METHODOLOGICAL EVALUATION OF RESPONDENT DRIVEN SAMPLE IN ISTAT LGB EXPERIMENTAL SURVEY

Eugenia De Rosa, Francesca Inglese, David Trambusti

Abstract. The Respondent Driven Sampling (RDS) technique was first applied by Istat (the Italian National Institute of Statistics) in the 2022 "Survey on Labour Discrimination against LGB (lesbians, gay and bisexuals) people not in Civil Union". The RDS is a valuable approach for studying populations that are difficult to reach, such as LGB people, thanks to its robust theoretical basis. However, the validity of the samples it produces depends on strict assumptions about network structure, the recruitment process, and the sample/population. Furthermore, its implementation is particularly sensitive to operational constraints, including privacy concerns. This work provides a methodological evaluation of the RDS sample obtained in the aforementioned survey. The aim is to identify critical issues related to design choices, implementation limitations and other factors, such as network characteristics and recruitment dynamics. The analyses focused on sample convergence and dependence on seeds, along with potential sources of recruitment bias, including network bottlenecks and homophilic behaviour among participants. The results highlighted several factors that undermined the inferential validity and representativeness of the sample. However, the Italian experience demonstrates the RDS's ability to engage with populations that are usually under-represented in probability-based surveys. It also contributes to the wider debate within official statistics on the use and enhancement of non-probability sampling methods, and on combining these with probability-based techniques.

1. Introduction

In 2022, Istat implemented for the first time the Respondent-Driven Sampling (RDS) technique in its web-based version (WebRDS) (Wejnert and Heckathorn, 2008) within the framework of the *Survey on Labour Discrimination against LGB people not in Civil Union* (De Rosa and Inglese, 2018).

RDS represents an advanced variant of snowball sampling and, under specific theoretical conditions, offers significant inferential advantages, as it can asymptotically approximate a simple random sample. It is a network-based sampling, that integrates peer-to-peer recruitment with a Markov process model: the sampling begins with the identification of initial participants (*seeds)* from within the target population. Seeds are tasked with recruiting individuals from their own social networks and belonging to the same population, who subsequently recruit further participants, generating chains of recruitment (Heckathorn, 1997; 2002).

RDS is particularly valuable for studying hard-to-reach populations, especially when conventional sampling methods risk excluding marginalised groups or are ineffective. Nevertheless, the validity of RDS relies heavily on strict assumptions regarding network structure, recruitment processes, and the sample/population. Specifically, the network must be connected and free of bottlenecks, characterised by reciprocal ties. Recruitment must be random among peers, with a fixed and limited number of recruits. For statistical inferences to be valid, the sample must reach equilibrium, and the size of the network must be accurately reported and remain stable. In practice, these conditions are often only partially satisfied.

The recruitment process is susceptible to bias from significant variations in personal network size, compounded by structural bottlenecks and behavioural homophily. Furthermore, researchers' limited control over sample composition can result in a final sample that does not sufficiently capture the heterogeneity of the target population. Consequently, a rigorous evaluation of the realized RDS samples is essential to assess data quality and ensure reliable inference. Measurement errors in the self-reporting of personal network size (degree) constitute a critical issue, as they can compromise the validity of the weighting process.

This article analyses the RDS data from the Istat survey of LGB people. Section 2 outlines the context of the survey and explains how RDS was implemented. Section 3 delineates the methodological framework for evaluating sample quality and presents a selection of key findings. The final section offers concluding remarks and proposes areas for potential intervention to improve the design and implementation of RDS in future studies.

2. The application of WebRDS in the Survey on Labour Discrimination against LGB people (not in Civil Union)

Between 2018 and 2023, Istat collaborated with UNAR (National Anti-Discrimination Office) to address the lack of statistical information on LGBT+ populations. The joint project, "Labour Discrimination against LGBT+ People and Diversity Policies in Enterprises", included three targeted surveys. The first two surveys were designed to reach two complementary target groups (mainly LGB people in Civil Union and LGB people not in Civil Union) combining standard and non-standard sampling techniques (De Rosa, 2024; De Rosa and Inglese, 2024). The third survey addressed trans and non-binary people.

The WebRDS approach was cautiously applied to the survey addressed to LGB people not in Civil Union (De Rosa and Inglese, 2018). Initial recruitment via RDS was insufficient to reach the desired sample size, so after several weeks and with a limited number of established recruitment chains, a convenience sampling strategy was introduced to complete the data collection.

The choice of the RDS technique was preceded by a formative study conducted to assess whether the target population was sufficiently networked. This involved a review of academic and grey literature, analysis of existing data, and qualitative research through interviews with stakeholders, key informants, and experts in LGBT+ issues and discrimination. Consultations with LGBT+ associations further explored levels of community belonging, inter-association networking, and informal networks relevant for recruitment feasibility.

The WebRDS design, detailed also in the Data Protection Impact Assessment (Art. 35, GDPR), included: partnership with LGBT+ associations, seeds selection by the association, anonymous web-based self-administration questionnaire on discrimination, and peer recruitment (De Rosa *et al*., 2020). Around 50 LGBT+ associations supported the survey, signing data protection agreements with Istat and in charge of seed identification. Each association selected up to 10 individuals belonging to the population of interest (LGB) seeds based on socio-demographic grid (that included sex, age, region and sexual orientation). Additional selection criteria requested these individuals had strong social connectivity and high motivation to support the study.

For data collection, each association was assigned a unique survey link, enabling monitoring of referral chains without disclosing individual identities. Respondents entered the survey via an “accession module,” which provided project information and determined eligibility (e.g., aged 18+, living in Italy, not in Civil Union, and personally knowing the recruiter). Eligible individuals submitted an email address to receive the survey link. Eligibility was confirmed in the initial section of the main questionnaire through a question on current sexual orientation; those who identified as “Other” or selected “Prefer not to say” were screened out. To preserve confidentiality, data from the accession module and the survey were stored on separate servers, and email addresses were encrypted. Network size questions essential for RDS estimation were included in the main questionnaire, asking how many homosexual or bisexual people the respondent knew, and how many they had contacted in the past month. Upon completing the questionnaire, respondents were invited to recruit up to four LGB people from their personal networks by sharing a system-generated referral link, available directly on the screen and sent via email. Recruitment could occur via email, SMS, or messaging apps. No incentives were offered, neither monetary nor of any other nature. Recruitment chains were tracked using anonymous unique codes to monitor the structure of referral waves.

During the data collection process, which started last week of January, Istat researchers closely monitored recruitment indicators such as chain length, demographic composition, and association participation. Despite these efforts, by late April it became evident that the RDS was not functioning effectively. Consequently, from April 26, 2022, the survey was opened to a convenience sample.

The final sample comprised 1,159 LGB individuals: 730 recruited via RDS and 429 via convenience sampling. The main findings of the study were published in May 2023 (Istat-UNAR, 2023).

3. Quality assessment of RDS sample: analysis and main results

Validating an RDS sample requires verifying its quality against the method's underlying assumptions. Given that recruitment occurs through social networks, a comprehensive evaluation is essential to identify and differentiate all potential sources of error that could increase sampling variance and introduce bias.

Convergence analysis is a key step in this process. This verifies whether the sample has reached saturation, which occurs when the composition of the observed characteristics stabilises and becomes independent of the initial seeds. Failure to achieve converge undermines the validity of the inference and the reliability of the results. However, convergence alone does not guarantee that the sample is representative, as structural or behavioural biases may still occur during recruitment (Wejnert, 2009; McCreesh *et al*., 2012; Yamanis *et al*., 2013). Furthermore, variability in network size can introduce recruitment bias by generating unequal inclusion probabilities. Individuals with larger networks are more likely to be recruited, which can lead to the over-representation of well-connected subgroups. Conversely, individuals with smaller networks - often from marginalised or socially isolated groups - are less likely to be included and may be under-represented, thus limiting the diversity captured in the sample.

Convergence must, therefore, be supported by additional diagnostic indicators, such as recruitment bottleneck measures and homophily indices. These tools help determine whether the sample is deep and diverse enough to reflect the target population. Thus, the validity of the RDS assumptions is crucial for data interpretation: if they are met, the results have inferential value and can be generalised; otherwise, the results remain descriptive and not representative.

3.1. RDS sample: preliminary results and seeds contribution

The RDS sample, derived from the Istat survey, comprises 730 individuals who display certain characteristics. More than 40% of respondents are seeds, but only 34% of these generate chains. These chains tend to be short, averaging just 1.76 waves, with only one reaching nine waves. Around 370 recruits (88% of the total recruits) are concentrated in the first three waves.

This preliminary overview highlights the key issues in the RDS procedure. The final composition is significantly affected by the non-random selection of seeds and their uneven contribution to the recruitment process. Furthermore, the limited expansion capacity of the recruitment chains implies partial coverage of the target population, which may indicate a structural failure in the RDS recruitment process, as there are fewer than the four to five waves typically considered to be adequate.

The differential contribution of each seed to the overall recruitment effort was analysed, taking into account the percentage distribution of seeds per recruitment wave generated. Recruitment capacity varies considerably among seeds: 20.3% generated only one recruitment wave, while just 5% produced chains extending between the third and fifth waves (Table 1).

Table 1 – *Distribution of seeds in the recruitment process by waves generated*.

Wave	0	1	2	3	4	5	9
% of seeds	65.9	20.3	8.4	1.9	1.6	1.6	0.3

Furthermore, to evaluate the effectiveness of the initial seeds, an overall seed productivity index was calculated using the formula:

$$\text{Productivity index } = \frac{n_{RDS}}{s_{RDS}} - 1 \tag{1}$$

where, n_{RDS} is the final sample of respondents (730) and s_{RDS} is the number of initial selected seeds (311). The seed productivity index is 1.35, indicating low recruitment efficiency. When non-productive seeds - those that did not recruit any participants - are excluded, the index rises to nearly 4. This suggests that recruitment may have been driven by a small number of highly productive seeds with specific, distinctive traits.

Subsequent analysis evaluates the contribution of the initial seeds to recruitment, employing both descriptive statistical test[1] and a multivariate model.

Firstly, the distribution of 106 productive and 205 non-productive seeds were analysed by network size. Due to substantial heterogeneity in the self-reported values, the variable was recoded into quartile-based categories. A chi-square test confirms a statistically significant difference in network size between productive and non-productive seeds (test statistic: 18.2; p-value: 0.0004). The proportion of productive seeds with a network size greater than 20 is nearly twice as high as that of non-productive seeds in the same category. Conversely, the proportion of non-productive seeds is higher for small networks and decreases for larger networks (Table 2).

[1] Asterisks indicate the level of statistical significance (p-value): * for p-value < 0.05, ** for p-value < 0.01, *** for p-value < 0.001. The method used in all contingency tables to obtain this result is the bootstrap simulation.

Table 2 – *Distribution of productive and non-productive seeds by network size class (% by column).*

Network size	Productive seeds	Non-productive seeds
≤ 5	10.4 **	24.4 *
< 5 - ≤ 10	18.9	25.4
< 10 - ≤ 20	29.2	28.8
> 20	41.5 **	21.5 *

Secondly, a logistic regression model was applied to further investigate the determinants of recruitment success, The dependent variable was defined as binary, taking the value 1 if the seeds were productive, and 0 otherwise. The model included the following auxiliary variables: employment status (employed, unemployed, inactive); geographical area (five municipality categories); the logarithm of network size; and participation (at the time of the survey) in LGBT+ associations or groups (yes or no).

The model achieved a classification accuracy of 69.5% and an area under the ROC curve (AUC) of 69%, indicating an acceptable level of discriminative power. Only three variables were found to be statistically significant predictors of seed productivity ($p < 0.05$): employment status, log-transformed network size, and participation in LGBT+ associations or groups (Table 3).

Table 3 – *Odds ratio of the logit model.*

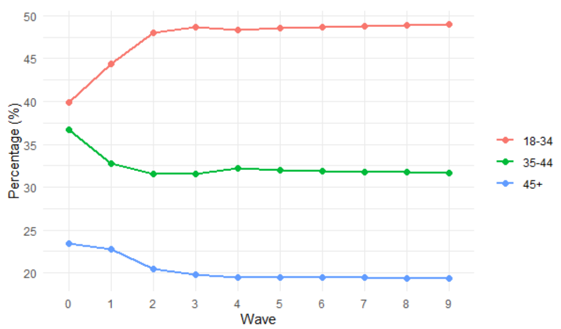
Variable	Odds ratio	p-value	significance
(Intercept)	0.11	0	***
Employment status:			
employed (reference)	–	–	
unemployed	0.2	0.032	*
inactive	1.38	0.445	
Municipality:			
≤ 5000 (reference)	–	–	
5,001-20,000	1.7	0.319	
20,001-50,000	0.5	0.259	
50,001-150,000	0.78	0.632	
> 150,000	1.37	0.513	
Log(network size)	1.33	0.025	*
Being part of associations or other groups	2.31	0.014	*

As shown in table 3, unemployment significantly reduces the likelihood of successful recruitment, making unemployed individuals around five times less likely to be recruited. Conversely, active engagement in LGBT+ associations or groups more than doubles the likelihood of successful recruitment. Having a larger social network also has a positive, albeit more modest, influence. No statistically significant association was observed between municipality category and seeds productivity.

3.2. RDS validation: sample convergence and recruitment analyses

To examine sampling dynamics, a convergence analysis of the cumulative frequencies of participants by age class across successive waves was carried out. An equilibrium plot (Figure 1) was used for this purpose to assess the stability of the sample over time. Considering a variation of ±2% between waves, it appears that all categories reach equilibrium by the second or third wave. However, this apparent stability does not guarantee true convergence of the RDS sample, since most respondents are concentrated in the first three waves (88% of recruits), and the proportions flatten quickly. Therefore, this result may be interpreted as reflecting the initial composition of the sample rather than as evidence of genuine convergence in the RDS process with respect to the age variable.

Figure 1 – *Equilibrium plot by age class.*



This interpretation is reinforced by comparing the seeds and recruits. The chi-square test reveals significant differences in age composition between the two groups (test statistic: 19.14; p-value: 0.0003). Younger respondents are over-represented among recruits (55.8%) compared to the initial group of seeds (39.9%). Conversely, all older age groups are under-represented among recruits. This decline is particularly pronounced for the 35–44 and 45–54 age groups (Table 4).

Table 4 – *Distribution of seeds and recruits by age group (% by column).*

Age class	Seeds	Recruits
18-34	39.9 **	55.8 **
35-44	36.7 *	27.9
45-54	18 *	11.5 *
55+	5.5	4.8

The recruitment process was analysed to identify potential sources of bias related to the high variability in network size, bottlenecks in network structure, and participants' homophilous behaviour.

Network size was assessed to determine whether individuals with larger or smaller personal networks were disproportionately represented within the sample. The analysis revealed systematic imbalances in the composition of the sample with respect to this variable, which appear to be closely associated with two structural dimensions: respondents' age and municipality of residence.

A chi-square test confirms that network size varied significantly with age (test statistic: 18.23; p-value: 0.033). Younger people predominate in small networks, while older people tend to be more visible in larger ones (Table 5).

Table 5 – *Distribution of RDS sample by network size and age group (% by row).*

Network size	Age class			
	18-34	35-44	45-54	55+
≤ 5	52.9	31.1	13.6	2.4 *
< 5 - ≤ 10	54	25.7	13.9	6.4
< 10 - ≤ 20	48.6	34.1	14	3.4
> 20	38.6 *	36.7	15.8	8.9 *

RDS preferentially captured certain subgroups. In essence, an individual's likelihood of being included in the sample was not random but was significantly influenced by their network size, which in turn correlates strongly with their age and where they live.

Larger municipalities have more extensive social networks, whereas very small networks prevail in municipalities with fewer than 5,000 inhabitants (test statistic: 40.26; p-value: 0.000065). This result highlights the difficulty of recruiting in small municipalities, where recruitment chains are quickly exhausted, and the risk of oversampling "super-spreaders" in large cities. Respondents living in larger municipalities tend to report wider and more heterogeneous personal networks, while small municipalities are characterized by a predominance of very limited networks (Table 6).

Table 6 – *Distribution of RDS sample by network size and municipality (% by row).*

Network size	Municipality (for number of inhabitants)				
	≤ 5000	5,001-20,000	20,001-50,000	50,001-150,000	> 150,000
≤ 5	13.1 *	21.4	17 **	22.3	26.2 ***
< 5 - ≤ 10	9.6	13.4	9.6	27.8 *	39.6
< 10 - ≤ 20	7.3	16.2	10.6	20.1	45.8
> 20	7	15.8	6.3 *	18.4	52.5 **

Structural barriers in the recruitment network were examined using a Bottleneck Indicator (BI) calculated for three age class, based on their representation in the sample generated from productive seeds only (Table 7). For each seed, the BI takes into account three components: (i) the size of the recruitment chain, (ii) the proportion of the target group within the chain, and (iii) the absolute deviation of this proportion from the group's share in the overall sample| pij - pj |. The bottleneck indicator is then computed as the weighted mean of these absolute deviations, following the formula below:

$$BI = \sum_{i=1}^{S} w_i \sum_{j=1}^{H} \left| p_{ij} - p_j \right|, \quad (1)$$

where: S = number of chains (seed), i=1,…,S; H = number of groups, j=1,…,H; p_{ij}= proportion of group j in chain i; p_j= proportion of group j in the total sample; w_i= weight of the chain i.

Table 7 - *Recruitment bottlenecks - proportion of age groups.*

Age group	Proportion in RDS sample	Bottleneck indicator (weighted with number of recruits)
18-34	52.9 %	30.9 %
35-44	29.9 %	24.1 %
45+	17.2 %	20.1 %

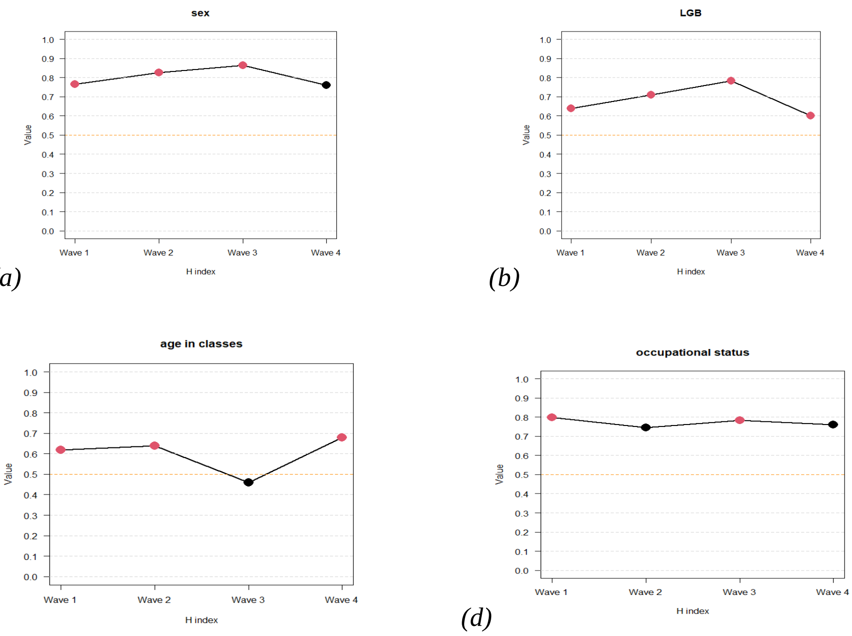
The results in table 7 indicate the presence of structural barriers in the network. The 18–34 age group, despite being the largest in the sample (52.9%), shows a relatively high BI (30.9%), suggesting that younger respondents are concentrated in specific chains rather than evenly distributed. The 35–44 age group has a moderate BI (24.1%), indicating a more uniform distribution along the chains. The 45+ group, although representing only 17.2% of the sample, exhibits a BI of 20.1%, revealing that older participants tend to cluster in certain chains, pointing to structural constraints in reaching this group. Overall, these findings highlight that structural barriers in the network affect both the largest and smallest age groups, while the intermediate group experiences a more balanced distribution. Such barriers may hinder the even diffusion of certain groups and restrict population mixing. This affects the representativeness of the sample and introduces substantial selection bias.

The homophilic tendencies of the RDS participants were evaluated by analysing recruiter-recruit similarity to understand the extent to which participants' recruited individuals similar to themselves. The homophily index (H) was calculated for different distributions - sex, sexual orientation, age class, employment status, educational level, income class, family size and municipality - on the transition matrix M of dimension k×k:

$$H = \frac{\sum_{i=1}^{k} m_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}} \tag{2}$$

In the formula, *i* and *j* denote row and column respectively; the numerator is the sum of the frequencies on the main diagonal of the transition matrix and the denominator is the sum of the total frequency of the matrix. The analysis revealed a significant homophily effect in the recruitment patterns of the following variables: sex, sexual orientation, age, employment status and level of education. This suggests the formation of homogeneous social clusters, which limits the diversity of the sample and likely results in the over-representation of certain subgroups. Significant results are shown in Figure 2, where red dots indicate significance ($p < 5\%$) in the chi-square test or Fisher's exact test if not performable due to a lack of data, while black dots indicate non-significance. In wave 3, approximately 90% of the sample recruited individuals of the same sex, over 80% recruited individuals of the same sexual orientation, and over 70% recruited individuals of the same occupational status. A significant result emerges for the age variable in waves 1 – 2.

Figures 2 - *H index for sex and sexual orientation (LGB group) (Figure (a) and (b)); H index for age in 3 classes and occupational status (employed vs not) (Figure (c) and (d)).*



(a) *(b)*

(c) *(d)*

4. Final remarks

The RDS sample from the Istat-UNAR survey revealed several critical limitations. The non-random selection of seeds resulted in significant imbalances, and the sample did not reach stationarity, with the initial selection of seeds continuing to influence its final composition. Recruitment chains rarely reached sufficient depth, which limited coverage and left substantial portions of the target population under-represented. Structural bottlenecks and homophilic behaviours restricted recruitment, resulting in uneven representation. Recruitment process integrity was compromised by the fact that participants often failed to recruit the maximum number of peers. The limited effectiveness of seeds selection was likely due to several factors: many nominated seeds did not participate, suggesting self-selection and suboptimal choices by associations. Privacy constraints prevented the research team from training or monitoring seeds, and procedural limitations - such as recruitment via email only and the absence of incentives - further weakened the development of recruitment chains. In future studies, seed selection should be considered a fundamental design element of RDS and supported by robust communication strategies, such as engaging LGBTQ+ influencers. The formative study is equally important and should not be considered marginal; prior data and formative findings are essential for understanding the target population and designing network models that promote diverse recruitment. Finally, the high variability found in the self-reported network size variable highlights the potential for measurement error, and underlines the need to test and refine these questions to ensure consistent and accurate interpretation by respondents.

References

DE ROSA E., INGLESE F. 2018. Diseguaglianze e discriminazioni nei confronti delle persone LGBT: quale contributo della statistica ufficiale?, *Rivista italiana di economia demografia e statistica RIEDS*, Vol. LXXII, No. 4, pp.77-88.

DE ROSA E., DE VITIIS C., INGLESE F., VITALINI A. 2020. Il web-respondent driven sampling per lo studio della popolazione LGBT+, *Rivista italiana di economia demografia e statistica RIEDS*, Vol. LXXIV-1 January-March 2020, pp. 66-73.

DE ROSA, E. (Eds) 2024. *Discriminazioni lavorative nei confronti delle persone LGBT+ e le diversity policy" Istat: Roma*. https://www.istat.it/wp-content/uploads/2024/07/Discriminazioni-lavorative-persone-LGBT-diversity-policy-Ebook.pdf

DE ROSA E., INGLESE F. 2024. Sampling and civil society engagement in surveying hard to reach populations, *Rivista italiana di economia demografia e statistica RIEDS*, Vol. LXXVIII-3 July-September 2024, pp. 135-147.

ISTAT-UNAR. 2023. *L'Indagine Istat-UNAR sulle discriminazioni lavorative nei confronti delle persone LGBT+ (non in unione civile o già in unione)*. Anno 2022.

HECKATHORN D. D. 1997. Respondent-driven sampling: A new approach to the study of hidden populations, *Social Problems*, Vol. 44, No. 2, pp. 174-199.

HECKATHORN D. D. 2002. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations, *Social Problems*, Vol. 49, No. 1, pp. 11-34.

MCCREESH N., FROST S. D. W., SEELEY J., KATONGOLE J., TARSH M. N., NDUNGUSE R., WHITE R. G. 2012. Evaluation of respondent-driven sampling, *Epidemiology*, Vol. 23, No. 1, pp. 138-147.

WEJNERT C., HECKATHORN D. D. 2008. Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research, *Sociological Methods & Research*, Vol. 37, No. 1, pp. 105-134.

WEJNERT C. 2009. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. *Sociological Methodology*, Vol. 39: pp. 73-116.

YAMANIS T. J., MERLI M. G., NEELY W. W., TIA, F., MOODY J., TU X., GAO E. 2013. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociological Methods & Research*, Vol. 42, No. 3, pp. 392-425.

Eugenia DE ROSA, Istat, eugenia.derosa@istat.it
Francesca INGLESE, Istat, fringles@istat.it
David TRAMBUSTI, Istat, david.trambusti@istat.it