Estimating the Homeless Population through Indirect Sampling and Weight Sharing Method

Claudia De Vitiis, Stefano Falorsi, Francesca Inglese, Monica Russo

Abstract The Italian National Institute of Statistics carried out the first survey on the homeless population. The survey aims at estimating the unknown size and some demographic and social characteristics of this population. The sample strategy for the survey refers to the theory of indirect sampling, based on the use of a sampling frame indirectly related to the target population. Following the indirect sampling approach, the estimation is performed through the "weight share method", based on the links connecting the frame of services with the population of homeless.

1 Introduction

The Italian National Institute of Statistics carried out the first survey on the homeless population, aiming at estimating the unknown size and some demographic and social characteristics of the population constituted by all persons who do not have a residence and resort to some of the services provided to persons in difficulty. The methodological strategy to investigate the homeless population could not follow the standard approaches of the surveys of the official statistics which are usually based on the use of population lists. Because of the lack of a sampling frame, it was necessary to resort to an indirect approach, identifying the population units through the services provided to them. Then, the chosen strategy is to construct firstly an archive of all centres providing services and then to select an indirect sample of persons from the users of these services, to be collected during an appropriate period of time.

The overall survey design (De Vitiis *et al.*, 2011) consisted of three steps: the first step built a complete list of the existing centres providing the services to persons in difficulty (ISTAT, 2010); the second step was a census survey of the centres; the third step was a sample survey conducted at the centres to capture the homeless population, carried out at the end of year 2011 and still in process at the moment of writing. The list of centres providing meals and night accommodation represented the sampling frame for the third phase. As the homelessness phenomenon is mainly concentrated in big towns, the geographical coverage of the survey has been restricted to the set of the main municipalities, for a total of 158 municipalities.

The sample strategy for the survey at the third step refers to the theory of indirect sampling (Lavallée, (2007)), based on the use of a sampling frame indirectly related to the target population. In this context, the sampling frame is represented by the service providers and, consequently, the target population is restricted to the people using services. Following the indirect sampling approach, the estimation is performed through the "weight share method", based on the links connecting the frame of services with the population of homeless.

The adopted approach represents an important innovation for the Italian official statistics because of two main reasons: the homeless population is surveyed at national level for the first time on the whole Italian territory and a new methodological instrument, such as the indirect sampling, is experimented for a large scale survey.

In this paper, the sampling strategy based on the approach of indirect sampling is discussed in section 2, while in section 3 the sample design defined for the third phase of the survey is described together with some numerical results.

2 Indirect Sampling in the context of homeless survey

The approach of indirect sampling is useful when a sampling frame of the target population, U^A, is not available, but it is possible to use a list referred to a different population, U^B, related to the target one. In this context, the sampling strategy consists in selecting a sample from U^B and producing the estimates of the parameters of interest referred to U^A taking into account the links between the two populations.

In the homeless survey U^A units are homeless persons, k indicates a person that receives at least one service during a defined period of time J, U^B units are the services, i, provided to persons in the considered centres during the same period of time. Hence, services are the units through which persons can be reached, given the one-to-one correspondence between the two populations in a specific point of time, during which a person could not receive two different services. A generic parameter of interest, defined on the target population, can be expressed as:

$$Y = \sum_{k \in U^A} y_k = \sum_{i \in U^B} \frac{y_k}{r_{k \equiv i}(J)} \;, \tag{1}$$

where $r_{k \equiv i}(J)$ is the number of services provided to person k (given the correspondence between person and service) during the period J in all centres in the survey field. The most relevant parameter of interest defined on the target population is its unknown size N (obtained defining $y_k = 1$,∀k), together with some other parameters referred to characteristics of the homeless persons. As the sampling frame is constituted of the list of centres providing services, each centre is a cluster of services, while the sampling units are the triplets (centre, point of time, service). The points of time are defined as a specific lunch time or dinner time for centres providing meals and a specific night for those providing accommodation.

In the indirect sampling framework the calculus of sampling weights focuses on the relationships between the units from the sampling frame and those from the target population. In fact, to assign a sampling weight to each interviewed person, it is necessary to start from the weight of the sampled services, adopting the estimation method known as weight sharing method (Lavallée, 2007). The estimator, expressed in terms of the weight $\tilde{w}_k$, is given by:

$$\hat{Y}_J = \sum_{k\in s} y_k \tilde{w}_k, \qquad \tilde{w}_k = \frac{1}{r_k(J)} \sum_{K(i)=k} w_i$$

, (2)

where w_i represents the sampling weight associated to the selected service, related to individual k belonging to sample s. To ensure a correct sharing of the weights, the map of the links between persons and centres has to be known: for each interviewed person the list of all visited centres had to be collected for a fixed period of time. The length of the observation period for each person has been set in a week, so that the estimate of the number of links refers to an average week, which has to be expanded to obtain a monthly estimate. The survey instrument to collect appropriately all the information to map the links is a daily diary in which the places where the person ate and slept in the seven days preceding the interview are collected.

It is useful to remark that the use of indirect sampling in the homelessness context introduces a risk of bias due, on one hand, to the presence in the centres of persons in economic difficulty but living in regular households and, on the other hand, to the fact that not all the target population is surveyed, because only users of those services are sampled. In order to reduce the risk of bias, in the collection phase the eligible units are previously identified and the survey time period was set to be long enough to ensure that most homeless people uses services at least once: one month has been proved to be a good choice (Ardilly and Leblanc, (2001)).

3 The Sampling Design for the Homeless Survey

The sample design defined for the third survey step considered all the soup kitchens and night shelters; being their number (about 800) too small to give reason for a sample selection, it was decided to consider all of them. In this way, a one stage stratified random sampling design was defined in which each centre represents a stratum. Centres providing meals both at lunch and dinner time have been considered twice, because the same homeless person can be find in both occasions. The random selection concerned both the time dimension and the centre users: each centre was visited in prefixed points of time, randomly selected among the opening hours during the 30-day reference period; for each selected couple (centre, point of time), the interviewer selected randomly a sample of persons among the centre users by means of a systematic procedure. The size of this sample is predefined accordingly to a proportional allocation fixed sample fraction.

The overall number of interviews was set to 5.400. This sample size has been evaluated in terms of expected sampling error for the estimate of the population size N. The formula of the sampling variance of the population size estimator has been derived in the context of the indirect sampling in order to obtain an evaluation of the expected sampling error (De Vitiis *et al* 2011). The variance of the weight share method estimator depends on the variability of the number of links (Lavallée, 2007). As the homeless phenomenon is completely unknown, an evaluation of the sampling variance has been obtained assuming a pessimistic distribution of the number of links.

The distribution of the overall sample among the 850 centres was performed in proportion to the number of monthly services provided to homeless persons, estimated from data collected in the second phase of the survey. Specific survey occasions were assigned at random to each centre, setting firstly the number of times the centre had to be visited by the interviewers (in the range 1-15 among 30 days, in relation to the total sample assigned to the centre), then the number of interviews to be conducted each time (in the range 1-12) and eventually assigning at random the specific dates. The interviewers were instructed to select the sample persons, at each specific occasion, accordingly to a systematic procedure from the list of users of the centre, if possible, or from the line of persons waiting to access to the service.

The survey phase produced a very satisfying overall result: in fact, as shown in Table 1, the percentage of realised interviews was above 90%, being the most part of non-response due to refusing or out-of-target centres.

Table 1: Planned and realised sample

	Services	Interviews
Planned sample size	850	5414
Refusals	37	208
Out of target	43	215
Realised sample size	770	4991

At present, weights for the respondent persons have to be evaluated. The main issues to deal with in this phase are treatment of non-response and reconstruction of the number of links, crucial for obtaining unbiased weights. While service non-response will be treated using usual reweighting methods, link non-response represents a very delicate concern. In fact, as highlighted in Xiaojian and Lavallée (2009), the main risk when using the weight share method is producing overestimation due to link non-response. Therefore, specific non-response adjustments will be studied for overcoming the lack of information regarding the relationship between units in the sampling population U^B and unit in the target population U^A.

References

1. Ardilly, P., Le Blanc, D.: Sampling and weighting a survey of homeless persons: a French example. Survey Methodology, Vol. 27, n.1, 109--118 (2001)
2. De Vitiis, C., Falorsi, S., Inglese, F., Russo, M.: Indirect sampling in the First Italian Survey on Homeless Population, Paper presented at the ITACOSM Conference 2011, Pisa, 27-29 June 2011
3. Istat: I servizi alle persone senza dimora: primi risultati Anno 2010,, Statistiche in breve, 13 Dicembre 2010.
4. Lavallée, P.: Indirect Sampling, Springer, New York (2007)
5. Xiaojian, X., Lavallée, P.:Treatments for link nonresponse in indirect sampling. Survey Methodology, Vol. 35, n.2, 153--164 (2009)