The Weight Sharing Method

Cyril Favre-Martinoz & Emmanuel Gros

***Résumé* — The aim of this methodological note is to provide a brief description of weight sharing and the contexts in which the method is applied. The position of weight sharing in relation to other post-collection processing methods is also discussed.**

I. Theoretical Framework and Contexts for Application

During statistical surveys, we are sometimes confronted with situations in which the observation unit either differs from the sampling unit or can be surveyed by drawing different sampling units. This is the framework for any survey for which the sample of desired final units results [1] rom the selection of one or more samples of intermediate units linked to the final units.

In such situations, the weight sharing method is the procedure used to calculate reference weightings and results in an unbiased estimator, under the sole condition that any final unit is linked to at least one intermediate unit.

In practice, the weight sharing method is mostly used in three specific contexts :

- lwhere the sample of final units was selected by means of ***indirect sampling (cf. II-A)*** : this is the most natural application framework for the method, which was developed specifically with this context in mind ;
- where there are ***multiple sampling frames (cf. II-B)***, *i.e.* where the sample of final units results in the concatenation of several samples selected from several sampling frames that are joined to one another ;
- during the use of ***samples that are fully or partially panellised (cf. II-C)*** :cross-sectional use of a panel, cross-sectional or longitudinal use of a rotating sample.

II. Description of the Method

A. Indirect Sampling

In order to illustrate the indirect sampling method, we will look at the classic "parents-children" example. We want to produce estimates for a population of children (population of interest), knowing that only a sampling frame of parents is available. We therefore select a sample of parents using a probability sampling method and then survey all of the children of the parents surveyed. This situation is illustrated in Figure 1. The lines between the two bases represent parent-child relationships.

1. This is the result of either not having a sampling frame from which a sample of final units can be directly selected or of a complex sampling process (panelling, for example).

More generally, indirect sampling consists of selecting a sample s^A from within a population U^A of size N^A in order to produce an estimate for a target population U^B of size N^B, basing this on the links that exist between the two populations. All of the units sampled indirectly within the population U^B that have at least one link to one of the units sampled are elements of the set Ω^B. In order to estimate the total Y^B based on measured values y_i on the basis of the set Ω^B, it is standard practice to use an estimator in the form of :

$$\hat{Y}^B = \sum_{i \in \Omega^B} w_i y_i$$

where w_i is the estimation weight for the unit i of Ω^B. A traditional way of defining a set of weights producing an unbiased estimate is to choose the weight as the opposite of the probability of inclusion. Unfortunately, in the case of indirect sampling, it is often very complicated or even impossible to determine the probabilities of inclusion of the units belonging to the sample Ω^B.Generally, only the sampling weight $d_j = 1/\pi_j$ of the unit j belonging to the sample s^A is available, defined as the opposite of the probability of inclusion π_j. In order to produce an unbiased estimator, we must therefore turn to weight sharing by defining a system of links L_{ij} between two units i and j belonging to populations U^B and U^A respectively. So, if there is a link between the unit i and the unit j, L_{ij} will be equal to 1, otherwise it will be 0. The estimator resulting from this method is then written as follows :

$$\hat{Y}^B = \sum_{i \in \Omega^B} w_i y_i$$

where $w_i = \sum_{j \in s^A} d_j \frac{L_{ij}}{L_i}$ and $L_i = \sum_{j=1}^{N^A} L_{ij}$.

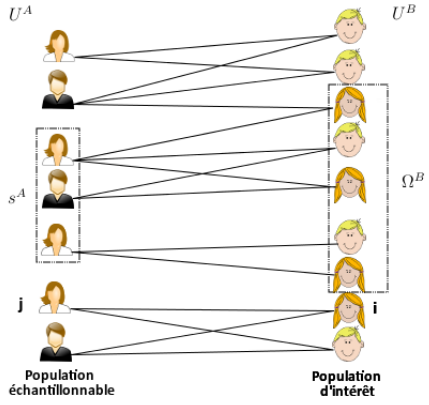

Fig. 1. Populations of parents and children with links between the two

Note that the total number of links L_i corresponds to the total number of links within the unit i i with the initial sampling frame. This total number of links L_i allows the

weight associated with the unit i to be adjusted : the higher the number of links L_i, the more likely the unit is to be selected ; it is therefore logical that its weight resulting from the weight sharing method w_i will reduce. It is important to note that this number of links L_i is counted across the entire population U^A (and not just within the sample s^A). Therefore, in the specific case of the indirect selection of children via their parents, it is necessary for each child belonging to the sample Ω^B to be able to indicate which members of the entire population U^A are their parents.

Finally, even though the weights resulting from the weight sharing method guarantee an unbiased estimate, they are not necessarily optimal in terms of variance. Results concerning the optimality of estimators resulting from the weight sharing method can be found in the article written by Deville and Lavallée (2006)

B. Multiple Bases

In order to compensate for a possible lack of coverage, it is standard practice to draw several samples from several sampling frames, the intersection of which is not necessarily empty. Some units falling within the scope of the survey may therefore be selected with a non-zero probability in each of the frames. The sampling weights used for the estimation must take account of this peculiarity. The units present within the various sampling frames therefore constitute an intermediate population $U^A$of size N^A which enables the whole of the population of interest U^B to be covered. This is the simple case summarised in Figure 2, where the sample results from two independent samples drawn from frames 1 and 2. One random sample s_1 of size n_1 is drawn from among the N_1 units within sampling frame 1, and one independent random sample s_2 of size n_2 is drawn from the N_2 units within sampling frame 2. The unit j within the sample s_1 has a weight $w_{j,1}$ orresponding to the opposite of the probability of inclusion of the unit j in frame 1. Similarly, the unit k within the sample s_2 has a weight $w_{k,2}$ equal to the inverse of the probability of inclusion in frame 2. If we were to take a naïve estimator $\hat{Y}^{HT} = \sum_{j\in s_1} w_{j,1} y_j + \sum_{k\in s_2} w_{k,2} y_k$ for the total Y this would overestimate the total Y due to the "double counts" resulting from the units positioned at the intersection of these two frames. Similarly to case (II-A), by designating the sample obtained by merging the two samples as Ω^B, and by removing the duplicates, it is possible to create an unbiased estimator by means of weight sharing as follows :

$$\hat{Y}^B = \sum_{i\in\Omega^B} \left(\sum_{j\in s_1\cup s_2} d_j \frac{L_{ij}}{L_i} \right) y_i \qquad (1)$$

where $d_j = w_{j,1} I_{j\in s_1} + w_{j,2} I_{j\in s_2}$ and $L_i = \sum_{j\in U_1\cup U_2} L_{ij}$.

We can relate this back to the previous "parents and children" case by considering the unit in frame 1 to be equivalent to the father and the unit in frame 2 to be equivalent to the mother. In this case, L_i corresponds to the number of frames in which the unit i could have been sampled. If we refer to Figure 2, the weight of the final sampling of the units sampled within the two sampling frames is equal to the sum of the weights of the unit in question within each frame, divided by two. The sampling weight of units belonging to only one of the two frames remains unchanged.

In cases involving multiple sampling frames, the application of weight sharing provides an unbiased estimator of the total ; however, that estimator is not necessarily optimal in terms of precision. More specifically, the estimator (1) belongs to a larger class of unbiased estimators taking the form :

$$\sum_{j\in s_1\cap\bar{U}_2} w_{j,1} y_j + \sum_{k\in s_2\cap\bar{U}_1} w_{k,2} y_k + \sum_{j\in U_1\cap U_2} \left[\Theta w_{j,1} I_{j\in s_1} + (1-\Theta) w_{j,2} I_{j\in s_2}\right] y_j$$

The optimal choice of the parameter Θ was examined in particular by Hartley (1962, 1974). The estimator (1) resulting from the weight sharing method corresponds to the choice of $\Theta = 1/2$. In practice, during the household surveys performed by INSEE, since the variance of the estimates is inversely proportional to the size of the samples, a parameter[2] $\Theta = \frac{n_1}{n_1+n_2}$ is chosen in order to limit the dispersion of the weights.

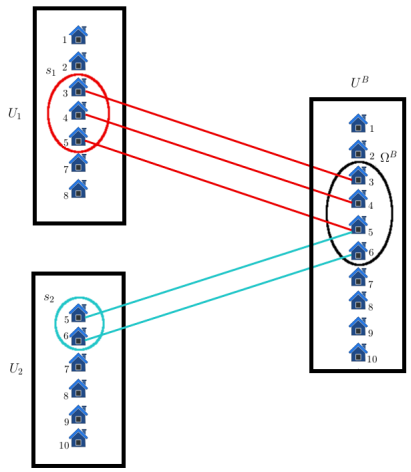

Fig. 2. Estimation in the presence of multiple sampling frames

C. Surveys Repeated Over Time and Weight Sharing

α. Cross-sectional use of a panel

A panel is a sample within which the units are surveyed at least twice over a given period : the sample is selected from the population on the initial date, then the units within this sample are followed for as long as is necessary for the purposes of the study. A panel is therefore a sample that represents the population on the date on which it was drawn. It therefore forms a fundamental part of a *longitudinal* approach consisting of measuring the evolution of a parameter over time.

At first glance, the use of a panel therefore seems incompatible with a *cross-sectional* approach that aims to estimate a parameter on the date on which the survey is conducted, since the panel does not represent the current population, but the population on the date on which the

2. In this expression, the size of samples n_1 and n_2 can sometimes be replaced by the number of respondents in each frame. This is the case in particular when the response rates for the two frames used differ from one another (for example where different collection modes are used for the two samples). Please note that the extension of the expression given for the parameter Θ cannot simply be extended to other sampling frames.

panel was drawn [3], and therefore does not cover units that entered the scope of the survey in question between the date on which the panel was drawn and the current date. Nevertheless, where there is a "natural" concept of grouping panel units, it is possible to obtain a cross-sectional sample through the clever use of indirect sampling.

If we take the case of a panel of individuals : these individuals are naturally grouped together in dwellings. Using the initial sample of "panel individuals", we will create a cross-sectional sample by surveying, on the current date, all individuals present in the dwellings containing at least one panel individual. This method of establishing the cross-sectional sample by means of indirect sampling will allow births to be taken into account and will therefore cover the population on the current date [4].The weight of the individuals within this cross-sectional sample will then be determined via the weight sharing method. Each individual i living in the same dwelling ℓ will therefore be assigned the same weight, $w_{i\ell}$, calculated as the sum of the sampling weights $d_{k\ell}$ of the panel individuals k residing within dwelling ℓ divided by the total number L_ℓ of individuals within dwelling ℓ on the current date that were able to be surveyed on the initial date, within the panel s_0, i.e. :

$$w_{i\ell} = \frac{1}{L_\ell} \times \sum_{k \in s_0, k \in \ell} d_{k\ell}$$

β. Longitudinal and cross-sectional uses of a rotating sample

As we have already seen, a panel primarily responds to a longitudinal approach, which aims to measure the evolution of a parameter over time. Although the indirect sampling method mentioned previously allows for cross-sectional use on the basis of a pure panel when associated with the selection of an additional sample, this method is simply a stopgap and can be complex to implement, particularly with regard to the selection of the additional sample.

Since our aim is to reconcile the objectives of cross-sectional and longitudinal use, we will therefore favour the use of a rotating sample. A rotating sample is a sample that brings together panels drawn on different dates ; it has a constant and limited lifespan, since the system was designed in such a way that one panel enters the sample and one panel leaves the sample during each survey campaign. The diagram in Figure 3 (inspired by those included in [4] in chapter IV.3.3) provides a summary of the situation for a rotating sample renewed by one quarter.

This rotating sample can be used for both longitudinal and cross-sectional purposes :

- in order to estimate the evolution of a parameter between two dates – in this case, for example, between t+2 and t+3 –, we will draw upon the longitudinal sample – made up, in this case, of the three shaded panels in Figure 3 which are the only ones that were surveyed in both t+2 and t+3 ;
- in order to estimate a parameter on a given survey date – for example in t+3 in this case –, we will draw upon the cross-sectional sample – in this case established by combining the four grey panels in Figure 3, all four of which were surveyed in t+3.

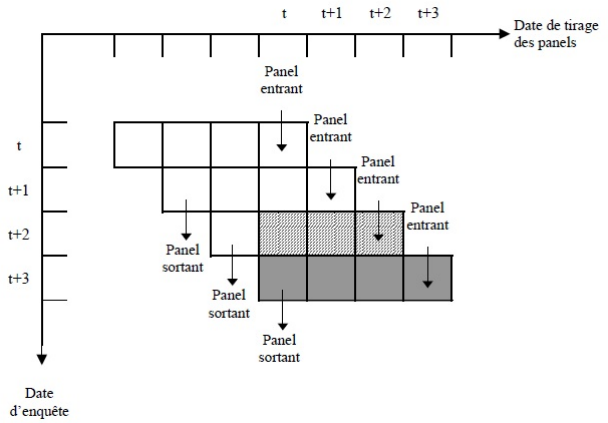

Fig. 3. Rotating sample renewed by one quarter

Therefore, in this configuration, the longitudinal and cross-sectional samples are both established by combining different panels, with each panel representing the population on the date on which it was drawn. The weights associated with these longitudinal and cross-sectional samples will once again be determined by means of weight sharing [5] :

- for the longitudinal sample, the weight of an individual will be the same as its sampling weight within the panel **via which they were selected** divided by the number of panels from which the individual could have been sampled ;
- or the cross-sectional sample, an initial weight sharing procedure must be performed in accordance with the method described in II-C-α, on a panel by panel basis. We then perform weight sharing for a second time, which consists of dividing the weight assigned to each individual during the first weight sharing procedure by the number of panels via which the household that they reside in could have been sampled.

III. Determining the Links

Determining the links is a crucial element of the weight sharing method. Indeed, the quality of the method, and in particular its unbiased nature, is dependent on the links between the sampling units and the observation units being correctly evaluated. Furthermore, it is essential that each of these units within the frame of interest have at least one link with the units that are able to be sampled in order to

3. In general, minus the units that are known to have exited the scope of the survey between the date on which the panel was originally drawn and the current date of the survey.

4. In practice, this approach still results in a gap in coverage, as the individuals living in dwellings in which there may not be a panel individual – immigrants living in a dwelling that only houses immigrants, for example – are not included in the survey. This residual gap in coverage can be addressed by selecting an additional sample drawn directly from the current population.

5. The results are presented here in a slightly simplified context, where the probability (very small in practice) of an individual being selected in more than one panel is not taken into account. Reference is made to Chapter IV.3.3 of [4] for details of the calculations and general formulae.

guarantee the unbiased nature of the estimators resulting from weight sharing.

The links can be determined in a number of different ways depending on the context in which the weight sharing method is used. In the classic case of indirect sampling (cf. II-A) and in the case of panellised samples (cf. II-C), a specific question is added to the questionnaire. For example, in the case of panels, the selected individual is asked whether they were included in the scope of the survey on the dates on which the panellised samples were drawn. Where multiple sampling frames are present, it is sometimes possible to perform matching between the frames. This will make it possible to identify which sampling frame(s) the units selected indirectly belong to.

IV. Post-collection Processing

This section merely provides an outline of the post-collection processing applied to samples on which weight sharing has been performed.

A. Weight Sharing and Non-response Adjustment

We distinguish between two, fundamentally different types of non-response in the case of weight sharing (in addition to the classic partial non-response) :

- the total non-response of the unit : this is generally handled upstream of the weight sharing process. A re-weighting procedure is first performed on the sample(s) before sharing the weight based on the respondents within the sample(s) with their corrected non-response weights. The processing of the total non-response of a unit for a panel or a rotating sample is more complex and is described in [4]in Chapter IV.3 ;
- link non-response : this is a partial non-response relating to the variable(s) within the questionnaire that enable the links to a responding unit to be determined. There are several methods, described in [5], that allow this thorny issue, which only affects samples involving weight sharing, to be handled. For example, each link variable can be modelled, for the sample of respondents, based on auxiliary variables using logistic regression and that logistic regression model can then be applied to the non-respondents in order to assign the missing links.

B. Weight Sharing and Margin Calibration

The interaction between weight sharing and margin calibration, and in particular the order in which operations are performed, will depend on the auxiliary information that is available :

- if the margins relate exclusively or predominately to the target population of the final units, the margins will be calibrated after weight sharing has taken place, based on the sample of final units that responded, following correction for non-response and weight sharing ;
- if the margins relate exclusively or predominately to the population(s) of intermediate units, the margins will be calibrated before weight sharing takes place, based on the sample(s) of intermediate units that responded, following correction for non-response ;
- if we have margins relating to the populations of both intermediate and final units, it is possible *via* an *ad hoc* modification of the calibration variables for the final units, to perform calibration only on the sample of intermediate units. In this case, we will therefore perform this specific calibration prior to weight sharing, based on the sample(s) of intermediate units that responded, following correction for non-response. This calibration procedure, which is more general, but still more complex than its predecessors, is described in [6] in paragraph 7.2.

Here, too, the issue of margin calibration in the case of a panel or a rotating sample is specific and more complex and also described in [4] in chapter IV.3.

References

[1] Deville, J.-C., Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, Vol. 32, No 2, p. 185.

[2] Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association. pp. 203-206.

[3] Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhya*, Vol. 36, No 997, p. 118.

[4] Ardilly, P. (2006). *Les techniques de sondages*. Éditions Technip, Paris.

[5] Xiaojian X., Lavallée, P. (2009). Traitements de la non-réponse de lien dans l'échantillonnage indirect. *Techniques d'enquête*, Vol. 35, No 2, pp. 165-177.

[6] Lavallée, P. (2007). *Indirect sampling*. Springer, 2007, New York.

Département des méthodes statistiques
Version n° 1, diffusée le 10 octobre 2017