Statistical Journal of the IAOS 39 (2023) 865–876
DOI 10.3233/SJI-230087
IOS Press

Unbiased estimation strategies for respondent driven sampling

Piero Demetrio Falorsi[a,*], Giorgio Alleva[a] and Francesca Petrarca[b]
[a]*Sapienza University of Rome, Rome, Italy*
[b]*University of Roma Tre, Rome, Italy*

Abstract. In this paper, we focus on respondent-driven sampling (RDS), which is a valuable survey methodology to estimate the size and the characteristics of hidden or hard-to-measure population groups. The RDS methodology makes it possible to gather information on these populations by exploiting the relationships between their components. However, RDS suffers from the lack of an estimation methodology that is sufficiently robust to accommodate the varying conditions under which it is applied. In this paper, we address the estimation problem of the RDS methodology and, by approaching it as a particular indirect sampling technique, we propose three unbiased estimation methods as possible solutions.

Keywords: Hard to reach populations, snowball sampling, network sampling, GWSM estimator

1. Introduction

In this paper, we focus on respondent-driven sampling (RDS), which is a valuable survey methodology for both national and international organisations to estimate the size and characteristics of hidden (e.g., homeless people, undocumented immigrants) or hard-to-measure population groups (e.g., minorities, indigenous people).

The principle of "leaving no one behind" is at the heart of the 2030 Agenda and a key requirement for many Sustainable Development Goals (SDG) indicators is to be available for the most vulnerable and marginalised population groups. Nevertheless, halfway through the implementation of the 2030 Agenda, most SDG indicators are still not available at the needed level of disaggregation to monitor the socioeconomic conditions of hidden and hard-to-count population groups. As a result, it is neither possible to produce reliable structural data on the needed disaggregation dimensions nor to monitor the developments of emerging phenomena that need to be approached with targeted evidence-based policy interventions.

The RDS methodology makes it possible to gather information on these populations by exploiting the relationships between their components. Moreover, the effectiveness of the RDS can be further increased by employing an integrated approach in which the RDS is used in conjunction with other information sources, such as administrative or geographical data.

The RDS is a network-based sampling technique [1, 2] that was developed first by Eckathorn [3]. RDS has been the favourite survey method for sampling populations that are difficult to reach due to the potential of a viable sampling technique with reasonable inferential approaches. As a result, since its establishment, it has been employed in countless investigations of these populations across many countries [4]. RDS starts with a small sample of individuals ("seeds") with which the researchers are familiar. Each participant is then given a small number of coupons with unique identifiers to distribute to their contacts in the target population, enrolling them in the study and growing the sample size until the sample includes the desired number of respondents. The RDS process stops either when, in the selection process, we encounter only units already identified in the previous steps or at a predetermined data-collection step (e.g., the fifth step). Picture 1.1 below illustrates an example of a network sampling process articulated into three steps. The blue lines are the links

*Corresponding author: Piero Demetrio Falorsi, Sapienza University of Rome, Mobile 00393298604119. Via di Monserrato 111, 00186, Rome, Italy. E-mail: piero.falorsi@gmail.com.

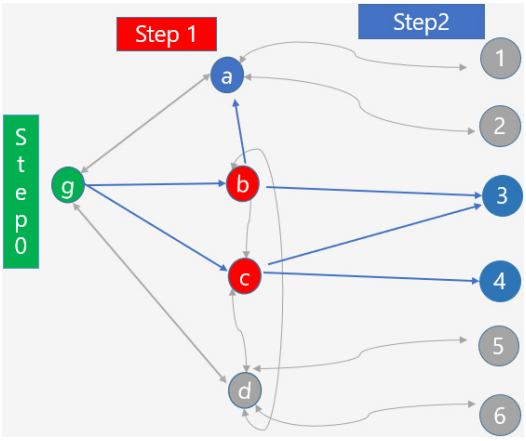


Fig. 1. Example of network sampling process*. *The thin grey lines represent links existing in the population but not observed in the sample.

observed in the sample. Up to and including step 2, participants g, b, c, a, 3, and 4 are kept in the sample. Participants d, 1, 2, 5, and 6 are not observed.

We may view RDS as a specific extension of the extensively used group of convenience sampling methods known as "snowball sampling networks" [5] which are frequently employed as a last option when a traditional sample frame is not available [6]. Compared to those of more traditional snowball sampling, RDS offers two key benefits. First, respondents receive few coupons. This enables statistical inference to be more appropriately defined and makes it more plausible to approximate the final sample as a probability sample. Second, asking respondents to pass coupons to their contacts in a potentially stigmatised community reduces potential confidentiality issues. Due to this innovation, RDS is a compelling method for gathering data from marginalized and difficult-to-reach populations.

However, the RDS methodology suffers from the lack of an estimation methodology that is sufficiently robust to accommodate the varying conditions under which it is applied. Although it is quite robust for estimating mean and proportion values [7], the accuracy of the total estimates depends on several features including the nature of the network connecting the individuals in the population and the lack of a rigorous sampling approach to select the sampling units.

In this paper, we address the estimation problem and by approaching the RDS methodology as a particular indirect sampling technique [8], we propose three unbiased estimation methods as possible solutions. In particular, the first method assumes a random sampling of the initial individuals. In contrast, the second method, which considers purposive sample selection, creates a nonbiased estimation if the initial sample of respondents falls into all the clusters of networks that characterise the population of interest. Finally, leveraging the generalised capture-recapture estimation approach [9], we propose an estimator that accounts for the noncoverage of two independent indirect samplings.

The paper is organised as follows. In Section 2, we summarise the traditional methodology of the RDS methodology, illustrating the data collection technique and the Volz and Heckathorn estimator [10], which has been very successful in practical applications due to its lack of computational complexity. Section 3 introduces the basic symbology, and we show how the RDS can be seen as a particular specification of indirect sampling in which each survey wave represents the indirect basis for the subsequent RDS phase. In Section 4, we expand the sampling aspects in the RDS. Section 5 introduces the three estimators. Section 6 concludes the work, and we begin to outline a strategy to overcome information gaps for SDG indicators for hard-to-reach populations, focusing on indigenous peoples.

2. Data collection and estimation in the classical RDS approach

RDS is frequently carried out by using techniques suggested by Salganik et al. [11] and outlined in protocols such as those proposed by White et al. [4] and Johnston [12].

A preliminary sample of typically 2 to 10 seeds is chosen. Aiming to represent all the key socioeconomic subpopulations that researchers anticipate may exist in the target population, seeds are selected to be as varied as possible. The rationale of this derives from the fact that each subpopulation may represent a separate network (or a cluster) of target individuals. If we select a seed in a given subpopulation, we can explore the network of related individuals. In contrast, if we do not select any individual in a subpopulation, the specific cluster of individuals cannot be observed in the RDS process. Therefore, picking up in the initial sample all possible distinct networks increases the possibility of constructing unbiased inferences on the target population.

The enumerators should include community opinion leaders in the initial seeds. Hence, their acceptance and support of the survey method may likely inspire widespread involvement from other target population members. This buy-in is crucial in target populations

that are unlawful or stigmatised, especially if the population has any prior exposure to risky research practices. Following an interview, the seeds are given some coupons, each with a unique identification number, to spread to other population members. This number was used to reconcile the practical need to prevent the early termination of the sample trees with the inferential aim of limiting the branching of the sample. Members of the population who receive coupons visit a *study centre* where they are directly interviewed or given an interview appointment. Three coupons are likewise supplied to subsequent replies; this process continues until the sample size is approximately reached and the coupons are tapered or discontinued. Participants are paid for their time spent taking the survey. Additionally, for each successful recruiting, participants receive rewards. In the survey, the number of target population contacts of each respondent must be measured. This is typically done by asking questions that narrow the recruit's references to the precise definition of the target population. Interviewers must also verify membership in the target population. Researchers must also assure participants do not participate in the survey more than once. Study staff are familiar enough with the target population in many settings to notice repeat participation attempts. In other cases, repetition is prevented by collecting non-identifiable but unique information about participants, as in Johnston [13].

The RDS methodology can be applied alternatively to the entire population of individuals or by considering only the subpopulation at risk of belonging to the target population. For instance, if the target population coincides with forced labour people, we may observe people working in sugarcane.

Contrary to what was previously believed, Eckathorn [3] used a Markov modelling of the peer recruitment process to demonstrate that bias from the convenience sample of beginning participants from which the sample started gradually diminished as the sample increased wave by wave. By using the model, they demonstrated how the sample approached an equilibrium independent of the beginning location or independent of the convenience sample of seeds from which it started as it expanded wave by wave. The conclusion was that this sampling technique may become reliable if there were enough waves, meaning that any seed selection can eventually yield the same equilibrium sample composition. However, the researchers did not show how an unbiased estimate can be derived.

Eckathorn [14] introduced the first RDS population estimator based on the essential idea that in RDS, relationships tend to be reciprocal. This implies that if person A knows person B, then B knows A.

The estimator bases its validity on the principle of network balance between population subgroups. Up to a constant factor, the volume of network connections to and from each group can be approximated. For each pair of groups, this results in a set of balance equations that may be used to solve for the relative size of each group. Volz and Eckathorn [10] proposed a slightly biased estimator. In the following we call this estimator the VH estimator. The VH estimator is based on the following hypotheses [15]: 1. The network size is large compared to that of the realised sampling, including the initial seeds and the respondents recruited by the RDS process. 2. Homophily is weak enough, where homophily is the tendency for nodes to preferentially form network contacts with others like themselves. 3. Reciprocity of contacts. 4. With-replacement sampling. 5. Enough sample waves. 6. Accurate measurement. 7. The recruitment in the subsequent waves of the RDS process is random.

The first three hypotheses relate to the nature of the contact network, while Hypotheses 4 to 7 relate to sampling. Hypothesis 4 is the most critical and may introduce a certain level of bias, as the sampling process adopted assumes a link between the persons recruited.

Focusing on the first three assumptions, let us consider the example below in Fig. 2, where we assume that people of the target population belong to three disjoint clusters. If the starting sample includes only persons in one group, the traditional RDS can estimate only the total number of persons in that cluster suffering from a substantial undercoverage problem. Since people of the target population are often grouped geographically, observing each cluster's units in the starting sample is appropriate.

The VH estimator has had great application success in the practice of real investigation due to its great computational ease. Subsequently, other estimators have been proposed in the literature (see among others [16,17]) each of which overcomes some of the limitations associated with the assumptions made with the VH estimator. These estimators have higher computational complexity, and introduce some modelling assumptions on the cluster variables or the nature of the contact network.

To illustrate the classical VH estimator, let us consider the case where the total of a characteristic

$$Y = \sum\nolimits_{k \in U} y_k \tag{1}$$

in the population of interest U (of N units) is to be estimated, where y_k is the value of y for unit $k \in U$.

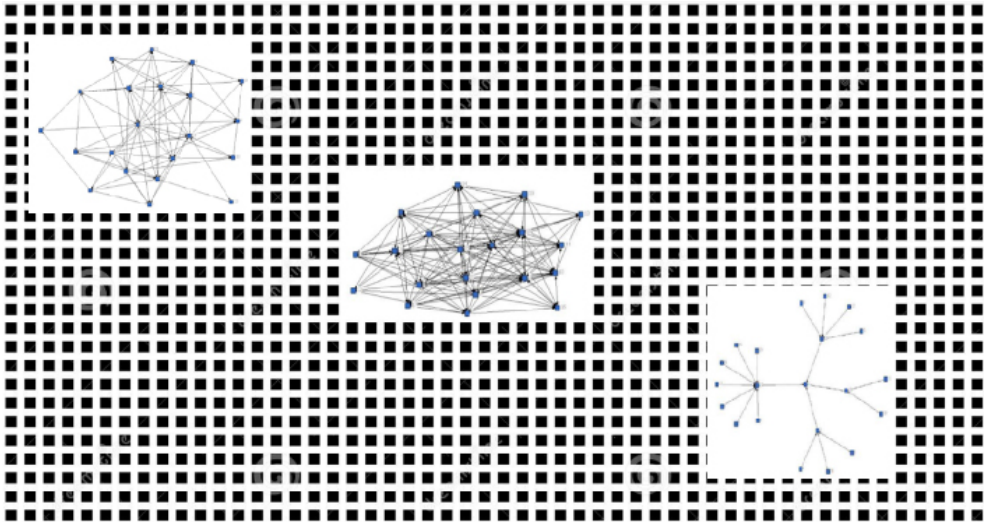


Fig. 2. Example of target population divided in disjoint clusters.

Let S be the sample of different units at the end of the RDS process. Let α_k be the number of times unit k is observed in S, where

$$m = \sum\nolimits_{k\in S} a_k.$$

The probability of selection of unit k is supposed to be proportional to their contacts

$$p_k = L_k/N\bar{L},$$

where

$$L_k = \sum\nolimits_{j\in U} \lambda_{j,k}, \bar{L} = \frac{L}{N}, \text{ and } L = \sum\nolimits_{k\in U} L_k$$

where $\lambda_{j,k}$ is the link (0,1) variable between individuals j and k. Let

$$\hat{p}_k = L_k/N\hat{\bar{L}}$$

be the *Hansen Hurwitz* (HH) [18] estimate of p_k, where

$$\hat{\bar{L}} = \left(\sum\nolimits_{\ell\in S} \frac{a_\ell L_\ell}{mp_\ell}\right) \Big/ \left(\sum\nolimits_{\ell\in S} a_\ell/mp_\ell\right)$$

is the HH estimate ratio of the average number of contacts in the population. The numerator on the right-hand side of the previous equality is given by:

$$\sum\nolimits_{\ell\in S} \frac{a_\ell L_\ell}{mp_\ell} = \sum\nolimits_{\ell\in S} \frac{a_\ell L_\ell}{mL_\ell} L$$

$$= \frac{1}{m} L \sum\nolimits_{\ell\in S} a_\ell = L.$$

The denominator can be expressed as

$$\sum\nolimits_{\ell\in S} \frac{a_\ell}{mp_\ell} = \sum\nolimits_{\ell\in S} \frac{a_\ell L}{mL_\ell}$$

$\hat{\bar{L}}$ and $\hat{p}_k$ are, therefore, given by:

$$\hat{\bar{L}} = \frac{L}{\sum_{\ell\in S} \frac{a_\ell L}{mL_\ell}} = \frac{m}{\sum_{\ell\in S} a_\ell \frac{1}{L_\ell}},$$

$$\hat{p}_k = L_k \Big/ N\frac{m}{\sum_{\ell\in S} a_\ell \frac{1}{L_\ell}}.$$

The *VH* estimator of Y is

$$\hat{Y}_{VH} = \sum\nolimits_{k\in S} \frac{a_k y_k}{m\hat{p}_k} = \sum\nolimits_{k\in S} y_k w_{VH,k} \tag{2}$$

where

$$w_{VH,k} = \frac{a_k}{m\hat{p}_k}. \tag{3}$$

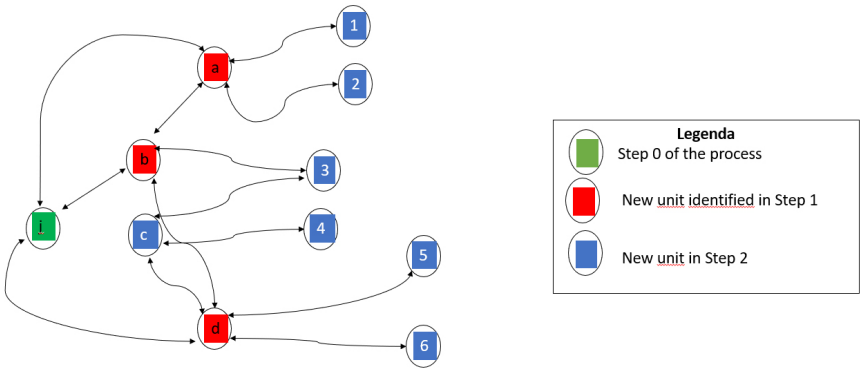
is the sample weight assigned to unit k.

3. Totals of interests and a formalisation of the RDS as a particular case of indirect sampling

The RDS can be formalised as an indirect sampling scheme. In this type of sampling, there is an initial U^A population of N^A units from which the research starts, and a U^B population of N^B units that constitute the study's target population.

In our case, $U^B \equiv U$ means that it coincides with the target population U, which implies $N^B = N$.

The specific unit j of the initial population U^A consists of the unit j itself and all its contacts. Let j and k be the labels identifying the population units in U^A and U^B, respectively. Let $\lambda_{j,k}$ be the link variable between units $j \in U^A$ and $k \in U^B$, where $\lambda_{j,k} = 1$ if j is directly linked to k. We have $\lambda_{j,k} = \lambda_{k,j}$ and $\lambda_{j,j} = 1$. Let



Step	Total
0	$Y_{S_0,0} = y_j$
1	$Y_{S_0,1} = y_j + y_a + y_b + y_d$
2	$Y_{S_0,2} = y_j + y_a + y_b + y_d + y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_d$

Fig. 3. Example of construction of the totals $Y_{S_0,R}$.

$$\bar{y}_j^A = -\sum_{k\in U^B} \frac{\lambda_{j,k}}{L_k^B} y_k \tag{4}$$

be the value of the characteristic of interest for unit $j \in U^A$, where

$$L_k^B = \sum_{j\in U^A} \lambda_{j,k}.$$

In the initial population U^A, each unit in contact with unit j contributes to the y value of that unit in a weighted manner, where the reciprocal of the total number of contacts gives the weight.

The two populations U^A and U^B have the same numbers of units: $N^A = N^B = N$. We have

$$\begin{aligned} N^A &= \sum_{j\in U^A} \sum_{k\in U^B} \frac{\lambda_{j,k}}{L_k^B} \\ &= \sum_{k\in U^B} \sum_{j\in U^A} \frac{\lambda_{j,k}}{L_k^B} \\ &= \sum_{k\in U^B} 1 = N^B = N. \end{aligned} \tag{5}$$

Moreover, the target parameter may also be expressed as the sum of the $\bar{y}_j^A$ values over the population U^A:

$$Y = \sum_{j\in U^A} \bar{y}_j^A. \tag{6}$$

Indeed, it is

$$\begin{aligned} Y &= \sum_{j\in U^A} \bar{y}_j^A \\ &= \sum_{j\in U^A} \sum_{k\in U^B} \frac{\lambda_{j,k}}{L_k^B} y_k \\ &= \sum_{k\in U^B} \sum_{j\in U^A} \frac{\lambda_{j,k}}{L_k^B} y_k = \sum_{k\in U^B} y_k. \end{aligned} \tag{7}$$

In addition to the total Y, another total that plays a crucial role in the RDS methodology is the aggregate, $Y_{S_0,R}$. That is, the total of the variable y where starting from the sample S_0 (which constitutes step 0 of the process), additional units are considered in subsequent steps through all their contacts. The total considers this aggregate after the step of this process. We can consider this as a search process on a graph.

Figure 3 illustrates this process starting from sample S_0, which includes only unit j.

To clarify how this total can be constructed, let us compute it step by step. To distinguish among the search processes of the different steps, let j_r (with $j_r \in U^A$ and $r = 0, 1, \ldots, R-1$) be the subscript of the population unit involved in step r of this search process on a graph.

At step 0, the total $Y_{S_0,0}$ is simply the sum of the values y_j for the units included in the sample S_0:

$$Y_{S_0,0} = \sum_{j_0\in S_0} y_j. \tag{8}$$

In step 1, we observe the links of units in S_0. We compute $Y_{S_0,1}$ by adding to the total $Y_{S_0,0}$ the values of y of the new units individuated in the search process starting from the units in S_0. We can formalise this process as

$$Y_{S_0,1} = \sum_{j_0\in S_0} \sum_{k\in U^B} y_k \frac{\lambda_{j_0,k}}{L_k^{S_0}}, \tag{9}$$

where $L_k^{S_0} = \sum_{j_0\in S_0} \lambda_{j_0,k}$.

To clarify the notation, we note that the j_0 unit is also one of the k units. We also note that $L_k^{S_0}$ is the total number of links to unit k (identified in step 1 of the process) that can be computed starting from the units in S_0.

We can reverse the order of the sums and formulate as

$$Y_{S_0,1} = \sum\nolimits_{k\in U^B} y_k \sum\nolimits_{j_0\in S_0} \frac{\lambda_{j_0,k}}{L_k^{S_0}} \tag{10}$$

where $L_k^{S_0} = \sum_{j_0\in S_0} \lambda_{j_0,k}$.

We note that Eq. (10) avoids the multiple counting of a unit linked to different elements of the initial sample, as illustrated in Fig. 4, where unit b is connected to both units j and a of S_0. Indeed, it is

$$\sum\nolimits_{j_0\in S_0} \frac{\lambda_{j_0,b}}{L_b^{S_0}} = \frac{\lambda_{a,b}}{L_b^{S_0}} + \frac{\lambda_{b,b}}{L_b^{S_0}} + \frac{\lambda_{j,b}}{L_b^{S_0}} = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1.$$

At **step 2,** we have

$$Y_{S_0,2} = \sum\nolimits_{j_0\in S_0} \sum\nolimits_{j_1\in U^A} \sum\nolimits_{k\in U^B} y_k \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_0}} \frac{\lambda_{j_1,k}}{L_k^B}. \tag{11}$$

We note that the last summation is over U^B, i.e., the target population. The middle summation is on U^A, i.e., on the initial population. The first summation is limited to the initial sample from which the research starts. This kind of organisation of the order of summations also appears in the following formulas. The last summation is always on U^B, and the first is on S_0. In contrast, the intermediate summations are always on the starting population U^A.

Reversing the order of the summations, we have

$$Y_{S_0,2} = \sum\nolimits_{k\in U^B} y_k \sum\nolimits_{j_0\in S_0} \sum\nolimits_{j_1\in U^A} \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_0}} \frac{\lambda_{j_1,k}}{L_k^B}. \tag{12}$$

Let us note that in this case, the unit $k \in U^B$ is counted only once, avoiding the multiple counting of a unit linked to different elements of the initial sample and its links. This is illustrated in Appendix 1.

...

Continuing recursively the above process, at step R we have

$$Y_{S_0,R} = \sum\nolimits_{j_0\in S_0} \sum\nolimits_{j_1\in U^A} \cdots \sum\nolimits_{j_{R-1}\in U^A} \sum\nolimits_{k\in U^B} y_k \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_0}} \frac{\lambda_{j_1,j_2}}{L_{j_2}^B} \times \ldots \times \frac{\lambda_{j_{R-2},j_{R-1}}}{L_{j_{R-1}}^B} \frac{\lambda_{j_{R-1},k}}{L_k^B} = \sum_{k\in U^B} y_k \sum_{j_0\in S_0} \sum_{j_1\in U^A} \cdots \sum_{j_{R-1}\in U^A} \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_0}} \frac{\lambda_{j_1,j_2}}{L_{j_2}^B} \times \cdots \times \frac{\lambda_{j_{R-2},j_{R-1}}}{L_{j_{R-1}}^B} \frac{\lambda_{j_{R-1},k}}{L_k^B}. \tag{13}$$

Even in this case, we avoid the multiple counting of a unit linked to different elements collected in the R steps of the RDS process.

Based on the above expressions, we note the following.

- Each step of network sampling can be formalised as an indirect sampling mechanism.
- In a given step of the RDS process, the participants from which the search starts constitute the source list U^A, and their links are the target population U^B.
- In the subsequent step, the people found in the target population U^B become the people of the initial population U^A, from which a new search starts.
- This switch in the role of the sample participants, from the target population U^B to the initial population U^A of the next step, occurs at each wave of the RDS search chain.

Remark 1. A path in graph theory is a finite or infinite sequence of edges that joins a sequence of vertices. We note that if R is greater than the maximum of the minimum paths between any pair of nodes in each cluster of the units of S_0, then the following implications follow.

- $Y_{S_0,R}$ represents the total of the y variable related to all the groups to which the elements of S_0 belong.
- if S_0 does not include all clusters characterising the population of interest, then $Y_{S_0,R} < Y$.

We obtain these important outcomes illustrated in remarks 2 and 3 below by interpreting the result of Remark 1 in an alternative way.

Remark 2. $Y_{S_0,R} = Y$ if

- R is greater than the maximum of the minimum paths between any pair of nodes in each cluster of the units of S_0.
- The S_0 sample must include people from all clusters with people from the target population.

Remark 3. We consider the case of two interconnected units in which they know each other. However, these units may belong to two separate clusters if the RDS search rules call for stopping the search if a connection is identified with a person living in a geographic location distinct from that of the original contact. In

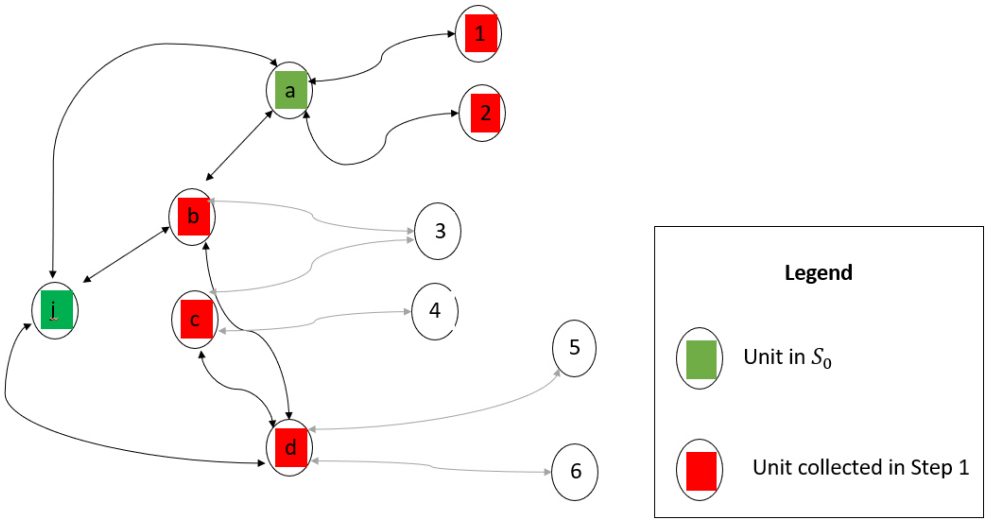


Fig. 4. Example of Steps 0 and 1.

this case, to ensure the equality $Y_{S_0,R} = Y$, in addition to the two conditions in Remark 2, there is the additional condition that the S_0 sample must include people from all geographic locations with people in the target population.

4. Sampling the RDS research chain

Unlike in the previous section, in the RDS process, not all links of a unit kept in the process are observed, but only a randomly selected sample of them is observed. The RDS process starts at step $r = 0$, with sample S_0 (which may be randomly selected or not), and in subsequent steps $r = 1, 2, \ldots, R$, we form samples $S_1 \subset S_2 \subset \ldots \subset S_r \subset \ldots \subset S_R$, each incorporating the sample from the previous step. To illustrate the formation of the generic sample S_{r+1}, we introduce additional symbology below. Let us consider the $j_r \in S_r$ unit and denote by

$$L_{j_r}^{S_r} = \sum\nolimits_{k \in U^B} \lambda_{j_r,k} \delta_k (S_r) = \sum\nolimits_{k \in S_r} \lambda_{j_r,k}$$

the total number of contacts of the unit that have been selected in sample S_r, where $\delta_k(A) = 1$ if unit k belongs to set A and $\delta_k(A) = 0$ otherwise. For the same unit j_r, let

$$L_{j_r}^{C_r} = L_{j_r}^{B} - L_{j_r}^{S_r}$$

be the number of contacts that have not been selected in sample S_r, and can be selected in sample S_{r+1}. For each unit j_r included in the S_r sample, we select, independently of the other units included in S_r, $m_{j_r} + 1$ units for the S_{r+1} sample where $m_{j_r} = \mathit{Min}(m, L_{j_r}^{C_r})$, being m is a fixed number (e.g., $m = 2$ or 3) that remains unchanged in the different steps of the RDS. Of these $m_{j_r} + 1$ units, one is the j_r unit itself, and the other m_{j_r} units are selected with a simple random sampling without replacement (SRSWOR) out from the $L_{j_r}^{C_r}$ units. The conditional probability that unit j_{r+1} is selected in sample S_{r+1}, given $j_r \in S_r$

$$\tau_{j_{r+1}|j_r \in S_r} = \begin{cases} 1 & \text{if } \left(\lambda_{j_r,j_{r+1}} = 1\right) \text{ and } \left[\delta_{j_{r+1}}(S_r) = 1\right] \\ \frac{m_{j_r}}{L_{j_r}^{C_r}} & \text{if } \left(\lambda_{j_r,j_{r+1}} = 1\right) \text{ and } \left[\delta_{j_{r+1}}(S_r) = 0\right] \\ 0 & \text{if } \lambda_{j_r,j_{r+1}} = 0 \end{cases}.$$

Remark 4 on the feasibility of the selection process. The illustrated selection process avoids the dead-end loops typical of graph sampling. However, to make it feasible, it is essential to know the $L_{j_r}^{C_r}$ quantity, obtained as difference of two quantities, $L_{j_r}^{B}$ and $L_{j_r}^{S_r}$. The value of $L_{j_r}^{B}$ can be requested directly from respondent j_r. Remembering that the relationships explored in RDS have the character of reciprocity, L_{j}^{B} corresponds to the total number of people who unit j_r knows and can point to in turn. Operationally, the $L_{jr}^{S_r}$ quantity can be obtained in alternative ways. Suppose nonidentifiable but unique information about contacts of units included in the S_r sample [13] is available in the data-collection APP used by the interviewer. In that case, a specific software application can be launched that identifies units not included in S_r and proceeds to select m_{jr} units to be included in sample S_{r+1} ran-

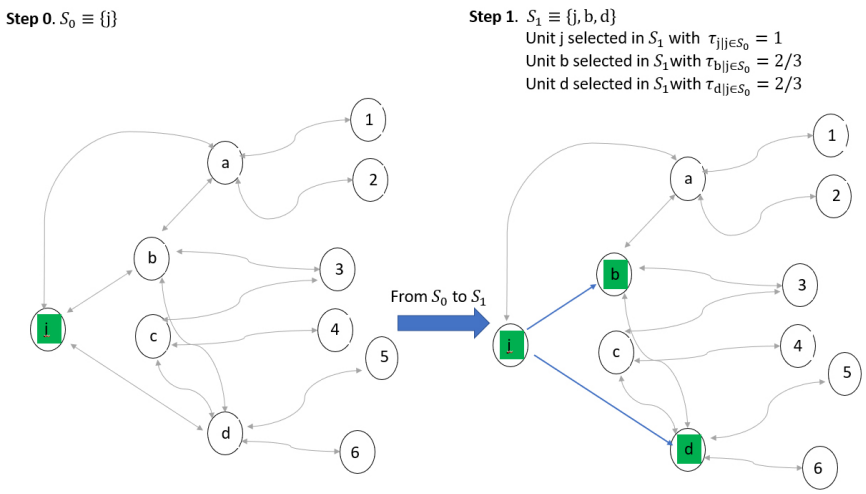


Fig. 5. Example of the formation of sample S_1 from sample S_0 with $m = 2$.

domly. Alternatively, the same software application can be run by the *study centre* (see Section 2 above) that supports the survey operations, and the results can be reported and provided in real-time to the interviewer who makes the S_{r+1} sample selection. Depending on the specific conditions of the survey, other feasible operational mechanisms can be defined.

Remark 5 on the research chain for a subpopulation. We consider the case of constructing the RDS sampling search chain only on the units of a subpopulation, for example, only on the people of a particular class-age. We denote by x_j a dichotomous variable that takes value 1 if unit j belongs to the subpopulation and takes value 0 otherwise. In this case, the link variables are defined as

$$\lambda_{(x)j,k} = \lambda_{j,k} x_j x_k.$$

The values L_k^B are modified accordingly.

5. Estimators

Next, we present three estimators. The first assumes a random selection of the S_0 sample, and the second adopts the traditional RDS methodology while considering a non-probabilistic S_0 sample. The third estimator is developed under a capture-recapture approach [12] while allowing for the smoothing of the coverage problems that may affect both of the first two estimators.

5.1. S_0 selected with a random sampling

Let us suppose a random sample S_0 of fixed size n_0 is selected from U^A without replacement and with inclusion probabilities π_{j0}, where

$$\pi_{j_0} > 0 \text{ for } j_0 = 1, \ldots, N^A \text{ and}$$
$$\sum\nolimits_{j_0 \in U^A} \pi_{j_0} = n_0. \tag{14}$$

To facilitate the understanding of the calculation method, we construct the estimator step by step. In each step, we obtain a sampling unbiased estimate of the total Y. However, as the steps of the RDS process progress, the estimate is based on a more significant number of observations.

At **step 0**, we have the classical HT estimator:

$$\hat{Y}_0 = \sum\nolimits_{j_0 \in S_0} y_{j_0} \frac{1}{\pi j_0} = \sum\nolimits_{j_0 \in S_0} y_{j_0} w_{j_0}, \tag{15}$$

where $w_{j_0} = 1/\pi_{j_0}$ is the sampling weight.

In **Step 1**, we have

$$\hat{Y}_1 = \sum\nolimits_{j_0 \in S_0} \sum\nolimits_{k \in S_1} y_k \frac{\lambda_{j_0,k}}{L_k^B} \frac{1}{\pi_{j_0}} \frac{1}{\tau_{k|j_0 \in S_0}} \tag{16}$$

As illustrated in Appendix 2, denoting $E(\cdot)$, the sampling expectation operator, we have $E(\hat{Y}_1) = Y$, meaning that $\hat{Y}_1$ is a sampling unbiased estimate of Y. Reversing the order of sums, we can express $\hat{Y}_1$ in the classical weighted form as

$$\hat{Y}_1 = \sum\nolimits_{k \in S_1} y_k w_{k_1} \tag{17}$$

where

$$w_{k_1} = \sum\nolimits_{j_0 \in S_0} \frac{\lambda_{j_0,k}}{L_k^B} \frac{1}{\pi_{j_0}} \frac{1}{\tau_{k|j_0 \in S_0}}.$$

Estimator $\hat{Y}_1$ can also be formulated referring to Eq. (6) as:

$$\widehat{Y}_1 = \sum\nolimits_{j_0 \in S_0} \frac{1}{\pi_{j_0}} \widehat{\bar{y}}_{j_0}^A \tag{18}$$

where $\hat{\bar{y}}^A_{j_0} = \sum_{k\in S_1} y_k \frac{\lambda_{j_0,k}}{L^B_k} \frac{1}{\tau_{k|j_0\in S_0}}$ is the unbiased estimate of $\bar{y}^A_{j_0}$

In Step 2, taking Eqs (16), (17), and (18) of step 1 as a reference, the unbiased estimator $\hat{Y}_2$ can be expressed according to these three alternative ways

$$\hat{Y}_2 = \sum_{j_0\in S_0}\sum_{j_1\in S_1}\sum_{k\in S_2} y_k \frac{\lambda_{j_0,j_1}}{L^B_{j_1}}\frac{\lambda_{j_1,k}}{L^B_k} \frac{1}{\pi_{j_0}}\frac{1}{\tau_{j_1|j_0\in S_0}}\frac{1}{\tau_{k|j_1\in S_1}},$$

$$\hat{Y}_2 = \sum_{k\in S_2} y_k w_{k_2},$$

$$\hat{Y}_2 = \sum_{j_0\in S_0}\sum_{j_1\in S_1} \frac{1}{\pi_{j_0}}\frac{\lambda_{j_0,j_1}}{L^B_{j_1}}\frac{1}{\tau_{j_1|j_0\in S_0}}\hat{\bar{y}}^A_{j_1},$$

where

$$w_{k_2} = \sum_{j_0\in S_0}\sum_{j_1\in S_1} \frac{\lambda_{j_0,j_1}}{L^B_{j_1}}\frac{\lambda_{j_1,k}}{L^B_k}\frac{1}{\pi_{j_0}} \frac{1}{\tau_{j_1|j_0\in S_0}}\frac{1}{\tau_{k|j_1\in S_1}},$$

and $\hat{\bar{y}}^A_{j_1} = \sum_{k\in S_2} y_k \frac{\lambda_{j_1,k}}{L^B_k}\frac{1}{\tau_{k|j_1\in S_1}}$ is the unbiased estimate of $\bar{y}^A_{j_1}$. We have $E\left(\hat{Y}_2\right) = Y$.

Recursively using the previous procedure, at the ultimate step R of the RDS process, we have:

$$\hat{Y}_R = \sum_{j_0\in S_0}\sum_{j_1\in S_1}\cdots\sum_{j_{R-1}\in S_{R-1}} \sum_{k\in S_R} y_k \times \frac{\lambda_{j_0,j_1}}{L^B_{j_1}} \times \ldots \times \frac{\lambda_{j_{R-1},k}}{L^B_k} \frac{1}{\pi_{j_0}}\frac{1}{\tau_{j_1|j_0\in S_0}} \times \ldots \times \frac{1}{\tau_{k|j_{R-1}\in S_{R-1}}},$$

$$\hat{Y}_R = \sum_{k\in S_R} y_k w_{k_R},$$

$$\hat{Y}_R = \sum_{j_0\in S_0}\sum_{j_1\in S_1}\cdots\sum_{j_{R-1}\in S_{R-1}} \frac{\lambda_{j_0,j_1}}{L^B_{j_1}} \times \ldots \times \frac{\lambda_{j_{R-1},k}}{L^B_k}\frac{1}{\pi_{j_0}}\frac{1}{\tau_{j_1|j_0\in S_0}} \times \ldots \times \frac{1}{\tau_{j_{R-1}|j_{R-2}\in S_{R-2}}}\hat{\bar{y}}^A_{j_{R-1}},$$

where

$$w_{k_R} = \sum_{j_0\in S_0}\sum_{j_1\in S_1}\cdots\sum_{j_{R-1}\in S_{R-1}} \frac{\lambda_{j_0,j_1}}{L^B_{j_1}} \times \ldots \times \frac{\lambda_{j_{R-1},k}}{L^B_k}\frac{1}{\pi_{j_0}}\frac{1}{\tau_{j_1|j_0\in S_0}} \times \ldots \times \frac{1}{\tau_{k|j_{R-1}\in S_{R-1}}}$$

and

$$\hat{\bar{y}}^A_{j_{R-1}} = \sum_{k\in S_{R-1}} y_k \frac{\lambda_{j_{R-1},k}}{L^B_k}\frac{1}{\tau_{k|j_{R-1}\in S_{R-1}}}.$$

In Appendix 2, we see $E\left(\hat{Y}_R\right) = Y$

Remark 6 on estimating the sampling variance. We can approximate the RDS design with multistage sampling with replacement, where each step may be viewed as a specific sampling stage, and the replacement refers to a single unit. In that way, we may derive an estimate of the sampling variance [19][(Formula 11.35)];

$$v\left(\hat{Y}_R\right) = \frac{1}{n_0\left(n_0-1\right)}\sum_{j_0\in S_0}\left(\frac{1}{z_{j_0}}\hat{Y}_{j_0} - \hat{Y}_R\right)$$

where

$$\hat{Y}_{j_0} = \sum_{j_1\in S_1}\cdots\sum_{j_{R-1}\in S_{R-1}}\sum_{k\in S_R} y_k \frac{\lambda_{j_0,j_1}}{L^B_{j_1}} \times \ldots \times \frac{\lambda_{j_{R-1},k}}{L^B_k}\frac{1}{\pi_{j_0}}\frac{1}{\tau_{j_1|j_0\in S_0}} \times \ldots \times \frac{1}{\tau_{k|j_{R-1}\in S_{R-1}}},$$

and $z_{j_0} = \pi_{j_0}/n_0$.

Remark 7 on the estimator for a subpopulation. As illustrated in Remark 5, the link variables are defined as $\lambda_{(x)j,k} = \lambda_{j,k}x_j x_k$, and the variables L^B_k are modified accordingly. Moreover, the target variable y_k is modified as $y_{(x)k} = y_k x_k$.

Remark 8 on type of estimator. Considering the above expressions, we can see how the $\hat{Y}_R$ estimator can be seen as a particular case of the generalised weight share method (GWSM) estimator.

Remark 9 on the starting sampling. The sampling design should maximise the number of observed individuals of the target population in the sample S_0 by adopting proper choices in the first and ultimate stages (or phases) of the sampling process. First-stage selection tends to oversample the areas where the researchers have some a priori information of a high concentration of the target population. Final-stage sampling tends to oversample the target people by modelling the inclusion probabilities on variables predictive of the phenomenon available in the sampling frames adopted to select the final-stage units.

5.2. S_0 selected with a nonrandom sampling

If the sample S_0 is selected nonrandomly, we can obtain a nonbiased estimate only of the total $Y_{S_0,R}$. We illustrate this case by referring only to step R and the weighted form of the estimator. An unbiased (see Appendix 2) estimator of the total $Y_{S_0,R}$ is

$$\hat{Y}_{S_0,R} = \sum\nolimits_{k \in S_R} y_k w_{k_R,S_0}$$

where

$$w_{k_R,S_0} = \sum\nolimits_{j_0 \in S_0} \sum\nolimits_{j_1 \in S_1} \cdots \sum\nolimits_{j_{R-1} \in S_{R-1}} \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_1}} \times \cdots \times \frac{\lambda_{j_{R-1},k}}{L_k^B} \frac{1}{\tau_{j_1|j_0 \in S_0}} \times \ldots \times \frac{1}{\tau_{k|j_{R-1} \in S_{R-1}}}.$$

We note that the formulation of $\hat{Y}_{S_0,R}$ is similar to that of the $\hat{Y}_R$ estimator, except that in $\hat{Y}_{S_0,R}$, the first weighting factor $(1/\pi_{j_0})$ of $\hat{Y}_R$ is equal to 1 , and $L_{j_1}^{S_0}$ replaces the $L_{j_1}^B$ factor.

Remark 10 on type of estimator. We can straightforwardly see how the $\widehat{Y}_{S_0,R}$ estimator can be seen as a particular case of the GWSM estimator.

Remark 11 on the starting sampling. To ensure that estimator $\hat{Y}_{S_0,R}$ is an unbiased estimate of the total Y, i.e., that condition $E(\hat{Y}_{S_0,R}) = Y$ is met, the initial sample S_0 should respect the three conditions illustrated in remarks 2 and 3.

5.3. Generalised capture-recapture estimator

Even if the S_0 sample is randomly selected, the first estimator $\hat{Y}_R$ may be biased. Indeed, undercoverage may occur if respondents do not trust the interviewers and tend to hide their status.

Likewise, if the S_0 sample is nonrandomly chosen, even the second estimator $\hat{Y}_{S_0,R}$ can undercover the total Y if the following conditions are not met: (i) R is greater than the maximum of the minimum paths between any pair of nodes in each cluster of the units of S_0, and (ii) S_0 does not include all clusters characterising the population of interest.

The generalised capture-recapture (*CReG*) estimator allows us to overcome the abovementioned undercoverage by leveraging a capture-recapture perspective. A comprehensive treatment of this estimator and how it mitigates undercoverage deserves much more space in this article than can be devoted. The interested reader can undoubtedly look to the extensive work reported in [9].

Let us consider two independent surveys based on the RDS methodology, and we suppose they are articulated in R steps. The first starts from an initial random sampling, while the second starts from a nonrandom sample. Furthermore, we suppose that the two sample selections are independent. The CReG estimator of Y may be expressed as

$$\hat{Y}_{CReG} = \frac{\hat{Y}_R \hat{Y}_{S_0,R}}{\hat{Y}_{intersect,R}}$$

where

$$\hat{Y}_{intersect,R} = \sum\nolimits_{k \in S_{R,\,intersect}} w_{k_R} w_{k_R,S_0} y_k$$

where $S_{R,intersect}$ is the intersection sample between the two samples that are generated from random and nonrandom sampling after R steps.

A useful approach is applying the estimator CReG on the two nonrandom starting samples but with a different mechanism of undercoverage of the two respondent groups.

6. Conclusions

The disaggregation of data for SDG indicators on hidden or hard-to-count population groups presents several critical issues that are difficult to overcome to produce reliable official statistics at the national and international levels. In this context, it is impossible to estimate the characteristics of these groups through models as in other situations.

Considering, for example, indigenous populations, data availability varies widely from country to country. Few countries provide up-to-date and high-quality data. Many other countries have only data that are scattered over time. Or the data they provide is not supported by a sufficiently robust methodology, both for precisely identifying the subpopulation group of interest and for the sampling technique adopted.

Given the current context of producing official statistics on the subject, it is unrealistic that this lack of data on indigenous peoples is going to improve soon.

Therefore, it becomes necessary to define and apply an implementation program that can improve this situation relatively quickly.

This implementation program should be based on the following pillars:

1. The first pillar is to develop a valuable data collection strategy for sample surveys that is capable of measuring the number of people belonging to the indigenous population in a given area. This strat-

egy needs to cover various aspects, including the formulation of the questionnaire for identifying persons belonging to the indigenous population and the characteristics of special sampling techniques that can maximise the efficiency of surveys aimed at obtaining data on these hidden or hard-to-measure population groups. Specifically, the data collection strategy consists of technical manuals, open software modules, ad hoc training materials, etc. In short, anything that enables or helps conduct surveys or specific survey modules to estimate the size of indigenous populations. Regarding the questionnaire, the data collection strategy should develop a set of standard questions on the key characteristics of the specific population of interest and not adopt generic questions on whether specific individuals belong to an indigenous population group.

2. The second pillar is to adapt the data collection strategy to ongoing survey programs. For example, the indigenous module may be applied to a large-scale national survey conducted by a national statistics office. Another example can be to promote the application of the indigenous sampling module to international surveys, such as the World Bank's LSMS survey. Regarding the sampling aspects, it is helpful to consider an overall sampling strategy to maximise the number of observed individuals of the target population in the sample by combining the first and final sampling stages. First-stage methods should tend to oversample the areas where the researchers have some a priori information on the geographical concentration of the target population. Final-stage sampling should oversample the target population by adopting strategies, such as respondent-driven sampling (RDS), that leverage the existing hidden relationships among the individuals of the target populations.
3. The third pillar is adopting estimation methods that allow unbiased estimates of phenomena of interest in target populations. In this paper, we have considered the RDS, widely used for observing hidden or rare populations but which lacks an estimation methodology that is sufficiently robust to accommodate the varying conditions under which it is applied. We have proposed three unbiased estimators. The first assumes a random selection of the starting sample, and the second considers a nonprobabilistic starting sample. The third estimator is developed under a capture-recapture approach and allows for smoothing of the coverage problems that may affect the estimators. We have studied their sampling expectation and have indicated the conditions that may be fulfilled to guarantee their unbiasedness concerning the target totals.

References

[1] Gile K, Beaudry IS. Handcock MS, Miles Q. Methods for Inference from Respondent-Driven Sampling Data. Annu Rev Stat Appl. 2018; 5(4): 1-429.

[2] Heckathorn DD, Cameron J. Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling. Annual Review of Sociology. August 2017.

[3] Heckathorn DD. Respondent driven sampling: a new approach to the study of hidden samples. Soc Probl. 1997; 44(2): 174-99.

[4] White RG, Hakim AJ, Salganik MJ, Spiller Spiller MV, Johnston LG, Kerr L, Kendall C, Drake A, Wilson D, Orroth K, Egger M, Wolfgang H. Strengthening the Reporting of Observational Studies in Epidemiology for respondent-driven sampling studies: "STROBE-RDS" statement. Journal Clinical Epidemiology. 2015 Dec; 68(12): 1463-71. doi: 10.1016/j.jclinepi.2015.04.002. Epub 2015 May 1.

[5] Goodman LA. Snowball sampling. Ann Math Stat. 1961; 32: 148-70.

[6] Handcock MS, Gile KJ. Comment: on the concept of snowball sampling. Sociol Methodol. 2011; 41(1): 367-71.

[7] Heckathorn DD, Jeffri. Finding the Beat. Using respondent-driven sampling to study jazz musicians. Poetics. 2001; 28: 307-329. Elsevier.

[8] Lavallée P. Indirect Sampling. Springer. New York, 2007.

[9] Lavallée P, Rivest LP. Capture – Recapture Sampling and Indirect Sampling. Journal of Official Statistics. 2012; 28(1): 1-27.

[10] Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. J Official Statistics. 2008; 24: 79.

[11] Salganik M.J., Douglas D. Heckathorn. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology. 2004; 34: 193-239.

[12] Johnston LG. Conducting respondent driven sampling studies in diverse settings: a manual for planning RDS studies. Cent Dis Control Prev, Atlanta, GA, 2007.

[13] Johnston LG. Introduction to HIV/AIDS and sexually transmitted infection surveillance. Module 4. Introduction to respondent-driven sampling. World Health Organ., Geneva. http//www.lisagjohnston. com/respondent-driven-sampling/respondent-driven-sampling, 2013.

[14] Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc Probl. 2002; 49: 11-34.

[15] Heckathorn DD. Assumptions of RDS: analytic versus functional assumptions. Presented at CDC Consult. Anal. Data Collect. Respond.-Driven Sampl., Atlanta, GA, 2008.

[16] Gile K, Handcock MS. Network model-assisted inference from respondent-driven sampling data. J R Stat Soc A. 2015; 178(3): 619-39.

[17] Verdery AM, Merli MG, Moody J, Smith J, Fisher JC. Respondent-driven sampling estimators under real and theoretical recruitment x conditions of female sex workers in China. Epidemiology. 2015a; 26: 661.

[18] Hansen MH, Hurwitz WN. On the theory of sampling from finite populations. *Ann. Math. Stat.*. 1943; 14(4): 333-62.
[19] Cochan WG.. Sampling Techniques. J Wiley New York, 1943.

Appendix 1: Equation (7)

We have

$$Y_{S_0,2} = \sum_{k\in U^B} y_k \sum_{j_0\in S_0} \sum_{j_1\in U^A} \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_0}} \frac{\lambda_{j_1,k}}{L_k^B} = \sum_{k\in U^B} y_k \sum_{j_0\in S_0} \frac{\lambda_{j_0,j_1}}{L_{j_1}^{S_0}} \sum_{j_1\in U^A} \frac{\lambda_{j_1,k}}{L_k^B}$$

being

$$\sum_{j_1\in U^A} \frac{\lambda_{j_1,k}}{L_k^B} = \begin{cases} 1 \text{ if } \lambda_{j_1,k}=1 \\ 0, \text{ otherwise} \end{cases},$$

$$\sum_{j_0\in S_0} \frac{\lambda_{j_0,j_1}}{L_{j_1}^B} = \begin{cases} 1 \text{ if } \lambda_{j_0,j_1}=1 \\ 0, \text{ otherwise} \end{cases}.$$

Appendix 2: Unbiasedness of $\hat{Y}_1$

We have

$$E\left(\hat{Y}_1\right) = \sum_{j_0\in U^A} \frac{1}{\pi_{j_0}} E\left[\delta_{j_0}\left(S_0\right)\right] \sum_{k\in U^B} y_k \frac{\lambda_{j_0,k}}{L_k^B} \left(\frac{1}{\tau_{k|j_0\in S_0}}\right) E\left[\delta_k\left(S_1\right) \mid j_0\right] = \sum_{j\in U^A} \frac{\pi_{j_0}}{\pi_{j_0}} \sum_{k\in U^B} y_k \frac{\lambda_{j_0,k}}{L_k^B} \left(\frac{\tau_{k|j_0\in S_0}}{\tau_{k|j_0\in S_0}}\right) \sum_{j_0\in U^A} \sum_{k\in U^B} y_k \frac{\lambda_{j_0,k}}{L_k^B} = \sum_{k\in U^B} y_k,$$

since

$$\sum_{j_0\in U^A} \frac{\lambda_{j_0,k}}{L_k^B} = 1.$$

Unbiasedness of $\hat{Y}_{S_0,R}$

$$E\left(\hat{Y}_R\right) = \sum_{j_0\in U^A} \sum_{j_1\in U^A} \times \cdots \times \sum_{j_{R-1}\in U^A} \sum_{k\in U^B} \left(\frac{\lambda_{j_0,j_1}}{L_{j_1}^B} \times \ldots \times \frac{\lambda_{j_{R-1},k}}{L_k^B}\right) \times \left(\frac{E\left[\delta_{j_0}\left(S_0\right)\right]}{\pi_{j_0}} \frac{E\left[\delta_{j_1}\left(S_1\right) \mid j_0\in S_0\right]}{\tau_{j_1|j_0\in S_0}} \times \ldots \times \frac{E\left[\delta_k\left(S_R\right) \mid j_{R-1}\in S_{R-1}\right]}{\tau_{k|j_{R-1}\in S_{R-1}}}\right) y_k$$

$$= \sum_{j_0\in U^A} \sum_{j_1\in U^A} \cdots \sum_{j_{R-1}\in U^A} \sum_{k\in U^B} \left(\frac{\lambda_{j_0,j_1}}{L_{j_1}^B} \times \ldots \times \frac{\lambda_{j_{R-2},j_{R-1}}}{L_k^B} \frac{\lambda_{j_{R-1},k}}{L_k^B}\right) y_k = \sum_{j_0\in U^A} \frac{\lambda_{j_0,j_1}}{L_{j_1}^B} \sum_{j_1\in U^A} \frac{\lambda_{j_1,j_2}}{L_{j_2}^B} \times \ldots \times \left(\sum_{k\in U^B} \sum_{j_{R-1}\in U^A} \frac{\lambda_{j_{R-1},k}}{L_k^B} y_k\right)$$

$$= \sum_{j_0\in U^A} \frac{\lambda_{j_0,j_1}}{L_{j_1}^B} \sum_{j_1\in U^A} \frac{\lambda_{j_1,j_2}}{L_{j_2}^B} \times \ldots \times \sum_{j_{R-2}\in U^A} \frac{\lambda_{j_{R-3,R-2}}}{j_{R-2}} Y = Y.$$

Unbiasedness of $\hat{Y}_{S_0,R}$

$$E\left(\hat{Y}_{S_0,R}\right) = \sum_{j_0\in S_0} \sum_{j_1\in U^A} \times \cdots \times \sum_{j_{R-1}\in U^A} \sum_{k\in U^B} \left(\frac{\lambda_{j_0,j_1}}{L_{j_0}} \times \ldots \times \frac{\lambda_{j_{R-1},k}}{L_k^B}\right) \times \left(\frac{E\left[\delta_{j_1}\left(S_1\right) \mid j_0\in S_0\right]}{\tau_{j_1|j_0\in S_0}} \times \ldots \times \frac{E\left[\delta_k\left(S_R\right) \mid j_{R-1}\in S_{R-1}\right]}{\tau_{k|j_{R-1}\in S_{R-1}}}\right) y_k$$

$$= \sum_{j_0\in S_0} \sum_{j_1\in U^A} \times \cdots \times \sum_{j_{R-1}\in U^A} \sum_{k\in U^B} \left(\frac{\lambda_{j_0,j_1}}{L_{j_1}} \times \cdots \times \frac{\lambda_{j_{R-1},k}}{L_k^B}\right) y_k = Y_{S_0,R}$$