Mambilla Journal of Science and the Environment Vol. 9, Issue 2, Page No. 102-113

MAMBILLA JOURNAL OF SCIENCE AND THE ENVIRONMENT

journal homepage: https://oer.tsuniversity.edu.ng/index.php/mjse

EVALUATION OF THE EFFECT OF DEGREE DISTRIBUTION ON RESPONDENT-DRIVEN SAMPLING (RDS) ESTIMATOR AMONG INJECTION DRUG USERS

[1,2*]Anjikwi, Y., [2]Jibasen, D., [3]Dike, I.J., [2]Torsen, E.

[1]Department of Agricultural Economics, University of Maiduguri, Borno-Nigeria.
[2]Department of Statistics, Modibbo Adama University, Yola-Nigeria.
[3]Department of Operations Research, Modibbo Adama University, Yola-Nigeria.

ARTICLE INFO

Article history:
Received 01 August 2025
Received in revised form 24 August 2025
Accepted 27 August 2025

Keywords:
Respondent-Driven Sampling, Proportion, Social-Networks, Without Replacement.

ABSTRACT

The Respondent-driven sampling (RDS) method has become a successful approach for recruiting hard-to-reach populations, leading to a significant rise in its use worldwide. However, making reliable inferences from RDS data poses challenges due to assumptions associated with its statistical methodology. Most studies focus on population prevalence estimates but often neglect the network size(degree). This study used simulations and real-world data to examine the effects of degree distributions in networked populations, aiming to evaluate the $RDS_{proposed}$ estimator in comparison to existing RDS estimators: Naïve, Salganik and Heckathorn (SH-RDS), Volz and Heckathorn (VH-RDS), and Gile Sucessive Sampling estimator (G-SS). The findings revealed that, with degree distribution from 51 to 100, SH-RDS shows consistency in gender proportions, while variations were noted in the Naïve, $RDS_{proposed}$, VH-RDS, and G-SS estimator. For degrees ranging from 251 to 300 and beyond, the proposed RDS and G-SS estimators shows a decline in variance, while SH-RDS variance increased, indicating a potential decline in performance under these conditions. Among higher degree distributions (>300), the $RDS_{proposed}$ estimator emerged as the most effective, with a variance of 1.11, compared to 1.95 for VH-RDS and 1.67 for G-SS. The $RDS_{proposed}$ estimator also shows a substantial decrease in variance from 2.43 to 1.11 as the sample size and degree increased. The results for IDU in Eastern Europe and in a Caribbean nation agreed with the results of simulated data. It was recommended that, degree distribution must be carefully investigated before analyzing RDS data.

1. Introduction

Respondent-driven sampling (RDS) is an advanced sampling technique that leverages social networks to recruit participants from hard-to-reach populations, which often lack a structured sampling framework (Falorsia *et al*., 2023). This novel method has been effective especially in connecting with groups that are frequently marginalized or neglected in conventional research approaches. In medical research, RDS has mainly been applied to recruit individuals from various high-risk demographics, such as intravenous drug users. These men engage in sexual relationships with other men, urban jazz musicians, and homeless individuals (Fellows, 2019).

Additionally, RDS is essential in helping recruit individuals for key health interventions, enabling researchers to reach those who might otherwise be disengaged from health services (Krivitsky *et al.* 2022). Its adaptability goes beyond healthcare; RDS has proven to be an effective method for engaging migrant populations in a variety of diverse environments (Górny & Napierała, 2016). The method's distinctive capability to produce dependable population estimates with favorable statistical properties, along with its practical applicability for real-world implementation, has led to a notable increase in the number of RDS studies carried out around the world in recent years (Jonsson *et al.* 2019). This rise highlights the growing significance of the method in enhancing the understanding of hidden populations and meeting the unique needs.

RDS has been utilized in more than 460 studies across 69 different countries. A distinctive feature of this approach is that the participants themselves are tasked with recruiting additional individuals. The recruitment process begins with initial participants chosen by the researchers, known as “seeds,” who are subsequently interviewed (Rozemberczki, *et al*, 2021).

* Corresponding author. Tel.: +2348036267312
E-mail address: y.anjikwi@gmail.com

These participants receive a batch of numbered and uniquely coded referral coupons to pass on to others and are offered an incentive for the participation in the research. The chain of recruitment continues until the targeted sample size is achieved. This recruitment system is structured to enable the calculation of relevant probabilities and to identify the connections between recruiters and recruits. This allows for the assessment and adjustment of recruitment biases during the analysis. Additionally, data regarding the personal network size of each individual is gathered, facilitating weighted analysis to balance the oversampling of respondents with larger social networks (Giles, 2024).

The RDS is composed of two main elements: data gathering and inference. While this approach is relatively recent, large samples of injection drug users have been successfully gathered through RDS in more than 69 nations. Nonetheless, the concern surrounding RDS pertains to the inferential aspect, which is based on six key assumptions: (1) 'seeds' are selected with a probability that corresponds to the network size, referred to as degree, (2) individuals within the target population uphold a mutual relationship with the contacts, (3) any individual can be accessed by another through a chain of network connections, (4) sampling is conducted with replacement, (5) the network size of individuals, or degree, is accurately assessed, and (6) referrals are random. However, in practice, many of these assumptions are not upheld, leading to bias in all RDS estimates produced.

Many RDS studies primarily concentrate on generating estimates of population prevalence without thoroughly examining the extent of the degree distribution within the relevant network structure of the population, the data quality, and any assumption violations (Spiller *et al*. 2018). While the RDS assumptions discussed here might be suitable for certain hard-to-reach groups, they largely do not reflect actual circumstances. For example, 'seeds' are selected based on a probability proportional to the degree, and degree is accurately measured (assumptions 1 and 5, respectively), which relies on the respondent's memory and can be challenging to obtain precise figures in specific situations, ultimately affecting whether the seed is selected in a manner proportional to the degree (Avery *et al.* 2021).

This study introduces an RDS estimator designed to address the bias created by the breach of assumptions 4 and 5 mentioned earlier through sampling without replacement, aiming to identify the concealed population proportion of injection drug users. The objective of the study is to enhance existing research on the validity of estimators by incorporating actual real-world reported network degrees within extensive simulated networks and evaluating the performance of the estimator on categorical outcomes using both real and synthetic networks.

2. Methods

Numerous studies Heckathorn (1997, 2002), Salganik and Heckathorn (2004), Volz and Heckathorn 2008, and Gile (2011) on RDS tend to focus mainly on estimating the prevalence of populations, often neglecting a detailed analysis of the degree distribution within the pertinent network structure of that population, data quality, and potential violations of assumptions (Spiller *et al.* 2018). This study offers a concise review of the commonly used estimators for sample prevalence of populations, with a particular focus on the relationships to the proposed estimator.

2.1 The Naïve Estimator

According to Heckathorn (1997), the Naïve estimator is denoted as

$$\hat{\mu}_A^N = \frac{n_a}{n} \tag{1}$$

Where,

$\hat{\mu}_A^N$ is the population proportion, n_a is the sample size of group A, and n is the total sample.

Likewise, if the actual sampling probabilities are represented by $\pi_i,\ i = 1, \dots, N$ then $\hat{\mu}_A^N$, then it serves as a generalized Hansen-Hurwitz estimator of the target parameter.

$$P_A^N = \frac{\phi_A}{\phi_A + \phi_B} \tag{2}$$

Where, P_A^N is the population proportion, ϕ_A number of recruitment in group A, ϕ_B. The number of recruits in group B.

Equation (2) indicates that when equal sampling probabilities are observed for individuals in both group A and group B, it acts as a generalized Hansen-Hurwitz estimator for the parameter of interest.

The Assumption of Naïve Estimator

The assumption of Naïve Heckathorn (1997) are

1. Respondents recruit peers from the social contacts with equal probability.
2. Sampling is done with replacement.
3. The degree of respondents is normally distributed.
4. The social network of the population is undirected.
5. The population forms a connected network.

2.2 SH-RDS Estemator

SH-RDS, as suggested by Salganik and Heckathorn (2004), is an estimator for population proportions that utilizes

sampling with replacement and is given as

$$\widehat{PP}_A = \frac{\hat{D}_B.\hat{C}_{B,A}}{\hat{D}_A.\hat{C}_{A,B}+\hat{D}_B.\hat{C}_{B,A}} \quad (3)$$

$$\widehat{PP}_B = \frac{\hat{D}_A.\hat{C}_{A,B}}{\hat{D}_A.\hat{C}_{A,B}+\hat{D}_B.\hat{C}_{B,A}} \quad (4)$$

where

$R_A = \sum_{i \in A} d_i$

$\hat{C}_{A,B} = \frac{R_{AB}}{R_{AA.}+R_{AB}}$

$\hat{C}_{B,A} = \frac{R_{BA}}{R_{BB.}+R_{BA}}$

$\widehat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}$

$\widehat{D}_B = \frac{n_B}{\sum_{i=1}^{n_B} \frac{1}{d_i}}$

The Assumption of the SH-RDS estimator

The assumption of Salganik and Heckathorn (2004) is

1. Respondents recruit peers from the social contacts with equal probability.
2. Each recruitment consists of only one peer.
3. Sampling is done with replacement.
4. The degree of respondents is reported without error.
5. The social network of the population is undirected, and
6. The population forms a connected network.

2.3 VH-RDS Estimator

VH-RDS, suggested by Voltz and Heckathorn (2008), serves as a population proportion estimator that employs sampling with replacement, and was given as

$$\widehat{PP}_{A_V} = \left(\frac{n_A}{n}\right)\left(\frac{\widehat{D}_u}{\widehat{D}_{A_V}}\right) \quad (5)$$

The variance of VH-RDS was given as

$$\hat{V}_{HH}(\langle\hat{y}\rangle) = \frac{1}{n(n-1)}\sum_s \left(\frac{\widehat{D}_u}{\hat{d}_i} - \langle\hat{y}\rangle\right)^2 \quad (6)$$

Where

y_i is an indicator function that a value $I_i(i) = \begin{cases} 1, & i \in A \\ 0, & otherwise \end{cases}$

$$\widehat{D}_{A_V} = \frac{N_u}{\sum_{i=1}^{N_u} \frac{1}{d_i}} \quad (7)$$

The Assumption of the VH-RDS Estimator

The assumptions of Volz and Heckathorn (2008) are

1. Recruitment is random. When recruiting others, respondents select uniformly at random from the network.
2. Each recruitment consists of only one peer.
3. Sampling is done with replacement.
4. Respondents accurately report the degree in the network.
5. Network connections are reciprocal.
6. The population forms a connected network.
7. Convergence. Recruitment is modeled as a Markov process (MP), where the state of the MP is the last individual recruited.

2.4 Gile's Successive Sampling (G-SS) Estimator

The G-SS estimator suggested by Gile (2011) serves as an estimator for population proportions when sampling without replacement. The inclusion probability estimate linked to a certain unit, referred to as unit i, was derived as

$$\tilde{\pi}_{SS} = \frac{U_i + 1}{M + 1} \quad (8)$$

Where U_i is the number of times unit i is ample in M trials. He proposes using these estimated probabilities in the standard Horvitz-Thompson estimator:

$$T_{H-T} = \sum_{j:S_{j=1}} \frac{z_j}{\tilde{\pi}_j} \quad (9)$$

Applied the resulting mapping $\hat{\pi}$ to estimate $\hat{\mu}_{SS}$ via the generalized Horvitz-Thompson estimator as

$$\hat{\mu}_{SS} = \frac{\sum_{j=1}^{N} \frac{S_j Z_j}{\hat{\pi}(d_j)}}{\sum_{j=1}^{N} \frac{S_j}{\hat{\pi}(d_j)}} \quad (10)$$

The assumptions of Gile's Successive Sampling (G-SS) Estimator

The assumptions of Gile (2011) are

1. Recruitment is random. When recruiting others, respondents select uniformly at random from the network.
2. Each recruitment consists of only one peer.
3. Sampling is done without replacement.
4. Respondents accurately report the degree in the network.
5. Network connections are reciprocal.
6. The population forms a connected network.
7. The population size N is known.

2.5 Proposed RDS Estimator

This section presents a detailed introduction to the proposed RDS estimator, along with a comprehensive formulation of its variance. The discussion is grounded in the assumption of sampling without replacement, which is crucial for accurately understanding the sampling dynamics in this context. The study delved into the underlying theoretical framework that supports the estimator, as well as the mathematical derivations that lead to the variance expression. This exploration aims to provide a robust basis for the evaluation and application of the RDS estimator in various research settings.

Assumptions of the Proposed Model.

The assumptions of the proposed model are:

1. Respondents recruit peers from the social contacts with equal probability (randomly).
2. Each recruitment involves selecting only one peer, and throughout the sampling period, no more than one peer can be recruited.
3. Sampling is conducted without replacement.
4. The reported degrees of the respondents are subject to negligible error.
5. The network is directed, and
6. The population forms a connected network.

Estimation of inclusion probability

The concept of Horvitz-Thompson estimation was adapted to create a sampling method known as the PPSWOR. A node was included in an RDS sample with a probability that is proportional to its degree. As a result, the likelihood of selecting a unit given that another unit was chosen, as put forth by Lawson and Ponkaew (2019), was defined using sampling without replacement.

$$\mathrm{P}_i = \begin{cases} \frac{\delta_i}{\sum_{\mathrm{j=1}}^{\mathrm{N}} \delta_j - \sum_{\mathrm{j=1}}^{\mathrm{k-1}} \delta_{R_j}}, & \mathrm{j} \notin (\mathrm{R}_1, \ldots, \mathrm{R}_{\mathrm{k-1}}) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

$$P(i|j) = \frac{\delta_i}{(1 - \delta_j)} \quad (12)$$

$P(i|j)$ = conditional probability of getting the set of units that was drawn, given that the j^{th} unit was drawn first. The modified Horvitz and Thompson (H-T) inclusion probability of i^{th} unit and population estimators proposed by Naser *et al. (*2018) was given as

$$\hat{\pi}_{H-T} = \frac{P(i|j)y_i + P(j|i)y_j}{\mathrm{P}_i} \quad (13)$$

And

$$Y = \sum_{i=1}^{N} \frac{z_i y_i}{z_i p_i} \tag{14}$$

As a population estimator. Since $\{\pi\}$ is usually unknown, this work used the proposed Naser *et al. (*2018) approach to approximate the probability of inclusion $\{\pi\}$. This was done by substituting p_i in equation (11) into equation (13) and obtaining the proposed inclusion probability:

$$\hat{\pi}_{proposed} = \frac{\frac{\delta_i}{(1-\delta_j)} y_j + \frac{\delta_j}{(1-\delta_i)} y_i}{\frac{\delta_i}{\sum_{j=1}^{N} \delta_j - \sum_{j=1}^{k-1} \delta_{R_j}}} \tag{15}$$

The proposed mean degree of the sample can be estimated as two ratios of the Horvitz-Thompson estimator, as defined by Lawson (2017).,

$$\hat{D}_{A_{H-T}} = \frac{\sum_{i=1}^{N} \delta_i z_i y_i / p_i}{\sum_{i=1}^{N} z_i / p_i} \tag{16}$$

Therefore, substituting p_i in equation (11) into equation (16) gave the proposed mean degree of the RDS sample as

$$\begin{aligned}
\hat{D}_{A_{proposed}} &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\delta_i z_i y_i / \left(\frac{\delta_i}{\delta_j - \delta_{R_j}}\right)\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(z_i / \left(\frac{\delta_i}{\delta_j - \delta_{R_j}}\right)\right)} \\
&= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\delta_i z_i y_i \left(\delta_j - \delta_{R_j}\right)/\delta_i\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(z_i \left(\delta_j - \delta_{R_j}\right)/\delta_i\right)} \\
&= \frac{\sum_{i=1}^{N}(\delta_i z_i y_i/\delta_i)}{\sum_{i=1}^{N}(z_i/\delta_i)} \\
\hat{D}_{(i)_{proposed}} &= \frac{\sum_{i=1}^{N} z_i\, y_i}{\sum_{i=1}^{N}(z_i/\delta_i)}
\end{aligned} \tag{17}$$

Where can i take on the value 1 for presence and 0 for absence

$$\hat{D}_{(1)_{proposed}} = \frac{\sum_{i=1}^{N} z_i\, y_i}{\sum_{i=1}^{N} z_i \delta_i^{-1}}, if\ \ z_i = 1 \tag{18}$$

and

$$\hat{D}_{(0)_{proposed}} = \frac{\sum_{i=1}^{N}(1-z_i) y_i}{\sum_{i=1}^{N}(1-z)_i \delta_i^{-1}}, \quad if\ \ z_i = 0 \tag{19}$$

Considered an RDS recruitment can be defined by a two-by-two recruitment matrix, R.

$$R = \begin{bmatrix} R_{ii} & R_{ij} \\ R_{ji} & R_{jj} \end{bmatrix} \tag{20}$$

The total of whole cells equals the sample size N.

$$N = R_{ii} + R_{ij} + R_{ji} + R_{jj} \tag{21}$$

The recruitment between the groups S_{ij} and S_{ii} with the group is defined as

$$S_{ii} = \frac{R_{ii}}{R_{ii}+R_{ji}}, S_{ij} = \frac{R_{ij}}{R_{ij}+R_{ii}}, S_{jj} = \frac{R_{jj}}{R_{jj}+R_{ij}}, S_{ji} = \frac{R_{ji}}{R_{ji}+R_{jj}}$$

A more general way to develop a Horvitz-Thompson type estimator for S_{ij} is to use the edges' inclusion probabilities:

$$\hat{S}_{(z_i=1,z_j=0)} = \frac{\sum_{i:z_i=1}\sum_{j:z_j=0} y_{ij}}{\sum_{i:z_i=1}\sum_{j:z_j=0} y_{ij} + \sum_{i:z_i=1}\sum_{j:z_j=1, i\neq j} y_{ij}} \tag{22}$$

$$\hat{S}_{(z_i=0,z_j=1)} = \frac{\sum_{i:z_i=0}\sum_{j:z_j=1} y_{ij}}{\sum_{i:z_i=0}\sum_{j:z_j=1} y_{ij} + \sum_{i:z_i=0}\sum_{j:z_j=0, i\neq j} y_{ij}} \tag{23}$$

Now that we have the $\hat{S}_{(z_i=1,z_j=0)}$, $\hat{S}_{(z_i=0,z_j=1)}$, $\hat{D}_{(1)}$, and $\hat{D}_{(0),}$ it is safe to estimate the population proportion of the yes $z_i = 1$ and absence $z_i = 0$ of population as characteristics as :

$$\widehat{RDS}_{proposed_1} = \frac{\hat{D}_{(0)}.\hat{S}_{WT(0,1)}}{\hat{D}_{(0)}.\hat{S}_{WT(0,1)} + \hat{D}_{(1)}.\hat{S}_{WT(1,0)}} \tag{24}$$

$$\widehat{RDS}_{proposed_0} = \frac{\widehat{D}_{(1)}.\hat{S}_{WT(1,0)}}{\widehat{D}_{(1)}.\hat{S}_{WT(1,0)} + \widehat{D}_{(0)}.\hat{S}_{WT(0,1)}} \quad (25)$$

2.6 Estimation of Variance of Proposed Model

Suppose that a sample S of size n is selected with a PPSWOR from a finite population $\cup\, of\ size\ N = \{1,2,..,N\}$. Let $\hat{Y} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{D_{(i)}.S_{(i,j)}}{D_{(i)}.S_{(i,j)}+D_{(j)}.S_{(j,i)}}$s the population proportion of the study variable y. Let z_i denote the response indicator variable of y_i, where $z_i = 1$ if y_i is observed, 0 otherwise, let $p_i = p(z_i = 1)$. Let $\hat{Y} = \sum_{i\epsilon s} M_i$ denote the population proportion estimator, where M_i is the function y_i, $M_{ij} = f(\ y_i)$, M_{ij} means person i and j are friends, hence i is observed. In line with Lawson (2017), the study put forward a variance estimator; the study suggested population proportion variance $\hat{V}(\hat{Y})$ is defined in the following way.

$$\hat{V}(\hat{Y}) = \frac{n}{n-1}\left[\sum_{i\epsilon s} \widehat{M}^2{}_i - \frac{1}{n}\left(\sum_{i\epsilon s} \widehat{M}_i\right)^2\right] \quad (26)$$

From the proposed estimator in equation (24), the study can write

$$\hat{Y} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{D_{(i)}.S_{(i,j)}}{D_{(i)}.S_{(i,j)} + D_{(j)}.S_{(j,i)}} = \sum_{i\epsilon s} M_i \quad (27)$$

$$M_{ij} = f(y_i) = \frac{D_{(i)}.S_{(i,j)}}{D_{(i)}.S_{(i,j)} + D_{(j)}.S_{(j,i)}}$$

The variance of $V(\hat{Y})$ is defined,

$$V(\hat{Y}) = EV_S(\hat{Y}) = EV_S(\sum_{i\epsilon s} M_i)$$

Therefore, the variance of $V(\hat{Y})$, is defined as

$$\hat{V}(\hat{Y}) = \frac{n}{n-1}\left[\sum_{i\epsilon s} E(M_i^2) - \frac{1}{n}\left(\sum_{i\epsilon s} E(\,M_i)\right)^2\right] \quad (28)$$

Considered $E(M_i)$ and $E(M_i^2)$ in equation (28)

$$E(M_i) = E\left(\frac{D_{(i)}.S_{(i,j)}}{D_{(i)}.S_{(i,j)} + D_{(j)}.S_{(j,i)}}\right) = \frac{D_{(i)}.E\left(S_{(i,j)}\right)}{D_{(i)}.E\left(S_{(i,j)}\right) + D_{(j)}.E\left(S_{(i,j)}\right)} = \frac{D_{(i)}}{D_{(i)} + D_{(j)}} \quad (29)$$

Let $M_i^2 = \frac{D_{(i)}^2 S_{(i,j)}^2}{D_{(i)}^2 S_{(i,j)}^2 + 2D_i S_{i,j} D_j S_{j.i} + D_{(j)}^2 S_{(j,i)}^2}$

$$E(M_i^2) = E\left(\frac{D_{(i)}^2 S_{(i,j)}^2}{D_{(i)}^2 S_{(i,j)}^2 + 2D_i S_{i,j} D_j S_{j.i} + D_{(j)}^2 S_{(j,i)}^2}\right)$$

$$= \frac{D_{(i)}^2 E\left(S_{(i,j)}^2\right)}{D_{(i)}^2 E\left(S_{(i,j)}^2\right) + 2D_i E\left(S_{i,j}\right) D_j E\left(S_{j.i}\right) + D_{(j)}^2 E\left(S_{(j,i)}^2\right)}$$

$$= \frac{D_{(i)}^2}{D_{(i)}^2 + 2D_i D_j + D_{(j)}^2} \quad (30)$$

Substituting $E(M_i)$ in equation (31) and $E(M_i^2)$ in Equation (32) into Equation (30), the proposed variance is given as

$$\hat{V}(\hat{Y}) = \frac{n}{n-1}\left[\frac{D_{(i)}^2}{D_{(i)}^2 + 2D_i D_j + D_{(j)}^2} - \frac{1}{n}\left(\frac{D_{(i)}}{D_{(i)} + D_{(j)}}\right)^2\right] \quad (31)$$

2.7 Simulation study

To evaluate the robustness of the proposed estimator, the simulated and computed the outcome variable related to gender for various sample sizes (1-50, 51-100, 101-150, 151-200, 201-250, 251-300, and above 300) selected from the overall population of 10000. A total of seven simulations were conducted, beginning with a sample that included a participant with a 1-50degree distribution. For each stage, the population proportions of IDU were determined for male and female participants, along with the variance. The second stage consisted of participants with 1-50 and 51-100 degree distributions, and this process continued until the study included participants with degree distributions exceeding 300.

3. Results and Discussion

3.1 Simulated results for the Effect of degree distribution for proposed and existing RDS estimators

Table 1 presents a comprehensive overview of a simulated population, highlighting various degree distributions among participants. The analysis revealed intriguing patterns in demographic estimates based on a degree distribution that spans from 1 to 50. Specifically, for the Naive method, the estimates show a slightly higher proportion of males (0.682) compared to females (0.318) and $RDS_{proposed}$ recorded a proportion of females (0.308) and males (0.692). In the case of the SH-RDS method, the gender distribution shifts slightly, with females making up (0.315) and males (0.685). This trend continues with the VH-RDS approach, where the estimates reveal that females represent (0.313) while males account for (0.687). Finally, the G-SS method yields the closest gender distribution, showing a slight decline in female participation to (0.310) and a corresponding increase in male representation to (0.690). These findings illustrate subtle yet significant variations in gender distribution across different estimation methods within the simulated population. Participants included in the study were those possessing a degree score between 51 to 100. The analysis of the results reveals specific findings for various methodologies employed. For the Naive approach, the distribution of participants was noted as follows: females constituted 0.308) while males represented (0.692). In the $RDS_{proposed}$ estimator, the gender breakdown shows females at (0.317) and males at (0.684). When assessing the SH-RDS, female participation remained consistent at (0.317), while male participation slightly decreased to (0.683). The VH-RDS produced similar results, with females at (0.316) and males at (0.685). Finally, in the G-SS approach, female participants accounted for (0.312) while the male counterparts stood at (0.688). These results illustrate the gender distribution across various sampling techniques, highlighting the nuances in participant representation. It's noteworthy to observe that the majority of the estimated proportions exhibited a remarkable consistency across various estimators. However, there were distinct deviations associated with the $RDS_{proposed}$ and G-SS, which stood out in contrast to the overall stability.

Table 1: Estimates for the proportion and variance of degree distribution for proposed and existing RDS estimators

Degree distribution		Naïve		$RDS_{proposed}$		SH-RDS		VH-RDS		G-SS	
		EST.	σ^2	EST.	σ^2	EST.	σ^2	EST.	σ^2	EST.	σ^2
1-50	Female	0.318	3.02	0.308	3.05	0.315	1.13	0.313	2.72	0.31	2.99
	Male	0.682		0.692		0.685		0.687		0.69	
51-100	Female	0.308	3.12	0.317	2.71	0.317	1.42	0.316	2.45	0.312	2.81
	Male	0.693		0.684		0.683		0.685		0.688	
101-150	Female	0.323	3.22	0.322	2.45	0.321	1.96	0.323	2.22	0.316	2.58
	Male	0.677		0.678		0.679		0.677		0.684	
151-200	Female	0.326	3.25	0.344	2.22	0.321	2.22	0.335	2.01	0.317	2.38
	Male	0.674		0.656		0.679		0.665		0.683	
201-250	Female	0.338	3.71	0.336	2.04	0.312	2.49	0.314	1.98	0.319	2.10
	Male	0.662		0.664		0.688		0.686		0.681	
251-300	Female	0.346	4.06	0.351	1.75	0.314	2.98	0.317	1.86	0.316	1.89
	Male	0.654		0.649		0.686		0.683		0.684	
>300	Female	0.354	4.12	0.308	1.65	0.313	3.53	0.318	1.97	0.314	1.78
	Male	0.646		0.692		0.687		0.682		0.686	

As illustrated in Figure 1, when a new group of participants whose degree range spanned from 1 to 150 is introduced, the results of this introduction unveiled fluctuations in the variance observed across the various estimators. Notably, the Naïve estimator experienced an increase, rising from 3.02 to 3.22 while the $RDS_{proposed}$ decrease from 3.05 to 2.45. In contrast, the proposed estimator exhibited a marked decrease, dropping from 2.45 to an impressive 2.01. Among the estimates, the SH-RDS, while witnessing a rise in variance from 1.13 to 1.96, still retained its position as the smallest variance overall. Turning the attention to VH-RDS, a reduction in variance is observed as well, with figures falling from 2.72 to 2.22. Similarly, G-SS shows a decline from 2.99 to 2.58, reinforcing the trend of variance reduction among certain estimators. In a parallel procedure, the participant pool is further expanded by including individuals with degrees ranging from 151 to 250. The outcomes of this addition revealed that both VH-RDS and the $RDS_{proposed}$ estimator continued this trend of decreased variance, dropping from 2.01 to 1.98 and from 2.22 to 2.04, respectively.

Moreover, the reduction in variance was particularly noticeable in G-SS, which fell from 2.58 to 2.10. In contrast, the Naïve

estimator and SH-RDS saw the variances increase significantly, climbing from 3.25 to 3.71 and from 2.22 to 2.49, respectively. This detailed analysis underscores the complex and evolving nature of variance among these estimators as the participant degrees fluctuated throughout the study.

As anticipated, the variance of the proposed RDS VH-RDS, and G-SS estimators demonstrates a clear decline with the inclusion of participants whose degrees fall within the range of 251 to 300 and beyond. Specifically, the variance associated with the proposed RDS estimator decreased significantly from 1.75 to 1.65, indicating a more precise estimate as the sample size expands. Similarly, the G-SS estimator shows a reduction in variance from 1.89 to 1.78. In contrast, the VH-RDS variance exhibited an increase, rising from 1.86 to 1.97, suggesting that this estimator may not perform as effectively under these conditions.

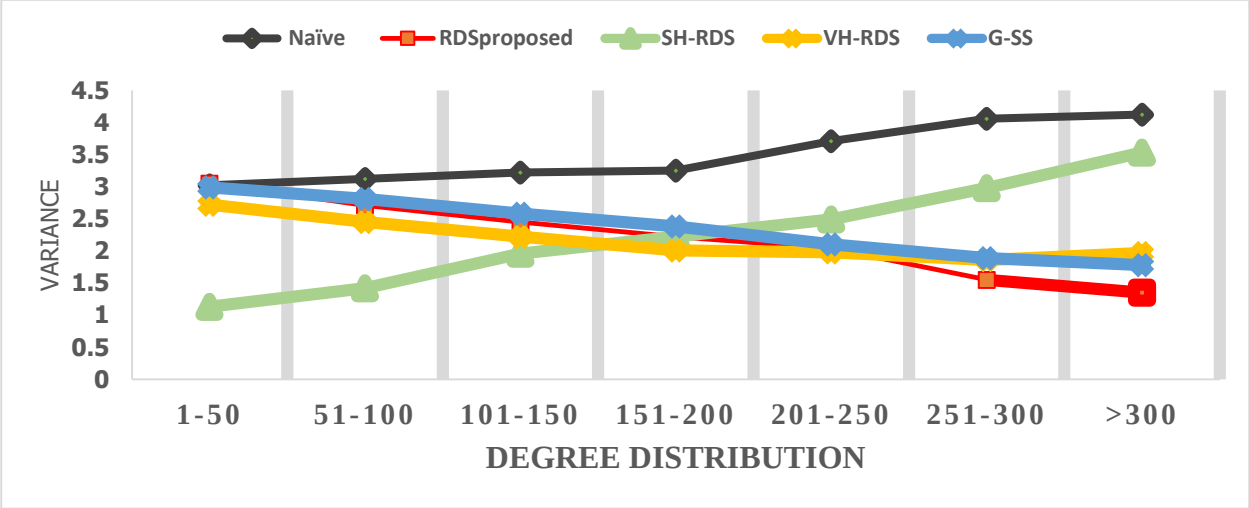

Figure 1: Simulated variance estimate of degree distribution for $RDS_{proposed}$ and Existing RDS estimators

This data is visually represented in Figure 1, reinforcing the statistical trends observed. Overall, the proposed estimator demonstrated exceptional performance, particularly in contexts where the degree distribution within the population was notably high. This observation is underscored by the consistently reliable estimates produced across a variety of degree sizes, emphasizing the robustness and adaptability of the estimator under diverse conditions. Moreover, it is worth noting that the SH-RDS estimator exhibited reliability when applied to populations with a low degree distribution. In practical terms, utilizing the SH-RDS estimator in these scenarios may result in more accurate and consistent outcomes, effectively reducing the likelihood of bias that can accompany the use of more complex estimators in such specific contexts. Conversely, when faced with a high degree of distribution, the VH-RDS and G-SS estimators present viable alternatives, providing options for researchers seeking reliable sampling methods tailored to the characteristics of the populations.

3.2 Application of Real-world data

3.3 Injecting Drug Users (IDU) in an Eastern European City

The first example comes from a comprehensive survey conducted in 2007 that focused on injection drug users (IDUs) in several major cities of a former Soviet bloc nation. This country is facing a severe HIV epidemic primarily driven by IDU activities, with alarming prevalence rates exceeding 50% in various urban centers. To effectively study this vulnerable population, researchers employed RDS, which allowed for an in-depth understanding of these interconnected groups. In this particular city, the estimated number of IDUs is around 1,200 (Robineau *et al*. 2020). This study systematically partitioned the dataset based on the degree distribution into specific ranges: 1-50, 51-100, 101-150, 151-200, 201-250, 251-300, and above 300. This categorization aimed to rigorously evaluate the performance of both the proposed and existing Respondent-Driven Sampling (RDS) estimators.

In the initial phase of the analysis, this study focused on the subset of the dataset with a degree distribution between 1 and 50. The findings from this analysis revealed that the proposed estimator yields results that are remarkably similar to those produced by the established VH-RDS and G-SS estimators, especially within larger population segments. As highlighted in Table 2, the proportion estimate from the SH-RDS estimator stands at 0.123, which closely corresponds with the estimates from VH-RDS at 0.122 and G-SS at 0.123.

In contrast, the proposed estimator produced an estimate of 0.111, while the Naïve estimator yielded a slightly higher estimate of 0.117. This indicates that although the proposed method shows promise, particularly in smaller degree distributions, it does demonstrate a slight divergence from the traditional estimators, emphasizing the nuanced performance

differences across varying population scales. The results substantiate the efficacy of the proposed estimator and its potential applicability in RDS-based research contexts.

Table 2: Estimates of proportion and variance of IDU in an Eastern European City for proposed and existing RDS estimators

Estimator	Naïve		$RDS_{proposed}$		SH-RDS		VH-RDSI		G-SS	
Parameter	EST.	σ^2	EST.	σ^2	EST.	σ^2	EST.	σ^2	EST.	σ^2
1-50	0.117	2.32	0.111	2.78	0.123	1.43	0.122	2.42	0.123	2.32
51-100	0.124	3.21	0.123	2.43	0.120	1.66	0.121	2.23	0.124	2.17
101-150	0.135	4.32	0.121	2.21	0.121	1.98	0.123	2.15	0.121	2.12
151-200	0.154	4.65	0.122	1.65	0.124	2.23	0.126	2.06	0.125	2.08
201-250	0.185	5.3	0.122	1.23	0.133	2.76	0.14	1.98	0.129	1.97
251-300	0.195	5.42	0.122	1.15	0.142	2.89	0.152	1.95	0.129	1.94
300	0.14	5.65	0.122	1.11	0.147	3.1	0.159	1.95	0.131	1.67

Figure 2 shows that, upon the incorporation of an additional dataset comprising 51 to 100 observations, the results revealing notable shifts in both the proportion and variance of the estimators. Specifically, the variance estimator exhibited a decrease, with its value transitioning from 2.78 to 2.43, indicating a reduction in variability within the dataset. In contrast, the Naïve estimator experienced an increase in its value, rising from 2.32 to 3.21, which suggests a growing level of uncertainty associated with this method. Furthermore, the SH-RDS estimator demonstrated a modest increase, moving from 1.43 to 1.66, yet it retained its status as the most effective estimator in this analysis. The VH-RDS estimator, on the other hand, shows a decline, decreasing from 2.42 to 2.23, which may indicate a diminishing performance in accurately capturing the underlying characteristics of the data. The G-SS estimator shows improvement, with its value decreasing from 2.32 to 2.17, suggesting it became more reliable as the dataset expanded. As the study continued to augment the dataset with varying degree distributions, both the VH and G-SS estimators exhibited progressively superior performance, highlighting the adaptability and effectiveness in handling diverse data structures. This reinforces the importance of choosing the right estimation strategy based on the specific characteristics of the data being analyzed.

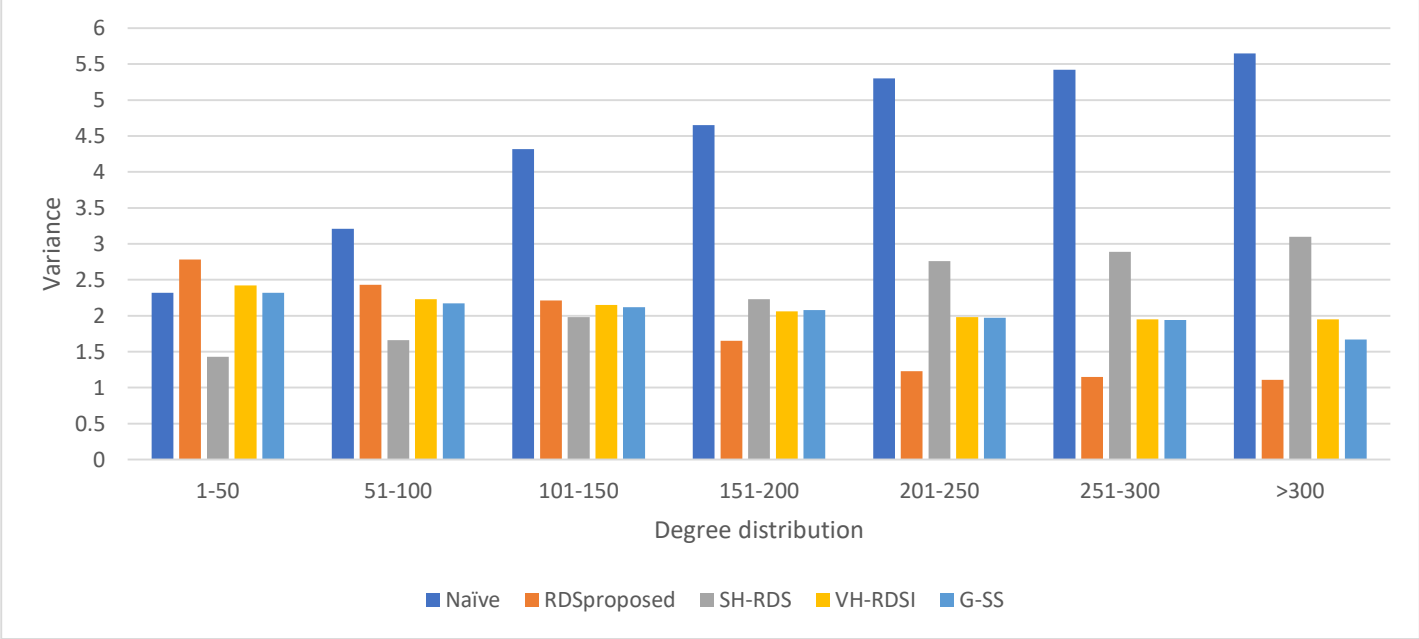

Figure 2: Real world variance estimate of degree distribution for $RDS_{proposed}$ and Existing RDS estimators

In a comparative analysis of different estimation techniques, the Naïve estimator exhibited a notable lack of effectiveness, falling short of expectations. In stark contrast, the SH-RDS estimator showcased a commendable level of performance, particularly when a dataset characterized by a degree distribution that spanned from 101 to 300 is examined. Notably, when degree distribution spanned from 101 to 300, the variance associated with the $RDS_{proposed}$ estimator experienced a significant reduction, dropping from 2.21 to a more favorable 1.15. Alternatively, the VH-RDS estimator demonstrated a modest decline in variance, decreasing from 2.15 to 1.95, while the G-SS estimator's variance showed a slight reduction from 2.12 to 1.94.

As the study expanded the dataset to include participants with a degree distribution exceeding 300, the SH-RDS estimator clearly emerged as the superior choice, eclipsing both the VH-RDS and G-SS estimators in terms of effectiveness. In this final evaluation, the variance for the $RDS_{proposed}$ estimator was a remarkable 1.11, illustrating in Figure 2 its precision, whereas the variances of the VH-RDS and G-SS estimators were recorded at 1.95 and 1.67, respectively. This evidence underscores the SH-RDS estimator's robustness and its capacity to deliver more reliable estimates under the tested conditions.

3.4 Drug Users in a Caribbean City

The second example comes from a small Caribbean nation that conducted extensive research on risk behaviors among injecting drug users (IDUs) and sex workers (SWs) across four major cities in 2008. This thorough study aimed to gather valuable insights into the dynamics of drug use within these populations. Notably, the research did not exclusively focus on individuals identified solely as injecting drug users; it also included other groups, fostering a more comprehensive understanding of the interconnected issues related to substance use and sexual health in the region. This inclusive approach enabled a more nuanced analysis of the challenges faced by these vulnerable communities (Ott *et al.* 2017). During the research, three coupons were initially distributed to the majority of respondents, ultimately generating a robust dataset consisting of 301 participants. From this group, only the 285 individuals who provided complete and reliable data were included, ensuring the integrity and depth of the findings.

The analysis presented in Table 3 provides a comprehensive overview of the performance of various estimators in measuring the IDU proportion across different degree distributions. The $RDS_{proposed}$ estimator achieved an IDU proportion of 0.022. In comparison, the Naïve estimator yielded a higher IDU proportion of 0.037, indicating potential biases in its estimator. The SH-RDS estimator recorded an IDU proportion of 0.032, while the VH-RDS shows a relatively lower value of 0.014. Finally, the G-SS estimator reported an IDU proportion of 0.024 for the sample group with a degree of 1-50.

Upon incorporating additional data corresponding to a degree range of 51-100, a systematic analysis revealed a notable reduction in the variance of the proposed estimator, specifically decreasing from 3.63 to 3.55. This indicates an enhancement in the estimator's precision as more data points were introduced. Conversely, the variance for the SH-RDS method increased from 1.65 to 1.98, suggesting that the inclusion of more complex data may have introduced additional variability in its estimates. On the other hand, the VH-RDS method experienced a slight decline in variance, moving from 2.34 to 2.26, while the G-SS estimator also saw a decrease from 2.98 to 2.67. These results not only highlight the relative strengths and weaknesses of each method but also suggest a consistent pattern similar to the findings from the initial example. Overall, this analysis underscores the importance of selecting the appropriate estimation method based on the characteristics of the dataset and the specific research objectives.

Table 3: Estimates of proportion and variance of IDU in a Caribbean City for proposed and existing RDS estimators

Estimator	Naïve		Proposed		SH-RDS		VH-RDS		G-SS	
Parameter	EST.	σ^2	EST.	σ^2	EST.	σ^2	EST.	σ^2	EST.	σ^2
1-50	0.037	2.54	0.022	3.63	0.032	1.65	0.014	2.34	0.024	2.98
51-100	0.024	3.46	0.027	3.55	0.032	1.98	0.019	2.26	0.025	2.67
101-150	0.021	3.43	0.029	2.96	0.033	2.12	0.088	2.21	0.029	2.32
151-200	0.014	3.65	0.032	2.32	0.054	2.34	0.032	2.13	0.031	2.19
201-250	0.081	4.12	0.033	2	0.042	2.41	0.032	2.11	0.031	2.00
251-300	0.091	4.32	0.033	1.95	0.062	2.53	0.035	2.09	0.032	1.99
300	0.02	4.35	0.033	1.76	0.077	2.65	0.039	1.99	0.033	1.96

The study observations indicate a noteworthy relationship between the size of the degree distribution and the performance of the estimator. As the degree distribution expanded, the study recorded a significant improvement in the $RDS_{proposed}$ estimator performance, reflected by a marked reduction in variance, which decreased from 3.63 to an impressive 1.76. This negative trend was similarly evident in the VH-RDS and G-SS estimators, both of which demonstrated a decline in variance specifically, the VH-RDS dropped from 2.34 to 1.99, while the G-SS shifted from 2.98 to 1.96.

Conversely, the Naïve and SH-RDS estimators exhibited a troubling reversal of this trend. Instead of improving, the variance escalated with the increasing degree distribution size. The Naïve estimator's variance rose significantly from 2.54 to a striking 4.35, and the SH-RDS estimator's variance also increased, moving from 1.65 to 2.65. This contrasting behaviour highlights the complex dynamics at play within the estimation approaches as the distribution size changes.

When comparing these estimates to those from a previous example, the current estimates are notably smaller in both absolute and relative terms. This confined with Avery and Rotondi (2023), who stated that, although the proportion estimates, discrepancies remain well within the estimator's margin of uncertainty, it is important to highlight that the estimates obtained from RDS can be valuable in various contexts. Moreover, in this specific analysis, the point estimate of prevalence consistently exceeds 0.33 across all population sizes examined, emphasizing the robustness of these findings.

4. Conclusion

it was concluded that SH-RDS shows consistency result with less than 100 degree distribution , while Naïve, $RDS_{proposed}$, VH-RDS, and G-SS estimator performed fairly. For degrees ranging from 251 to 300 and beyond, the proposed RDS and G-SS estimators shows a decline in variance, while SH-RDS variance increased, indicating a potential decline in performance under these conditions. Among higher degree distributions (>300), the $RDS_{proposed}$ estimator emerged as the most effective, compared VH-RDS and G-SS. The $RDS_{proposed}$ estimator also shows a substantial decrease in variance as the sample size and degree increased.

Reference

Abdesselam, K., Ashton V., Linda P., Parminder D., Franco M., & Ann M. J. (2020). The Development of Respondent-Driven Sampling (RDS) Inference: A Systematic Review of the Population Mean and Variance Estimates. Drug and Alcohol Dependence 206:107702

Avery, L. & Rotondi, M. (2023). Evaluation of Respondent-Driven Sampling Prevalence Estimators Using Real-World Reported Network Degree. *Sociological Methodology* 53(2) 269–287.

Avery, L., Alison M., Sarah F., & Rotondi, M. (2021). A Review of Reported Network Degree and Recruitment Characteristics in Respondent Driven Sampling Implications for Applied Researchers and Methodologists. *PLoS ONE* 16(4).

Falorsia. P, D, Allevaa, G, and Petrarcab. F, (2023). Unbiased estimation strategies for respondent-driven sampling. *Statistical Journal of the IAOS* (39) 865–876.

Fellows, I. E. (2019). Respondent-Driven Sampling and the Homophily Configuration Graph. *Statistics in Medicine* 38(1):131–50.

Gile K J. (2024). Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation. *Stat.ME 1-36.*

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):135–46.

Górny A., & Napierała J. (2016). Comparing the effectiveness of respondent-driven sampling and quota sampling in migration research. *Int J Soc Res Methodol*. 19(6):645–61.

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

Heckathorn, D. D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social problems* 49, 11–34.

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

Jonsson, J., Stein, M., Johansson, G., Bodin, T., & Stro¨mdahl S. (2019). A performance assessment of web-based respondent-driven sampling among workers with precarious employment in Sweden. *PLoS ONE* 14(1): e0210183.

Krivitsky, P. N., Martina, M., & Michał B. (2022). Impact of survey design on estimation of exponential-family random graph models from egocentrically-sampled data. *Social Networks* 69, 22–34.

Lawson, N. (2017). Variance estimation in the presence of nonresponse under probability proportional to size sampling. 116-119. In the 6th Annual International Conference on Computational Mathematics, Computational Geometry & Statistics (CMCGS 2017) and 5th Annual International Conference on Operations Research and Statistics, 6-7.

Lawson, N., & Ponkaew, C. (2019). New variance estimator for unequal probability sampling without replacement in the presence of non-response. *The Journal of Applied Science* Vol. 18 No. 2: 1-10

Lee S., Ong A. R. & Elliott M. (2020). Exploring Mechanisms of Recruitment and Recruitment Cooperation in Respondent Driven Sampling. *J of Stat.* 36(2): 339–360.

Naser, A.A., Ahamed, M. Q., & Reed. H. A (2018). New procedure for selecting a sample with unequal probability without replacement. *Far East Journal of Mathematical Science*. 107(1):231-239

Ott, M. Q., Gile, K. J., Harrison, M. T., Johnston, L. G., & Hogan, J. W. (2017). Reduced bias for respondent-driven sampling: accounting for non-uniform edge sampling probabilities in people who inject drugs in Mauritius. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1-34.

Robineau, O., Marcelo, F. C. Gomes, C. K., Ligia, K., André, P., & Pierre-Yves B. (2020). Model-based respondent-driven sampling analysis for HIV prevalence in Brazilian MSM. *Scientific Reports 10:2646*

Rozemberczki, B., Allen, C., & Sarkar, R. (2021). Multi-scale Attributed Node Embedding. *Journal of Complex Networks* 9(2): cnab014.

Spiller, M. W., Gile, K. J., Handcock, M.S., Mar, C. M., & Wejnert, C. (2018). Evaluating Variance Estimators for Respondent-Driven Sampling. *Journal of Survey Statistics and Methodology*. 6(1):23–45.

Zins, S., & Jan P. B. (2020). Considering interviewer and design effects when planning sample sizes. *Survey Methodology* 46.1: 93-119