Mambilla Journal of Science and the Environment Vol. 9, Issue 2, Page No. 28-39

MAMBILLA JOURNAL OF SCIENCE AND THE ENVIRONMENT

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

RESPONDENT-DRIVEN SAMPLING MODEL FOR SAMPLING AND ESTIMATION OF HIDDEN POPULATION

[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 28 July 2025
Received in revised form 14 August 2025
Accepted 20 August 2025

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

ABSTRACT

The Respondent-Driven Sampling (RDS) has emerged as an effective method for recruiting hidden populations, contributing to a notable increase in its global usage. Its ability to provide reliable population estimates, supported by established statistical models like the Naïve estimator, Salganik and Heckathorn RDS estimator (SH-RDS), Volz and Heckathorn RDS estimator (VH-RDS), Gile Sucessive Sampling estimator (G-SS), enhances its robustness. A simulation study involving networked populations was conducted, comparing the performance of a proposed RDS estimator with that of existing ones across various sample sizes and degree distributions. The results revealed that a simulation from a population of 500 with a samplze of 150 showed the following gender proportions. The $RDS_{proposed}$, yielded 0.3033 (females) and 0.6967 (males), Naïve estimator provided 0.3133 (females) and 0.6867 (males), SH-RDS reported 0.3109 (females) and 0.6890 (males), VH-RDS gave 0.3103 (females) and 0.6897 (males), and G-SS produced 0.3059 (females) and 0.6941 (males). The $RDS_{proposed}$, estimator had a design effect of 2.36, indicating inefficiency, while the Naïve estimator was at 2.33. The SH-RDS method performed best with a design effect of 1.42. When the sample size increased to 200, $RDS_{proposed}$, VH-RDS, and G-SS showed lower design effects, with VH-RDS leading and expanding sample sizes from 350 to 475, which led to significant improvements in $RDS_{proposed}$, with design effect of 1.23. Real-life analysis from Project 90 (5,475 individuals) confirmed the $RDS_{proposed}$ method outperformed traditional methods in high degree distributions, reducing design effects. No estimator accurately featured drug dealers or non-White ethnicities. It was recommended that RDS research develop an estimator that can perform well in every degree distribution setting.

1. Introduction

The hidden or difficult-to-reach population generally refers to a small, frequently marginalized group of people who operate outside standard data collection methods, making them challenging to identify and reach. These populations usually face significant risks associated with public awareness, which can compromise their safety and confidentiality (Zins & Jan 2020). Examples of such groups include injecting drug users (IDU) who may navigate life in the shadows due to the stigma associated with drug use, victims of human trafficking who endure exploitation in silence, Men who sex with other men (MSM) may be anxious about societal backlash; individuals facing homelessness frequently navigate unstable living situations; and migrants may be without official recognition or supportive communities. Members of these communities often grapple with social stigmatization, which can intensify their desire for privacy and complicate any efforts to reach out and collect meaningful data (Sarah *et al.*, 2022).

In many cases, the absence of a reliable and comprehensive sampling frame compels researchers to turn to non-probability sampling techniques. Consequently, they frequently employ convenience sampling methodologies,

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

including snowball sampling, time-location sampling, and respondent-driven sampling(RDS), which incentivizes participants to refer others. RDS is an advanced sampling technique that utilizes social networks to recruit participants from hard-to-reach populations that typically do not have a formal sampling framework. This innovative approach has proved to be particularly effective for engaging with populations that are generally marginalized or overlooked in traditional research methods. In the field of medical research, RDS has primarily been used to recruit people from various high-risk groups, including those who use intravenous drugs, men who engage in sexual relationships with other men, city jazz musicians, and individuals experiencing homelessness (Card *et al.,* 2017; Heckathorn and Cameron, 2017; Lyons *et al.,* 2017; Sypsa *et al*., 2017; White *et al.,* 2015).

RDS starts with the selection of specific individuals known as "seeds," who are taken from a convenience sample of the broader target population. These seeds are asked to participate in a survey, which can be administered online or offline to improve accessibility. Following their participation, the seeds are prompted to invite a limited number of their contacts, called "peer-recruited participants," who are also part of the targeted population. To effectively oversee and manage the recruitment process, RDS utilizes a coupon system. RDS is backed by several well-recognized statistical models that bolster the robustness and reliability of the findings derived from this method. These include the Naïve estimator introduced by Heckathorn (1997), the initial RDS Heckathorn (RDS-HK1) model developed by Heckathorn (2002), the Salganik and Heckathorn RDS estimator (SH-RDS) created by Salganik and Heckathorn (2004), the Volz and Heckathorn RDS estimator (VH-RDS) designed by Volz and Heckathorn (2008), and the Gile Sucessive Sampling estimator (G-SS) put forward by Gile (2011).

The SH-RDS and VH-RDS estimators are particularly sensitive to unequal edge sampling probabilities. These estimators will likely underestimate the proportion of people with the particular characteristic; their effectiveness is significantly undermined by their dependence on the degree counts reported by respondents, which are notoriously prone to inaccuracies and bias in the RDS process (Lu *et al.,* 2013). This methodological choice imposes substantial constraints on the estimator's ability to accurately identify and estimate sizable hidden population groups that are frequently marginalized and difficult to access in traditional survey methodologies (Malmros *et al.,* 2016). This intricacy results in the overrepresentation of individuals who possess higher degrees within social networks, a phenomenon thoroughly examined by Giles (2024). A notable limitation inherent to the SH-RDS and VH-RDS estimator is their reliance on replacement sampling. The SH-RDS and VH-RDS estimator requires many waves of sampling to justify their reliance on a stationary distribution. In practice, the number of waves is small (almost always fewer than 20, and often 5 or fewer). This work considers a without-replacement process for which stationarity does not apply. Specifically, utilizing the groundbreaking model proposed by Naser *et al.* (2018) that modifies the core concepts of probability proportional to size without replacement (PPSWOR) (Horvitz & Thompson, 1952). To develop an RDS estimator using without replacement sampling, determine the hidden population proportion of dichotomous variables, such as male and female, and evaluate the performance of the proposed RDS estimator with simulation and real-world data.

While estimation techniques have been adapted for categorical results, the validation of RDS estimators has predominantly concentrated on binary outcomes, likely due to the scarcity of large networks with categorical results available for validation. The National Longitudinal Study of Adolescent to Adult Health, Harris and Udry (2021) and data from Project 90, as referenced by Avery *et al. (*2021) and Rozemberczki, Allen, and Sarkar (2021), have been widely utilized for RDS validation (Abdesselam 2019; Abdesselam *et al*. 2020; Fellows 2019; Spiller *et al*. 2018). Nevertheless, as Spiller *et al.* (2018) pointed out, these datasets differ significantly from the hidden population networks typically associated with RDS coupons. These populations are quite limited in size (N = 1,249 in the National Longitudinal Study of Adolescent to Adult Health, N = 1,259 in Facebook College), and rather than using reported network degree, a truncated proxy was employed, which further limits the range of degrees. In conclusion, our aim is to advance the current research on estimator validity by integrating observed real-world reported network degrees into larger simulated networks, and to assess estimator performance on categorical outcomes using both actual and synthetic networks.

2. Methods

Various estimators for RDS have been suggested and put into practice, including those by Heckathorn (1997, 2002), Salganik & Heckathorn (2004), Volz & Heckathorn (2008), and Gile (2011). Nonetheless, studies Gile, (2024), Zins and Jan (2020) show that no single estimator outperforms all others in every situation. This review provides a brief evaluation of some of the most frequently utilized estimators for sample proportion, particularly emphasizing their connections to a proposed estimator.

2.1 The Naïve Estimator

The naive estimator was proposed by Heckathorn (1997) and is simply the proportion of infected individuals found in the sample. The estimator is given by

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

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 Estimator (1997) is

i. Respondents recruit peers from their social contacts with equal probability.
ii. Sampling is done with replacement.
iii. The degree of respondents is normally distributed.
iv. The social network of the population is undirected.
v. The population forms a connected network.

2.2 SH-RDS Estemator

SH-RDS, as proposed by Salganik and Heckathorn (2004), is a population proportion estimator using with replacement sampling as

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

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

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

vi. Respondents recruit peers from their social contacts with equal probability.
vii. Each recruitment consists of only one peer.
viii. Sampling is done with replacement.
ix. The degree of respondents is reported without error.
x. The social network of the population is undirected., and
xi. The population forms a connected network.

2.3 VH-RDS Estimator

VH-RDS, as proposed by Voltz and Heckathorn (2008), is a population proportion estimator using with replacement sampling 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

i. Recruitment is random. When recruiting others, respondents select uniformly at random from their network.
ii. Each recruitment consists of only one peer.
iii. Sampling is done with replacement.
iv. Respondents accurately report their degree in the network.
v. Network connections are reciprocal.
vi. The population forms a connected network.
vii. 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

G-SS, as proposed by Gile (2011), is a population proportion estimator using without replacement sampling. The estimate of inclusion probability associated with a specific unit, denoted as unit i, was obtained 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

i. Recruitment is random. When recruiting others, respondents select uniformly at random from their network.
ii. Each recruitment consists of only one peer.
iii. Sampling is done without replacement.
iv. Respondents accurately report their degree in the network.
v. Network connections are reciprocal.
vi. The population forms a connected network.
vii. The population size N is known.

Design Effect

According to Zins and Jan (2020), the design effect is determined by comparing the variance from an estimator that relies on a typically complex sampling design to the variance of a different estimator that uses a simple random sample (SRS) with the same sample size. By making this comparison, researchers can assess the extra variability introduced by the sampling design and can evaluate the efficacy and efficiency of different survey approaches. The design effect can be expressed in the following manner:

$$D_{eff,p}(\hat{\theta}) = \frac{Var(\hat{\theta}_w)}{Var(\hat{\theta}_{srswor})} \quad (11)$$

Where $Var(\hat{\theta}_w)$ is the variance of the estimator used in RDS, and $Var(\hat{\theta}_{srswor})$ is the variance of the estimator that uses simple random sampling.

2.5 Proposed RDS Estimator

This section presents the proposed RDS estimator developed based on the assumption of samples without replacement. It must be emphasized that the uniform sampling approach does not yield the steady-state probability distribution for the random walker, primarily due to the potential inaccessibility of certain nodes within the network.

Assumptions of the Proposed Model

The assumptions of the proposed model are:

i. Respondents recruit peers from their social contacts with equal probability(random).
ii. Each recruitment consists of only one peer (throughout the sampling period, you cannot recruit more than one).
iii. Sampling is done without replacement.
iv. The degree of respondents reported has a negligible *error*.
v. The network is directed.
vi. The population forms a connected network.

Estimation of inclusion probability

The idea of Horvitz-Thompson estimation was extended to an RDS estimator called the PPSWOR. Since a node was recruited into an RDS sample with a probability proportional to its degree. Therefore, the probability of sampling i unit given that *j* unit was selected as proposed by Lawson and Ponkaew (2019) using without replacement sampling, and was defined as.

$$\mathrm{P}_i = \begin{cases} \dfrac{\delta_i}{\sum_{\mathrm{j}=1}^{\mathrm{N}} \delta_j - \sum_{\mathrm{j}=1}^{\mathrm{k}-1} \delta_{R_j}}, & \mathrm{j} \notin (\mathrm{R}_1, \ldots, \mathrm{R}_{\mathrm{k}-1}) \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

$$P(i|j) = \frac{\delta_i}{(1 - \delta_j)} \quad (13)$$

$P(i|j)$ = 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

$$\hat{\pi}_{H-T} = \frac{P(i|j)y_i + P(j|i)y_j}{\mathrm{P}_i} \quad (14)$$

And

$$Y = \sum_{i=1}^{N} \frac{z_i y_i}{z_i p_i} \quad (15)$$

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 (12), into equation (14) and obtain the proposed inclusion probability:

$$\hat{\pi}_{proposed} = \frac{\frac{\delta_i}{(1-\delta_j)}y_j + \frac{\delta_j}{(1-\delta_i)}y_i}{\frac{\delta_i}{\sum_{\mathrm{j}=1}^{\mathrm{N}} \delta_j - \sum_{\mathrm{j}=1}^{\mathrm{k}-1} \delta_{R_j}}} \quad (16)$$

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

$$\widehat{D}_{A_{H-T}} = \frac{\sum_{i=1}^{N} \delta_i z_i y_i / p_i}{\sum_{i=1}^{N} z_i / p_i} \quad (17)$$

Therefore, substituting p_i in equation (12), into equation (17) gave the proposed mean degree of RDS sample as

$$\widehat{D}_{A_{proposed}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \left(\delta_i z_i y_i / \left(\frac{\delta_i}{\delta_j - \delta_{R_j}} \right) \right)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \left(z_i / \left(\frac{\delta_i}{\delta_j - \delta_{R_j}} \right) \right)}$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\delta_i z_i y_i\left(\delta_j - \delta_{R_j}\right)/\delta_i\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(z_i\left(\delta_j - \delta_{R_j}\right)/\delta_i\right)}$$

$$= \frac{\sum_{i=1}^{N}(\delta_i z_i y_i/\delta_i)}{\sum_{i=1}^{N}(z_i/\delta_i)}$$

$$\widehat{D}_{(i)_{proposed}} = \frac{\sum_{i=1}^{N} z_i y_i}{\sum_{i=1}^{N}(z_i/\delta_i)} \quad (18)$$

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

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

and

$$\widehat{D}_{(0)_{proposed}} = \frac{\sum_{i=1}^{N}(1-z_i)y_i}{\sum_{i=1}^{N}(1-z)_i \delta_i^{-1}}, \quad if\ \ z_i = 0 \quad (20)$$

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

$$R = \begin{bmatrix} R_{ii} & R_{ij} \\ R_{ji} & R_{jj} \end{bmatrix} \quad (21)$$

The total of whole cells equals the sample size N.

$$N = R_{ii} + R_{ij} + R_{ji} + R_{jj} \quad (22)$$

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:

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

$$\hat{S}_{(z_i=0,z_j=1)} = \frac{\sum_{i:z_i=0}\sum_{j:z_j=1} y_{ij}\,/\hat{\pi}_{proposed}}{\sum_{i:z_i=0}\sum_{j:z_j=1} y_{ij/\hat{\pi}_{proposed}} + \sum_{i:z_i=0}\sum_{j:z_j=0,i\neq j} y_{ij}\,/\hat{\pi}_{proposed}} \quad (24)$$

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

$$\widehat{RDS}_{proposed_1} = \frac{\widehat{D}_{(0)}.\hat{S}_{WT(0,1)}}{\widehat{D}_{(0)}.\hat{S}_{WT(0,1)}+\widehat{D}_{(1)}.\hat{S}_{WT(1,0)}} \quad (25)$$

$$\widehat{RDS}_{proposed_{\mathbf{0}}} = \frac{\widehat{D}_{(1)}.\hat{S}_{WT(1,0)}}{\widehat{D}_{(1)}.\hat{S}_{WT(1,0)}+\widehat{D}_{(0)}.\hat{S}_{WT(0,1)}} \quad (26)$$

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.

Following Lawson (2017), we propose a variance estimator; our proposed population proportion variance $\hat{V}(\hat{Y})$ is defined as follows.

$$\hat{V}(\hat{Y}) = \frac{n}{n-1}\left[\sum_{i\epsilon s}\hat{M}^2{}_i - \frac{1}{n}\left(\sum_{i\epsilon s}\hat{M}_i\right)^2\right] \quad (27)$$

From the proposed estimator in equation (25), we can write

$$\hat{Y} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{D_{(i)}.S_{(i,j)}}{D_{(i)}.S_{(i,j)}+D_{(j)}.S_{(j,i)}} = \sum_{i\epsilon s} M_i \quad (28)$$

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

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

$$\hat{V}(\hat{Y}) = \frac{n}{n-1}\left[\sum_{i\epsilon s} E(M_i^2) - \frac{1}{n}(\sum_{i\epsilon s} E(\,M_i))^2\ \right] \quad (29)$$

Considered $E(M_i)$ and $E(M_i^2)$ in equation (30)

$$E(M_i) = E\left(\frac{D_{(i)}.S_{(i,j)}}{D_{(i)}.S_{(i,j)} + D_{(j)}.S_{(j,i)}}\right) = \frac{D_{(i)}.E(S_{(i,j)})}{D_{(i)}.E(S_{(i,j)}) + D_{(j)}.E(S_{(i,j)})} = \frac{D_{(i)}}{D_{(i)} + D_{(j)}} \quad (30)$$

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

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

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

$$= \frac{D_{(i)}^2}{D_{(i)}^2 + 2D_i D_j + D_{(j)}^2} \quad (31)$$

Substituting $E(M_i)$ in equation (30) and $E(M_i^2)$ in Equation (31), into Equation (29), the proposed variance is given as

$$\hat{V}(\hat{Y}) = \frac{n}{n-1}\left[\frac{D_{(i)}^2}{D_{(i)}^2 + 2D_i D_j + D_{(j)}^2} - \frac{1}{n}\left(\frac{D_{(i)}}{D_{(i)} + D_{(j)}}\right)^2\right] \quad (32)$$

We can omit the term $\frac{n}{n-1}$ in Equation (32), when it is closer to 1, then we can write $\hat{V}(\hat{Y})$ in the following form.

$$\hat{V}(\hat{Y}) = \left[\frac{D_{(i)}^2}{D_{(i)}^2 + 2D_i D_j + D_{(j)}^2} - \frac{1}{n}\left(\frac{D_{(i)}}{D_{(i)} + D_{(j)}}\right)^2\right] \quad (33)$$

2.7 Simulation study

To assess the robustness of the proposed estimator, $RDS_{proposed}$, we simulated and calculated the outcome variable of gender for a series of samples (n=150, 200, 250, 300, 350, 400, 450, 475) drawn from the overall population of 500. For each sample, we meticulously computed the proportions of male and female participants, along with important statistical metrics such as the design effect, standard error, and confidence interval.

3. Results and Discussion

3.1 Results of the Simulated study on the Dichotomy variable in the hidden population

The findings detailed in Table 1 reveal insights obtained from a comprehensive simulation comprising 2000 iterations, conducted across a population of 500 individuals with 150 samples drawn from this population. The proposed $RDS_{proposed}$ estimator yielded estimated proportions of 0.3033 for females and 0.6967 for males. In contrast, the Naïve estimation approach provided slightly different results, with calculated proportions of 0.3133 for females and 0.6867 for males.

Additionally, other estimators demonstrated comparable estimates: the SH-RDS produced proportions of 0.3109 for females and 0.689 for males, whereas the VH-RDS yielded estimates of 0.3103 for females and 0.6897 for males. Lastly, the G-SS estimator reported proportions of 0.3059 for females and 0.6941 for males. These results illustrate the nuanced variations in estimated proportions across different sampling methodologies, emphasizing the importance of method selection in obtaining accurate demographic estimations.

In a parallel simulation involving a larger sample size of 200 participants, the results are detailed in Table 1. The naïve estimator yielded values of (0.3143 for females, 0.6857 for males), while the SH-RDS produced estimates of (0.3139 for females, 0.6861 for males), and the advanced VH-RDS offered (0.3123 for females, 0.6877 for males). Notably, these estimates closely align with those derived from the smaller sample of 150 participants, suggesting a level of consistency across sample sizes. In contrast, the proposed estimator $RDS_{proposed}$ provided values of (0.3013 for females, 0.6987 for males), and the G-SS estimator generated (0.3088 for females, 0.6912 for males). These values exhibit a slight deviation when comparing the results from the larger 200-sample cohort to the previous 150-sample analysis.

Table 1**:** Estimates of gender proportion and design effect, for proposed and existing RDS estimators

Samples	Gender	Naïve		$RDS_{proposed}$		SH-RDS		VH-RDS		G-SS	
		EST.	DE	EST.	DE	EST.	DE	EST.	DE	EST.	DE
150	Female	0.3133	2.33	0.3033	2.36	0.3109	1.42	0.3103	2.17	0.3059	2.68
	Male	0.6867		0.6967		0.6891		0.6897		0.6941	
200	Female	0.3143	2.43	0.3013	2.32	0.3139	1.51	0.3123	2.1	0.3088	2.33
	Male	0.6857		0.6987		0.6861		0.6877		0.6912	
250	Female	0.3162	2.67	0.3153	2.28	0.3149	1.63	0.3163	1.97	0.3096	2.26
	Male	0.6838		0.6847		0.6851		0.6837		0.6904	
300	Female	0.3191	2.78	0.3373	2.2	0.3146	1.98	0.3282	1.94	0.3101	2.23
	Male	0.6809		0.6627		0.6854		0.6718		0.6899	
350	Female	0.3293	2.98	0.3273	2.02	0.3032	2.16	0.3049	1.9	0.3098	2.05
	Male	0.6707		0.6727		0.6968		0.6951		0.6902	
400	Female	0.3396	3.05	0.3443	1.5	0.3071	2.36	0.31	1.82	0.3096	1.68
	Male	0.6604		0.6557		0.6929		0.69		0.6904	
450	Female	0.3493	3.76	0.303	1.3	0.3084	2.41	0.3136	1.81	0.3097	1.56
	Male	0.6507		0.697		0.6916		0.6864		0.6903	
475	Female	0.3596	4	0.3347	1.23	0.3112	2.94	0.3195	1.8	0.3104	1.49
	Male	0.6404		0.6653		0.6888		0.6805		0.6896	

An analysis of the gender population proportion patterns detailed in Table 1 reveals distinct variations in the estimates produced by the proposed $RDS_{proposed}$ estimator and the G-SS estimator. These estimates fluctuate slightly within the range of 200 to 350 samples, stabilizing between 400 and 475 samples. This consistency is likely influenced by the without-replacement sampling methodology implemented in both the proposed $RDS_{proposed}$ and G-SS estimators. In contrast, the estimates generated by the SH-RDS and VH-RDS methods exhibit more pronounced variations, shifting from 150 to 350 samples, and then ranging from 350 to 475 samples as sample size increases. The Naïve estimator presents the most considerable fluctuation, with estimates spanning from 200 to 475 samples. These variations underscore a critical concern associated with the Naïve, SH-RDS, and VH-RDS estimators: they tend to overestimate the representation of individuals with higher degrees (those considered the most popular) within the population. This bias arises because the sampling mechanisms inherent in these methods disproportionately favour individuals with extensive social networks, leading to potential misrepresentation in the overall estimates. Consequently, the intended goals of the proposed $RDS_{proposed}$ estimator are particularly relevant in addressing this issue of overrepresentation, ensuring a more accurate and equitable reflection of the entire population's characteristics. The analysis of the design effects, as illustrated in Figure 1, demonstrates varying performance levels among the different estimators. The proposed estimator achieved a design effect of 2.36, indicating a high degree of inefficiency relative to the small sample size. The Naïve estimator followed closely behind, with a design effect of 2.33. Meanwhile, the VH-RDS and G-SS estimators recorded design effects of 2.17 and 2.68, respectively, at a sample size of 150. In stark contrast, the SH-RDS estimator exhibited a significantly lower design effect of 1.42, highlighting its relative efficiency in this context. Further examination of Figure 1 reveals that across sample sizes ranging from 150 to 200, the design effect estimates for the Naïve, $RDS_{proposed}$, VH-RDS, and G-SS estimators remained quite comparable. This stability suggests a consistent performance among these approaches within this sample range. However, the SH-RDS estimator diverged from this trend; its design effect increased from 1.42 at 150 samples to 1.51 at 200 samples, illustrating a slight deviation, yet it remains the most effective option among the estimators presented. This consistent trend underscores the varying capacities of different estimators in terms of design effects, directly impacting their applicability in research scenarios.

As illustrated in Figure 1, the simulation was conducted with sample sizes of 200 and 350 participants. The findings revealed a notable decline in the design effect across several estimators, specifically $RDS_{proposed}$, VH-RDS, and G-SS. The design effect decreased from 2.32 to 2.02, while the VH-RDS estimator showed a similar reduction from 2.1 to 1.9. Meanwhile, the G-SS estimator also demonstrated a decrease, falling from 2.33 to 2.05. At these sample stages, the VH-RDS estimator exhibited superior performance compared to the other estimators. In contrast, the results

for the $RDS_{proposed}$, naïve, SH-RDS, and G-SS estimators did not match the efficacy observed in VH-RDS, underscoring the effectiveness of the VH-RDS approach in capturing more reliable and accurate data outcomes in this sample.

As the sample size ranges from 350 to 475 participants, both the proposed $RDS_{proposed}$ estimator and the G-SS estimator demonstrate superior performance. Specifically, the design effect is significantly reduced from 2.02 to 1.50 for the proposed $RDS_{proposed}$ estimator and from 2.05 to 1.68 for the G-SS estimator. This improvement can be attributed to the implementation of sampling without replacement, which effectively mitigates issues related to double-counting and the overestimation of individuals possessing higher degrees of connectivity within the sample. As the sample size continues to increase, the robustness of the proposed estimator remains evident. For instance, at a sample size of 475, a marked reduction in the design effect of the proposed $RDS_{proposed}$ estimator is observed, highlighting its continued efficacy even in larger sample contexts. This consistent performance underscores the estimator's reliability and its capability to yield more accurate representations of the population being studied, ultimately enhancing the overall quality of the data collected.

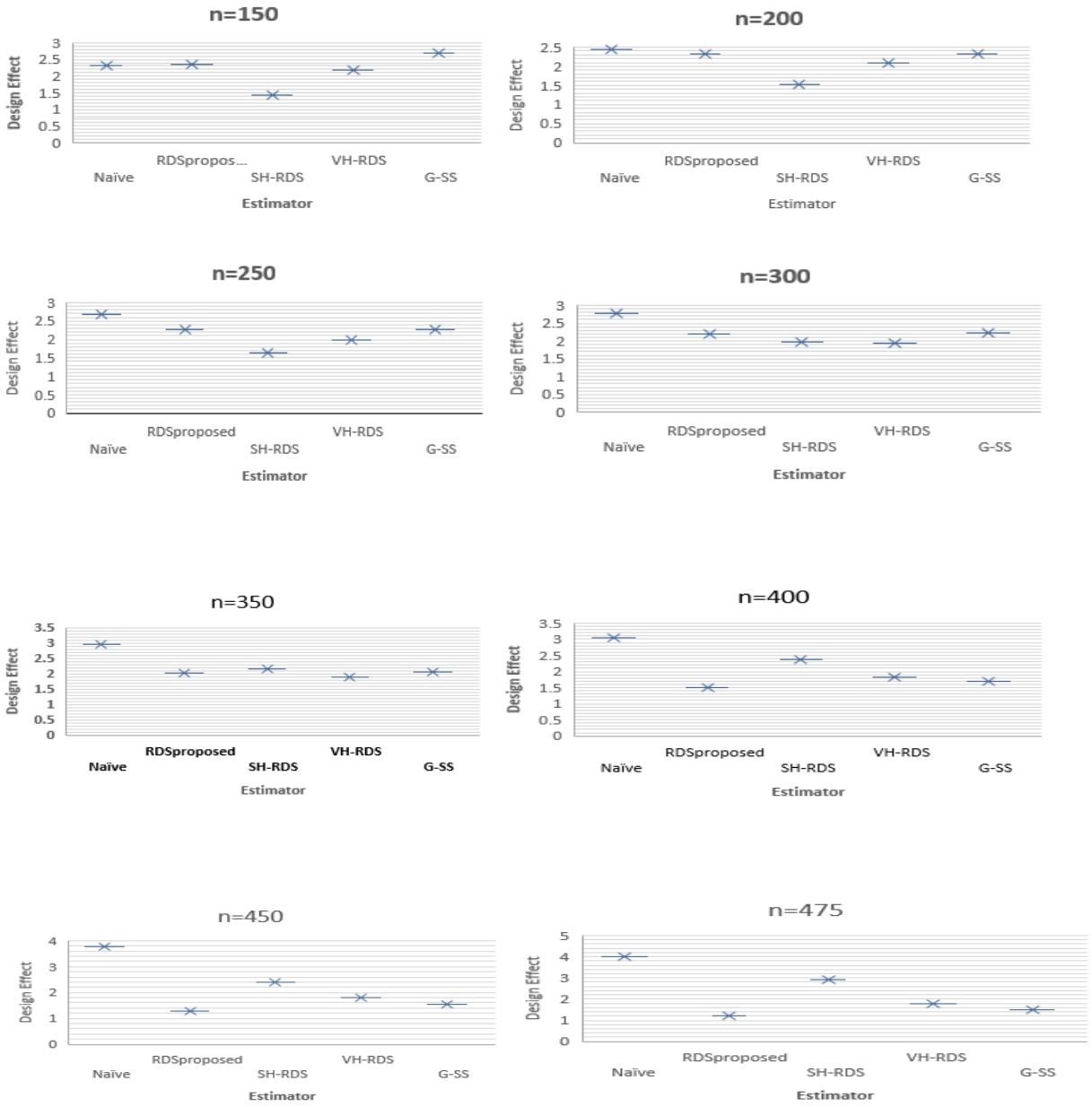

Figure 1: Estimates of gender Design effect for proposed and existing RDS estimators (n=150 to 475)

3.2 Application of Real-life Datasets

This section details the application of the proposed $RDS_{proposed}$ estimators to various real-life datasets, specifically focusing on three distinct surveys: Project 90, Facebook data, and injection drug users data. In this analysis, we anticipate encountering four distinct types of degree distributions among the populations: equally distributed, moderately distributed, very highly distributed, and very low distributed. To evaluate the efficacy of the proposed estimators, we will compare their performance against Naïve, SH-RDS, VH-RDS, and G-SS estimators. The entire analytical process, including the implementation of these estimators and comparative assessments, was meticulously coded in the *R* programming environment.

3.3 Estimating Project 90 Data

This example was drawn from the Project 90 dataset, which encompasses detailed information on 15 key characteristics for a total of 5,475 individuals. In prior analyses, the emphasis has primarily been on examining the single largest connected network within this dataset (Harris and Udry 2021; Avery *et al.*, 2021; Abdesselam *et al*. 2020). However, in this study, we broaden our approach by including all individuals in the dataset to ensure a more comprehensive analysis. The findings from the Project 90 dataset reinforced our simulation results, particularly highlighting the significant impact of degree distribution on the performance of various estimators.

The results presented in Table 2 offer detailed sample estimates regarding the characteristics of the Project 90 Population, revealing significant variations across different traits. The proposed $RDS_{proposed}$ estimator demonstrates remarkable performance advantages and enhanced accuracy compared to traditional methods such as the naïve, SH-RDS, VH-RDS, and G-SS estimators. This superior performance is particularly evident in scenarios where the degree distribution is markedly high. As a result, the $RDS_{proposed}$ estimator effectively reduces design effects across the majority of traits assessed, highlighting its robustness as an analytical tool within this population study.

However, an exception exists for the sex work client trait, where the G-SS estimator is favored, contingent upon accurate knowledge of the population size. Additionally, for the housewife demographic, the VH-RDS estimator outperforms the proposed estimator, highlighting its specific strengths in that context. Moreover, the naive sample proportion proves to be an inadequate.

Table 2: Estimating the Proportion and Design Effect of Project 90 Data Population Characteristics

	Status	Naïve		$\boldsymbol{RDS_{proposed}}$		RDS-I		RDS-II		G-SS	
		Est	DE	Est	DE	Est	DE	Est	DE	Est	DE
Equally	drug cook	0.010	4.3	0.010	1.49	0.015	3.8	0.010	1.24	0.010	1.36
	Homeless	0.010	3.24	0.010	1.35	0.010	2.83	0.020	1.32	0.010	1.31
	Retired	0.040	3.23	0.030	1.25	0.010	2.23	0.040	1.21	0.030	1.23
	Thief	0.020	5.4	0.030	1.69	0.030	4.3	0.030	1.83	0.020	1.25
	Pimp	0.030	3.54	0.020	1.56	0.030	2.96	0.020	1.76	0.020	1.25
	Housewife	0.040	3.42	0.070	1.32	0.070	2.36	0.050	1.47	0.060	1.29
	Disabled	0.040	5.4	0.030	1.79	0.040	4.8	0.050	2.54	0.040	1.48
	sex work client	0.070	3.21	0.070	1.42	0.090	2.43	0.080	2.34	0.090	1.36
	drug dealer	0.060	3.51	0.040	1.29	0.045	2.87	0.050	2.2	0.060	1.23
	Gender	0.440	4.53	0.450	1.51	0.450	3.82	0.420	3.05	0.450	1.64
	sex worker	0.070	4.65	0.060	1.39	0.060	3.83	0.050	3.12	0.050	1.87
	Unemployed	0.170	4.23	0.180	1.28	0.150	3.23	0.180	3.16	0.160	1.42
Moderate	drug cook	0.010	4.52	0.010	1.54	0.015	3.8	0.010	3.05	0.010	2.04
	Homeless	0.010	4.72	0.010	1.41	0.010	3.33	0.020	3.12	0.010	1.92
	Retired	0.039	4.32	0.029	1.34	0.010	3.43	0.040	3.23	0.030	1.82
	Thief	0.020	5.21	0.030	1.43	0.030	4.66	0.030	4.05	0.020	2.7
	Pimp	0.030	4.32	0.020	1.65	0.030	3.83	0.020	3.12	0.020	2.42
	Housewife	0.039	4.65	0.070	1.76	0.070	3.93	0.050	3.9	0.060	2.48
	Disabled	0.040	4.3	0.030	1.21	0.040	3.8	0.050	4.05	0.040	3.87
	sex work client	0.070	3.32	0.070	1.89	0.090	1.83	0.080	3.12	0.090	3
	drug dealer	0.056	3.65	0.060	1.95	0.045	1.23	0.048	3.9	0.060	3.67
	Gender	0.440	5.43	0.413	1.43	0.450	4.98	0.430	4.05	0.449	2.25
	sex worker	0.069	4.32	0.080	1.96	0.060	3.93	0.046	3.12	0.050	2.48
	Unemployed	0.180	4.21	0.178	1.76	0.149	3.63	0.176	3.9	0.159	2.84
Very high	drug cook	0.010	5.5	0.010	1.37	0.015	4.8	0.010	4.04	0.010	3.04

	Homeless	0.010	5.71	0.010	1.25	0.010	4.33	0.020	4.11	0.010	2.92
	Retired	0.039	5.31	0.029	1.19	0.010	4.43	0.039	4.22	0.030	2.82
	Thief	0.020	6.19	0.030	1.27	0.030	5.66	0.029	5.03	0.020	3.7
	Pimp	0.029	5.3	0.019	1.42	0.030	4.82	0.020	4.12	0.020	3.41
	Housewife	0.039	5.65	0.070	1.56	0.070	4.93	0.050	4.9	0.059	3.47
	Disabled	0.039	5.28	0.029	1.04	0.039	4.79	0.050	5.05	0.039	4.86
	sex work client	0.070	4.32	0.072	1.68	0.089	2.83	0.082	4.12	0.089	4
	drug dealer	0.055	4.65	0.062	1.73	0.045	2.23	0.048	4.9	0.067	4.67
	Gender	0.430	6.41	0.413	1.27	0.449	5.98	0.429	5.05	0.448	3.25
	sex worker	0.069	5.32	0.080	1.74	0.064	4.93	0.046	4.12	0.050	3.48
	Unemployed	0.190	5.21	0.178	1.56	0.149	4.63	0.176	4.9	0.159	3.84
Very low	drug cook	0.008	6.48	0.008	2.36	0.013	5.8	0.009	5.01	0.009	4.03
	Homeless	0.009	6.69	0.009	2.21	0.009	5.32	0.017	5.1	0.009	3.91
	Retired	0.034	6.31	0.025	2.19	0.009	5.43	0.035	5.21	0.026	3.81
	Thief	0.017	7.17	0.025	2.23	0.026	6.65	0.026	6.03	0.018	4.69
	Pimp	0.026	6.3	0.017	2.42	0.026	5.82	0.018	5.12	0.018	4.41
	Housewife	0.035	6.64	0.062	2.56	0.062	5.93	0.044	5.89	0.053	4.47
	Disabled	0.034	6.26	0.026	2.04	0.045	5.79	0.044	6.05	0.035	5.86
	sex work client	0.062	5.32	0.064	2.68	0.079	3.82	0.073	5.12	0.079	4.99
	drug dealer	0.049	5.64	0.055	2.73	0.050	3.22	0.043	5.9	0.059	5.67
	Gender	0.500	7.39	0.480	2.27	0.490	6.98	0.490	6.03	0.510	4.24
	sex worker	0.061	6.31	0.071	2.74	0.057	5.93	0.041	5.1	0.044	4.47
	Unemployed	0.169	4.3	0.158	1.49	0.132	3.8	0.156	1.24	0.141	1.36

estimator for RDS samples. This inadequacy is particularly pronounced when there is a significant variation within the group exhibiting the trait of interest. Our findings align closely with earlier studies Avery *et al.* (2021) regarding the diverse degree distribution; however, we present an intriguing observation of reduced design effect, indicating a noteworthy decrease in variability. We posit that this enhancement is a more accurate reflection of real-world performance. Consistent with the studies conducted by Gile and Handcock (2015) and Fellows (2019), our research reveals that the VH-RDS and G-SS methodologies exhibit superior accuracy and coverage compared to their predecessors, the SH-RDS and Naïve estimators.

4. Conclusion

Based on this finding, it was concluded that both the SH-RDS and VH-RDS are reliable estimators, particularly when the sample size approaches approximately fifty percent of the underlying population. It was observed that the Naïve estimator exhibited superior performance only in situations where the participants' degree distribution adhered to a normal pattern. The proposed RDS and G-SS demonstrated enhanced effectiveness when the degree distribution among participants was notably high. Additionally, this work concluded that the $RDS_{proposed}$, estimator outperforms all comparable estimators under conditions of a very high degree of distribution.

Reference

Abdesselam, K. (2019). Network Distribution and Respondent-Driven Sampling (RDS) Inference about People Who Inject Drugs in Ottawa, Ontario. PhD thesis, University of Ottawa, Ottawa, Canada.

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., 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)

Card, K.G., Lachowsky, N.J., & Cui, Z. (2017). Exploring the role of sex-seeking apps and websites in the social and sexual lives of gay, bisexual, and other men who have sex with men: a cross-sectional study. *Sex Health.* 14:229 -237.

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.

Harris, K. M., & Udry, R. J. (2021). National Longitudinal Study of Adolescent to Adult Health (Add Health), *Chapel Hill.* 1994–2018

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.

Heckathorn, D.D., & Cameron, C.J. (2017). Network sampling: from snowball and multiplicity to respondent-driven sampling. *Annu Rev Sociol*; 43(1):101-119.

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.

Lawson, N. (2017). Variance estimation in the presence of nonresponse under probability proportional to size sampling. pp. 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

Lyons, C.E., Grosso, A., & Drame, F.M. (2017). Physical and sexual violence affecting female sex workers in Abidjan, Côte d'Ivoire: prevalence, and the relationship with the work environment, HIV, and access to health services. *J Acquir Immune Defic Syndr*.;75:9–17.

Lu X., Malmros J., Liljeros F., & Britton T. (2013). Respondent-driven sampling on directed networks. Electron. *J. Stat*. 7, 292–322.

Malmros J., Masuda N., & Britton T. (2016). Random Walks on Directed Networks: Inference and Respondent-driven Sampling. *Journal of Official Statistics*, 32(2).

Naser, A.A., Ahmed, 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

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

Salganik, M. J., & Heckathorn, D.D. (2004). Sampling and estimation in Hidden population using respondent-driven sampling. *Sociological Methodology*, Vol. (34);193-239

Sarah, R., Michelle, A. D., · Jean C. D., · Yea-Hung, C., & Meghan, D. M. (2022). Respondent-Driven Sampling: A Sampling Method for Hard-to-Reach Populations and Beyond. *Current Epidemiology Reports* (2022) 9:38 47.

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.

Sypsa, V., Psichogiou, M., & Paraskevis, D. (2017). Rapid decline in HIV incidence among persons who inject drugs during a fast-track combination prevention program after an HIV outbreak in Athens. *J Infect Dis.* 215: 1496–505.

Volz, E., & Heckathorn, D. (2008). Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics*. 24(1):79–97.

White, R. G., Hakim, A. J., Salganik, M. J., Spiller, M. W., Johnston, L. G., Kerr, L., & Orroth, K. (2015). Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies: STROBE-RDS statement. *Journal of Clinical Epidemiology*, *68*(12), 1463–1471.

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