Survey Methodology, June 2001
Vol. 27, No. 1, pp. 109-118
Statistics Canada, Catalogue No. 12-001

Sampling and Weighting a Survey of Homeless Persons: A French Example

Pascal Ardilly and David Le Blanc [1]

Abstract

In 2001, the INSEE conducted a survey to better understand the homeless population. Since there was no survey frame to allow direct access to homeless persons, the survey principle involved sampling the services they received and questioning the individuals who used those services. Weighting the individual input to the survey proved difficult because a single individual could receive several services within the designated reference period. This article shows how it is possible to apply the weight sharing method to resolve this problem. In this type of survey, a single variable can produce several parameters of interest corresponding to populations varying with time. A set of weights corresponds to each definition of parameters. The article focuses, in particular, on "an average day" and "an average week" weight calculation. Information is also provided on the use data to be collected and the nonresponse adjustment.

Key Words: Weight sharing; Incomplete frame; Homeless persons.

1. Introduction

In 2001, INSEE conducted a survey to better understand the homeless population. This was the first representative survey of this type in France (A survey of this type was conducted in the United States in 1991 by *Research Triangle Institute* (RTI) in the Washington metropolitan area (RTI 1993)). The survey principle was to reach homeless persons through the services provided to them, specifically, overnight accommodation and meals. Obviously, a person could use one or more of the services of the survey frame during the reference period considered, which creates a problem when weighting the survey's individual data files. In this article, we will show how the weight sharing method can be applied to this problem. In this type of survey, unlike most traditional household surveys, a single variable can produce several parameters of interest corresponding to different population concepts: the ones used most often by practitioners are the " average day" and "average week" parameters. A set of weights corresponds to each definition of parameters. We will provide precise definitions of these concepts and will focus in particular on the practical calculation of the corresponding weights. The article is laid out as follows: we will begin by stating the objectives of the survey, identifying its reference population and describing its sample design. We will then introduce the parameters of interest and derive the estimators of these parameters using the weight sharing method. We will describe the practical application of "average day" and "average week" weight calculations. Lastly, we will discuss practical considerations related to the nonresponse adjustment.

2. "Homeless" Survey

2.1 Objectives of the Survey

The purpose of the survey conducted by the INSEE in February 2001 was to obtain a better understanding of the "homeless" population. This population is normally defined by default as all persons who do not have a fixed residence. It is a population that is not captured by traditional household surveys conducted by the Institut since such surveys have an accommodation survey frame. Since there was no sampling frame for this population, the survey principle involved reaching the target population through the services provided to persons in difficulty, specifically accommodation and meals. These service are provided at certain times that vary depending on their nature: meals are provided every day at noon and in the evening, while overnight accommodation is provided once a day.

This indirect sampling introduces two biases into the population initially targeted and the population actually surveyed. First, the entire target population is not surveyed: only those members who use the services in the survey field are potentially sampled. Second, the population actually surveyed contains individuals who do not belong to the population initially targeted to the extent that the services provided primarily for homeless persons are also used by persons who live in a regular household but who are in a vulnerable situation (this is especially true in the case of meals). Throughout this article, while keeping this distinction in mind, we will however sometimes use the expression "homeless" to designate the persons using the services in the survey field.

1. Pascal Ardilly and David Le Blanc, Institut National de la Statistique et des Études Economiques, 18 boulevard Adolphe Pinard, 75675, Paris, Cedex, France. E-Mail: pascal.ardilly@insee.fr, leblanc@ensae.fr.

2.2 Reference Population

The main feature of the services surveyed is that they are provided in specific locations; this location is accordingly called a *centre*. Several types of services correspond to a given centre. The statistical unit sampled, which we will call a *service*, will be defined as a quadruplet (service, day, time interval, person): it consists of a given type of service in a given centre, on a given day, in a given interval of time, to a given person. Of course, a person could receive several services on the same day, let alone in a given week or during the survey month.

The survey *reference period* covers one month (January 15 to February 15, 2001). The total number of days in the survey reference period is designated as J, denoted by the index j.

The *geographic field* of the survey is that of population centres with more than 20,000 inhabitants.

The *services in the survey field* are those that are provided by one of the two types of services retained – meals and accommodation – when they are provided at least one day during the survey reference period.

The *reference population*, designated as $P(J)$, consists of persons who receive at least one service in the survey field during the reference period.

This population of interest depends fundamentally on the reference period. Its size increases with the length of that period, but "more slowly" than the time: in actual fact, certain people are found in the centres every day. In reality, the change in $P(J)$ in relation to J is complex because there are two separate phenomena coming into play that would appear to have different characteristic times:

- at any given time, the "homeless population" only occasionally visits the centres in the frame: to claim to cover that population, it would be necessary to survey over a period of time that would ensure that all persons in this population had used the services at least once (this period is not known but it is acknowledged in France, "according to the experts", that the population not covered during one full month of winter is negligible).
- the "homeless" population is self-renewing over time. Year to year, there are no doubt numerous persons coming into and going out of this population, linked to demographic change or economic or structural changes in society (persons coming into and going out of vulnerable situations).

The question of how to determine J ultimately comes down to knowing whether interest is mainly in a concept of homeless "at a given moment" (J is relatively short) or a concept of homeless over a long period of time (J relatively long). The approach adopted by the INSEE is a compromise between the two.

2.3 Sample Design

The survey's sample design has three stages: selection of population centres, selection of centres and time intervals, and selection of services.

2.3.1 Selection of Population Centres

The first stage of the sample design consists of selecting the population centres, based on a size criterion defined as a combination of the population of the population centres and the ability to provide services so that they could be identified in the records of associations and of the Ministère de la Santé. This first selection stage was carried out several months before the other two. This screening was necessary because the exhaustive census of the centres and the data related to them (type of service provided, average capacity, days open, ...) was then carried out in the selected population centres. This operation was done twice: a detailed survey the year before the data collection and an update just before the start of the data collection. This process produced a survey frame of centres. This frame has a fundamental role: persons who used only non-identified centres were not be sampled.

2.3.2 Selection of Centres, Days and Time Intervals

For practical reasons, it was not possible to survey all of the centres and to keep an interviewer on site at a given centre the entire day. Nor was it possible to interview everyone in a centre. It was therefore imperative to sample:

- centres in the selected population centres (index c).
- survey days during the collection period (index j).
- intervals of time during the survey days (index t).
- persons within one of the selections (centre, day, time interval).

For theoretical reasons, *time intervals were defined in such a way that an individual could not receive two different services during a single time interval* (for example, one of these time intervals was the period from 11:00 a.m. to 2:00 p.m.). It was not reasonably possible to measure the links to the survey frame unless the persons interviewed could easily identify in time and space the services they received during the survey period. In the case of centres offering meals, one time interval covered the noon meal and one time interval covered the evening meal. It was assumed that an individual could use only one centre during the time interval corresponding to the noon meal, otherwise it would be necessary to ask the individual if he had already received a meal somewhere else or if he had eaten twice in the same centre. It was also determined that the length of an interval ensuring use of only one service was also the length of time that an interviewer could reasonably be asked to remain on site interviewing (two to three hours maximum). (Note that daytime accommodation is not part of the services included in the survey field. This restriction of the field reflects two concerns. First, it would be very difficult to divide the day into time intervals of

three or four hours and to determine the links using this breakdown (the memory effort required of the person interviewed would be significant and did not seem reasonable to the survey's designers). Second, it is very difficult to predict the use of these services. We wanted to avoid having a team of interviewers go to a site and not be able to conduct any interviews because of lack of use.)

In actual fact, there is no fundamental difference between the sampling of the centres and the sampling of the periods of time: the relevant units to be considered are the triplets (c, j, t) that correspond to the overlap between a centre, a day and a time interval. Some of the boxes in the "time" and "centres" cross-tabulation table can be eliminated automatically prior to the selection, either because the centre is closed during the time slot considered, or because there is clearly not enough use. (In the latter case, caution must be exercised with respect to the possible restriction of the field should it be found that persons use only this centre and only attend during this time slot. If the latter are atypical, biases will be introduced into the estimations.)

The selection method used was a random selection of the triplets (centres, days, time intervals) in proportion to the size of the centres obtained during the centre census. (In practice, in order to avoid difficulties with centre officials, time intervals were grouped together when a centre was sampled more than four times during the survey period.) Centres were stratified by type. (For accommodation services, centres were stratified by the criteria of men only/women only/mixed accommodation.) However, since this "precautionary" stratification does not apply directly to the observation units, it is useful only if the behaviour of the individuals differs significantly by the type of centre in which they are found.

2.3.3 Selection of Services

This last stage of the sample design consisted in completing the sampling of services, that is, in selecting individuals in a selected centre on a given day during a given time interval. The data collected during the census of the centres were not generally enough to constitute a survey frame of services. Some accommodation centres had lists: this was the more positive scenario where persons could be selected using these lists. However, at the majority of centres (for example, a soup kitchen), it was not even known how many people would show up in a given time interval: it was therefore not possible to develop a survey frame of services. Sampling of the services was done on an equal probabilities basis. As is traditional in multiple stage surveys, selecting a constant number of services (last stage) ensures constant probabilities of selection and thereby limits the risk of expanding sampling variances.

In practice, the selection method used varies from one type of centre to the next, depending on the topography of the sites; existing list, waiting list, arrivals spaced over time, population "grouped" in no order at a single site at the same time, *etc.* It also takes into consideration the maximum number of interviews that can reasonably be done by the interviewer or interviewers during the survey's time interval, and the fact that it is not desirable to keep the sampled persons too long after the closing of a centre or after meal service has stopped because of the risk of increasing the nonresponse rate.

In all instances, a "counter" counts the number N of services provided during the sampling period. This is crucial to determining the selection probability of the sampled services. At the same time, the counter carries out a standard systematic selection (ideally, the selection should be done by another person (or "sampler") to avoid measurement errors in the use. For budget reasons, it was not possible to resolve this problem) using the following method:

- in centres where a list was available, n services were selected, n being set before the survey;
- in centres without a list, services were selected with a fixed f sampling ratio. f is determined based on the number of expected services $\tilde{N}$ and the number of services that we wanted to sample $\tilde{n}$ in order to ensure equal selection probabilities. In these cases, the size of the sample was not known in advance.

3. Parameters of Interest

The quantities of interest are essentially totals or ratios. We want to estimate a total in relation to a variable y defined for the population $P(J)$,

$$Y_J = \sum_{k \in P(J)} y_k. \qquad (1)$$

One specific example of these totals is the size of $P(J)$, $N_J = \text{card}\,(P(J)) = \sum_{k \in P(J)} 1$.

We also want to estimate the average of y in the reference population,

$$\bar{Y}_J = \frac{Y_J}{N_J} = \frac{1}{N_J} \sum_{k \in P(J)} y_k. \qquad (2)$$

For example, y can be the nationality of the individual, the age at which he completed his education, or the number of centres that he visited the day of the interview.

We then have to distinguish between two types of variables:

- variables that are fixed during the survey reference period (such as, age at time of completion of education);
- variables that vary during the survey reference period $(y_k = y_k(j))$. The number of centres visited on the day of the survey fall into this category.

We will begin with the variables that are fixed during the survey reference period. Section 6 looks briefly at those variables that change during that period.

4. Estimation of a Total or Ratio in Cases where the Variable of Interest is Constant During the Survey Period

For the convenience of the discussion, we will not present explicitly all of the selection stages. Instead, we will use as an example a population centre sampled at the first selection stage.

We note:

C: all centres in the population centre open at least one day during the survey period, denoted by index c.

$\prod_{c,\,j,\,t}$: all services provided in centre c on day j during time interval t, denoted by index i.

$\prod_{j,\,t}$: all services provided in the population centre on day j during time interval t.

$P_{c,\,j,\,t}$: all persons who visit centre c on day j during time interval t, denoted by index k.

$P_{j,\,t}$: all persons who visit a centre in the population centre on day j during time interval t.

Based on the definition of the time intervals, we find that for each individual $k \in P_{j,\,t}$, there is one and only one service i. Thus, there is a one-to-one correspondence between $P_{j,\,t}$ and $\prod_{j,\,t}$. In other words, for every couple (j, t), the $P_{c,\,j,\,t}$ are separate. On the other hand, $P_{c,\,j,\,t}$ and $P_{c^*,\,j^*,\,t^*}$ can have a non-empty intersection, when $t \neq t^*$.

The population of interest is therefore written

$$P(J) = \bigcup_{c,\,j,\,t} P_{c,\,j,\,t} = \bigcup \left(\coprod_{c \in C} P_{c,\,j,\,t} \right).$$

The central point of the reasoning consists in expressing the total of one variable of the population of *individuals* (which is our total of interest) as the total of another variable of the population *of services* (which are the sampled units), since estimation of the latter does not pose any particular problem. To obtain this result, we can use direct reasoning or apply the weight sharing method, either of which may seem more natural.

Using direct reasoning, we define the application K, which links to each service i received during reference period J in all of the centres in the survey frame the individual who received that service.

$$K: \{\text{services}\} \to \{\text{individuals}\}.$$
$$i \to K(i)$$

The population of interest $P(J)$ is represented by K of $\prod(J)$, all services provided during the reference period in all centres in the survey field. For each $k \in P(J)$, we define $r_k(J) = \text{card}(K^{-1}(k))$, the number of services provided to individual k during period J in all centres in the survey field, which we will also call the "number of links".

This gives us the fundamental equation:

$$Y_J = \sum_{k \in P(J)} y_k = \sum_{i \in \prod(J)} \frac{y_{K(i)}}{r_{K(i)}(J)}. \qquad (3)$$

Since variable y takes the same value for all services i "pointing" to individual k, such that $K(i) = k$, the right-hand side can be written

$$\sum_{k \in P(J)} \left[\sum_{i \in \prod(J);\, K(i) = k} \frac{y_k}{r_k(J)} \right] = \sum_{k \in P(J)} \frac{y_k}{r_k(J)} \left[\sum_{i \in \prod(J);\, K(i) = k} 1 \right].$$

But the quantity in the square brackets is the number of services provided to individual k during period J, or $r_k(J)$, which proves the equation.

We can then see $y_{K(i)}$ as attached to corresponding service i and write y_i in place of $y_{K(i)}$, and $r_i(J)$ in place of $r_{K(i)}(J)$. By using $z_i = y_i / r_i(J)$, $Z = \sum_{i \in \prod(J)} z_i$, we get $Z = Y_J$.

Formula (3) is none other than the weight sharing formula. The above reasoning is actually the reasoning underlying this method. (Only the expressions change; the weight sharing method describes the links between the sampled population and the population of interest by a matrix rather than an application, a single unit of the sampled population being able to "point" to several units of the population of interest.) The principle of this latter method is set out in Appendix 1.

4.1 Estimation of a Total

Let us now assume that we have a sample $s_\prod$ of services to which a set of weights is linked $(w_i)_{i \in s_\prod}$. We assume these weights are unbiased (this is the inverse of the probabilities of inclusion of services in the sample). $s_\prod$ implicitly defines a sample of individuals s_P, which is actually all of the individuals who receive the sampled services. The weight sharing formula (see Appendix 1) ensures that the estimator

$$\hat{Y}_J = \sum_{s_P} y_k \tilde{w}_k$$

is unbiased, where we write for every $k \in s_P$:

$$\tilde{w}_k = \frac{1}{r_k(J)} \sum_{s_\prod;\, K(i) = k} w_i. \qquad (4)$$

Formula (4) simply states that an individual's weight is equal to the sum of the weights of the services that were used to "catch him", divided by the number of links with the survey frame, $r_k(J)$. In this way, it is possible to *work directly on the individuals sampled*: for each individual k, we calculate the weight $\tilde{w}_k$, and we estimate the total Y_J by $\hat{Y}_J$.

Figure 1 gives a fictitious sampling example. The service universe contains 13 services, provided to 8 persons. 6 services are sampled. The sample of individuals contains 5 persons, individual number 2 having been "caught" by two different services. Using formula (4), the weights of the individuals sampled will be equal to:

$$\tilde{w}_1 = w_1, \tilde{w}_2 = \frac{1}{2}(w_2 + w_8), \tilde{w}_3 = w_{10}, \tilde{w}_6 = w_7, \tilde{w}_7 = \frac{1}{3}w_9.$$

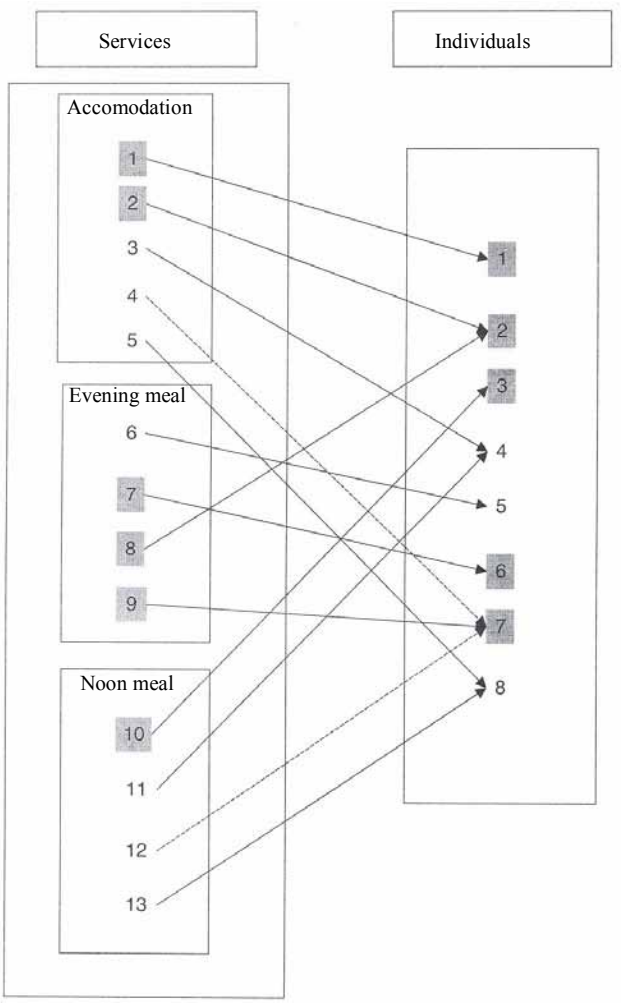

Figure 1. The arrows represent the links between the services and the individuals. The shaded services were sampled. They point to shaded individuals. Dotted lines represent the links reported by individual 7, which were not used to include the individual in the sample.

If the services all have the same weight equal to 13/6 (for example, if the services had been selected by simple random sampling), the number of persons having used services during the survey is estimated by:

$$\hat{Y}_J = \sum_{s_P} \tilde{w}_k = \frac{13}{6}\left[1 + \frac{1}{2}.2 + 1 + 1 + \frac{1}{3}\right] = \frac{169}{18} \simeq 9.39.$$

In this case where the variable being considered does not vary during the survey period, identifying the persons using the services does not affect the estimator bias. Consider an individual "caught" by two different services with weights w_1 and w_2. In practice, this could produce two cases:

- it is determined that this is the same individual; the weighting associated with this individual will be equal to $(w_1 + w_2)/r_{k(J)}$, and the expression corresponding to the individual in the estimator will be equal to $y_k\ (w_1 + w_2)/r_{k(J)}$.
- it is not determined that the individual has already been interviewed: two different individuals are counted; the weights associated with these individuals will be equal to $w_1/r_{k(J)}$ and $w_2/r_{k(J)}$, and the expression corresponding to these two pseudo-individuals in the estimator will still be equal to $y_k(w_1 + w_2)/r_{k(J)}$.

Of course, this presumes that the information provided by the same person surveyed in two different locations/on two different days is the same, which is far from given.

However, identifying individuals can be important in order to limit nonresponse (see section 7).

4.2 Estimation of a Ratio

Let us now suppose that we are interested in the estimation of the average $\overline{Y}_J$ (see Formula (2)). $\overline{Y}_J$ Can be estimated by the Hájek estimator,

$$\hat{\overline{Y}}_J = \frac{\hat{Y}_J}{\hat{N}_J}$$

where $\hat{N}_J = \sum_{k \in s_P} \tilde{w}_k$.

4.3 Variance Calculation

The variance of the estimators presented above is calculated in the classic manner provided that the reasoning is based on services. The calculation is still complex because it is a multi-stage design with unequal probabilities. To avoid underestimating the true variance, it is essential that all services be retained in cases where several sampled services point to a single individual.

4.4 Comparison with Other Estimating Methods

Having introduced "weight sharing" estimators, it is appropriate to consider an alternative estimating method where we will try to estimate directly the selection probabilities of individuals in the sample. (The weight sharing estimator is not a classic Horvitz-Thompson estimator : the weights of that estimator clearly depend on the complete service sample (see formula (4)). This method can appear more natural. However, we must make two comments:

– it is not reasonably possible to obtain the selection probabilities of physical persons without relying on the services that the individual receives, based on the information provided by the latter when visiting the various centres. Based on the previous expressions, we get:

$$\text{Prob}\,(k \in s_P) = \text{Prob}\left(\bigcup_{i \in \prod(J);\, K(i)=k} i \right).$$

The Poincaré formula enables us to express this probability from single, double, triple, *etc* probabilities of inclusion of services. Except for the single inclusion probabilities, these are complex probabilities derived as they are from selections of unequal and without replacement probabilities. We cannot therefore hope to obtain a calculable expression for $(k \in s_P)$. In contrast, the weight sharing method is very simple to apply:

– in a more structured manner, a problem comes from the fact that the selection probabilities of unsampled services are not known in advance because of the multi-stage sample. At the earlier stages, the selection probabilities depend on the previous selection. In our case, we do not know the use of the centres that are not surveyed. To obtain the selection probability of an individual, we must know the inclusion probabilities of *all* services that the individual receives. On the other hand, one of the strengths of the weight sharing method is that the weights of units obtained indirectly (in this case individuals) can only depend on the weights of units sampled directly (services). Lavallée (1995) points out this advantage of the method.

5. Estimation Difficulties and Practical Solutions in the Case of a Constant Variable

In the formulae that we have presented, knowing the links between individuals and the services universe is critical. However, these quantities are not known for several reasons:

– a theoretical reason: because the data collection is spread over time, and an individual interviewed at the start of the period cannot anticipate the services that he will use after the interview date (Note that data collection must necessarily be spread over time to ensure good coverage of the target population; synchronous collection, even if technically possible, would not capture the whole target population but only the persons using the services on that date);

– practical reasons: because the memory of the person interviewed becomes questionable after a few days, and because detection by the interviewer or the designer of the survey of the services provided in centres not belonging to the survey frame is very difficult.

In practice, it is therefore impossible to estimate without bias a total of interest over the period of the survey (one month) without making assumptions at the outset (see Section 5.3).

5.1 "Average Day" and "Average Week" estimations

This forces us to look at quantities that bring into play links over a short period, for example, a day or week. The population of persons who use the services in the survey field on a given day j is $P_J = \cup_{c,\,t} P_{c,\,j,\,t}$. Let us now introduce the following quantities that relate to day j:

$$\Theta_j = \sum_{k \in P_j} y_k$$

$$N_j = \sum_{k \in P_j} 1 = \text{card}\,(P_j).$$

If $\tau = \text{card}(J)$ is the number of days in the survey reference period, we define the following parameters of interest:

– the total of y in the population of persons who use the services in the survey field on an "average" day, as follows:

$$\Theta = \frac{1}{\tau} \sum_{j=1}^{\tau} \Theta_j. \qquad (5)$$

A specific case is the number of persons who use the services in the survey field on an "average" day, $\bar{N} = 1/\tau \sum_{j=1}^{\tau} N_j$.

In the same way, the average of y in the population of persons who use the services in the survey field on an "average" day is defined as:

$$\psi = \frac{\Theta}{\bar{N}} = \frac{\sum_{j=1}^{\tau} \Theta_j}{\sum_{j=1}^{\tau} N_j}. \qquad (6)$$

Defining totals or averages for a given week or an "average week" follows the same principle.

We can estimate these parameters by simply adapting the formulae in the previous section, noting that the $r_k(J)$ *must be replaced by the number of services in the survey field that the person sampled received on the day (or week) of the survey.*

Note that s_j is the sample of persons interviewed on day j, $r_k(j)$ the number of services in the universe received by individual k on day j only, and $s_k(j)$ the services sampled on day j that link to individual k.

$$\Theta_j \text{ will be estimated by } \hat{\Theta}_j = \sum_{k \in s_j} y_k \tilde{w}_k,$$

$$\text{where } \tilde{w}_k = \frac{1}{r_k(j)} \sum_{i \in s_k(j)} w_i.$$

Here, the weights of the individuals depend on the day j. (But *not* the weights of the services, w_i, which are set one time for all (if there are no nonresponses, this would be the inverse of the selection probabilities of services)). The following analogy is useful to convince oneself of the difference between Θ et Y_J : Consider a service window where everyone who comes must fill out a file. Y_J corresponds to an approach where the person fills out a file the first time that he arrives at the window and does not fill one out on subsequent visits; the "average day" case corresponds to an approach where everyone who arrives at the window fills out a file, regardless of whether he has come to the window on some other day or not. At the end of a week, for example, the analysis of the characteristics of the persons who filled out the files will be very different in the two cases: in the second case, *persons who come to the window often will be over-represented compared to the first case*. It is possible to formalize this approach. We refer interested readers to Ardilly and Le Blanc (1999).

5.2 Practical Estimation of the Links with the Survey Frame

Even if we restrict ourselves to estimating "average week" and "average day" quantities, it is not generally possible to determine the links with the survey frame on a given day (much less a given week or over the whole of the survey period).

5.2.1 "Average Day" Estimation

To share the weights, we must estimate the links relating to the survey day; the situation that presents the most problems is that of persons interviewed at noon in a centre that provides meals; we do not know which centres (meals and/or accommodation) these persons will use that same evening. One option not retained by the INSEE survey designers is to include in the questionnaire questions of the type "Where will you eat (or sleep) this evening?". The answers can be used to determine the links. Of course the issue is whether the answers to these questions reflect the true links and whether the nonresponse rate for the question would be too high. From a more statistical standpoint, (hypothesizing that there is a certain regularity of behaviour) we could use information relating to the same time interval on the day before the survey. The corresponding links are undoubtedly reasonable approximations of the actual links. The practical problem relates to the possible difference in use of the centres depending on the day of week: for example, some centres are not open on weekends and others are open only on specific days.

5.2.2 "Average Week" Estimation

To share the weights, we retain all the links relating to the week. Clearly, the first option described in 5.2.1 cannot be used. For a given week estimations, we can use, as an approximation of the services used on day j following the interview date, the services used by the individual on day $(j-7)$. This is consistent if we assume that there is a certain pattern to the services used depending on the day of the week. This approach would mean that the calendar week would be replaced in estimators by a sliding week, that is, the last seven days beginning on the date of the interview. This is the option that was used for the survey, the questionnaire having been designed to collect the links over the 7 days preceding the interview.

5.3 Estimation Over the Whole of the Survey Period

It may seem that estimating totals and averages for the population $P(J)$ is one of the survey's objectives. This estimation calls on the links between individuals sampled and the services in the survey field during the whole of the data collection period, which are not known. This means that we have to model the evolution of the links beyond a week or, what amounts to the same thing, model the use behaviour of the individuals in the centres.

The solution is not simple. For example, the hypothesis that comes to mind is

$$\forall k,\ r_k(J) = A.r_k(S) \qquad (7)$$

where A is the number of weeks of the survey and $r_k(S)$ is the number of links for individual k with the services of the survey field during a week S, leads to estimators for the whole of the period that are identical to the estimators for an average week. In effect, an "average week" estimator weights individual k by

$$\sum_{i \in s_k(J)} \frac{w_i}{A.r_k(S_i)}$$

where S_i is the week during which he received service i and $s_k(J)$ is the sampling of services that link to individual k, whereas a theoretical "whole period" estimator weights the individual k by

$$\sum_{i \in s_k(J)} \frac{w_i}{r_k(J)}.$$

Equation (7) is therefore an adequate condition of equality of these estimators. This condition is satisfied in particular when for any j and any k

$$r_k(J) = \text{card}(J).\, r_k(j) \qquad (8)$$

that is, when the number of daily links does not depend on j.

This hypothesis is definitely too strong. To expand on this point, we will have to use the data provided by the survey itself on the behaviour of the individuals with respect to use of the centres.

The most sought after figure of the survey – in the French context – is undoubtedly an estimate of the size of the "homeless" population, that is, an estimation of the size of $P(J)$. In addition to the issues regarding counting the links that have already been discussed extensively, this estimation runs up against several inadequacies in the survey frame as well as the indirect nature of the sampling.

- The risk of overlooking certain structures when identifying the centres is significant. Even with an exhaustive inventory, the gap between when the inventory is established and the survey itself takes place makes it likely that new unidentified centres will appear in the survey frame. This can introduce a bias to the extent that some individuals who might use these structures would not use any other service in the survey frame. (We might also expect those in charge of certain centres to refuse to cooperate: for the INSEE survey, there was virtually no refusal by the institutions (less than 1% refusal rate). This was due largely to consideration awareness building at the time the centres were identified and just before the survey.) Further, the lack of bias depends on a correct calculation of the links; use of centres not included in the frame should not be counted in these links.
- Individuals who use the centres only outside the "classic" hours (those in which we have the means to count the services) are outside the survey frame. (Counting them would create significant on-site implementation problems.)
- Another source of bias can come from the careful counting of the total number of services provided in the centres during the survey, these numbers being used to calculate the probability of a service being sampled. For budget reasons, one person only counted the services and did the sampling, a situation that could create problems of rigour in the sampling if there is confusion in the field.
- In terms of the concepts, the only remaining problem was that the survey had to take place over a month and that the target population may have changed during that period.

The estimation of the size of the population is therefore particularly fragile. For this reason, we can expect any errors to be larger for the totals than for the averages.

6. Estimation in the Case of Variables of Interest that are not Constant Over the Survey Period

Some of the survey's variables of interest depend on the observation date and therefore are not constant over the survey period. This can be the case with answers to questions dealing with the day before the interview, for example "How many meals did you have yesterday?", "How many times did you sleep in the street last week?", etc. The questions on links also fall into this category. It is therefore important to determine the extent to which we can adapt the earlier formalism to estimations involving this type of variable. In other words, where y is such a variable of interest.

If we go back to expression (3), it is easy to see that the constancy of y_k during the survey period is the condition that makes it possible to factor y_k and to reveal the links $r_k(J)$. From this we can deduce that *the above type of calculation is always valid for estimations covering shorter periods than the period for which the* y_k *are constant*.

This means that for variables that are constant for a day, we can appropriately use the "average day" estimators. For variables that are constant over the week, we can use the "average day" or "average week" estimators.

7. Adjustment for Total Nonresponse

To describe the operation fully, we still need to explain how to move from a set of inclusion probabilities (and thus initial weights of services included in the sample) to a set of weights on respondent services. Some people will agree to the interview, others will not. We will refer to services in the first case as respondent services and those in the second case as nonrespondent services. The usual adjustment methods for total nonresponse can be applied. We suggest a nonresponse adjustment by homogeneous subgroup (for a description of the method, see for example Hambaz and Legendre 1999).

In reality, the main problem relates to the fact that there is no survey frame of individuals and thus no advance information on nonrespondents. In a world that is likely very heterogeneous, this is a considerable handicap. We therefore have to model the service response behaviour. We know from the test surveys of the INED (Institut National des Etudes Démographiques) that nonresponse varies widely depending on the type of centre (Firdion and Marpsat 1997). Other variables in the survey frame can be used to build homogeneous groups (day of the week, period of the day, groups of population centres, ...).

A reweighting of the respondent services produces weights for the respondent services of the type

$w_i = 1/\delta_i\pi_i$, where

π_i is the probability of inclusion of service i in the sample

δ_i is the probability estimated after the fact that service i will result in a response.

This provides us with a set of weights for the respondent services.

In fact, some of the nonresponses come from the fact that the same individual is sampled several times: obviously, an individual who is sampled twice might respond the first time but not the second . (The frequency of occurrence of this event was not known at the time of writing this paper.) The second selection therefore produces a "false non-response". If this is not detected, the total nonresponse adjustment procedure leads to an incorrect reweighting, when the true value can be obtained from a questionnaire that has already been completed. To avoid this problem, the interviewer tries to find out the reason for the refusal and must check off a specific box when the individual states that he has already been interviewed. In this situation, the

interviewer collects some information, including the first name and the date of birth, that can be used to link this questionnaire to the questionnaire that has already been completed. (The ideal situation would be to have an identifier for the respondents. This approach was not used because of confidentiality requirements and consideration of the reaction of the persons interviewed to such a measure.) However, in the field, it can be difficult to obtain a reason for refusal. Even if a reason is given, problems can occur. (It is hard to verify that a person who states that he or she has already been interviewed has in fact been interviewed. Even if the person is showing goodwill, he may have been interviewed a few days earlier for a completely different survey than the INSEE survey.)

8. Conclusion

In this article, we show how the weight sharing method can be used to weight the survey conducted by the INSEE in order to better understand homeless persons. The method has many advantages. It makes it possible to work on a file of individuals, that is, on the natural statistical units used in the definition of the parameters of interest. Simple to apply, it also makes it easy to move from one reference period to another ("average day", "average week" estimation). Operations following to the survey, such as the nonresponse adjustment and the calculation of variance can be carried out in a traditional framework because they are done on sampled units (services), for which the selection probabilities are known, and not on individuals, for which the selection probabilities are not known. We show that a crucial quality criterion of such a survey is reliable data collection on use of services by the persons interviewed. Without these data, it is not possible to weight the survey. The weight sharing method appears to be a good compromise for a survey in which the purpose is not simply to count a population but to better understand it through the use of a questionnaire. Other alternative methodologies could be used for a survey aimed simply at determining the size of the homeless population. The first such methodology uses capture- recapture techniques to determine the size of animal populations (see for example, Pollock, Turner and Brown 1994). These techniques cannot be easily applied to a population that is often suspicious of any attempt to identify it, which they perceive negatively. Another technique is that of "snowball" sampling, which involves finding individuals of interest through the intermediary of individuals already sampled (Franck and Snijders 1994). It relies on a system of mutual knowledge of persons, who are probably illusive in the community. These methods always run up against the issue of the identifying individuals. In our case, the only places where it is possible to find the persons we are seeking are the centres: it is essential that we work through the centres.

Acknowledgements

The authors thank the journal's Editor and two anonymous referees whose comments helped improve both the content and layout of the article. Any errors that remain are entirely our responsibility.

Appendix 1: The Weight Sharing Method Applied to the Problem

This appendix briefly presents the principle of the weight sharing method. For a more complete discussion, the reader may consult Lavallée (1995) or Deville (1999) whose notations we have used.

1. We have a population U of n units, and a population V of m units. The units of U are services in the survey field. The units of V are persons who used at least one service during the survey period (otherwise expressed in the present case as $V = P(J)$ with the previous notations).
2. It is assumed that there are links between the units of the two populations. These links can be written in the form of a matrix
$$(r_{ik})\ 1 \le i \le n,$$
$$1 \le k \le m$$
where $r_{ik} = 1$ if unit k of V is linked to unit i of U, $r_{ik} = 0$ otherwise. In this case, the links connect the services to the persons who used these services: $r_{ik} = 1$ if person k used service i of U, $r_{ik} = 0$ otherwise.
3. All units of U have at least one link to a unit of V. Clearly, that is achieved here by definition of population V. Further, in this case, each unit of population U points to one and only one unit of V.

In general, we are interested in the total of a variable of interest y in V,

$$Y = \sum_{k \in V} y_k.$$

If, for example, we use $y \equiv 1$, the total of interest is the number of persons who used a service in the survey field during the month of the survey.

We can write

$$r_k = \sum_{i \in U} r_{ik}.$$

The identity $Y = \sum_{i \in U} \sum_{k \in V} (r_{ik} / r_k) y_k$ makes it possible to define for any $i \in U$ the variable $z_i = \sum_{k \in V} (r_{ik} / r_k) y_k$ which gives:

$$Z = \sum_{i \in U} z_i = \sum_{k \in V} y_k = Y.$$

Let us now assume that we have a sample s_U from the population U, which is associated with a set of weights $(w_i)_{i \in s_U}$. This sample implicitly defines a sample in V, s_V, specifically

$$s_V = \{k \in V;\ \exists i \in s_U,\ r_{ik} = 1\}.$$

We assume that we collected the r_{ik} for all $k \in s_V$, that is, that all links between individuals and the universe U are known (this point is fundamental).

The total $Z = Y$ is estimated by $\hat{Z} = \sum_{s_U} w_i z_i$.

And consequently, if the weights are unbiased (that is, set so that $\hat{Z}$ is without bias), $\hat{Y}$ estimates Y without bias.

We can rewrite $\hat{Z} = \sum_{s_U} w_i \sum_{k \in V} r_{ik}\ y_k / r_k = \hat{Y}$.

The second equation impacts only s_V by definition and therefore $\hat{Y} = \sum_{s_V} y_k (\sum_{s_U} w_i r_{ik} / r_k) = \sum_{s_V} y_k \tilde{w}_k$, where we have written for all $k \in s_V$:

$$\tilde{w}_k = \frac{1}{r_k} \sum_{s_U} w_i\, r_{ik}.$$

We can work directly on the individuals sampled. In our case, r_k is the number of links, that is, the number of services used by the person interviewed during the survey reference period. It is the quantity that is written $r_k(J)$ in the previous sections, the dependence on J being intended to remind that links affecting the weight can vary by the type of estimator ("average day", "average week") considered. This number is derived from the use data collected in the survey.

Appendix 2: Summary Table of Expressions

J	All days in the survey reference period
τ	$= \text{card}(J)$, number of days in the reference period
$P(J)$	population of interest, all persons who used at least one service in the survey field during the reference period
N_J	$= \text{card}(P(J))$, size of the population of interest
C	all centres in the population centre, denoted by index c
$\prod_{c,\ j,\ t}$	all services provided in centre c on day j during time interval t, denoted by index i
$\prod_{j,\ t}$	all services provided in the population centre on day j during time interval t
$P_{c,\ j,\ t}$	all persons who visit centre c on day j during time interval t, denoted by index k
$P_{j,\ t}$	all persons who visit one of the centres in the population centre on day j during time interval t
P_j	all persons who use services in the survey field on day j
y	variable of interest
Y_J	total of variable y in the reference population
$\bar{Y}_J$	average of y in the reference population
$\prod(J)$	all services provided during the reference period in all centres in the survey frame
$r_k(J)$	number of services provided to individual k during period J in all centres in the survey field, or "number of links"
s_Π	sample of services
w_i	weight associated with the services sample
s_P	sample of individuals, all individuals who received sampled services
$\tilde{w}_k$	weight associated with the sample of individuals
Θ_j	total of y in P_j
N_j	$= \text{card}(P_j)$
Θ	total of y "an average day"
$\bar{N}$	number of persons on "an average day"
ψ	$= \frac{\Theta}{\bar{N}}$, average of y "on an average day"
$r_k(j)$	number of services received by the individual k on day j only
s_j	sample of persons interviewed on day j
$s_k(j)$	all services sampled on day j that point to individual k
$s_k(J)$	all services sampled during period J that point to individual k

References

Ardilly, P., and Le Blanc, D. (1999). Enquête auprès des personnes sans-domicile : Éléments techniques sur l'échantillonnage et le calcul de pondérations individuelles, une application de la méthode du partage des poids. Working Paper, *INSEE* , F9903.

Chambaz, C., and Legendre, N. (1999). Calcul des pondérations dans le panel européen de ménages. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.

Deville, J.-C. (1999). Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes? suivi de : Comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistique*, *INSEE Méthodes*, 84-86.

Firdion, J.M., and Marpsat, M. (1997). Comptes rendus du groupe « pondérations » de l'enquête auprès des personnes sans-domicile, mimeo.

Franck, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

Pollock, K.H., Turner, S.C. and Brown, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20, 117 -124.

RTI (1993). Prevalence of drug use in the Washington DC metropolitan area, homeless and transient population: 1991. *Technical report*, 2.