Additional samples with overlapping and balancing conditions: theoretical aspects and application to students' assessment data

Marc Christine[1] ,Thierry Rocher[2]
[1]INSEE, Paris, France
[2]DEPP, Paris, France

This paper provides a theoretical frame and methods to solve a problem which occurs as soon as a first sample has been drawn at a given time and that one intends later to draw a 2nd sample in an updated sampling frame, linked in a way with the 1st one, but without any possibility of changing the conditions or results of the drawing of the former sample. The origin of this issue lies in PISA (Programme for International Student Assessment) surveys: the cycle 2012 was on the same main topic as in 2003 and it was necessary to make comparisons between both surveys. In this context, some countries wish to build the 2012 sample of schools with overlapping conditions with the 2003 sample. But it is also necessary to have the best representativeness for the new sample. This one can be met introducing balancing conditions when the new sample is drawn. Other constraints should be prescribed (fixed size, given inclusion probabilities...). The main tools used are first conditional successive samples and, secondly, balancing techniques. But it will be shown that only approached solutions can be reached, not only from a statistical point of view, but also from a computational one, to obtain numerical solutions. After developing the theoretical approach, results on French sampling frames of schools will be given.

Key words: balancing conditions, overlapping, inclusion probabilities

1 Introduction

The aim of this paper is to provide a theoretical frame and methods to solve a problem which occurs as soon as a first sample has been drawn at a given moment and that, some time after, one intends to draw a 2nd sample, with links (in a way which will be defined after) with the 1st one, but without any possibility of changing the conditions or results of the drawing of the former sample. The origin of this issue lies in PISA surveys: the next cycle 2012 will be on the same main topic as in 2003 (see for example : OECD, 2012). Therefore, it would be necessary to have the possibility of making comparisons between both surveys. One of the ways to perform it is to build the 2012 sample of schools with overlapping conditions with the 2003 sample. But it is also necessary to have the best representativeness for the new sample. This one can be met introducing balancing conditions when the new sample is drawn.

The problem can be written like this : given the fact that the former sample has been already drawn and cannot be changed, how to do to draw a new sample in an updated sampling frame, with given fixed size, with given inclusion probabilities, and subject both to overlapping conditions with the previous one and balancing conditions at present time t ? The theoretical approach developed here to solve this problem may apply to a wider range of issues. The most usual cases for this approach are : 2-phase samples, samples with overlapping conditions, samples drawn separately so as there is no overlapping at all (disjunction, negative coordination), samples with conditions of "representativeness" when one uses either both samples, or the 2nd one only, but with one of the constraints said above.

This problem has great interest when units are drawn with inequal probabilities. Therefore the frame of this work can be applied to business surveys, to the case of drawing

primary geographical units with different sizes, to the sampling of schools as in PISAĔ, all different surveys being characterized by the fact that their statistical units have generally a size factor which their inclusion probabilities depend on. The main tools for this approach are, first, the idea of conditional successive samples and, secondly, the balancing technique. Besides, the issue becomes much more complicated when the constraints on the samples are balancing conditions or can be turned to them. At last, we will see that only approached solutions can be reached, not only from a statistical point of view, but also from a computational one, to obtain numerical solutions.

2 General Framework

We denote U for the "universe", i.e. the refrence population or the sampling frame, index i for individuals (statistical units), N for the size of U, Y for the variable of interest (not random) and $T(Y)$ for the total of Y in the whole population.

A first sample S_1 has been drawn without replacement in U, characterized by :

- its size (might be random) : $n_1 = n(S_1)$
- the probabilities of inclusion : $Pi \in S_1 = \pi_i^1 \in [0,1]$ (defined *ex-ante*)
- balancing conditions on known variables X (might be vectorial) : $\sum_{i \in S_1} \frac{X_i}{\pi_i^1} = T(X)$

Balanced sampling allows to obtain a kind of representativeness regarding variables X ; it means that the sample is a reduced model of the universe : the estimation of the total of X based on the sample takes values which are exactly the same as the true values in the universe (Deville & Tillé, 2004).

Issue Sample S_1 has been drawn earlier in the past and its characteristics cannot be changed at present time. The result is a given sample. Then one intends to draw a 2nd sample S_2. If this new sample has no relationship with the 1st one, the easiest way is to draw the 2nd sample independently of the 1st one. This case is of no interest in that paper. On the contrary, if there is a link between both samples, whatever it is (overlapping conditions, disjunction, simultaneous use of both samples, global balancing conditions...), then the drawing of the 2nd sample should take into account the result of the 1st one. In those more complex cases, studied in this paper, this 2nd sampling procedure should be seen as conditional to the result of the 1st one.

For the drawing of the 2nd sample conditionally to the 1st one (without replacement), we have the following characteristics :

- the universe where S_2 is drawn may depend on S_1 : it will be called $U(S_1)$, with size $N_2(S_1)$. $U(S_1)$ is the sampling frame for S_2.
- its size may also depend on S_1 : $n_2(S_1)$.
- conditional inclusion probabilities for each unit : $Pi \in S_2/S_1 = \pi_i^{2/S_1} \in [0,1]$.
- if sample S_2 has fixed size n_2, not random, we have to take this condition into account at the moment the 2nd sample is drawn conditionally to the 1st one. It implies that conditional inclusion probabilities should verify :
  $\sum_{i \in U(S_1)} \pi_i^{2/S_1} = n_2$

- potential balancing conditions on variables called Z_i, written :
$$\sum_{i \in S_2} \frac{Z_i}{\pi_i^{2/S_1}} = \sum_{i \in U(S_1)} Z_i$$

This condition can be implemented from an algorithmic and statistical point of view, using the "CUBE Method" (Deville & Tillé, 2004). Its corresponds to the situation where a sample S_2 is drawn in a sampling frame $U(S_1)$, where S_1 is assumed to be fixed. The algorithm works whatever S_1 is random or not, according to the fact that S_1 is fixed and perfectly known when drawing S_2.

In this paper We assume that the reference universe U is not changing over time between the drawings of the two samples[1]. We wish to have an overlapping rate of $\alpha \in]0,1[$ between S_1 and S_2.

3 Conditions

We can show that the constraints on S_2 cannot be satisfied unless specific conditions on S_1 have been set before and that it implies some relationships between parameters.

The conditional probabilities π_i^{2/S_1} may be written as follow :

$$\pi_i^{2/S_1} = \begin{cases} a_i & \text{si } i \notin S_1 \\ b_i & \text{si } i \in S_1 \end{cases} = a_i \mathbb{1}_{i \notin S_1} + b_i \mathbb{1}_{i \in S_1} \text{ with } a_i \text{ et } b_i \in [0,1]$$

Then
$$\pi_i^2 = E(\pi_i^{2/S_1}) = a_i(1 - \pi_i^1) + b_i \pi_i^1$$

C1. How to satisfy simultaneously balancing conditions and given inclusion probabilities?

Wa can show that the 2nd sample will be balanced on variables V_i if and only if the 1st sample is balanced on $(b_i - a_i)\frac{\pi_i^1}{\pi_i^2} V_i$.

It means that for the sub-sample of S_2 drawn in S_1, conditional drawing is balanced on variables $Z_{1,i} = \frac{b_i}{\pi_i^2} V_i$.

For the sub-sample of S_2 drawn in $\complement S_1$, conditional drawing is balanced on variables $Z_{1,i} = \frac{a_i}{\pi_i^2} V_i$.

C2. How to take into account overlapping conditions between both samples ?

To obtain an overlapping rate of α between S_1 and S_2, we can show that it implies that S_1 is balanced on $b_i \pi_i^1$, that is $\sum_{i \in U} b_i \pi_i^1 = \alpha n_1$.

C3. How to ensure the fixed size n_2 for sample S_2 ?

To have a fixed n_2 for S_2, it is equivalent to a balancing condition for sample S_1 on variables $a_i \pi_i^1$, that is $\sum_{i \in U} a_i(1 - \pi_i^1) = n_2 - \alpha n_1$. This condition is satisfied.

[1] It might be false if some new units appear or others die. Further developments are in progress.

4 Approached solutions

The main difficulty in the previous solutions is that the first sample should verify specific balancing conditions. If they are not been planned, the constraints on the second sample will not be reached.

The idea is to change the final inclusion probabilities... :

- ... finding final new inclusion probabilities $\tilde{\pi}_i^2(S_1)$
- ... close to the π_i^2
- ... and keeping all the balancing conditions on the first sample.

More precisely, the idea is to find the $\tilde{\pi}_i^2$ (and thus the $\tilde{a}_i$ and $\tilde{b}_i$) which minimize an objective function :

$$min \sum_{i \in U} d(\tilde{\pi}_i^2, \pi_i^2) \text{ with } d \text{ a distance, e.g. euclidian or } \chi^2$$

And which satisfy the constraints on fixed size, on overlapping and on balancing conditions described before.

A particular case We consider for S_2 only the condition of fixed size and the overlapping condition relatively to S_1 and no other balancing condition.

To ensure a fixed size, we can take $b_i = \alpha \in [0, 1]$. Then, $a_i = \dfrac{\pi_i^2 - b_i \pi_i^1}{1 - \pi_i^1}$.

The case of take-all-strata ($\pi_i^1 = 1$) can be treated as well but is not presented here.

The maximization program is : $min \sum_{i \in U} (\tilde{\pi}_i^2 - \pi_i^2)^2$

with the constraints :

$$\begin{cases} \sum_{i \in U} \tilde{\pi}_i^2 = n_2 \\ \sum_{i \notin S_1} \dfrac{\tilde{\pi}_i^2 - \alpha \pi_i^1}{1 - \pi_i^1} = n_2 - \alpha n_1 \end{cases}$$

Solving this program (with lagrangian method) leads to :

$$\tilde{\pi}_i^2 = \pi_i^2 + \left[\frac{I_{i \notin S_1}}{1 - \pi_i^1} - \frac{1}{N} \sum_{j \notin S_1} \frac{1}{1 - \pi_j^1} \right] \frac{n_2 - \alpha n_1 - \sum_{j \notin S_1} \dfrac{\pi_j^2 - \alpha \pi_j^1}{1 - \pi_j^1}}{\sum_{j \notin S_1} \dfrac{1}{(1 - \pi_j^1)^2} - \dfrac{1}{N} \left(\sum_{j \notin S_1} \dfrac{1}{1 - \pi_j^1} \right)^2}$$

For individuals in S_1, probabilities π_i^2 are changed with a translation by a fixed value. For individuals outside S_1, the translation is depending on π_1^1. We can show that sign of these translations is undetermined.

Generalization The general case implies the additional balancing constraint on variables V_i :

$$\sum_{i \in S_1} \frac{1}{\pi_i^1}(b_i - a_i)\frac{\pi_i^1}{\tilde{\pi}_i^2}V_i = \sum_{i \in U}(b_i - a_i)\frac{\pi_i^1}{\tilde{\pi}_i^2}V_i$$

This new constraint implies an iterative solution to solve the minimization program, unless Cardan's formula could be used. Further developments are in progress to solve this problem.

Choice of estimators Once S_2 has been drawn, 4 possible estimators may be calculated to estimate the total $T(Y)$:

1. $\hat{T}_1(Y) = \sum_{i \in S_2} \frac{Y_i}{\pi_i^2}$: we use the target probabilities π_i^2 but S_2 has not be drawn with those probabilities.
2. $\hat{T}_2(Y) = \sum_{i \in S_2} \frac{Y_i}{\tilde{\pi}_i^2}$: we use the actual probabilities $\tilde{\pi}_i^2$ but they are random and they does not correspond to the sampling probabilities.
3. $\hat{T}_3(Y) = \sum_{i \in S_2} \frac{Y_i}{\pi_i^{*2}}$ with $\pi_i^{*2} = E(\tilde{\pi}_i^{2/S_1})$: this is the true Hovitz-Thomson estimator but the π_i^{*2} are difficult to be computed.
4. $\hat{T}_4(Y) = \sum_{i \in S_2} \frac{Y_i}{\tilde{\pi}_i^2}$: this estimator is unbiased but with random coefficients.

5 Simulations

First of all, evaluating the "effects" of the modifications of the π_i^2 is difficult to do analytically. Simulations on real data shall provide insights about this issue. Furthermore, the simulations will allow us to analyze the link between $\tilde{\pi}_i^2$ and α. What is the impact of the overlapping rate ? Also we need to check the conditions on probabilities (in particular, $\tilde{a}_i^2 \in [0, 1]$). An other issue is the choice of the distance d for the program. We solved the program for the χ^2 distance as well (not presented here). The simulations will be useful to evaluate the impact of the choice of the distance. Finally, the choice of the estimator is an issue we can also study with simulations.

We conducted two kinds of simulations based on the PISA sampling procedure applied in the French context between 2000 and 2009 :

1. The first set simulations consists of sampling many S_1, calculating the $\tilde{\pi}_i^2$, computing the distance with the π_i^2 and then assessing whether distance depends on the overlapping rate and on the kind of distance (euclidian or χ^2).
2. The second set of simulations consists of sampling many S_1 then sampling a S_2 for each S_1, estimating total known variables according 3 kinds of estimators, and then assessing the quality of the estimators (bias and precision) according to the overlapping rate and to the kind of distance.

First results show that the impact of the overlapping rate on the distance between $\tilde{\pi}_i^2$ and π_i^2 is not clearly established. Further investigations have to be made, using an other sampling frame based on national assessment programs because PISA target population implies a bounded overlapping rate. Furthermore, the bias seems to be negligible whatever the kind of estimators, and precision of $\hat{T}_1(Y)$ and $\hat{T}_2(Y)$ appears to be slightly better that the precision of $\hat{T}_4(Y)$.

6 Conclusions and future developments

The following conclusions may be done :

- There is a difficulty to assign conditions to a given sample taking into account a sample which has been drawn earlier. There is no exact solution, only approached solutions.
- One shall derive guidelines for future: a sample drawn at a given time must look ahead future drawings and constraints that will be applied to them, in order to incorporate them through appropriate balancing conditions.
- The simulations show that it is possible to implement the approach of "inverse balancing" and that properties of the method can be verified.
- Further simulations should be done in more general cases (including numerical computation).

References

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.
OECD (2012). *PISA 2009 Technical Report*. Paris : OECD.