

CHAPTER 1

INTRODUCTION

Sample surveys today make up a varied and often indispensable source of information. Whether at the level of governments, company managers, sociologists, economists, or ordinary citizens, surveys allow the informational needs necessary in taking a decision to be met. For example, to establish their policies concerning certain economic sectors, governments must have a picture of the situation before taking decisions concerning these sectors.

1.1 REVIEW OF SAMPLING THEORY AND WEIGHTING

Sample surveys are carried out by selecting samples of persons, businesses or other items (called *units*) that we survey in order to get the desired information. Sample selection is often done by randomly selecting certain units from a list that we call a *sampling frame*. This list, or sampling frame, is supposed to represent the set of units for which we are looking to produce information; this is what makes up the *target population*. The sample size can be determined prior to the selection (*fixed size sampling*) or at the time of the sampling itself (*random size sampling*). In this book, we will restrict ourselves to fixed size sampling which is, in practice, the most widespread.

Strictly speaking, fixed size sampling is described as follows. Consider $\mathbf{Y}_U = (y_1, \dots, y_N)$, the vector containing the values y_k for a population U of size N . For a survey on tobacco use, for example, the variable of interest y_k of \mathbf{Y}_U can be the number of cigarettes smoked by individual k during a given day. In general, we want to know the value for instance of the total $Y = \sum_{k=1}^N y_k$, or otherwise the mean $\bar{Y} = Y / N$. If the

size N of population U is known, the problem in determining the total Y or the mean \bar{Y} is the same. Going back to the previous example on tobacco, the total Y represents the total number of cigarettes smoked during the day, while the mean \bar{Y} represents the average number of cigarettes smoked by an individual.

To estimate the total Y (or the mean \bar{Y}) of population U , we select a sample s of size n . A *sampling design* \mathbf{p} is a function $\mathbf{p}(s)$ of the set Ξ of all samples s selected from U such that $\mathbf{p}(s) \geq 0$ and $\sum_{s \in \Xi} \mathbf{p}(s) = 1$. The function $\mathbf{p}(s)$ is in fact the probability of selecting sample s among all samples of Ξ . We assume that $\mathbf{p}(s)$ is known for the set Ξ ; this is what we call *probability sampling*. A well-known sampling design is *simple random sampling* (without replacement) where all possible samples of Ξ have the same chance of being selected. We have in fact $\mathbf{p}(s) = n!(N-n)!/N!$. By dividing the population U into subpopulations U_h called *strata*, where $U = \bigcup_h U_h$, we define *stratified simple random sampling* that consists of selecting a simple random sample in each of the h strata.

We define the *selection probability* (or *inclusion probability*) of unit k from population U by

$$\pi_k = \sum_{s \ni k} \mathbf{p}(s), \quad (1.1)$$

where the sum of (1.1) is carried out over all the samples of s from the set Ξ that contains unit k . We assume that $\pi_k > 0$ for all units k of population U , i.e., all units have a non-zero chance of being selected. For example, with simple random sampling, we get $\pi_k = n/N$, for $k=1, \dots, N$.

For each unit k of s , we measure the value of the variable of interest y_k . We can estimate the total Y with the following *Horvitz-Thompson estimator*:

$$\hat{Y}^{HT} = \sum_{k=1}^n \frac{y_k}{\pi_k}, \quad (1.2)$$

where the sum of (1.2) is carried out over all units k of sample s (Horvitz and Thompson, 1952).¹ We can show that the estimator \hat{Y}^{HT} is

¹ In this book, the sums will be based on a re-indexing of units. For example, for a sum over the population of size N and another over the sample of size n selected

unbiased for Y with respect to the sampling design, i.e., that if \hat{Y}_s^{HT} represents the value of \hat{Y}^{HT} obtained for sample s , we have:

$$E(\hat{Y}^{HT}) = \sum_{s \in \Xi} \mathbf{p}(s) \hat{Y}_s^{HT} = Y. \quad (1.3)$$

The mean of the values of \hat{Y}^{HT} weighted by the selection probability of sample s then corresponds to the true value of the total Y .

Consider t_k , an indicator variable where $t_k = 1$ if $k \in s$, and 0 otherwise. With this variable, we can rewrite the estimator \hat{Y}^{HT} under the form

$$\hat{Y}^{HT} = \sum_{k=1}^N \frac{t_k}{\pi_k} y_k. \quad (1.4)$$

Moreover, we note that

$$E(t_k) = 1 \times P(k \in s) + 0 \times P(k \notin s) = P(k \in s) = \pi_k. \quad (1.5)$$

From (1.4) and (1.5), we can prove the unbiasedness of the Horvitz-Thompson estimator in the following way:

$$\begin{aligned} E(\hat{Y}^{HT}) &= E\left(\sum_{k=1}^N \frac{t_k}{\pi_k} y_k\right) = \sum_{k=1}^N \frac{E(t_k)}{\pi_k} y_k \\ &= \sum_{k=1}^N \frac{\pi_k}{\pi_k} y_k = \sum_{k=1}^N y_k = Y. \end{aligned} \quad (1.6)$$

The formula for the *variance* of the estimator \hat{Y}^{HT} , with respect to the sampling design, is given by

$$Var(\hat{Y}^{HT}) = \sum_{k=1}^N \sum_{k'=1}^N \frac{(\pi_{kk'} - \pi_k \pi_{k'})}{\pi_k \pi_{k'}} y_k y_{k'} \quad (1.7a)$$

or, in an equivalent manner, by

$$Var(\hat{Y}^{HT}) = -\frac{1}{2} \sum_{k=1}^N \sum_{k'=1}^N (\pi_{kk'} - \pi_k \pi_{k'}) \left(\frac{y_k}{\pi_k} - \frac{y_{k'}}{\pi_{k'}} \right)^2 \quad (1.7b)$$

from the population, we will respectively use $\sum_{i=1}^N$ and $\sum_{i=1}^n$. This notation has been used in several books on sampling theory such as, among others, Cochran (1977) and Morin (1993).

where $\pi_{kk'}$ represents the joint selection probability of units k and k' . For the details in the proofs of (1.7a) and (1.7b), we can consult Särndal, Swensson and Wretman (1992).

We can also write the estimator \hat{Y}^{HT} given by (1.2) as a function of the *sampling weight* $d_k = 1/\pi_k$. We then have

$$\hat{Y}^{HT} = \sum_{k=1}^n d_k y_k . \quad (1.8)$$

In sampling theory, the sampling weight is the inverse of the selection probability π_k of unit k from sample s . The sampling weight of unit k corresponds to the expected number of units from population U represented by this unit. For example, if an individual has one chance out of four ($\pi_k = 1/4$) of being part of the sample, it will have a sampling weight of 4; we then say that this individual in the sample represents on average four individuals within the population. Let us note that the sampling weight d_k may possibly not be an integer.

It is possible to define in a general way an *estimation weight* w_k that we associate to unit k of sample s . This weight leads to the estimator

$$\hat{Y} = \sum_{k=1}^n w_k y_k . \quad (1.9)$$

The properties (bias and variance, for example) of this estimator depend upon the construction of the estimation weight w_k . In this book, we will focus on an estimation weight obtained by the generalisation of a method called weight share.

To learn more about sampling theory, the reader can consult books such as Cochran (1977), Grosbras (1986), Särndal, Swensson and Wretman (1992), Morin (1993), Ardilly (2006), and Lohr (1999).

1.2 CLUSTER SAMPLING

It often happens that sample surveys are performed in clusters. *Cluster sampling* is in fact a sampling design commonly used in practice. This technique of sampling is not suitable for the drawing of samples of units, but rather the selection of groups of units called *clusters* [or *primary sampling units* (PSU)]. The units in the clusters are called *secondary sampling units* (SSU). In cluster sampling, we survey for all the SSU belonging to the selected PSU. When we

survey only for a subsample of the SSU, within the selected PSU, we are instead speaking of *two-stage sampling*.

For social studies, several surveys are built in such a way that we sample households in order to survey for the set of individuals from these households. The households thus form clusters of individuals. This is particularly the case for the Labour Force Survey conducted by Statistics Canada (Singh *et al.*, 1990). For economic surveys, the sampling of enterprises is often done with the goal of obtaining information on their components, for instance, the establishments or the local units. Enterprises are therefore composed of clusters of establishments, or local units, which we survey in order to provide economic statistics, in particular for national accounts.

With cluster sampling, the survey statistician can hope for reductions in collection costs. Indeed, surveying for entire households, for example, allows the interviewer to considerably reduce his number of trips compared to sampling for the same number of persons, but in different households. Cluster sampling also allows for the production of results at the cluster level itself, on top of the units. For example, we can calculate the average income of the households.

Cluster sampling is presented in most books that deal with sampling theory. We assume that the population U consists of N clusters where each cluster i contains M_i units. This is illustrated in Figure 1.1. We select a sample s containing n clusters in population U according to a certain sampling design. We assume that π_i represents

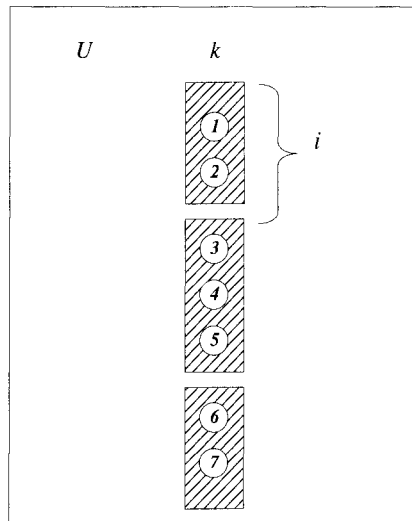


Figure 1.1: Cluster sampling

the selection probability of cluster i , where $\pi_i > 0$ for all clusters $i \in U$. As each cluster i of population U contains M_i units, we have in total $M = \sum_{i=1}^N M_i$ units in the population. We survey all units of clusters i for sample s . Each unit k of cluster i therefore has the same selection probability as the cluster, i.e., $\pi_{ik} = \pi_i$.

With cluster sampling, we are looking to estimate the total $Y = \sum_{i=1}^N \sum_{k=1}^{M_i} y_{ik}$ for a characteristic y . Considering the Horvitz-Thompson estimator (1.2), we can use the estimator $\hat{Y}^{CLUS,HT}$ given by

$$\hat{Y}^{CLUS,HT} = \sum_{i=1}^n \frac{Y_i}{\pi_i} \quad (1.10)$$

where $Y_i = \sum_{k=1}^{M_i} y_{ik}$. The superscript *CLUS* refers to the term *cluster sampling*. The variance of $\hat{Y}^{CLUS,HT}$ is given by

$$\text{Var}(\hat{Y}^{CLUS,HT}) = \sum_{i=1}^N \sum_{i'=1}^N \frac{(\pi_{i'} - \pi_i \pi_{i'})}{\pi_i \pi_{i'}} Y_i Y_{i'}. \quad (1.11)$$

We can rewrite estimator (1.10) in the following manner:

$$\begin{aligned} \hat{Y}^{CLUS,HT} &= \sum_{i=1}^n \frac{1}{\pi_i} \sum_{k=1}^{M_i} y_{ik} \\ &= \sum_{i=1}^n \sum_{k=1}^{M_i} \frac{y_{ik}}{\pi_{ik}} = \sum_{i=1}^n \sum_{k=1}^{M_i} d_{ik} y_{ik} \end{aligned} \quad (1.12)$$

where $d_{ik} = 1/\pi_{ik}$.

Estimator (1.10) can then be written as a function of units k for clusters i of sample s with sampling weight d_{ik} . In a general way, we can construct an estimation weight w_{ik}^{CLUS} and define an estimator of the form

$$\hat{Y}^{CLUS} = \sum_{i=1}^n \sum_{k=1}^{M_i} w_{ik}^{CLUS} y_{ik}. \quad (1.13)$$

The properties of this estimator depend upon the construction of the estimation weight w_{ik}^{CLUS} .

1.3 INDIRECT SAMPLING

To select in a probabilistic way the necessary samples for social or economic surveys, it is useful to have available sampling frames, i.e., lists of units meant to represent the target populations. Unfortunately, it may happen that no available sampling frame corresponds directly to the desired target population. We can then choose a sampling frame that is indirectly related to this target population. We can thus speak of two populations U^A and U^B that are related to one another. We wish to produce an estimate for U^B but unfortunately, we only have a sampling frame for U^A . We can then imagine the selection of a sample from U^A and produce an estimate for U^B using the existing links between the two populations. This is what we can refer to as *indirect sampling*.

For example, consider the situation where the estimate is concerned with young children (units) belonging to families (clusters) but the only sampling frame we have is a list of parents' names. The target population is that of the children, but we must first select a sample of parents before we can select the sample of children. Note that the children of a particular family can be selected through the father or the mother.

This is illustrated by Figure 1.2. In this example, the families are represented by the rectangles and we note that the children can come from different unions.

Another example of an application of indirect sampling is the situation where we wish to conduct a survey of enterprises (clusters) when we only have an incomplete sampling frame of establishments of these enterprises. For each establishment selected from the sampling frame, we want to sample the set of establishments (units) belonging to the same enterprise. The establishments that are not represented in the frame must be represented by those that are part of this frame (Lavallée, 1998b).

This example can be represented by Figure 1.3. Here we see that establishments **a**, **b**, **c**, **d**, and **e** are part of the sampling frame whereas establishments **f** and **g** are not part of it.

A third example is one where we are looking to conduct a survey on people (units) who live in dwellings (clusters). We have for this case a sampling frame of dwellings, but which is unfortunately not up-to-date. This sampling frame does not contain, among others, renovations affecting the division of buildings into apartments.

An example of this type of renovation is illustrated in Figure 1.4a. We note that dwellings **a**, **b**, **c**, **d**, and **e** have been transformed to get dwellings **a'**, **b'**, **c'**, and **d'**. By selecting a sample of dwellings from the sampling frame, we then go to new dwellings using the correspondence between the old and new dwellings. This correspondence is illustrated in Figure 1.4b.

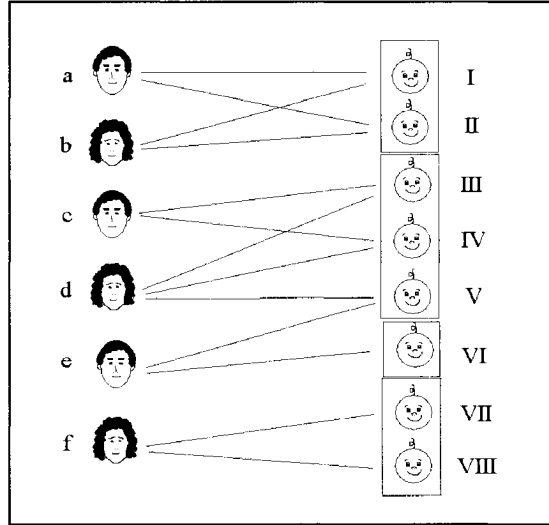


Figure 1.2: *Indirect sampling of children*

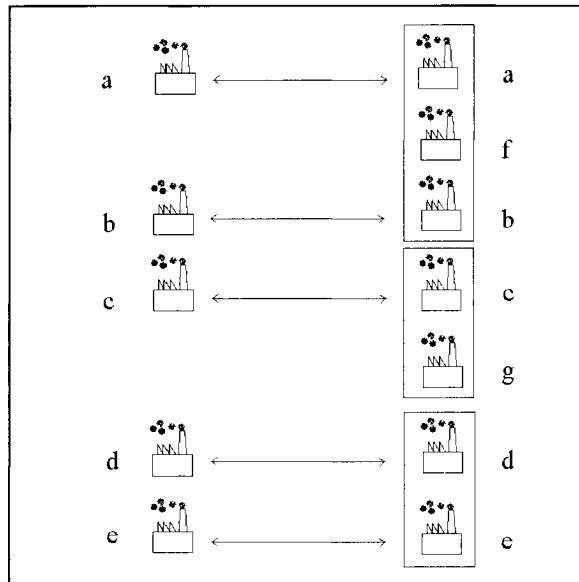


Figure 1.3: *Indirect sampling of*

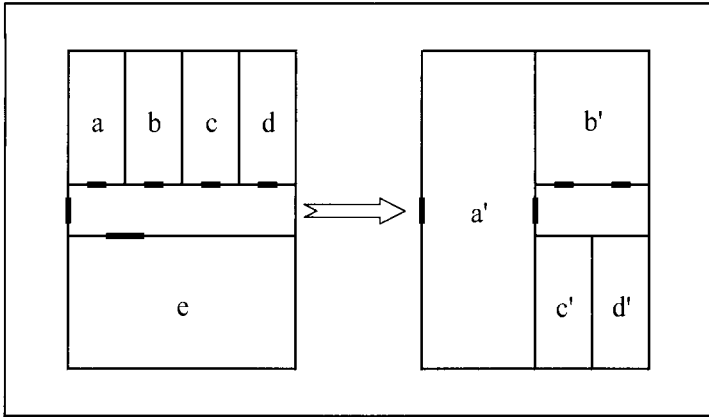


Figure 1.4a: *Indirect sampling of dwellings*

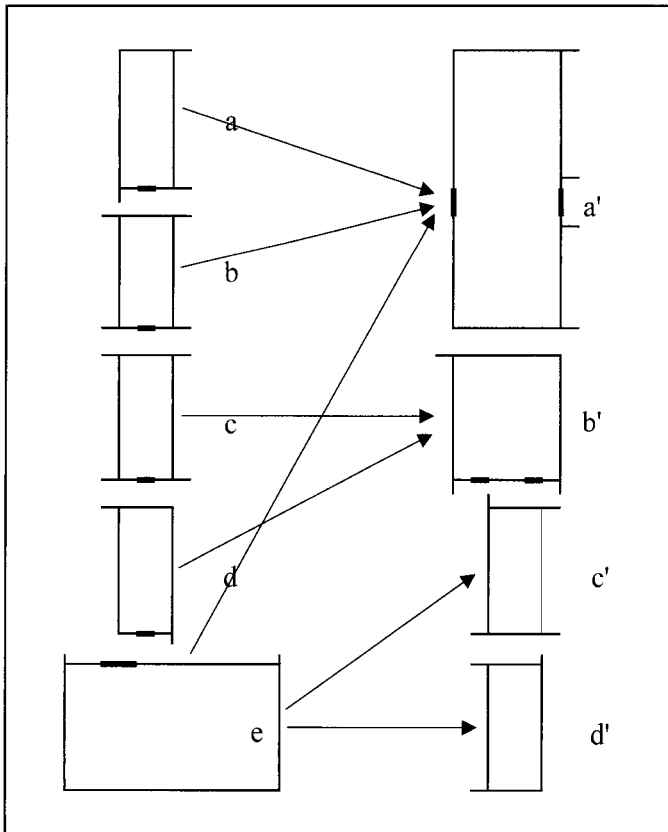


Figure 1.4b: *Indirect sampling of dwellings*

1.4 GENERALISED WEIGHT SHARE METHOD

The estimation of a total (or a mean) of a target population U^B of clusters using a sample selected from another population U^A that is related in a certain manner to the first can be a major challenge, in particular if the links between the units of the two populations are not one-to-one. The problem comes especially from the difficulty of associating a selection probability, or an estimation weight, to the surveyed units in the target population.

If we consider the example of families in Figure 1.2, it can be very difficult to associate a selection probability to each child of a selected family (or cluster). Indeed, we could have selected a family through one or more of the parents but, unfortunately, to know the selection probability of the family, and consequently of the children, we must know the selection probability of each parent, whether selected or not. In practice, this is not always the case, particularly if we used, for the selection of parents, a multi-stage design. In the example of selecting enterprises (or clusters of establishments) from the establishments (Figure 1.3), the problem is above all to associate an estimation weight to the new establishments (**f** and **g**) of the target population. In order to solve this type of estimation problem, we developed the *generalised weight share method* (GWSM).

The GWSM produces an estimation weight for each surveyed unit from the target population U^B . This estimation weight basically constitutes an average of the sampling weights of the population U^A from which the sample is selected. Lavallée (1995) presented for the first time the GWSM within the context of the problem of cross-sectional weighting for longitudinal household surveys. The GWSM is a generalisation of the weight share method described by Ernst (1989). We can also consider the GWSM as a generalisation of network sampling as well as adaptive cluster sampling. These two sampling methods are described by Thompson (1992) and by Thompson and Seber (1996).

This book is meant to be a detailed document on the GWSM encompassing the different developments carried out by the author on this method. The theory dealing with the GWSM is presented, in addition to different possible applications that bring out the appeal of this. In Chapter 2, we present a formal description of the GWSM and we describe its use. In Chapter 3, we give a literature review where we associate the GWSM with different sampling methods appearing in literature. We will see that the GWSM is a generalisation of methods

such as the fair share method and adaptive cluster sampling. In Chapter 4, we present theoretical results on the GWSM, for instance the unbiasedness of the method and the variance of estimates resulting from it. In Chapter 5, we examine other possible generalisations of the GWSM. For example, we describe how to extend indirect sampling from one stage to two stages. In Chapter 6, we look at one of the main applications of the GWSM, being that related to longitudinal surveys. In Chapter 7, we describe how we can try to improve the precision of estimates coming from the GWSM by using calibration. In Chapter 8, we deal with the practical case where non-response occurs during data collection. We see that we can correct the weights coming from the GWSM by calculating a response probability associated with the responding units. In Chapter 9, we discuss the case where the links between populations U^A and U^B were established from a process of probabilistic linkage. We then see that it is possible to modify the GWSM in order to adapt it to the situation where the links between the two populations are not deterministic. Finally, we end the book with a conclusion that emphasises new applications of the indirect sampling.