Estimating the Size of Clustered Hidden Populations

Laura Gamble[1] Katherine McLaughlin[2]

Abstract

Successive sampling population size estimation (SS-PSE) is a method used by government agencies and aid organizations around the world to estimate the size of hidden populations using data from respondent-driven sampling (RDS) surveys. SS-PSE addresses a specific need in estimation and helps us evaluate the vulnerability of areas to HIV and other epidemics by estimating the size of populations that are at higher risk of contracting and spreading HIV. However, SS-PSE relies on several assumptions, one of which requires the underlying social network of the hidden population to be fully connected. This research proposes two modifications to SS-PSE for estimating the size of hidden populations whose underlying social network is clustered.

Key Words: Respondent-driven sampling, successive sampling population size estimation, hidden populations, hard-to-reach populations, network sampling

1. Introduction

Hidden population size estimation is of great interest to aid organizations and public health agencies since many vulnerable populations at higher risk of contracting and spreading HIV and other infectious diseases are hidden. Size estimates for these at-risk populations can help in assessing the magnitude of the HIV epidemic, monitoring the trend of the epidemic in at-risk populations over time, and effectively allocating resources for the most vulnerable members of society (World Health Organization and UNAIDS 2010). The ability to estimate the size of a hidden population is important in global efforts to understand and address the HIV epidemic.

Successive sampling population size estimation (SS-PSE) is one commonly used method to estimate the size of hidden populations (Handcock, Gile, and Mar 2014). SS-PSE uses a Bayesian approach that combines prior expectations with data from respondent-driven sampling (RDS) surveys in order to make probabilistic statements about the unknown population size N. SS-PSE is most often used to estimate the size of vulnerable populations at higher risk of contracting and spreading HIV (Johnston et al. 2015; Weikum et al., 2019; McLaughlin et al. 2019), but it has been recently extended to other hidden populations of interest as well (Johnston et al. 2017; Wesson et al. 2018). SS-PSE is advantageous compared to other population size estimation methods, since it only requires a single data source and can be appended to existing studies with relative ease. In addition, the Bayesian framework allows for the incorporation of external information about the population size from local experts or previous studies.

Gathering information about hidden populations is often challenging in itself. RDS surveys, like those used by SS-PSE, provide one way to sample directly from a hidden population by utilizing the underlying social network of connections between hidden population members. Since RDS uses social connections in its sampling process, the structure of the underlying population network is influential both in implementing an RDS sample and in making inference using RDS data. In modeling the RDS process, SS-PSE relies on the assumption of a connected social network to make accurate inference about the population size. In the case of disconnected or weakly connected networks, the probability

[1]Westat, 1600 Research Blvd, Rockville, MD 20850

[2]Oregon State University, 1500 SW Jefferson Way, Corvallis, OR 97331

model used by SS-PSE may be inappropriate, which leads to bias in point estimation and misleadingly small estimates of variance.

If the connected network assumption can be relaxed, SS-PSE may be applicable for studying other hidden populations which are not as strongly connected as those used in classical applications. It could also be applicable in estimating the size of hidden populations in areas that are naturally divided by geography, socioeconomic variables, or language barriers. Finally, developing a modification to SS-PSE that allows for clustering in the underlying population could have applications in combining city-level population size estimates into a country-level estimate.

Section 2 gives a background discussion on RDS sampling and the existing SS-PSE method. Section 3 introduces two modifications to SS-PSE – the Posterior Sum and Clustered SS-PSE methods – which extend the practical application of existing methodology to clustered populations. Finally, Section 4 contains concluding remarks.

2. Hidden Population Size Estimation

Hidden or hard-to-reach populations are populations for which no sampling frame exists and public acknowledgment of membership is potentially threatening to members (Heckathorn 1997). They tend to be characterized by members' incentive to remain hidden, oftentimes because members practice stigmatized or illegal behaviors. Many hidden populations are also small in size relative to the general population. While the study of hidden populations is important for research regarding the HIV epidemic, these populations pose several unique problems for statistical inference. Since no sampling frame exists, it is difficult and often impossible to obtain a probability sample directly from a hidden population.

2.1 Respondent-Driven Sampling

RDS is an adaptive sampling method that produces a sample directly from a hidden population using the underlying social network of its members (Heckathorn 1997). The sampling process for RDS begins with a convenience sample of "seeds" who are chosen from the population of interest. Generally, these seeds are self-selected volunteers or people purposefully selected by researchers who are believed to be well connected in the target population. Once selected for the study, participants receive a set number of coupons they can use to recruit other members of the target population in their social circle. Coupons with unique identification codes allow researchers to track peer recruitment while maintaining confidentiality. In-person RDS studies then ask coupon recipients to report to a study center, where they participate in the survey and receive their own coupons to distribute. Recruitment continues in this way until the desired sample size is reached. The resulting sample is a collection of trees, or chains, from the underlying social network.

Participants fill out an anonymous survey asking about any variables of interest to the researchers. For at-risk populations to HIV, these surveys typically ask about risk behaviors and experiences of violence or discrimination, among other things (World Health Organization 2017). Participants are also asked how many people in the population of interest they know who also know them. This question is meant to measure each individual's degree in the underlying social network. For at-risk populations to HIV, a free and anonymous HIV test is often included, which acts both as a public service and an incentive for participation in the study.

Because of peer recruitment, design-based inclusion probabilities are not known before data collection, and must therefore be modeled from the observed data. Several such models have been proposed to estimate inclusion probabilities for RDS respondents (Salganik

and Heckathorn 2004; Volz and Heckathorn 2008), including the successive sampling (SS) model proposed by Gile (2011). The SS model is an improvement over older methods since it does not require assumptions that the population is very large relative to the sample size and that the sampling process is done with replacement.

2.2 Successive Sampling Population Size Estimation

SS-PSE is a hidden population size estimation method that only requires data from a single RDS sample. SS-PSE is a Bayesian method that combines each individual's network size and order of recruitment in the sample with prior information gained from local experts or previous studies to model the observed depletion of the population and produce a posterior distribution on the population size N (Handcock, Gile, and Mar 2014; Handcock, Gile, and Mar 2015).

Like other hidden population size estimators, SS-PSE has advantages and disadvantages. Since it only requires sample data about individuals' degrees and their sample order, adding an SS-PSE estimate to an existing RDS study is usually easy and low in cost. Additionally, since SS-PSE is a Bayesian method, it offers the opportunity to incorporate additional sources of information about population size. However, SS-PSE methods also require several strong assumptions. The performance of SS-PSE estimators relies heavily on the underlying population structure and the quality of the RDS data and the prior information used. Results from SS-PSE methods should be interpreted carefully and considered in conjunction with other population size estimates (McLaughlin et al. 2019).

SS-PSE has been used to estimate the size of populations at-risk of contracting HIV in a variety of contexts, including in Morocco (Johnston et al. 2015); Papua New Guinea (Weikum et al. 2019); Armenia (McLaughlin et al. 2019); and Bratislava, Bucharest, Verona, and Vilnius (Johnston et al. 2021). Additionally, it has been used to estimate the number of women in South Kivu Province of The Democratic Republic of Congo who have had sexual violence-related pregnancies (Johnston et al. 2017) and in the United States to estimate the number of transgender women living in San Francisco (Wesson et al. 2018).

To employ the Bayesian SS-PSE method, both a probability model on the observed data and prior distributions on the unknown parameters in that model are necessary. The following sections briefly describe the theoretical framework behind SS-PSE, which helps inform the extension for clustered populations in Section 3.

2.2.1 Probability Model of Observed Data

SS-PSE estimates the inclusion probabilities of an RDS sampling process using the SS approximation, which models the RDS process as a without-replacement random walk through the network. Inclusion probabilities are then calculated over all possible network configurations of a set degree distribution (Gile 2011).

Let $\mathbf{G} = (G_1, ..., G_n)$ be the ordered random vector of observed unit indices, where G_i has support $1, ..., N$. Let $\mathbf{g}$ be the vector of realized unit indices, and let $\setminus\mathbf{g}$ be the set of indices in the population not in $\mathbf{g}$. Let $\mathbf{u} = (u_1, ..., u_N)$ be the vector of unit sizes for each member of the population.

Under the SS model, the first unit is sampled with probability proportional to its unit size u_k:

$$P(G_1 = k) = \frac{u_k}{\sum_{j=1}^{N} u_j}$$

Subsequent units are then sampled with probability proportional to their unit size from the

remaining units:

$$P(G_i = k | G_1 = g_1, ..., G_{i-1} = g_{i-1}) = \begin{cases} \frac{u_k}{\sum_{j \notin \{g_1,...,g_{i-1}\}} u_j} & k \notin \{g_1, ..., g_{i-1}\} \\ 0 & \text{otherwise} \end{cases}$$

This gives the overall probability of observing a sample $\mathbf{g} = (g_1, ..., g_n)$ as

$$P(\mathbf{G} = \mathbf{g}) = \frac{N!}{(N-n)!} \prod_{k=1}^{n} \frac{u_{g_k}}{\sum_{j=1}^{N} u_j - \sum_{j=1}^{k-1} u_{g_j}}$$

Because of the sum $\sum_{j=1}^{N} u_j$, this probability model relies on the degree of all units in the population, some of which are not observed. For this reason, Handcock, Gile, and Mar (2014) add a super-population component to the probability model, wherein the distribution of degrees in the network is considered a random draw from a super-population of possible degree distributions, governed by unknown parameters.

Let $\mathbf{U} = (U_1, ..., U_N)$ be the random vector of unit sizes for all members in the population. We treat $\mathbf{U}$ as an i.i.d. sample from some super-population distribution $f(\cdot|\eta)$ supported on the natural numbers. Let $\mathbf{U}_{obs}$ and $\mathbf{u}_{obs}$ be the random and realized vectors of the observed unit sizes in the sample. Similarly, let $\mathbf{U}_{unobs}$ and $\mathbf{u}_{unobs}$ be the random and realized vectors of the unobserved unit sizes in the population. After adding the super-population component, the full probability model is now

$$\begin{aligned} P(\mathbf{G} = \mathbf{g}, \mathbf{U}_{obs} = \mathbf{u}_{obs} | \eta) &= \sum_{\mathbf{u}_{unobs} \in \mathcal{U}(\mathbf{u}_{obs})} P(\mathbf{G} = \mathbf{g}, \mathbf{U}_{obs} = \mathbf{u}_{obs}, \mathbf{U}_{unobs} = \mathbf{u}_{unobs} | \eta) \\ &= \frac{N!}{(N-n)!} \sum_{\mathbf{u}_{unobs} \in \mathcal{U}(\mathbf{u}_{obs})} \prod_{k=1}^{n} \frac{u_{g_k}}{r_k} f(u_{g_k} | \eta) \prod_{i \in \setminus \mathbf{g}} f(u_i | \eta) \end{aligned} \tag{1}$$

Where $\mathcal{U}(\mathbf{u}_{obs})$ is the set of all possible unobserved unit size vectors $\mathbf{u}_{unobs}$ given the observed unit sizes $\mathbf{u}_{obs}$, and $r_k = \sum_{j=1}^{N} u_j - \sum_{j=1}^{k-1} u_{g_j}$ is the sum of remaining degrees from the unsampled population units at step k of the sampling process.

Since any one of the $N - n$ entries in $\mathbf{u}_{unobs}$ can range over the support of $f(\cdot|\eta)$, this likelihood is generally very difficult to compute, so Handcock, Gile, and Mar (2014) suggest working under the Bayesian framework. They employ a four-component Gibbs sampler to estimate the joint posterior distribution of all model parameters conditioned on the sample data. This simulated distribution can then be marginalized to produce estimates of the population size N, the unit size distribution parameters η, and the full distribution of unknown unit sizes $\mathbf{U}_{unobs}$.

2.2.2 Prior Selection

Since SS-PSE is a Bayesian method, priors need to be selected for each of the model parameters. The unknown parameters in the probability model given in equation 1 are the population size N and the parameters governing the unit size distribution, η. Handcock, Gile, and Mar (2014) suggest modeling the sample proportion $\frac{n}{N}$ as Beta(α, β), where the hyper-parameters α and β can be set by using population size estimates from experts in the area or from information collected in previous studies. It is common practice to use a single point estimate $\hat{N}$ as the median of the prior and fit a distribution to that median using the additional restriction that $\alpha = \frac{\hat{N}}{\hat{N}-n}$ (Handcock and Gile 2022). Common choices for the unit size distribution $f(\cdot|\eta)$ are counting distributions that have Poisson-like behavior while also allowing for over and under dispersion, such as the Conway-Maxwell-Poisson.

2.2.3 *Assumptions*

The successive sampling model is an improvement over previous methods in that it does not require the population to be large, and it does not model the sampling process as with replacement (Gile and Handcock 2010). A complete discussion of SS assumptions can be found in Gile (2011). For our purposes, the SS model assumes that the graph is connected, meaning any member of the population can be reached by any other member through a path in the network. This is not true for clustered populations.

2.3 Clustered Populations

For the purposes of this paper, a clustered population is a disconnected population network such that any member of cluster i cannot reach any member of cluster j through a path in the network for all clusters $i \neq j$.

Such populations can arise within cities where there are strong geographical, social, or linguistic divisions. If this type of clustering in the underlying population is ignored, the unadjusted SS-PSE estimator performs poorly in a variety of ways (Gamble and McLaughlin 2023). Clustered populations are also a way to model a population of interest that is spread over multiple cities, where the cities act as clusters and the goal is an overall estimate of population size across cities.

3. Theoretical Justification for Two Novel Methods

In this section, we introduce two modifications to SS-PSE for estimating the size of clustered populations. The Posterior Sum method is theoretically straightforward but relies on prior information about the cluster level population sizes N_i, which may not always be available. The Clustered SS-PSE method is more theoretically complex since it introduces new parameters into the hierarchical Bayesian structure of the SS-PSE model, but it allows for estimation in settings where prior information about the cluster level N_i is unavailable. Both proposed methods require RDS chains from each cluster in the population, where cluster membership of each chain is known.

3.1 Posterior Sum Model

Consider a population of N units divided into m distinct clusters. Each cluster is connected, meaning any one member of cluster i can be reached by any other member in cluster i through a path on the network, but no two clusters are connected to one another. In other words, no member of cluster i can be reached by any member of cluster j for $i \neq j$. This implies that all participants recruited from the same seed are in the same cluster. A cluster can contain multiple seeds.

Assume we have taken RDS samples from each of the m clusters. Next, we run m SS-PSE methods on each of the m sample subsets with seeds in the same cluster. The results of these m SS-PSE methods are posterior distributions on the cluster sizes $P(N_1|D_1), ..., P(N_m|D_m)$, where D_i is the sample information from cluster i. To simulate the overall posterior for $N = N_1 + ... + N_m$, the Posterior Sum method takes the sum of random draws from the individual cluster posteriors until a sufficient sample from $P(N|D_1, ..., D_m)$ has been obtained. The median of this posterior sum distribution can be used as a point estimate for N, and its variance reflects the aggregated uncertainty of each cluster level estimate, which can be used to obtain credible intervals.

Since the Posterior Sum method relies on m separate SS-PSE models, it requires some prior information about the cluster level population sizes N_i, which may not be attainable

in every situation. It is also relatively sensitive to misspecification of those priors on N_i.

3.2 Clustered SS-PSE Model

The Clustered SS-PSE Model is proposed to extend the application of SS-PSE to clustered populations where no reliable prior information exists for the individual cluster sizes. The Clustered SS-PSE method introduces a new set of parameters to the Bayesian framework of SS-PSE that represent the cluster proportions: $p_i = \frac{N_i}{N}$ for each cluster i. Since $\mathbf{p}$ is a simplex of positive numbers that sum to 1, a weakly informative or non-informative prior can be imposed in situations where prior information about the cluster sizes N_i is unreliable.

The introduction of this new $\mathbf{p}$ parameter is the novel contribution of the Clustered SS-PSE Model, and it allows for the joint estimation of the population size and cluster proportions using data from all clusters. Including $\mathbf{p}$ changes both the probability model on the data and the full conditional posterior distributions of model parameters that are required in the Gibbs sampler. Below is a derivation of the new probability model on the data.

Consider a population of N units divided into m distinct clusters. Let $\mathbf{U}_i = (U_{i1}, ..., U_{iN_i})$ be the random vector of unit sizes associated with each population member in cluster i, where $\mathbf{U}_i$ is a random sample from the super-population unit size distribution $f(\cdot|\eta_i)$.

Assume we have taken RDS samples from each of the m clusters with sample sizes $n_1, ..., n_m$. Let $\mathbf{G}_i = (G_{i1}, ...G_{in_i})$ be the ordered random vector of observed unit indices from cluster i, where G_{ij} has support $1, ..., N_i$ and length n_i. Let $\mathbf{g}_i = (g_{i1}, ...g_{in_i})$ be the realized values of those unit indices from cluster i. Let $\backslash\mathbf{g}_i$ be the set of unit indices in cluster i that are not in $\mathbf{g}_i$. Let $\mathbf{U}_{i,obs} = (U_{g_{i1}}, ..., U_{g_{in_i}})$ be the random vector of observed unit sizes in cluster i, with realized values $\mathbf{u}_{i,obs} = (u_{g_{i1}}, ..., u_{g_{in_i}})$. Similarly, let $\mathbf{U}_{i,unobs}$ and $\mathbf{u}_{i,unobs}$ be the random and realized vectors of unobserved unit sizes in cluster i. Let $\mathbf{U}_{unobs}$ be the set containing $\mathbf{U}_{1,unobs}, ...\mathbf{U}_{m,unobs}$. Let D_i be the matrix of complete sample data from cluster i: $(\mathbf{G}_i, \mathbf{U}_{i,obs})$, and let D be the set containing $D_1, ..., D_m$.

3.2.1 Probability Model of Observed Data

Since the sample is composed of m independent RDS samples from each cluster, the probability of observing sample data D is the product of the cluster level probability models across all clusters.

$$P(D|\boldsymbol{\eta}, N, \mathbf{p}) = \prod_{i=1}^{m} P\left(D_i|\boldsymbol{\eta}, N, \mathbf{p}\right)$$

The probability model within each cluster can be rewritten as the sum of the joint probability $P\left(D_i, \mathbf{U}_{i,unobs}|\boldsymbol{\eta}, N, \mathbf{p}\right)$ over all possible unobserved unit sizes.

$$P(D_i|\boldsymbol{\eta}, N, \mathbf{p}) = \sum_{\mathbf{u}_{i,unobs} \in \mathcal{U}(\mathbf{u}_{i,obs})} P\left(D_i, \mathbf{U}_{i,unobs}|\boldsymbol{\eta}, N, \mathbf{p}\right)$$

Where $\mathcal{U}(\mathbf{u}_{i,obs})$ is the set of all possible unobserved unit sizes in cluster i given the observed unit sizes $\mathbf{u}_{i,obs}$. Next, the joint probability of the sample data and the unobserved unit sizes — $(D_i, \mathbf{U}_{i,unobs}) = (G_i, \mathbf{U}_{i,obs}, \mathbf{U}_{i,unobs})$ — can be rewritten using the definition of conditional probabilities.

$$P\left(D_i, \mathbf{U}_{i,unobs}|\boldsymbol{\eta}, N, \mathbf{p}\right) = P\left(G_i|\mathbf{U}_{i,obs}, \mathbf{U}_{i,unobs}\boldsymbol{\eta}, N, \mathbf{p}\right) P\left(\mathbf{U}_{i,obs}, \mathbf{U}_{i,unobs}|\boldsymbol{\eta}, N, \mathbf{p}\right)$$
$$= \left(\frac{(p_i N)!}{(p_i N - n_i)!} \prod_{k=1}^{n_i} \frac{u_{ig_{ik}}}{r_{ik}}\right) \left(\prod_{k=1}^{n_i} f(u_{ig_{ik}}|\eta_i) \prod_{j \in \backslash \mathbf{g}_i} f(u_{ij}|\eta_i)\right) \tag{2}$$

The first term of this equation is the probability of an observed sampling order given a set population of unit sizes $(\mathbf{u}_{i,obs}, \mathbf{u}_{i,unobs})$ (Handcock, Gile, and Mar 2014). The second term is the probability of that set population of unit sizes under the unit size distribution $f(\cdot|\eta_i)$. The notation $r_{ik} = \sum_{j=1}^{p_i N} u_{ij} - \sum_{j=1}^{k-1} u_{ig_j}$ is the sum of remaining degrees in cluster i from the unsampled population units at step k of the sampling process.

All together, this gives the probability model

$$P(D|\boldsymbol{\eta}, N, \mathbf{p}) = \prod_{i=1}^{m} \frac{(p_i N)!}{(p_i N - n_i)!} \prod_{k=1}^{n_i} \frac{u_{ig_{ik}}}{r_{ik}} f(u_{ig_{ik}}|\eta_i) \prod_{j \in \backslash \mathbf{g}_i} f(u_{ij}|\eta_i) \tag{3}$$

As with the unadjusted SS-PSE method, the likelihood that follows from this probability model is restrictively difficult to compute. However, a Gibbs sampler can be used to simulate the joint posterior distribution of all parameters in the model, which allows for inference on the parameters of interest—overall population size, cluster sizes, and the unit size distributions in each cluster (Gamble and McLaughlin 2023).

3.2.2 Prior Selection

The prior on N can be specified according to Handcock, Gile, and Mar (2014), as described in Section 2.2.2. In the simplest case, it is possible to use a single prior estimate for the total population size in order to fit a Beta(α, β) distribution to the sample proportion $\frac{n}{N}$. The prior on each η_i can also be specified as in the standard SS-PSE method.

The parameter $\mathbf{p}$ is the new component of this extended model. Since $\mathbf{p}$ is a vector of proportions that sum to 1, a Dirichlet$(\alpha_0, \boldsymbol{\alpha})$ distribution can be used, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$ is the mean vector of the prior $\pi(\mathbf{p})$ and α_0 is a global concentration parameter controlling the variance of $\pi(\mathbf{p})$. In other words,

$$\pi(\mathbf{p}) = \frac{\Gamma\left(\sum_{i=1}^{m} \alpha_0 \alpha_i\right)}{\prod_{i=1}^{m} \Gamma\left(\alpha_0 \alpha_i\right)} \prod_{i=1}^{m} p_i^{\alpha_0 \alpha_i - 1}. \tag{4}$$

Note that this parameterization is equivalent to a Dirichlet$(\alpha_0\alpha_1, ..., \alpha_0\alpha_m)$ distribution. The separation of the mean vector from the concentration parameter is useful for discussion on the properties of the Dirichlet family as a prior on $\mathbf{p}$.

By setting $\alpha_1 = ... = \alpha_m = \frac{1}{m}$ and $\alpha_0 = m$, it is possible to obtain a uniform prior over the unit hypercube in m dimensions. This property is useful for applications in which no prior information is known about the cluster sizes $p_i N$. However, the Dirichlet family also provides freedom to incorporate prior knowledge of the relative cluster sizes, if such information is known.

The global concentration parameter α_0 plays a key role in the behavior of the Dirichlet distribution. For example, Figure 1 shows the marginal densities of a Dirichlet$(\alpha_0, (0.5, 0.25, 0.25))$ distribution for values of α_0 from 1 to 10. All of these distributions have the same mean, as indicated by the vertical line. However, those with smaller concentration values are required to have larger variances, which results in the distribution having more density away from the mean. As a special case, when $\alpha_0 < \frac{1}{\alpha_i}$ for

any dimension i, the marginal density in that dimension will be convex. Because of this, it is recommended to select $\alpha_0 \geq \frac{1}{\min\{\boldsymbol{\alpha}\}}$, where $\boldsymbol{\alpha}$ is the mean vector of $\pi(\mathbf{p})$.

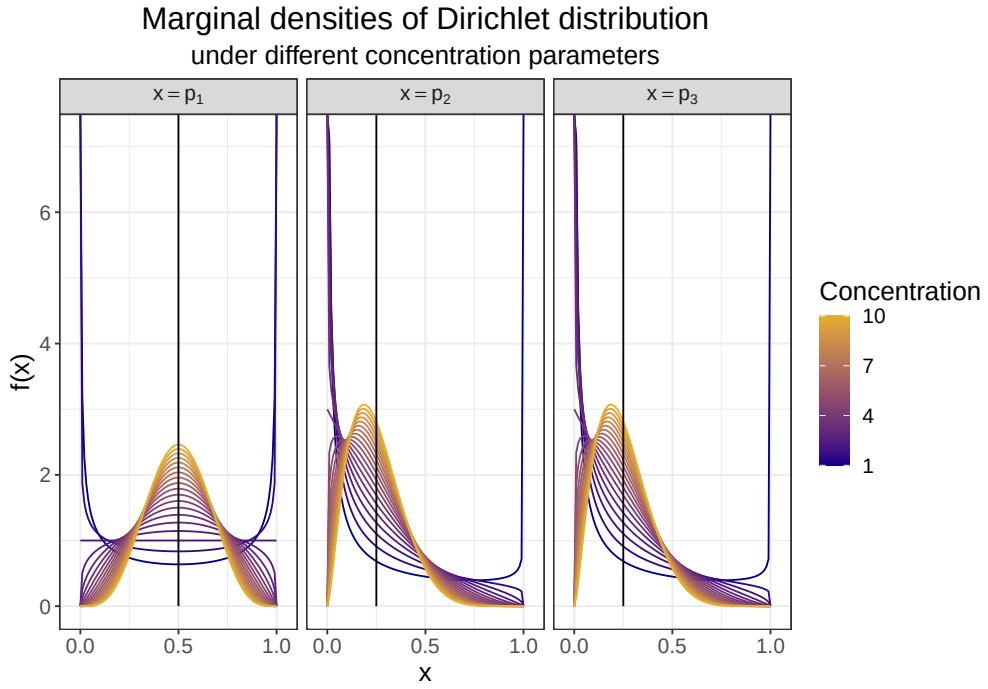

Figure 1: Marginal densities of Dirichlet distribution under different concentration parameters.

4. Discussion

Estimation methods for RDS data such as SS-PSE address the need for sampling from hard-to-reach populations, and their implementation helps to answer important questions about global health and well-being. This research extends the functional application of SS-PSE to populations whose underlying social network is disconnected. We propose two modifications to SS-PSE to allow for the estimation of population size when the underlying population is clustered.

The Posterior Sum method is a straightforward correction that combines the results from several different SS-PSE fits to obtain an estimate for the overall population size that incorporates variance from each population individually. This method performs well under most simulation settings considered, as long as the cluster level prior information is correct. In general, the Posterior Sum method relies heavily on the quality of prior information (Gamble and McLaughlin 2023).

The Clustered SS-PSE method works by introducing a new set of parameters into the SS-PSE model that represent the proportion of the population in each cluster. This method is useful in a larger variety of settings, due to the fact that it does not require prior information about the population size at the cluster level. The Clustered SS-PSE method also performs well under most simulation settings considered and is generally less sensitive to prior misspecification than the Posterior Sum method (Gamble and McLaughlin 2023).

Potentially the most important area of future work for these and other SS-PSE methods is in addressing the almost certain measurement error present in both the self-reported degree and observed sampling order. To this end, McLaughlin et al. (2015) have developed

a measurement error model for degree, which can improve estimates in situations where the self-reported network sizes are inaccurate. It would be useful to implement in the Clustered SS-PSE model. However, there has been little investigation into the effect of measurement error in observed sampling order. Since the order of enrollment is so integral to SS-PSE results, this is an area for future research.

Finally, the further investigation of these and all other SS-PSE methods on a wider variety of population structures is always of value. The performance of SS-PSE depends greatly on the structure of each sample drawn and the complex connections in the underlying populations. Any additional information about specific features that can affect SS-PSE would be of use to those implementing the methods in real populations.

REFERENCES

Gamble, L. J. and McLaughlin, K. R. (2023), "Estimating the Size of Clustered Hidden Populations," Manuscript submitted, Statistics Department, Oregon State University.

Gile, K. J. and Handcock, M. S. (2010, August), "Respondent-Driven Sampling: An Assessment of Current Methodology," *Sociological Methodology*, 40, 1, 285-327, DOI: 10.1111/j.1467-9531.2010.01223.x.

Gile, K. J. (2011), "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation," *Journal of the American Statistical Association*, 106, 493, 135–146, DOI:10.1198/jasa.2011.ap09475.

Handcock, M. S., Gile, K. J., and Mar, C. M. (2014), "Estimating Hidden Population Size Using Respondent-Driven Sampling Data," *Electronic Journal of Statistics*, 8, 1, 1491–1521, DOI: 10.1214/14-EJS923.

Handcock, M. S., Gile, K. J., and Mar, C. M. (2015), "Estimating the Size of Populations at High Risk for HIV Using Respondent-Driven Sampling Data," *Biometrics*, 71, 1, 258–266, DOI: 10.1111/biom.12255.

Handcock, M. S. and Gile, K. J. (2022, August), *sspse: Estimating Hidden Population Size using Respondent Driven Sampling Data*, Version 1.0.3. Available at https://cran.r-project.org/web/packages/sspse/sspse.pdf.

Heckathorn, D. (1997), "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," *Social Problems*, 44, 2, 174-199, DOI: 10.2307/3096941.

Johnston, L. G., McLaughlin, K. R., El Rhilani, H., Latifi, A., Toufik, A., Bennani, A., Alami, K., Elomari, B., and Handcock, M. S. (2015), "Estimating the Size of Hidden Populations Using Respondent-driven Sampling Data: Case Examples from Morocco," *Epidemiology*, 26, 6, 846–852, DOI: 10.1097/EDE.0000000000000362.

Johnston, L. G., McLaughlin, K. R., Rouhani, S. A., and Bartels, S. A. (2017), "Measuring a Hidden Population: A Novel Technique to Estimate the Population Size of Women with Sexual Violence-Related Pregnancies in South Kivu Province, Democratic Republic of Congo," *Journal of Epidemiology and Global Health*, 7, 1, 45-53, DOI: 10.1016/j.jegh.2016.08.003.

Johnston, L. G., McLaughlin, K. R., Gios, L., Cordioli, M., Staneková, D. V., Blondeel, K., Toskin, I., Mirandola, M., and SIALON II Network (2021), "Populations Size Estimations Using SS-PSE Among MSM in Four European Cities: How Many MSM are Living with HIV?" *European Journal of Public Health*, 31, 6, 1129–1136, DOI: 10.1093/eurpub/ckab148.

McLaughlin, K. R., Handcock, M. S., Johnston, L. G., Japuki, X., Gexha-Bunjaku, D., and Deva, E. (2015), "Inference for the Visibility Distribution for Respondent-Driven Sampling," *American Statistical Association*, Alexandria, VA.

McLaughlin, K. R., Johnston, L. G., Gamble, L. J., Grigoryan, T., Papoyan, A., and Grigoryan, S. (2019), "Population Size Estimations Among Hidden Populations Using Respondent-Driven Sampling Surveys: Case Studies From Armenia," *JMIR Public Health and Surveillance*, DOI:10.1177/1471082X211043945.

Salganik, M. J. and Heckathorn, D. D. (2004), "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," *Sociological Methodology*, 34, 193-239, DOI: 10.1111/j.0081-1750.2004.00152.

Volz, E. and Heckathorn, D. D. (2008), "Probability Based Estimation Theory for Respondent Driven Sampling," *Journal of Official Statistics*, 24, 1, 79-97.

Weikum, D., Kelly-Hanku, A., Hou, P., Kupul, M., Amos-Kuma, A., Badman, S. G., Dala, N., Coy, K. C., Kaldor, J. M., Vallely, A. J., and Hakim, A. J. (2019), "Kuantim mi tu ("Count me too"): Using Multiple Methods to Estimate the Number of Female Sex Workers, Men Who Have Sex With Men, and Transgender Women in Papua New Guinea in 2016 and 2017," *JMIR Public Health and Surveillance*, 5, 1, e11285, DOI: 10.2196/11285.

Wesson, P., Qabazard, R. F., Wilson, E. C., McFarland, W., and Raymond, H. F. (2018), "Estimating the Population Size of Transgender Women in San Francisco Using Multiple Methods, 2013," *International*

Journal of Transgender Health, 19, 1, 107-112, DOI: 10.1080/15532739.2017.1376729.

World Health Organization and UNAIDS (2010), *Guidelines on Estimating the Size of Populations Most at Risk to HIV*, Geneva, Switzerland: WHO Press.

World Health Organization (2017), *Biobehavioural survey guidelines for populations at risk for HIV*, Geneva, Switzerland: WHO Press.