A WEB SURVEY ON AN ELUSIVE POPULATION: A FOCUS ON INDICATORS TO MANAGE DATA COLLECTION PROCESS

Monica Perez, Linda Porciani, Federico De Cicco[1]

Abstract. In 2021, the Italian National Statistical Institute (Istat) started the study to carry out a survey about labor discrimination of LGB (lesbians, gays and bisexuals) persons without legal relationship[2]. The survey population is extremely sensitive and hard-to-count mainly for two reasons: it is (self) defined by sexual orientation and there is no list to count and identify the initial population. Moreover, labor discrimination is a topic generally underrepresented in any kind of subgroups of Italian population.

The survey fieldwork has been carried out from January to the end of May in 2022.

The project team chose the web Respondent Driven Sample (w-RDS) method as the most suitable way to sample the population. The web questionnaire has been developed following “a privacy by design” approach to preserve the privacy of respondents.

The main efforts of data collection design were devoted to the implementation of a system of ad hoc indicators able to: 1. monitor the quality of data collection process (how to define and calculate the response rate without an initial population?); 2. evaluate the goodness of the sample (how many and which respondents can guarantee the data quality criteria?); 3. allow decisions regarding the change of data collection techniques or data design *in itinere*.

The paper addresses methodological and organizational aspects of data collection design, focusing on system of monitoring survey indicators, for an experimental survey on LGB adult population resident in Italy. Lesson learned could be useful for RDS implementation in future surveys conducted by National Statistical Institute.

[1] The paper is the joint work of the authors. More in detail, the single paragraphs are as follows: par. 1 and 2 to Monica Perez; par. 2.1 to Federico De Cicco; par. 2.2 and 2.3 to Linda Porciani; par. 4 to the joint work of the authors.

[2] Italian legislation (law n. 76, 20 may 2016) recognized civil union among homosexual persons.

1. Introduction

The survey has been conducted by Istat within the project on "Access to work, working conditions, labour discrimination of LGBT+ people and diversity policies implemented at enterprises" carried out in collaboration with the National Office Against Racial Discrimination (UNAR). The objective of the study is to provide a picture of the perception and prevalence of forms of discrimination, threats and assaults that homosexual and bisexual people may have experienced according to sexual orientation in the daily life, mainly focusing on labour situation.

With reference to the part of the project aimed at investigating the experiences of the LGBT+ population, given the heterogeneity and plurality of this population which is extremely sensitive and elusive, we ideally divided the population into three subgroups that were investigated through three different surveys: (i) homosexual people who are in civil union or have been in civil union previously (selected by municipal lists as of January 1, 2020) (2020-2021); (ii) LGB people who are not in a civil union nor have been in a civil union previously (2022); (iii) trans and non-binary people (2023).

As for the survey on the LGB people without legal relationship (ii), which is the subject of this paper, the lack of a sampling frame led us to use the Respondent Driven Sampling method, which is a probabilistic sampling method based on social ties (De Rosa et al. 2020; Sheim et al. 2016; Vitalini 2012). RDS is similar to snowball sampling, a chain-referral sampling method where participants recommend other people they know belonging to the target population. The main difference between the two methods is that RDS is mathematically tweaked to add an element of randomness. RDS can be thought as a group of snowballs, each rolling down a hill in its own random direction.

This type of sampling model is usually associated with the web technique for data collection, the so called WebRDS. It has been used in surveys conducted to study the risk on gays of HIV transmission in Vietnam and Sweden (Bengtsson et al. 2012; Stromdahl et al. 2015); transgender women in San Francisco (Wesson et al. 2013); the risk on gays and bisexuals of sexually transmitted diseases in New Zeland (Ludlman et al. 2015).

Application of the RDS method as well as privacy of the respondents and the protection of data confidentiality due to the sensitiveness of the theme are crucial points that have numerous implications with repercussions on the survey design.

2. The project design

The Lgb survey, due to its experimental design, has implied innovations mainly in three phases of the Generic Statistical Business Process Model (GSBPM), that are interconnected:

a. Design phase: *Privacy by design*
b. Build and Collect phase: *Respondent Driven Sampling (RDS)*
c. Evaluate phase: *ad hoc monitoring process indicators*

Before going into the details of each of these steps, it may be helpful to specify that:

a) LGBT associations played an important role in the survey. The survey design based on 50 LGBT associations and 10 initial (potential) respondents per each (the so called "seeds") . The list of seeds was identified by the LGBT Associations among the people belonging to the associations themselves. The choice of the seeds, based on socio-demographic criteria provided by Istat, was a crucial point because the seeds must be capable of generating long chains of recruitment. Each LGBT association was provided with a different link to deliver to its own seeds, in order to trace the origin of the chains of the propagation network;
b) LGBT associations did not have to disclose the identity of selected seeds. Each association had to appoint the Data Processor according to art. 28 of GDPR;
c) each respondent was required to recruit other four individuals belonging to the target population and belonging to the circle of one's own acquaintances;
d) due to RDS method, questions concerning the respondent's acquaintance with the person who recruited him/her (reciprocity of ties) and the size of each respondent's social network need to be included. The former is necessary for calculating the balance condition in the recruitment process, the latter is an indispensable variable for estimating the probability of inclusion (par. 2.2).

2.1 Privacy by design

Given the sensitive topic of the survey privacy measures have been taken - both methodologically and organizationally - for data processing and storage, taking into account a privacy by design and privacy by default approach.

The survey design has been defined according to a specific Data Protection Impact Assessment (Art. 35 of EU Regulation 2016/679) and risk based approach, focusing on risks analysis and the measures to enhance data protection.

On the field, the privacy measures have been realized through a sequence of actions needed to access the questionnaire (Fig. 1). Each respondent receive a link to enter the “Accession module”, that is an introductive web page containing preliminary questions aimed at identifying the person's eligibility to be part of the survey sample. The “Accession module” describes the key points of the survey such as the aim, data collection mode, survey period, privacy legislation and, finally, some preliminary questions to ascertain eligibility of the person in the sample, and namely: (i) being aged 18 years or more ; (ii) being resident in Italy; (iii) knowing the person sending the link.

After the compilation of this module respondent needs to indicate an email address to receive a personal link to access the “Questionnaire”.

First questions of “Questionnaire” aim to complete the eligibility evaluation of the respondent, according to sexual orientation and marital status. These data are recorded in a different data server respect to the information provided compiling the “Accession module”. Moreover, the email address is stored in encrypted mode to minimize the risk of individual identification of respondent.

Once the questionnaire is completed, an “invitation” link - automatically produced by the data collection system - is immediately available to the respondent on the web page to be sent to other potential respondents identified by him/her to further propagate the respondents’ chain. The invitation link can be transmitted by e-mail, WhatsApp, SMS or other instantaneous messages services. Simultaneously, the system send the link to the e-mail address of the respondent, so that it remains available to recruited people even at a later time, after closing the web page.

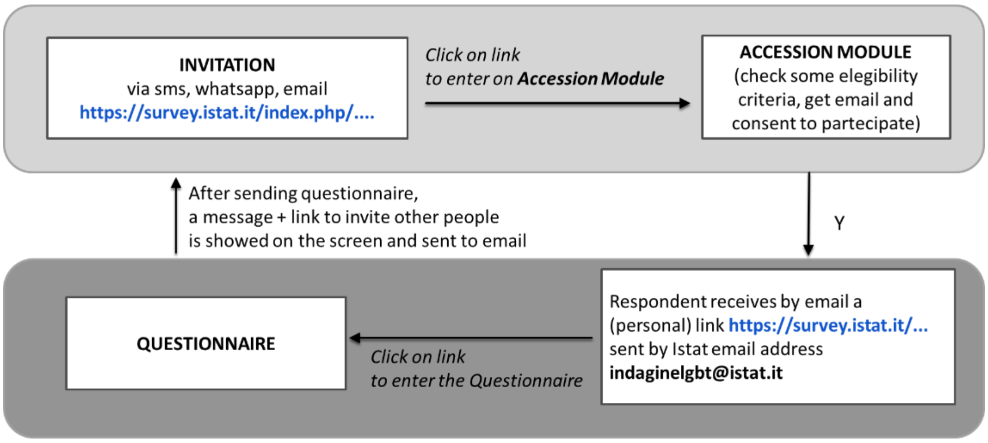
The initial respondent has two possibilities to propagate the questionnaire: i) copying the link showed on web page at the end of own questionnaire and immediately sharing it with other people (by instantaneous messages services); ii) using the link sent by email and sharing it later time.

The following steps are a part of the cycle (Figure 1) starting from again by the “Accesion module”. Each respondent should be a recruiter to build a good sample.

The protection of privacy has been guaranteed by the separation of Accession module and Questionnaire: the respondent-recruiter can not access personal data of invited respondents. Moreover, a check system has been implemented on the number of the sent invitation links, which has a maximum of 10 for the associations and 4 for respondents. Each link, both for Accession module and Questionnaire, is unique and it is highly unlikely to reproduce because of the token length based on the combination of 10^{26} order.

Data protection was an element of attention in each survey step. Coded variables have been used for associations, first respondents (seeds) and further respondents in order to avoid the identification of the subjects (even indirectly) end preserve data protection not only of the investigation process but also of the survey monitoring

Figure 1 – *Data collection design scheme*


INVITATION
via sms, whatsapp, email
https://survey.istat.it/index.php/....
Click on link
to enter on Accession Module
ACCESSION MODULE
(check some elegibility
criteria, get email and
consent to partecipate)
After sending questionnaire,
a message + link to invite other people
is showed on the screen and sent to email
Y
QUESTIONNAIRE
Click on link
to enter the Questionnaire
Respondent receives by email a
(personal) link https://survey.istat.it/...
sent by Istat email address
indaginelgbt@istat.it

2.2 Respondent Driven Sampling (RDS)

RDS strategy is helpful to reach a population without a sampling frame to select sampling units and, consequently, without the possibility to have a probabilistic sample.

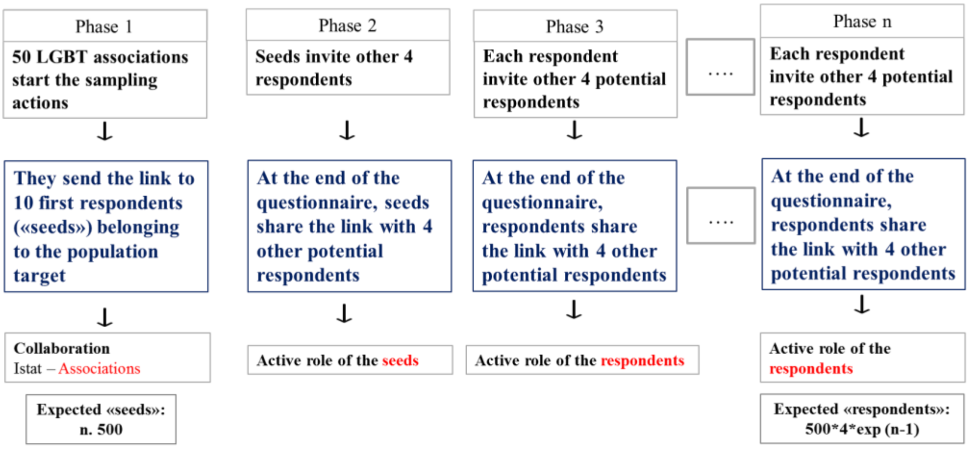
The sampling strategy based on RDS has a probabilistic approach. It combines the snowball technique - the sample is constructed by using sample units (individuals) provided by the initial recruiters (seeds) and subsequent recruits/recruiters (called nodes) - with a mathematical model that formalizes the recruitment process. Under certain conditions the recruitment process is a Markov chain (probabilistic process). (Salganik e Heckathorn, 2004; Volz e Heckathorn, 2008). The seeds need to be chosen non-randomly, based on the differentiation criterion and their ability to recruit. The recruitment process develops in waves that are generated starting from the initial recruiters until an equilibrium condition is reached, in which the probability of inclusion of the sample units stabilizes.

As mentioned above, in the experimental LGB survey, the seeds have been recruited by LGBT associations based on socio-demographic criteria. Afterward, the process of respondent-driven sampling started: at each wave, respondents were used to select or drive the next sampling wave by recruiting other individuals from the target population (Figure 2).

The data collected during the sampling process are used to make inferences about the structure of the social network to which they belong and to obtain unbiased estimates of the target population. Information on: (i) properties of the (responding) nodes; (ii) who recruits whom (recruitment matrix); (iii) size of the respondents' personal network (number/strength of ties) are basic elements for generating inferences on the characteristics of the population. Under specific assumptions, RDS estimators are asymptotically unbiased (Salganik e Heckathorn, 2004).

The assumptions under the mathematical model concern both network structure and sampling. In fact, the network of the target population must be sufficiently dense and connected so that each node is reachable from the other nodes. Furthermore, the network does not have to be too segmented, to prevent the chains from becoming trapped in subgroups. Such situation would not allow equilibrium to be achieved. Respondents have to mantain symmetrical relationships and must recognize each other as members of the reference population (unoriented network). The number of ties between members must be sufficiently high to support recruitment process (recruitment chains spanning multiple waves) to ensure that each member of the population has a non-zero probability of entering the sample. As far as the sampling, the hypotheses concern: the selection with re-entry of units, the accuracy of estimate of ties, the randomness of recruitment (Gile e Handcock, 2010; Xin, 2013).

Figure 2 – *RDS scheme in LGB labor discrimination survey*


Phase 1
50 LGBT associations start the sampling actions
They send the link to 10 first respondents («seeds») belonging to the population target
Collaboration Istat – Associations
Expected «seeds»: n. 500
Phase 2
Seeds invite other 4 respondents
At the end of the questionnaire, seeds share the link with 4 other potential respondents
Active role of the seeds
Phase 3
Each respondent invite other 4 potential respondents
At the end of the questionnaire, respondents share the link with 4 other potential respondents
Active role of the respondents
….
….
Phase n
Each respondent invite other 4 potential respondents
At the end of the questionnaire, respondents share the link with 4 other potential respondents
Active role of the respondents
Expected «respondents»: 500*4*exp (n-1)

2.3. *Monitoring survey process indicators*

The effectiveness of the method RDS and the duration of the survey depend on the propagation capacity of the network.

If for any reason a participant decides not to "propagate" because he/she becomes discouraged, loses confidence, loses referrals, that node does not produce offspring and the network reduces its propagation effectiveness by limiting the achievement of a satisfactory sample.

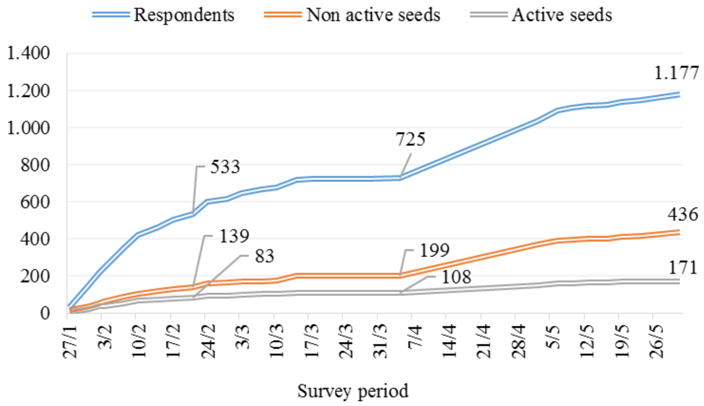
In order to monitor the survey, two sets of indicators have been elaborated: the first set refers to indicators to monitor the strength of «seeds» and network propagation (we call them "network propagation" indicators); the second set, is a set of indicators able to provide essential information about the typology of the respondents ("profile" indicators).

The "network propagation" indicators include the number of active/non active seeds; the number of active/non active respondents; the number of created chains.

The "profile" indicators monitor the number of total respondents (seeds + subsequent respondents) by sex, sexual orientation, age group and participation in LGBT association.

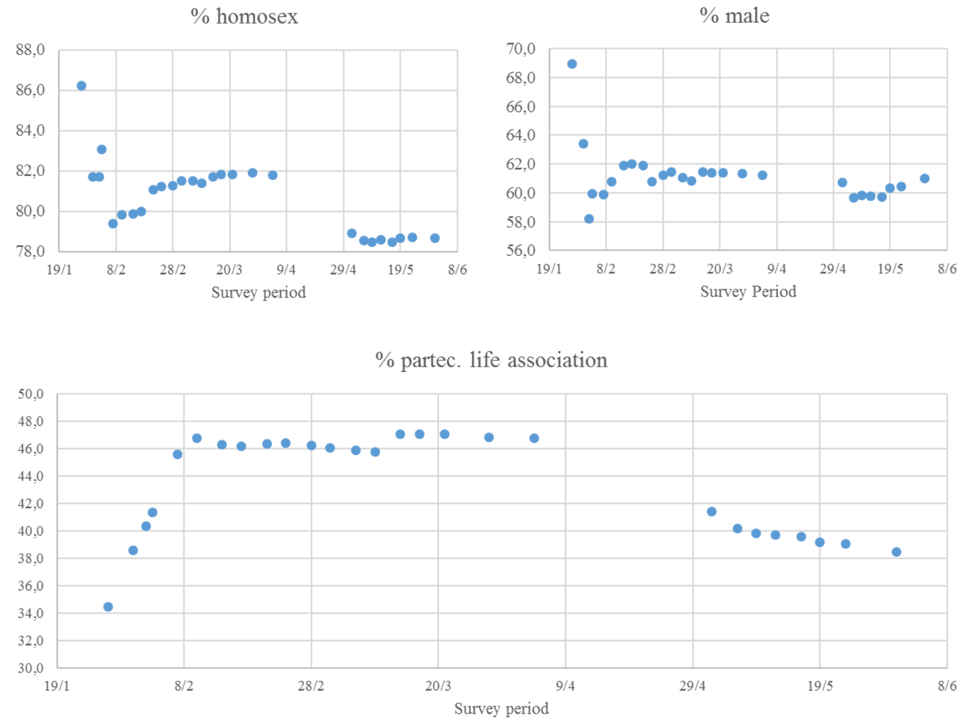
After a month of fieldwork, the indicators gave us some signal of criticalities of the network propagation for LGB population. On 50 involved associations, 24 percent were completely inactive (any seed has compiled the form); 76 percent were active (they have active seeds) producing 6.4 seeds on average (versus the expected 10). Nevertheless, and more seriously for the sample building, 62 percent of active seeds have compiled just its own questionnaire without any propagation activity (83 active seeds and 139 non active seeds), so they did not contribute to create respondent chain. The active seeds generated just 2.4 respondents on average versus the expected 4. Totally, the respondents were 533 after one month (Figure 3).

Figure 3 – *"Network propagation" indicators: respondents and non active/ active seeds (absolute values) during survey period.*



The "profile" indicators describe a homogeneous sample in terms of gender composition (more than 60 percent are men), sexual orientation (almost all units are homosexual) and participation in the life of LGBT associations (more than 40 percent of respondents have or have had experience in a thematic association): these characteristics are some of the criteria used to assess the quality of the probability sample (Figure 4).

Figure 4 – *"Profile" indicators (percentage values) during survey period.*



In order to increase the participation of all the expected seeds (10 per each association) and, above all, the generation of all the expected respondents (4 per each recruiter), reminder actions have been activated involving the LGBT associations (45 days after the start of the survey).

In spite of this, a small impact on the spread of the chains was observed. Both types of indicators remained stable (Figure 3 and Figure 4). After two months of data collection, it is only possible to observe an increase in the activity of the associations: from 76 percent to 90 percent of them had at least one active seed. However, the general scenario did not change, confirming the low participation of seeds in the survey (6.5 seeds per association on average), the low level of reproduction activity (64 percent of seeds have no offspring respondents) and the small size of the network

(2.4 respondents per seed on average). After 45 days, the total number of respondents is 725 (+36 percent). Analysis of process indicators suggested the change of data collection strategy, moving from RDS to simple snowball sampling method. This is made by the publication of the questionnaire link on the associations' web page. It implies no distinctions between a seed and a respondent, so the possibility to keep track of who referred who in the sample is loose and consequently the probabilistic approach is lost. The change of the sample strategy has been shared with LGBT associations, because it required their different participation in the survey process.

At the end of fieldwork period (lasted 4 months) the total respondents are 1,177: 54.3 percent are seeds (or first respondents in the second strategy), of which 28 percent are generative (Figure 3).

3. Final considerations

Istat has had its first experience of field application of RDS method for elusive population in LGB labor discrimination survey.

At the end of this innovative and experimental survey, the research group acquired expertise and suggestions for planning further surveys with similar characteristics.

Firstly, the need for ad hoc procedures to manage privacy issues was evident. In the LGB survey, a high privacy protection model was implemented through a two-step access to the questionnaire. The adopted data protection measures in such a survey could be more prominently promoted to enhance the involvement of distrustful people. However, challenges persist in ensuring a balance between privacy and data collection effectiveness.

Secondly, the LGB survey revealed a critical point in the activity of seeds and respondents. A low knowledge base among initial respondents (seeds) could be mitigated through an initial training activity between the research group and seeds, focusing on recruitment strategies. The low activity of respondents' propagation may also be influenced by various factors, such as the sensitive nature of the topic, the length of the questionnaire, and the intricacies of the recruitment process. These aspects warrant further detailed analysis and consideration. Moreover, exploring the potential for providing incentives to respondents could prove beneficial in enhancing participation rates in future Official Statistics surveys employing RDS.

Thirdly, there is a clear need to enhance the set of indicators with more in-depth elements on network propagation. Developing a robust evaluation benchmark for the quality of the data collection process is imperative. This includes not only understanding the breadth of the network but also its depth and the effectiveness of each step in the sampling process.

Fourthly, the representativeness of the sample is a significant challenge in RDS. Unlike traditional sampling methods, RDS does not guarantee a random sample from the target population. The method relies on the social networks of the initial participants (seeds) to recruit additional participants. This can lead to biases, especially if certain segments of the population are more connected or influential within the network. Ensuring that the seeds are diverse and well-connected within the target population can mitigate this issue to some extent. However, the degree to which the sample reflects the true population remains a challenge in RDS studies, and statistical adjustments or modeling techniques may be necessary to account for this bias.

References

BENGTSSON L., LU X., NGUYEN Q. C., CAMITZ M., LE HOANG N., NGUYEN T. A., LILJEROS F., THORSON A. 2012. Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam. Published: https://doi.org/10.1371/journal.pone.0049417

DE ROSA ET AL. 2020. Il Web-Respondent driven sampling per lo studio della popolazione LGBT+. Rivista Italiana di Economia Demografia e Statistica, Vol. LXXIV n.1, Gennaio-Marzo 2020.

GILE K. J., HANDCOCK M. S. 2010. ‘‘Respondent-Driven Sampling: An assessing of current methodology’’. Sociol Methodol. 40(1): 285-327.

LUDLAM A., SAXTON P., DICKSON N. P., ADAMS J. 2015. Respondent-driven sampling among gay and bisexual men: experiences from a New Zealand pilot study, BMC Res Notes, 8:549.

SALGANIK M. J., HECKATHORN D. D. 2004. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, Sociological Methodology, Vol. 34, pp. 193-239..

SCHEIM A. I., BAUER G. R., COLEMAN T. A. 2016. Sociodemographic Differences by Survey Mode in a Respondent-Driven Sampling Study of Transgender People in Ontario, Canada, Published Online: 1 Oct 2016 https://doi.org/10.1089/lgbt.2015.0046

STRÖMDAHL S., LU X., BENGTSSON L., LILJEROS F., THORSON A. 2015. Implementation of Web-Based Respondent Driven Sampling among Men Who Have Sex with Men in Sweden. Published: November 12, 2012 https://doi.org/10.1371/journal.pone.0049417

VITALINI A. 2012. L’uso delle reti sociali per la costruzione di campioni probabilistici. Roma: Aracne.

VOLZ E., HECKATHORN D. D. 2008. Probability-Based Estimation Theory for Respondent-Driven Sampling, Journal of Official Statistics. Vol. 24, No. 1, pp. 79–97

WESSON P., QABAZARD R. F., WILSON ERIN C., MCFARLAND W., FISHER R. H. 2013. Estimating the population size of transgender women in San Francisco using multiple methods, pp. 107-112. Published online: 28 Sep 2017 https://doi.org/10.1080/15532739.2017.1376729

XIN L. 2013. Respondent-Driven Sampling: Theory, Limitations & Improvements. Karolinska Institute. Printed by US-AB, Stockholm.

Monica PEREZ, Italian National Institute of Statistics, perez@istat.it
Linda PORCIANI, Italian National Institute of Statistics, porciani@istat.it
Federico DE CICCO, Italian National Institute of Statistics, fedecicc@istat.it