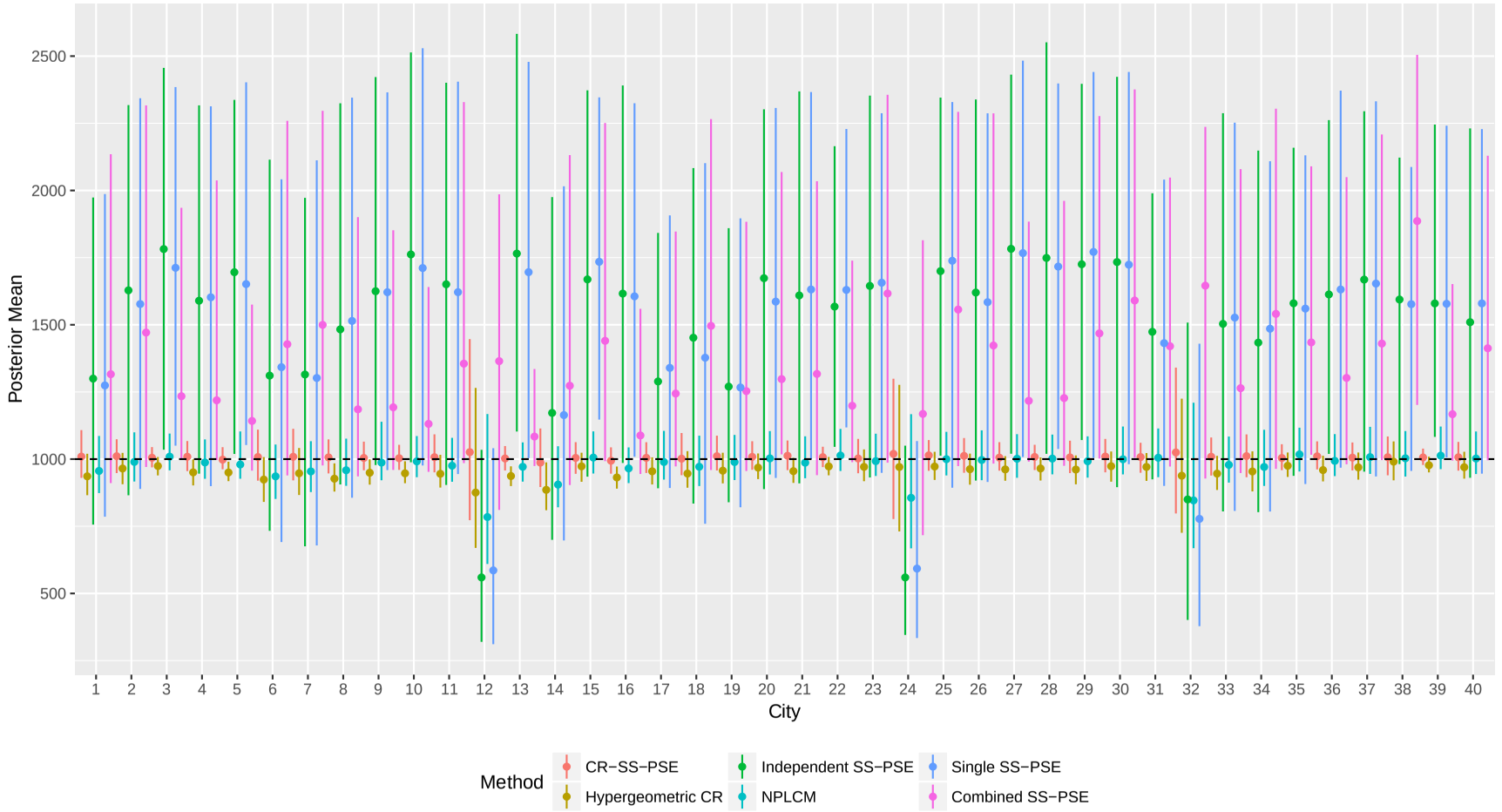



Figure 5.13: Posterior means from the 40 simulated cities with $N = 1000$. The median and middle 95% of the posterior mean are shown with the point and line for each method.

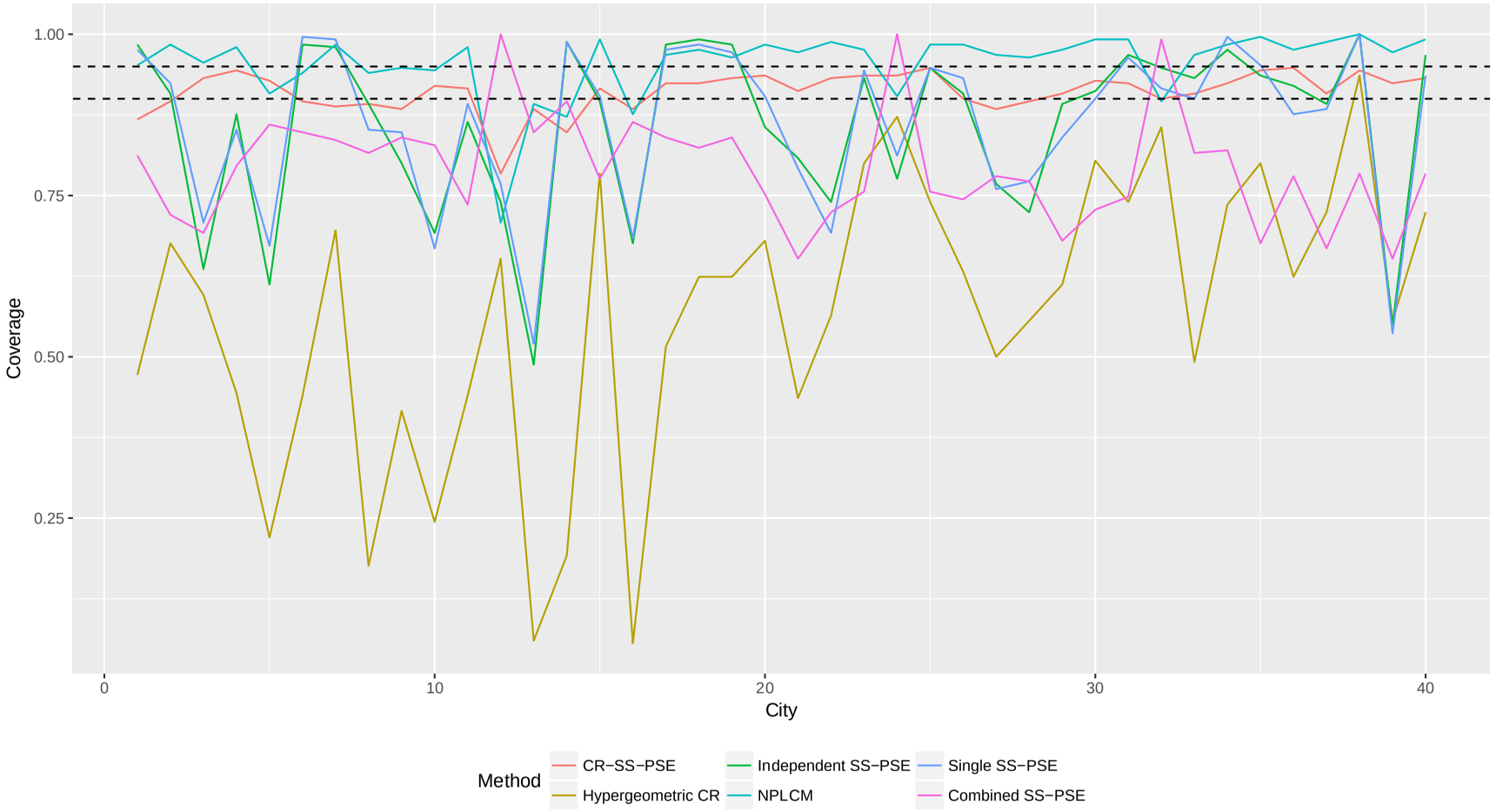



Figure 5.14: Coverage results from the 40 simulated cities for $N = 1000$. Dashed lines show 90% and 95% coverage.

### 5.9.3 Simulations with $N = 5000$

Figure 5.15 shows the results from the simulations with $N = 5000$. As before, CR-SS-PSE does well, with the median of the population size estimates generally being very close to the true population size of $N = 5000$. The SS-PSE methods generally do not perform well, with two of them underestimating significantly. NPLCM also tends to underestimate, though it does tend to be closer to the true value. Hypergeometric Capture-Recapture does relatively well, though not as well as CR-SS-PSE.

Looking at Figure 5.16, we see that Hypergeometric Capture-Recapture method again does very poorly on coverage, suggesting it does not do a good job of estimating the variance. There doesn't seem to be one method that dominates in terms of coverage, but notably, NPLCM does show good coverage despite tending to underestimate on the point estimate. Once again, CR-SS-PSE coverage rates are fairly consistently between 90% and 95%.

Figure 5.17 shows the CR-SS-PSE posterior means from all three population sizes ($N =$ 1000, 5000, 10000). For each of the different population sizes, CR-SS-PSE had good point and interval estimates, and though there were cases in which other methods may have performed just as well, they were not consistent across all three population sizes. We also note that the coverage rates for the CR-SS-PSE were relatively consistent, with lower spikes and drops compared to the other methods, and they were generally around 90% coverage.

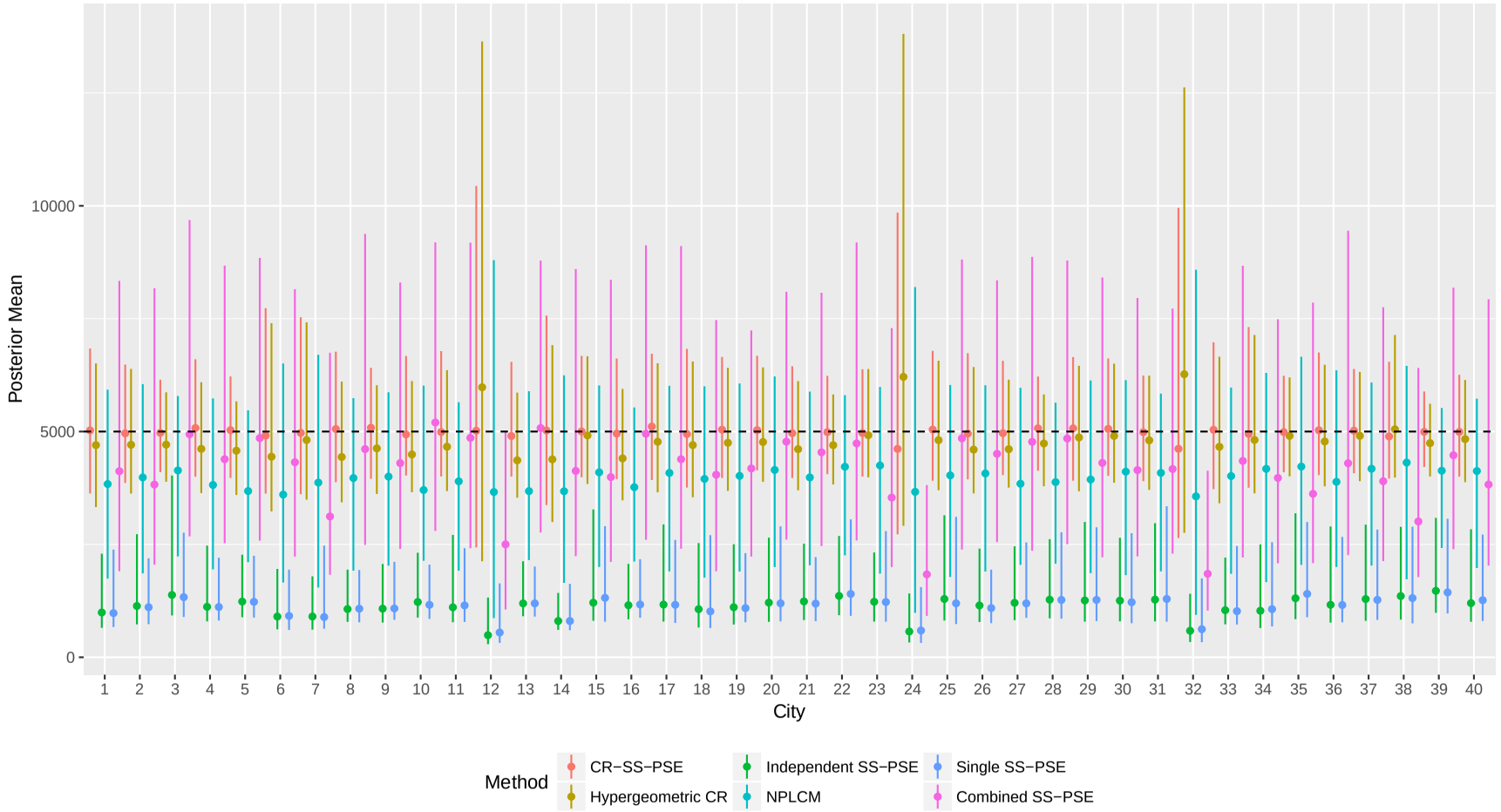



Figure 5.15: Posterior means from the 40 simulated cities with $N = 5000$. The median and middle 95% of the posterior mean are shown with the point and line for each method.

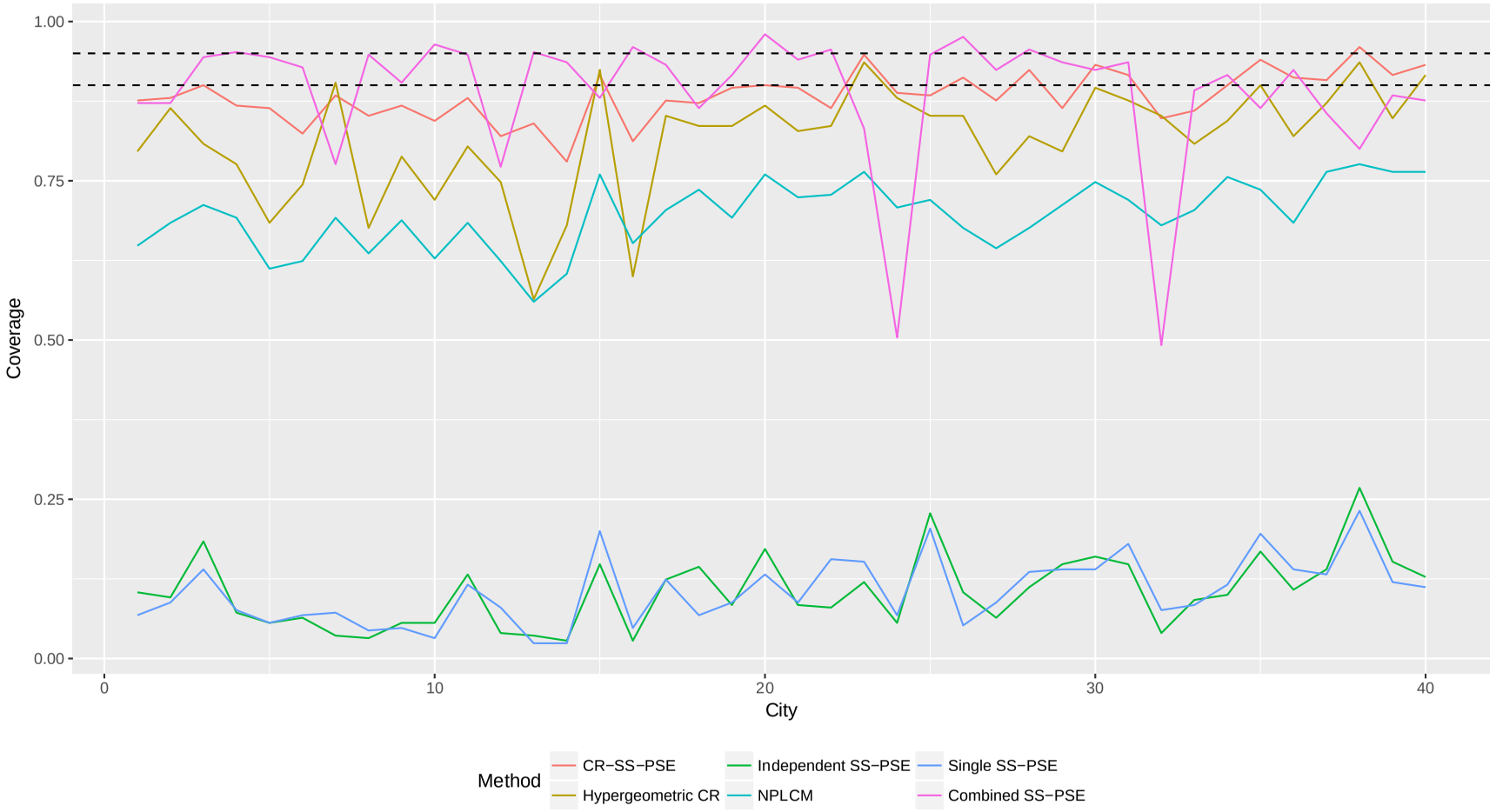



Figure 5.16: Coverage results from the 40 simulated cities for $N = 5000$. Dashed lines show 90% and 95% coverage.

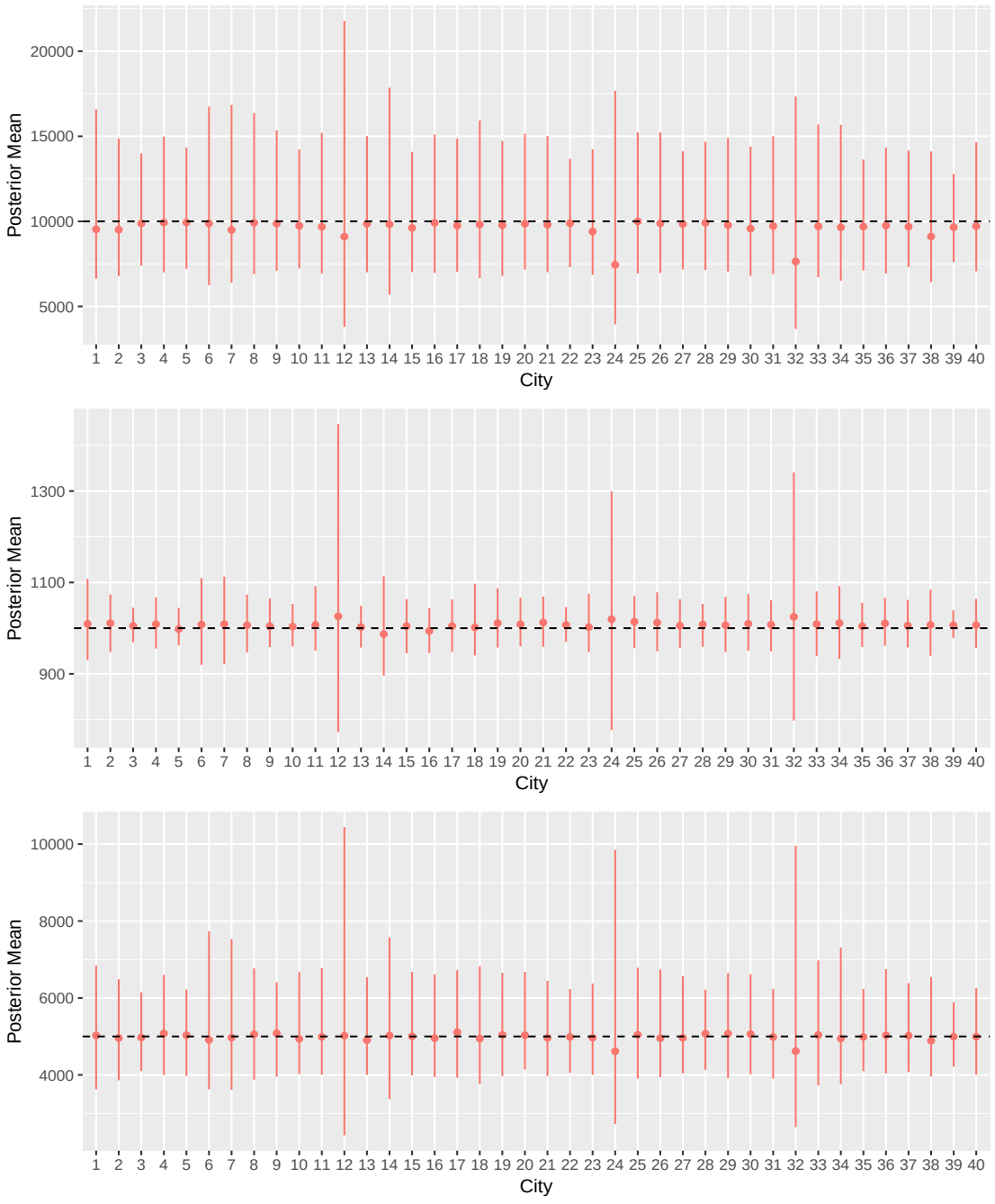



Figure 5.17: Posterior means for CR-SS-PSE for all three population sizes ($N = 10000, 1000, 5000$). The median and middle 95% of the posterior mean are shown. The dashed line shows the true population size.

## 5.10 Empirical Analysis of Frequentist Properties

In general, we see that the 95% SCI regions don't have 95% coverage. This does not come as a surprise, since the SCI regions do not have the same interpretation as confidence intervals. However, developing a method of generating intervals that act like 95% (or other) confidence intervals from the posterior distribution can still be useful for researchers, since they can use some of the same interpretations of confidence intervals with these Bayesian credible intervals.

For this reason, we use a process called calibrated Bayes (Little 2011, Rubin 1984). We will approach this empirically, varying the SCI levels and checking when the credible interval contains the population size 95% of the time. For the simulations above, we looked at the coverage rates of a range of SCI regions from 95% to 100% in increments of 0.2%. We calcuated the coverage using each of the probability regions in this range, and found where the coverage reached 95%. We also looked at how wide intervals could get if we increased the SCI level.

### 5.10.1 Faux Sycamore, Faux Madrona, and Add Health

Figure 5.18 shows the coverage rate of the SCIs by the width of the credible intervals for Faux Sycamore, Faux Madrona, and Add Health networks. A dashed line shows 95% coverage, and the point at which one of the solid lines crosses the dashed line represents the credible interval width at which the credible interval acts like a 95% confidence interval.

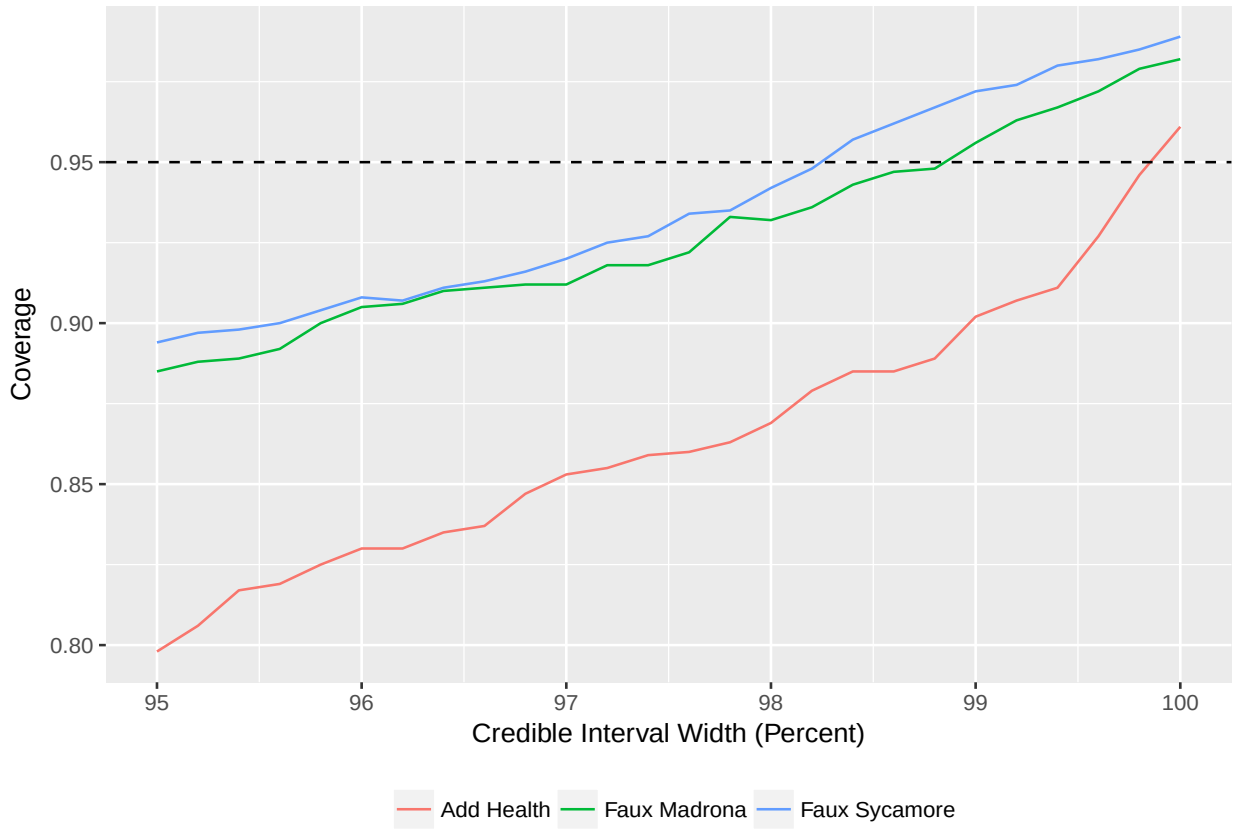



Figure 5.18: The coverage of SCIs by width for Faux Madrona, Faux Sycamore, and Add Health networks. The dashed line indicates 95% coverage.

### 5.10.2 Calibrated Bayes with NHBS Networks

Figure 5.19 shows the plot of SCI and coverage for each of the 40 cities. We see that for many of them, a fairly high SCI region is needed, with most of the lines only crossing the dotted line representing 95% coverage above around 99% SCI region. We note that there are some cases in which it may seem as though the coverage actually drops as SCI percentage goes up. This is due to the fact that we're using draws from the posterior distribution along with artificial multi-modality resulting from the discrete nature of $N$. This means that a slight shift in the SCI percentage can actually cause the interval calculated to shift rather than exhibiting a nesting behavior. If we were to observe the actual posterior distribution instead of using draws, we would expect the lines to be monotonically increasing, and we should treat what is shown in Figure 5.19 as approximations.

We also looked at how wide the intervals would get if we were to use a higher SCI region to determine the effect of using a wider SCI. Figure 5.20 shows the median upper and lower bound when using a 95% and 100% SCI region. The bounds tended to be around 1.5 times larger for the 100% region. Note that the $12^{th}$, $24^{th}$, and $32^{nd}$ SCIs get much wider compared to the other cities. This is due to the much smaller sample size in those samples — they had sample sizes ($n' = n''$) of 206, 210, and 211, respectively, while the others had sample sizes ranging from 423 to 700.

Figure 5.21 shows the results from the NHBS simulations using $N = 1000, 5000, 10000$. Generally, they all seem to require around at least a 99% credible interval in order to reach 95% coverage, and some don't reach 95% coverage even at the 100% SCI level. In Figure 5.22, we see that the bounds for 95% and 100% are generally relatively close, which suggests

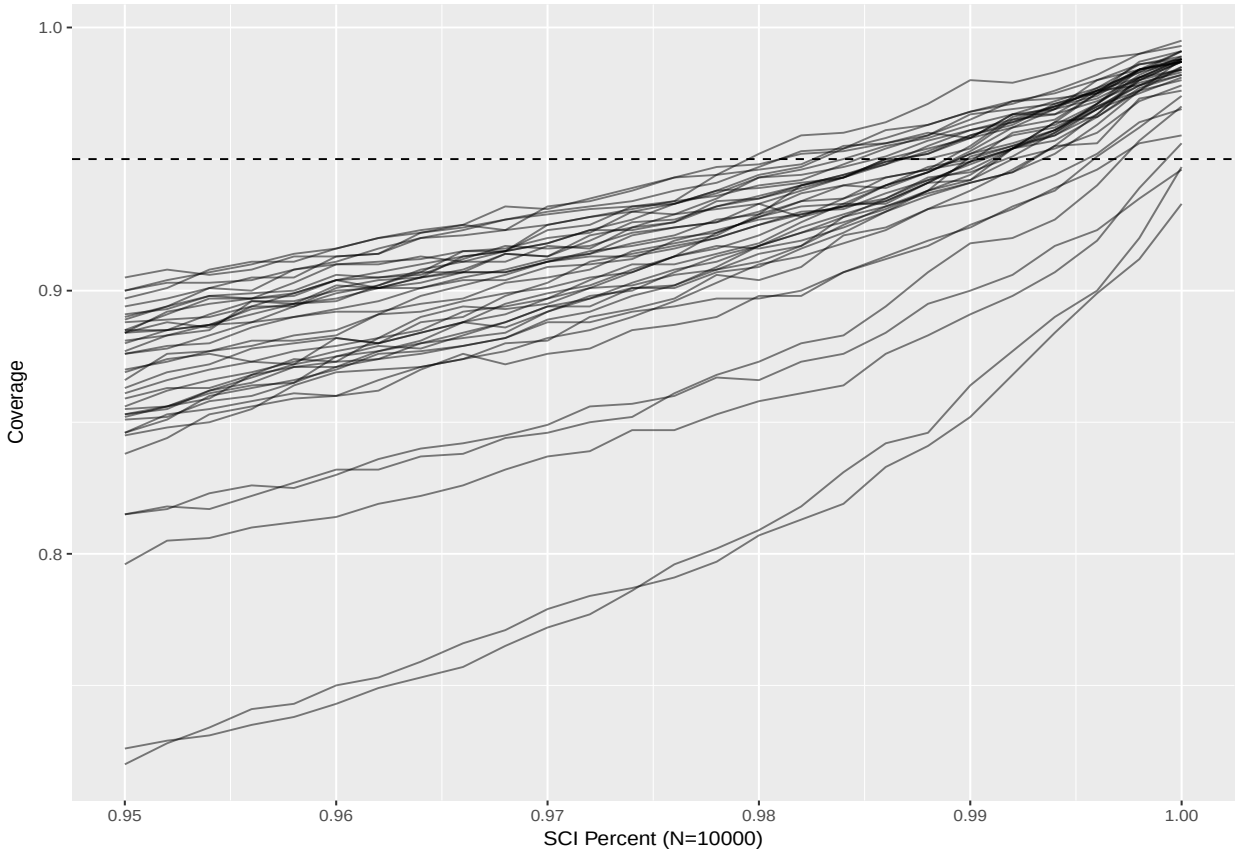



Figure 5.19: The coverage of SCIs by width for 40 Cities data for $N = 10000$. The dashed line indicates 95% coverage.

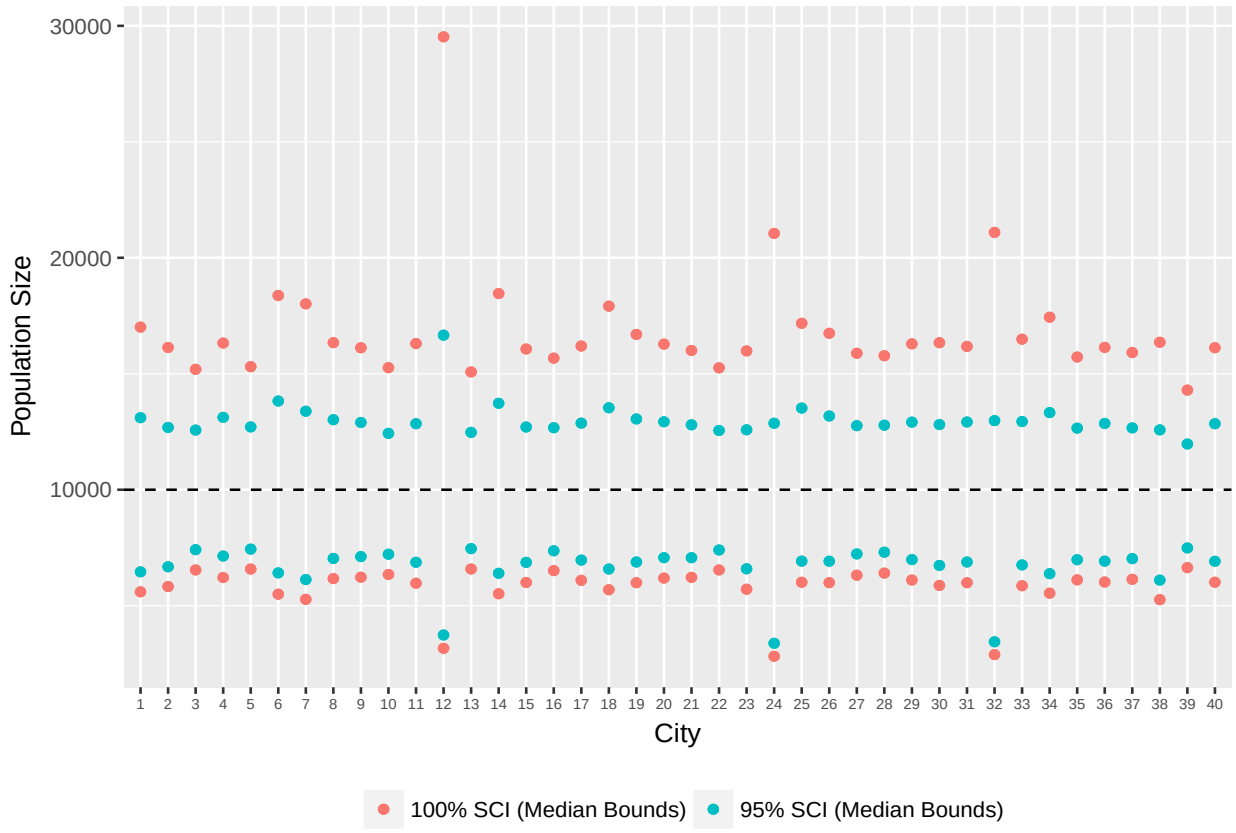



Figure 5.20: The median of the upper and lower bounds for the 95% and 100% SCI regions for each of the 40 cities for $N = 10000$.

increasing the SCI level would not result in wildly different intervals. We see the same trend of wider intervals for city 12, 24, and 32 due to the smaller sample size in each of the different population sizes.

Overall, based on the coverage rates and relative widths of the 99% or more SCIs for each of these cases, it seems reasonable to use something around a 99% SCI in order to get closer to an interval estimate that acts similar to a 95% confidence interval.

## 5.11 Conclusions

Table 5.5 shows a summary of the simulation results presented in this chapter based on the point estimates of each of the methods. Overall, we see that CR-SS-PSE consistently performs well in all scenarios. Though there are cases in which other methods perform just as well — for example, in the 40 PWID case with $N = 1000$, Hypergeometric Capture-Recapture has very similar results — CR-SS-PSE is the only one that consistently provides good estimates across different network structures and sample fractions.

We note that a shortcoming of both the Hypergeometric Capture-Recapture model and NPLCM is that both models rely on a fixed probability of capture (even if there are different probabilities for each unit) regardless of when they were sampled — that is, they do not take into account the without-replacement nature of RDS. This could explain why they both do relatively poorly, as well as why they generally tend to underestimate, since the probability of capture gets higher for each unit later on in the chain. On the other hand, the SS-PSE methods have extremely wide intervals due to the relatively low amount of information (since it uses only one list and doesn't use any recapture information). This means that though

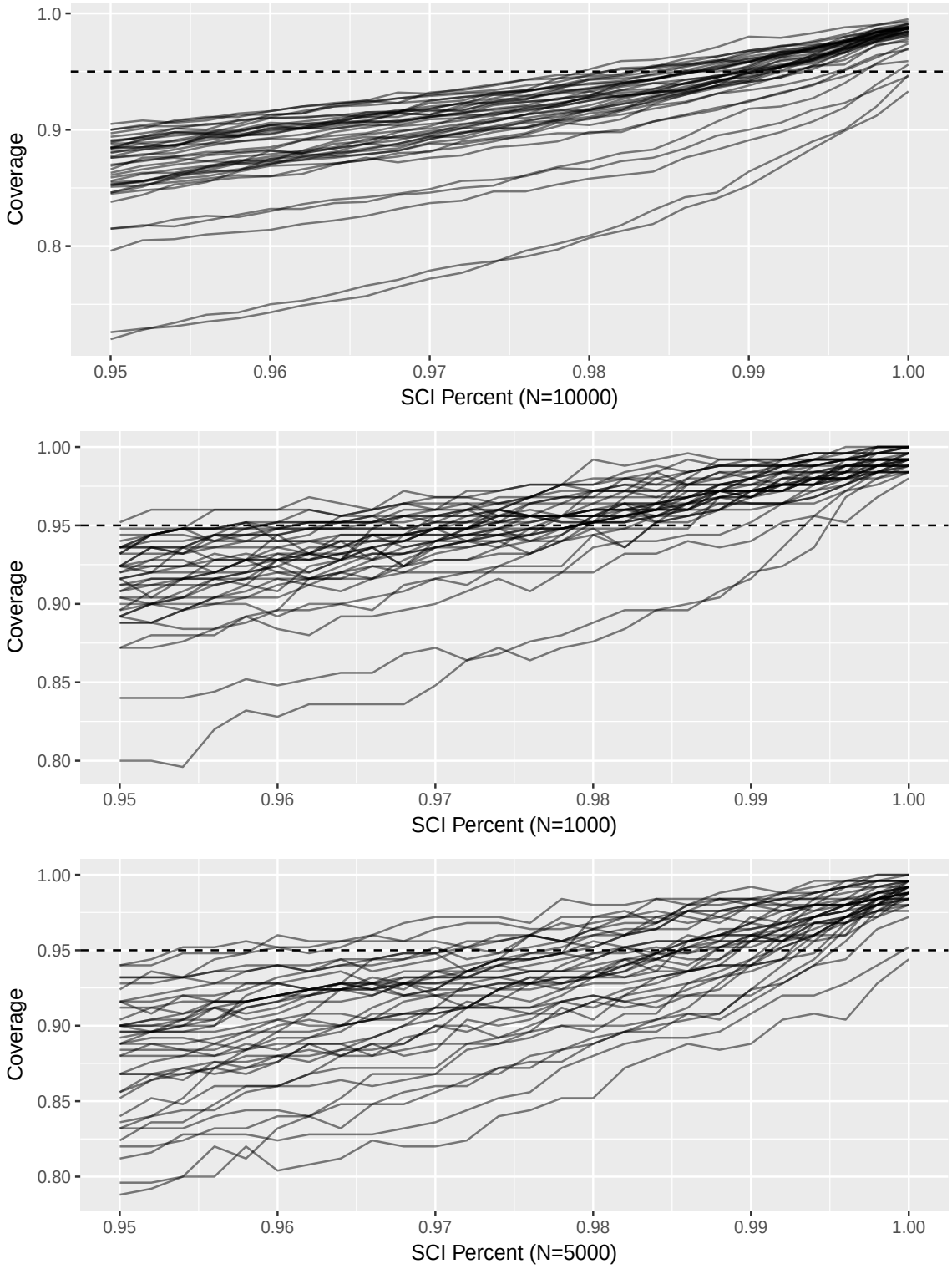



Figure 5.21: The coverage of SCIs by width. The dashed line indicates 95% coverage.

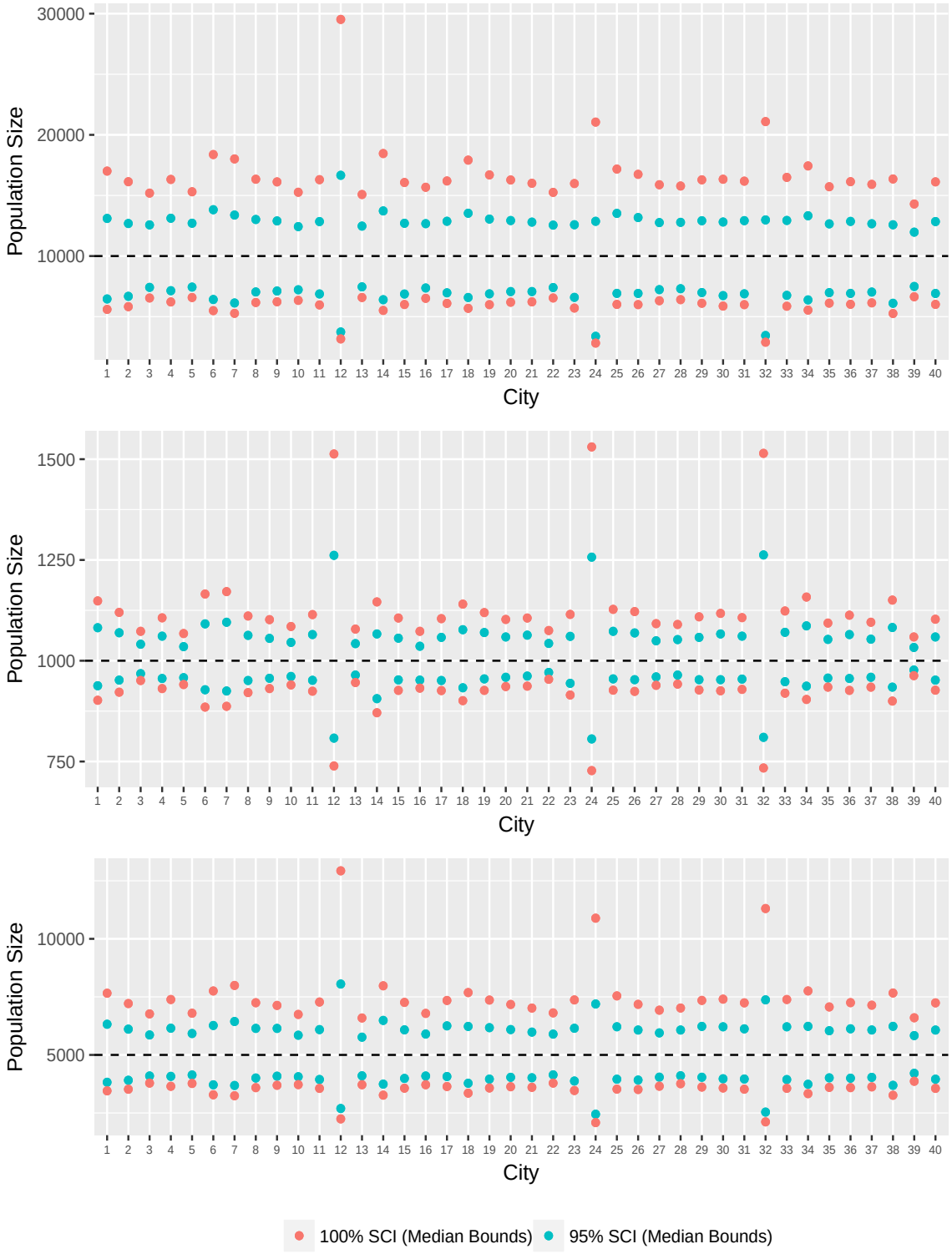



Figure 5.22: The median of the upper and lower bounds for the 95% and 100% SCI regions.

one might be confident that the true population size is within the interval, this interval is too wide to be of much practical use in many cases.

CR-SS-PSE, however, combines the benefits of both types of methods and gives good point estimates with intervals that are not too wide. Along those lines, we also used a calibrated Bayes approach to design a method for creating interval estimates for the population size that act like 95% confidence intervals by using the 99% credible interval.

| Network | CR-SS-PSE | SS-PSE | Indep. SS-PSE | Combined SS-PSE | HCR | NPLCM |
|---|---|---|---|---|---|---|
| Faux Sycamore | Accurate | Under | Under | Over | Under | Under |
| Faux Madrona | Accurate | Under | Under | Over | Under | Under |
| Add Health ($n' = 100$) | Accurate | Under | Under | Under | Accurate | Under |
| Add Health ($n' = 250$) | Accurate | Under | Under | Accurate | Under | Under |
| Add Health ($n' = 500$) | Accurate | Under | Under | Over | Under | Under |
| Add Health ($n' = 1000$) | Accurate | Under | Under | Over | Under | Under |
| PWID 40 Cities ($N = 1000$) | Accurate | Over | Over | Over | Under | Accurate |
| PWID 40 Cities ($N = 5000$) | Accurate | Under | Under | Under | Under | Under |
| PWID 40 Cities ($N = 10000$) | Accurate | Under | Under | Under | Accurate | Under |

Table 5.5: Table of results for simulations in this chapter by whether they were **Accurate**, tended to **Over**estimate, or tended to **Under**estimate. Accurate results are shown in green while inaccurate (either over or underestimate) are shown in red.

# CHAPTER 6

# Future Work and Conclusion

In this dissertation, I have proposed a new method of estimating the size of a hidden population using multiple RDS surveys. This model, called Capture-Recapture Successive Sampling for Population Size Estimation, or CR-SS-PSE, has been shown to have good properties when applied to simulated networks, and has outperformed many existing methods. RDS has become more common, partially because of the relative ease with which RDS surveys can be implemented. CR-SS-PSE uses both information about the sampling order as well as recapture information to provide better estimates of the population size than existing methods.

Further improvements to CR-SS-PS could involve various changes to more closely match what actually happens in the RDS process. For example, we assume that respondents randomly choose from among their own networks to recruit. However, this may not be the case. In particular, we may see a sort of "status sorting," in which some sort of social class structure greatly affects the recruitments. That is, we might expect respondents with higher social status to recruit people with lower social status relatively easily, but for the reverse to be quite rare. Considering these sorts of factors that affect recruitment — perhaps by treating them as latent variables — could provide a better model for RDS recruitment.

Furthermore, we have assumed the self-reported unit sizes to be accurate. However, this may not be the case, due to faulty reporting, forgetfulness, or simply a human tendancy to prefer round numbers. Further work could be done to account for inaccurate reporting of unit sizes.

In addition, in CR-SS-PSE, we have assumed that we set a target sample size, and that target sample size is hit exactly. However, in practice, this may not be the case. For example, if the sample size is sufficiently large, the RDS process might stop earlier than our predetermined max sample size because the members of the population are being exhausted. Further work could be done to address this assumption.

Another area for possible future work involves the actual sampling scheme used. CR-SS-PSE uses a successive sampling approximation to the RDS process. We use this approximation because it more closely captures the without-replacement nature of the samples being taken in RDS surveys. However, this is simply because the respondent-driven samples currently being collected generally don't allow for repeated observations of the same unit in the same sample. In other words, respondents are typically turned away if they try to come in a second time.

Though we have developed a method that tries to model the RDS process as closely as possible in this way, we can also adjust the respondent-driven sampling scheme to fit estimation methods. One alternative to RDS might involve using the same sampling scheme, but allowing for repeated observations of the same respondent. Then, we would be able to treat RDS as a with-replacement process.

This alternate sampling method comes with other complications. For example, we must

ensure that respondents don't actively seek out the benefits of being in the study multiple times. In addition, two respondents would theoretically be able to recruit each other repeatedly, and even three or more respondents could form a loop. Steps must be taken to prevent this sort of behavior. In addition, despite the possible simplification from being able to treat the sampling process as with-replacement, there would still be issues of dependency, and it is possible that we would need a stronger version of the "homophily weak enough" assumption.

Bibliography

Abdul-Quader, A., Heckathorn, D., Sabin, K., and Saidel, T. Implementation and Analysis of Respondent Driven Sampling: Lessons Learned from the Field. *Journal of Urban Health*, 83(7), 2006.

Andreatta, G. and Kaufman, G. Estimation of Finite Population Properties When Sampling is Without Replacement and Proportional to Magnitude. *Journal of the American Statistical Association*, 81(395):657–666, 1986.

Bengtsson, L., Lu, X., Nguyen, Q., Camitz, M., Hoang, N., Nguyen, T., Liljeros, F., and Thorson, A. Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam. *PLoS ONE*, 7, 2012.

Berchenko, Y. and Frost, S. Capture-Recapture Methods and Respondent-Driven Sampling: Their Potential and Limitations. *Sexually Transmitted Infections*, 87:267–268, 2011.

Bernard, H., Hallett, T., Iovita, A., Johnsen, E., Lyerla, R., McCarty, C., Mahy, M., Salganik, M., Saliuk, T., Scutelniciuc, O., Shelley, G., Sirinirund, P., Weir, S., and Stroup, D. Counting Hard-to-Count Populations: The Network Scale-Up Method for Public Health. *Sexually Transmitted Infections*, 86:ii11–ii15, 2010.

Bickel, P., Nair, V., and Wang, P. Nonparametric Inference Under Biased Sampling from a Finite Population. *The Annals of Statistics*, 20(2):853–878, 1992.

Butts, C. A Relational Event Framework for Social Action. *Sociological Methodology*, 38 (155), 2008.

Centers for Disease Control and Prevention. HIV Infection and HIV-Associated Behaviors Among Injecting Drug Users — 20 Cities, United States, 2009. *Morbidity and Mortality Weekly Report*, 61:133–138, 2012.

Centers for Disease Control and Prevention. HIV Infection and HIV-Associated Behaviors Among Persons Who Inject Drugs — 20 Cities, United States, 2012. *Morbidity and Mortality Weekly Report*, 64:270–275, 2015.

Chapman, D. Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses. *University of California Publications in Statistics*, 1(7):131–160, 1951.

Chesson, J. A Non-central Multivariate Hypergeometric Distribution Arising from Biased Sampling with Application to Selective Predation. *Journal of Applied Probability*, 13(4): 795–797, 1976.

Coleman, J. Relational Analysis: The Study of Social Organizations with Survey Methods. *Human Organization*, 17:28–36, 1958.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. A Three-Sample Multiple Recapture Approach to Census Population Estimation with Heterogeneous Catchability. *Journal of the American Statistical Association*, 88:1137–1148, 1993.

Dunson, D. and Xing, C. Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal of the American Statistical Association*, 104:1042–1051, 2009.

Fienberg, S., Johnson, M., and Junker, B. Classical Multilevel and Bayesian Approaches

to Population Size Estimation Using Multiple Lists. *Journal of Royal Statistical Society*, 162:383–405, 1999.

Fog, A. Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in statictics, Simulation and Computation*, 37(2):258–273, 2008.

Freeman, L. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1): 35–41, 1977.

Freeman, L. Centrality in Social Networks: Conceptual Clarification. *Social Networks*, 1: 215–239, 1979.

Gile, K. *Inference from Partially-Observed Network Data*. PhD thesis, University of Washington, 2008.

Gile, K. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. *Journal of the American Statistical Association*, 106:135–146, 2011.

Gile, K. and Handcock, M. Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology*, 40(1):285–327, 2010.

Gile, K. and Handcock, M. Network Model-Assisted Inference from Respondent-Driven Sampling Data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):619–639, June 2015.

Good, P. *Permutation Tests*. Springer-Verlag New York, 2000.

Goodman, L. Snowball Sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 1961.

Haley, D. et al. Venue-Based Recruitment of Women at Elevated Risk for HIV: An HIV Prevention Trials Network Study. *Journal of Women's Health*, 23(6):541–551, 2014.

Handcock, M. and Gile, K. Comment: On the Concept of Snowball Sampling. *Sociological Methodology*, 2011.

Handcock, M. and Jones, J. Interval Estimates for Epidemic Thresholds in Two-Sex Network Models. *Theoretical Population Biology*, 70(2):125–134, 2006.

Handcock, M., Gile, K., and Mar, C. Estimating Hidden Population Size using Respondent-Driven Sampling Data. *Electronic Journal of Statistics*, 8:1491–1521, 2014.

Handcock, M., Gile, K., and Mar, C. Estimating the Size of Populations at High Risk for HIV Using Respondent-Driven Sampling Data. *Biometrics*, 71:258–266, 2015.

Handcock, M. S. and Gile, K. J. *sspse: Estimating Hidden Population Size using Respondent Driven Sampling Data*. Los Angeles, CA, 2015. URL `http://CRAN.R-project.org/package=sspse`. R package version 0.5-1.

Handcock, M. S., Fellows, I. E., and Gile, K. J. *RDS: Respondent-Driven Sampling*. Los Angeles, CA, 2016. URL `https://CRAN.R-project.org/package=RDS`. R package version 0.7-8.

Harris, K., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J. The

National Longitudinal Study of Adolescent to Adult Health: Research Design, 2009. URL http://www.cpc.unc.edu/projects/addhealth/design.

Heckathorn, D. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44:174–99, 1997.

Heckathorn, D. Snowball Versus Respondent-Driven Sampling. *Sociological Methodology*, 41 (1):355–366, August 2011.

Johnston, L., McLaughlin, K., El Rhilani, H., Toufik, A., Bennani, A., Alami, K., Elomari, B., and Handcock, M. Estimating the Size of Hidden Populations Using Respondent-Driven Sampling Data: Case Examples from Morocco. *Epidemiology*, 26(6):846–852, 2015.

Kendall, C., Kerr, L., Gondim, R., Werneck, G., Macena, R., Pontes, M., Johnston, L., Sabin, K., and McFarland, W. An Empirical Comparison of Respondent-driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveillance in Men Who Have Sex with Men, Fortaleza, Brazil. *AIDS and Behavior*, 12:S97–S104, 2008.

Killworth, P., McCarty, C., Johnsen, E., Bernard, H., and Shelley, G. Investigating the Variation of Personal Network Size Under Unknown Error Conditions. *Sociological Methods and Research*, 35(1):84–112, 2006.

Lansky, A., Drake, A., Wejnert, C., Pham, H., Cribbin, M., and Heckathorn, D. Assessing the Assumptions of Respondent-Driven Sampling in the National HIV Behavioral Surveillance System among Injecting Drug Users. *The Open AIDS Journal*, 6:77–82, 2012.

Little, R. Calibrated Bayes, for Statistics in General, and Missing Data in Particular. *Statistical Science*, 26(2):162–174, 2011.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance. *AIDS*, 19(2):S67–S72, 2005.

Malekinejad, M., Johnston, L., Kendall, C., Kerr, L., Rifkin, M., and Rutherford, G. Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. *AIDS and Behavior*, 12:S105–S130, 2008.

Manrique-Vallier, D. Bayesian Population Size Estimation Using Dirichlet Process Mixtures. *Biometrics*, 72:1246–1252, 2016.

Manrique-Vallier, D. and Fienberg, S. Population Size Estimation Using Individual Level Mixture Models. *Biometrical Journal*, 50:1051–63, 2008.

Nair, V. and Wang, P. Maximum Likelihood Estimation under a Successive Sampling Discovery Model. *Technometrics*, 31(4):423–436, 1989.

Paz-Bailey, G., Jacobson, J., Guardado, M., Hernandez, F., Nieto, A., Estrada, M., and Creswell, J. How Many Men who have Sex with Men and Female Sex Workers Live in El Salvador? Using Respondent-Driven Sampling and Capture-Recapture to Estimate Population Sizes. *Sexually Transmitted Infections*, 87:279–82, 2011.

Platt, L., Wall, M., Rhodes, T., Judd, A., Hickman, M., Johnston, L., Renton, A., Bobrova, N., and Sarang, A. Methods to Recruit Hard-to-Reach Groups: Comparing Two Chain Referral Sampling Methods of Recruiting Injecting Drug Users Across Nine Studies in Russia and Estonia. *Journal of Urban Health*, 83:i39–i53, 2006.

Rivest, L. and Baillargeon, S. Applications and Extensions of Chao's Moment Estimator for the Size of a Closed Population. *Biometrics*, 62:999–1006, 2007.

Rubin, D. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

Salganik, M. and Heckathorn, D. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34:193–239, 2004.

Salganik, M., Fazito, D., Bertoni, N., Abdo, A., Mello, M., and Bastos, F. Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence From a Multiple-Method Study of Heavy Drug Users in Curitiba, Brazil. *American Journal of Epidemiology*, 174:1190–1196, 2011.

Semaan, S. Time-Space Sampling and Respondent-Driven Sampling with Hard-To-Reach Populations. *Methodological Innovations Online*, 5(2):60–75, 2010.

Shmueli, G., Minka, T., Kadane, J., Borle, S., and Boatwright, P. A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:127–142, 2005.

Snijders, T., Pattison, P., and Handcock, M. New Specifications for Exponential Random Graph Models. *Sociological Methodology*, 36(1):99–153, 2006.

Spiller, M., Gile, K., Handcock, M., Mar, C., and Wejnert, C. Evaluating Variance Estimators for Respondent-Driven Sampling. *Journal of Survey Statistics and Methodology*, Forthcoming .

Spreen, M. Rare Populations, Hidden Populations, and Link-Tracing Designs: What and Why? *Bulletin of Sociological Methodology*, 36(1):34–58, 1992.

UNAIDS and Organization, W. H. Guidelines on Estimating the Size of Populations Most at Risk to HIV. Technical Report UNAIDS/00.03E, UNAIDS and World Health Organization, 2010.

Vincent, K. and Thompson, S. Estimating Population Size With Link-Tracing Sampling. *Journal of the American Statistical Association*, 0(0):1–10, 2016.

Volz, E. and Heckathorn, D. Probability-Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics*, 24:79–97, 2008.

Wallenius, K. *Biased Sampling: The Non-central Hypergeometric Probability Distribution*. PhD thesis, Stanford University, 1963.

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

Wattana, W., van Griensven, F., Rhucharoenpornpanich, O., Manopaiboon, C., Thienkrua, W., Bannatham, R., Fox, K., Mock, P., Tappero, J., and Levine, W. Respondent-Driven Sampling to Assess Characteristics and Estimate the Number of Injection Drug Users in Bangkok, Thailand. *Drug and Alcohol Dependence*, 90:228–233, 2007.

West, M. Inference in Successive Sampling Discovery Models. *Journal of Econometrics*, 75: 217–238, 1996.