

# On sampling elusive populations

## *Sul campionamento di popolazioni elusive*

Vijay VERMA

**Abstract** The paper describes some useful techniques for sampling elusive populations: populations for which complete frames are not available, which are small and rare, are mobile, and/or are reclusive to participation in surveys. These include sampling of informal sector establishments, multiplicity and multi-frame sampling, adaptive cluster sampling, capture-recapture and ‘time-location’ sampling, controlled selection, and snowball and respondent driven sampling.

**Abstract** Questo articolo descrive alcune metodologie utili nel campionamento di popolazioni elusive: popolazioni piccole e rare, per le quali non è disponibile una lista, che si spostano, e/o sono riluttanti a partecipare ad indagini. Queste includono il campionamento di attività nel settore informale, campionamenti ripetuti e multi-frame, campionamento adattivo a grappolo, ‘cattura e ricattura’ e ‘time-location’, selezione controllata, campionamento a valanga e respondent driven.

**Key words:** imperfect frames, elusive, rare, mobile, reclusive population, sampling units, analysis units, direct, multiplicity, link-trace sampling

## 1 Introduction: the concept of elusive populations

By elusive populations we mean populations for which - by virtue of their characteristics, or of the lack of suitable sampling frames, or difficulties in obtaining the required information - adequate samples cannot be defined, drawn or implemented using the normal procedures of general population sampling. Both stages – that of selecting a representative sample, and of successfully enumerating the units selected – can present severe problems and require special procedures. Problems can arise from various sources. For instance, certain categories of persons

may not be covered in the sampling frame; there may be difficulties in locating the selected units due, for instance, to their mobility or a lack of information about their whereabouts; the identification of persons of interest, even after they have been located, may require information beyond what is available; language, class, religious and other cultural differences can present obstacles; and of course, the individuals may not wish to be located, contacted or interviewed for various reasons.

Addressing sampling issues requires the identification of the special characteristics and circumstances of the populations to be sampled (see for instance, Kish, 1991). Characteristics of elusiveness include the population being: ill-defined; heterogeneous; small or rare; unevenly distributed with patchy concentrations; mobile and changeable over time; hidden, difficult to identify and locate; and reclusive, with sensitive characteristics. Elusive populations, seen from the point of view of the consequences for sampling methodology, have four aspects: (1) populations lacking an adequate sampling frame; (2) rare populations and traits; (3) mobile populations; and (4) reclusive populations. These aspects are distinct though often overlapping; furthermore, they tend to form a hierarchy: for instance, sampling reclusive populations often involves all the problems (1)-(4).

## **2 Sampling from imperfect frames: sampling small and informal sector establishments**

There are essential requirements for any sampling frame, as well as desirable properties which a frame should have in order to yield a reasonable probability sample. In practical sampling, a basic distinction is often made between area frames and list frames. Common problems with area frames include: failure to cover the population of interest exhaustively; errors and changes in area boundaries; inappropriate type and size of units; lack of information on size and other characteristics of the units; and high cost. Concerning list frames, we have a 'perfect list' when there is one-to-one correspondence between sampling and analysis units. In practice, lists are subject to imperfections, such as the presence in the list of blanks representing no real units, clusters of units, insufficient or incorrect information on unit characteristics or for locating them, duplications, under-coverage, and in the worst case we may have no frame at all.

Examples of sampling from imperfect frames are provided, for example, by surveys of sampling small and informal sector establishments. For large and medium-sized establishments, samples are often selected directly from lists. Special considerations arise in the design of samples for surveys of small and informal sector establishments. Such units, like households, are small-scale, numerous and widely but unevenly dispersed in the population. The commonly used samples for small and informal sector establishments are area-based and involve two (or sometimes more) stages of sampling. The main problem is to deal with units of many different types in the same survey. A special tool for this purpose is the construction and use of 'strata

of concentration' of different types and sectors of establishments to control distribution of the sample (Verma 2001).

### 3 Relationship between sampling units and analysis units

In sampling methodology, the nature of correspondence between sampling units and analysis units is of fundamental importance.

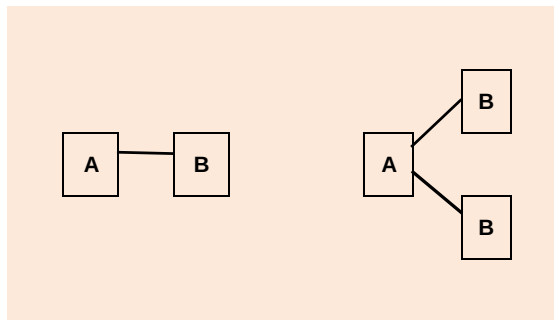
(1) Notwithstanding frame imperfections, in most applications of conventional sampling we have one-to-many/one/none correspondence between sampling and analysis units: any analysis unit is associated with at most one sampling unit. We term this *direct sampling*. By implication, its complement is 'indirect sampling' when such correspondence between sampling and analysis units is lacking. There are two forms of indirect sampling.

(2) A widespread situation encountered in surveying elusive populations is the presence of many-to-one and many-to-many links between sampling and analysis units - an analysis unit being associated with more than one sampling unit. This gives rise to *sampling with multiplicity*.

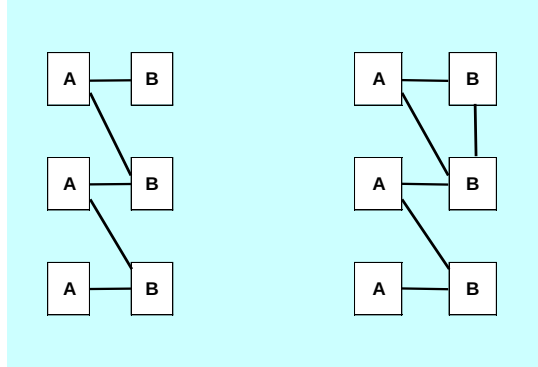
(3) Then there are situations when the sample has to be obtained by exploiting links between analysis units themselves, rather than primarily between analysis units and sampling units. This is *link-trace sampling*, also called *chain-referral sampling*.

We will use the term "multiplicity sampling" below to refer to the specific sampling procedures of the type developed originally by Sirken (1970). The term "sampling with multiplicity" has been used here to refer to the more general sampling procedure for situations when individual analysis units are linked to multiple sampling units. Multiplicity sampling, sampling from multiple frames and adaptive cluster sampling are specific applications of this general procedure.

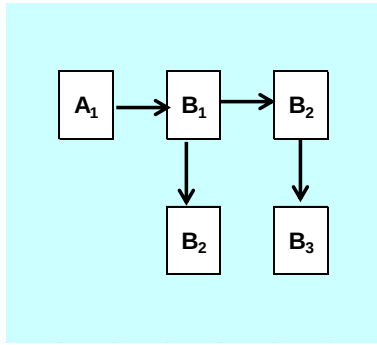
#### Direct sampling



'A' refers to sampling unit(s), and 'B' refers to associated analysis unit(s).

**Sampling with multiplicity**

Analysis units may also be linked, such that the selection of one brings the other(s) into the sample.

**Link-trace or chain-referral sampling**

$B_i$  stands for analysis unit entering into the sample in the “ $i^{\text{th}}$  recruitment wave”.

The multiplicity estimator is the common link between many of the sampling techniques. Different methods of estimation are possible in sampling with multiplicity, all with equal statistical validity but normally giving different numerical results (in terms of the estimates and their variances) in a particular application.

(1) The Horvitz-Thompson estimator is based on estimating the overall probability of the unit appearing in the sample. Let subscript  $j$  indicate a unit, and subscript  $k$  a particular ‘trial’, way, or source of its selection into the sample. Thus  $f_{jk}$  denotes the probability of appearing in the sample of unit  $j$  from source  $k$  (among a total of  $m_j$  sources for the unit). The probability that it appears in the sample is:

$$\pi_j = 1 - \prod_{k=1}^{m_j} (1 - f_{jk}) \quad w_j = 1/\pi_j$$

(2) The Hansen-Hurwitz estimator uses the expected number of times a unit is selected into the sample, taking into account its selection (and possibly multiple selections) from any of the sources. In this procedure,  $f_{jk}$  stands for the expected number of selections of unit  $j$  from source  $k$ . Assuming independence between multiple selections, the expected number of times a unit appears in the sample is:

$$p_j = \sum_{k=1}^{m_j} f_{jk} \quad w_j = (1/p_j)$$

(3) With multiplicity estimator of Sirken, the probabilities and weights are

$$p_{jk} = m_j f_{jk}, \text{ with } w_{jk} = \frac{1}{p_{jk}} = \frac{1}{m_j} \left( \frac{1}{f_{jk}} \right).$$

Estimator (2) replaces  $f_{jk}$  in (3) by its average value over the  $m_j$  listings representing unit  $j$  in the frame. The advantage of multiplicity estimator is that the probabilities and corresponding weights are determined separately for each appearance of the analysis unit in the sample. In fact it is not necessary to know whether different selections refer to the same analysis unit or to different units. That is, no matching of different appearances from different sources or selections of the same analysis unit is required, so long as its multiplicity ( $m_j$ ) is known.

## 4 Sampling rare populations

The characteristic feature of a rare population is that sampling the whole population with normal procedures does not yield a representative sample of adequate size for the subpopulation of interest because of its small size (Sudman, Sirken and Cowan, 1988). We need procedures for more intensive and targeted sampling in order to capture rare populations more effectively and efficiently. A successful sampling strategy involves identifying and making use of the pockets of concentration and patterns of distribution of the rare population of interest. There are five aspects of the strategy, which are distinct but are often used in combination for greater effectiveness: (1) locating concentrations of the rare population using existing large-scale sources; (2) partitioning the frame according to the degree of concentration of the rare population; (3) oversampling strata of concentration, making use of the patterns of concentration identified; (4) listing, screening and two-phase sampling; (5) using special procedures to increase selection probabilities of units in the rare population, such as multiplicity sampling, multi-frame sampling, and adaptive cluster sampling. Cumulation of data or estimates over samples is an additional possibility.

### 4.1 *Multiplicity sampling*

It is important to identify conditions which may make multiplicity rules preferable to conventional sampling procedures: consideration must be given to the joint effect of sampling, coverage and measurement errors in selecting the counting rules for multiplicity sampling (Sirken, 1970). Successful implementation of the multiplicity sampling procedure depends on salience of the social relationships on which its counting rules are based, and on salience of events and of the objects of the enquiry for the respondent. Reporting biases are often larger for multiplicity counting rules

than for ordinary unitary counting rules; another concern is the increased complexity. But the procedure may be more efficient in terms of costs and variance.

## **4.2    *Multi-frame sampling***

When no single sampling frame can provide a complete representation of the target population, the use of multiple frames can reduce coverage errors. Generally, the multiple sampling frames overlap and procedures are needed to deal with this, for example by constructing a new single frame without duplicates or by accounting for the duplicates in the estimation procedure. The multi-frame sampling technique has the potential to be usefully employed in population-based surveys especially when the situation has the following characteristics: the main frame for the survey is a frame of area units; the area frame provides reasonably complete coverage of the target population, even if only implicitly; and lists of analysis units can be found, each with a high concentration of particular subgroups in the rare population but not necessarily accounting for a large proportion of it. The lists may overlap, or may contain only incomplete information for the identification and linking of individuals. Appearance in multiple frames can be used to enhance unit selection probabilities.

## **4.3    *Adaptive cluster sampling***

Adaptive sampling (Thompson and Seber, 1996) is a technique designed to obtain more adequate and efficient samples for a population which is rare and very unevenly distributed. Starting from a conventional initial sample, the technique involves selecting an additional sample in the neighbourhood of points where a concentration of the population of interest is found during implementation of the initial sample. The technique is most effective when the population of interest tends to be concentrated in relatively few and large clusters but little information is available on the extent, location and patterns of its concentration.

In practical application, a number of additional technical aspects may need to be considered, such as: unequal unit selection probabilities; stratification; multistage sampling; multivariate criteria for adaptive sampling; adaptive sampling using ‘order statistics’. Then there are a number of implementation issues: whether or not to introduce adaptive sampling; criteria, definitions, rules for adaptive sampling; choosing size of the initial sample; controlling sample size, etc.

## **5    *Sampling mobile populations***

In the context of sampling, the concept of ‘mobile population’ is more general than simply not having a fixed place to be sampled from. It refers to situations when it is necessary or preferable to sample and enumerate units through their mobility (movement). The difficulties of enumerating such populations concern the identification of (i) who are the eligible respondents for the survey, and (ii) where and (iii) when to find them; also (iv) what information to ask them concerning their mobility, and (v) how to obtain the information. Another set of issues concerns (vi) how to use sample data to produce valid estimates for the population, and (vii) how to assess variances and biases to which those estimates are subject. Procedures are required to establish links between mobile units and fixed reference points through which a probability sample of the units could be obtained (e.g., Kalton, 1991). We need a framework to organise the variety of circumstances, problems and solutions encountered in sampling mobile populations. Four important concepts in the framework are: sampling locations, observation points, time segments, and ‘time-location primary sampling units’.

Time-location sampling is designed to monitor the flow of individuals through fixed locations during specified time segments. When the sample is taken at an ‘observation point’ at a given ‘time segment’, the probability of an individual being included in the sample is affected by the person’s frequency and pattern of visits to the locations and times covered in the survey population. It can require a lot of information to be collected in order to take fully into account the variations in selection probabilities; in practice, it is necessary to have alternative, simplified and more practical strategies.

## 5.1 *Capture-recapture sampling*

The capture-recapture sampling technique involves taking two (or more) independent samples from the same population and using the overlap found between the samples to estimate the selection probabilities applied to obtain those samples and the total population size (Seber, 1982). Capture-recapture applications in the social field are usually based – in developing countries in particular – on a combination of sample surveys and administrative sources.

The capture-recapture procedure is based on certain assumptions about the population and the manner in which the samples have been drawn. Such studies are normally designed with the aim of meeting the following basic assumptions: (1) the target population is ‘closed; (2) each source used is representative of the population; (3) the different sources are linked at the micro-level without ambiguity or errors. The commonly used models make some additional assumptions: (4) the different sources are independent; (5) all units have the same chance of being selected; and that (6) each source is an independent simple random sample.

We need procedures for the estimation of sample weights when the assumptions of the simple capture-recapture model are violated. Sample weights are needed to compensate for variations in unit selection probabilities, which can arise from survey

conditions (conditioning, unknown population size, incomplete frame, haphazard differences in coverage and ‘catchability’); and from survey design and implementation (the samples being dependent by design, variations in sampling rates, inherent differences in unit selection probabilities, non-response). Even though these assumptions are not always valid, fortunately (1) often the procedure is rather robust against departures from the assumed statistical model; and (2) statistical procedures have been developed to control the effect of certain departures from the original simple model (e.g., Cormack, 1989).

## **5.2 *Controlled selection and balanced sampling***

Surveys, in particular of mobile and other difficult-to-access populations, often have to be restricted to a limited area and to a small number of primary units. Controlled selection (Goodman and Kish, 1950) is a useful technique when one has to select a small sample of primary units, but at the same time ensure that it is ‘balanced’ and ‘representative’ of the population in terms of many characteristics (or control variables). The procedure controls the structure of the sample beyond what is possible with ordinary independent selection within strata, while conforming to the requirements of probability sampling. Controlled selection can also be viewed in the context of the modern theory of balanced sampling (Deville, and Tillé, 2004).

## **6 Reclusive populations**

The following approaches are based on link-trace or chain-referral sampling.

### **6.1 *Snowball sampling***

In the practical context of sampling reclusive populations, the term snowball sampling is used to refer to a convenience sampling mechanism in settings characterised by the lack of a serviceable sampling frame. This interpretation of the technique derives from Coleman (1958), and is quite different from the often quoted formal presentation of the technique in Goodman (1961).

With snowball sampling a unit of the target population can enter the sample through direct selection into the initial sample, or by being identified (‘named’) for inclusion by someone already in the sample. There are a number of parameters which define the design of a snowball sample: the number of waves, number of contacts to request, and criteria for including a participant in the sample. The primary advantage of the method is its potential for providing a means of accessing social groupings which are vulnerable or difficult to penetrate, and obtaining

respondents in settings where trust is required to initiate contact. The most important deficiency of the method concerns the selection bias which limits the validity of the sample. The snowball method is generally not an effective tool for producing reliable estimates of total population size or population aggregates of other variables, but it can capture diversity of the target population. The snowball sampling technique in principle is not complex, but its utility and success is determined by practical aspects of implementation (Biernacki and Waldorf, 1981).

## **6.2 Respondent-driven sampling**

The respondent-driven sampling (RDS) method is an improved variant of the usual snowball sampling (Heckathorn, 1997). Both the procedures are types of chain-referral sampling. In both cases, the process starts with a small number of peers, usually chosen non-randomly. However, in an effort to produce a closer approximation to probability sampling compared to ordinary snowball sampling, RDS incorporates features such as the direct recruitment of peers by their peers, a dual system of incentives (for participation and for recruiting), and restricted recruitment quotas (e.g. three recruits only).

The basic assumptions of the RDS model is the application of Markov chain theory to the procedure of recruitment into the sample. The theory implies that after a sufficient number of waves, the characteristics of the individuals in the final sample are independent of the seeds' characteristics. Under the hypothesis that the probability of selecting each participant is proportional to his/her social network size, supposedly unbiased survey estimates and estimates of their variance can be produced. These assumptions need empirical validation. Above all, as with the snowball sampling technique generally, the utility and success of RDS is determined by practical aspects of implementation. The RDS method has been widely used, especially in studies of drug abuse.

## **7 Concluding remark: the context of this presentation**

This paper draws on two works on sampling methods for surveys of child labour (Verma 2008; Verma 2013). Household-based sampling provides an efficient approach to estimating the prevalence and characteristics of the predominant forms of child labour for children living in private households. Surveying working children not living at home requires alternative approaches. The appropriate survey approach depends on the circumstances of child labour – specifically, on the relationship of the child and of child labour with the household. Several scenarios may be identified: (1) children living and working at home; (2) living at home but working away; (3) living at home and working away, but with insufficient information available on the location or type of work they do; (4) children not living at home, but with sufficient

connection to their household to be selected, identified and traced from there; (5) children not living at and insufficiently connected to the household, but working at some fixed, more or less identifiable location; (6) children not living at and insufficiently connected to the household, who are working at mobile, difficult to locate and identify, or even unknown places; and (7) working children approachable at best only through intermediaries – through adults or other children.

Situations (1)-(3), and possibly a part of (4) may be covered through household-based surveys. These cover a large majority of labouring children, but not necessarily of those living or working in the most adverse circumstances. Surveying situations (5)-(7), and possibly also (4), generally require data collection approaches other than the conventional household survey. Alternative approaches are also required when the objective is to focus on specific sectors or types of child labour activity, when their concentrations cannot be captured efficiently or the required information cannot be obtained with sufficient detail and accuracy through general household-based surveys.

## References

1. Biernacki, P., Waldorf, D: Snowball sampling: Problem and techniques of chain referral sampling. *Sociological Methods and Research*, 10(2), 141–63 (1981)
2. Coleman, J.: Relational analysis: The study of social organizations with survey methods. *Human Organization*, 17, 28–36 (1958)
3. Cormack, R.: Log-linear models for capture-recapture. *Biometrics*, 45, 395-413 (1989)
4. Deville, J., Tillé, Y.: Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912 (2004)
5. Goodman, L.: Snowball sampling. *Annals of Mathematical Statistics*, 20, 572-9 (1961)
6. Goodman, R., Kish, L.: Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-72 (1950)
7. Heckathorn, D.: Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44 (2), 174–99 (1997)
8. Kalton, G.: Sampling flows of mobile human populations. *Survey Methodology*, 17(2), 183-94 (1991)
9. Kish, L.: Taxonomy of elusive populations. *Journal of Official Statistics*, 7(3), 339-47 (1991)
10. Seber, G.: *The Estimation of Animal Abundance*. Caldwell NJ: The Blackburn Press (1982)
11. Sirken, M.: Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-66 (1970)
12. Sudman, S., Sirken, M., Cowan, D.: Sampling rare and elusive populations. *Science*, 240, 991-6 (1988)
13. Thompson, S., Seber, G.: *Adaptive Sampling*. New York: Wiley (1996)
14. Verma, V.: Sample designs for national surveys: surveying small-scale economic units. *Statistics in Transition*, 5(3), 367-82 (2001)
15. Verma, V.: *Sampling for Household-based Surveys of Child Labour*. Geneva: International Labour Organisation (ISBN 978-92-2-121503-5) (2008)
16. Verma, V.: *Sampling Elusive Populations: Applications to Studies of Child Labour*. Geneva: International Labour Organisation (ISBN 978-92-2-128321-8) (2013)