Calibration of Expansion Factors in a Multilevel Sampling survey

María Angélica Ferrandi[1], Nora Daruich[2], Sabrina Balbi[3], Juliana Merello[4]

[1]Inter-American Open University, Rosario, Argentina maferrandi@gmail.com
[2]Provincial Institute of Statistics and Surveys of Santa Fe, Rosario, Argentina
ndaruich@hotmail.com
[3]Provincial Institute of Statistics and Surveys of Santa Fe Rosario, Argentina
balbisabri@yahoo.com.ar
[4]Provincial Institute of Statistics and Surveys of Santa Fe, Rosario, Argentina
juliana_merello@yahoo.com.ar

Abstract

In large complex surveys using multilevel sampling with conglomerates, different levels of stratification, selection probabilities proportional to size, etc., one of the main problems relates to bias.

Complex sampling designs, such as those just mentioned, sometimes do not allow for bias control, because different selection probabilities are involved for sampling units, and samples are not self-weighed. All such bias-related difficulties impede a correct reading of estimations or make outcome comparison impossible. Consequently, the usual practice is to correct or adjust baseline expansion factors or weighs (usually, regression of selection probabilities) based on auxiliary known data or information preset from external sources or records.

The purpose of this paper is to make baseline weight adjustments by means of the "calibration of fixed marginal probabilities" technique in the "Uso de Tiempo" (Time Usage) survey, carried out in the City of Rosario, Santa Fe Province (Argentina 2010). This technique was applied as per Deville and Särndal's [1992] methods.

Based on a general approach, such technique introduces new weights as a result of adjusting or calibrating baseline weightings set at the design phase by solving out a problem of numerical minimization. The problem is defined by the selection of a distance between the new fixed marginal probabilities and the baseline ones, and the use of a set of restrictions on the auxiliary variables involved in the adjustment. The calibration performed kept sampling designs used in the survey. Therefore, it was applied to the baseline expansion factors for each sampling unit (households in the Time Usage Survey), which were corrected as per non-response, since this was the last sampling unit of the design.

Regarding auxiliary information used for calibration and adjustments of the sampling inner schema, data were retrieved from the "Censo Nacional de Población, Hogares y Viviendas 2001" (2001 National Survey on Population, Housing and Households).

Keywords: Calibration, Expansion Factors, Multilevel Sampling, Estimations.

1. Introduction

The "Uso de Tiempo" (Time Usage) survey, carried out in the City of Rosario, Santa Fe Province during May, June and July 2012, provided with representative information of people +15 years old from the City of Rosario, registering the sequencing and duration of daily activities carried out by one person during a specific period of time, usually 24 hours, such as jobs, either professionally or at home, education, free time, voluntarism, etc. for weekdays (from Monday to Friday), for weekends and for the whole week.

By means of a survey, the aim is analyzing the behavior of a particular feature and estimating its population total, but the amount of items observed will not be the same as its population total, and, to estimate such total, weightings will be used to get the first estimation.

By estimating a variable for which a population distribution is known, such as gender and age, and not taking them into account during the screening process, it is highly likely that, due to the sampling randomness or the lack of response, estimated distributions are erroneous.

The total of men and women in the sample were calculated, and they mismatched with the population distribution, so, if this survey fails to represent population by gender and age, it will not be valid for providing any other type of data either.

Baseline weights were adjusted using re-weighting techniques, such as calibration methods developed by Deville and Säfrndal (1992) and scheduling designed by Deville, Särndal and Sautory (1993) within the SAS software, which is statistics-specific, under the name CALMAR (CALage sur MARges).

2. Methods

The following paragraphs include a summarized idea of the theoretical approach explained by Deville, Särndal and Sautory (1993).

The technique of estimation by calibration was introduced by Deville and Särndal in 1992. The idea is to use auxiliary information to obtain a better estimate of a population statistic. First, consider a finite population U of size N with unit labels 1, 2, . . . ,N. Let yi, i = 1, . . . ,N be the study variable and xi, i = 1, . . . ,N be the k-dimensional vector of auxiliary variables associated with unit i.

Suppose we are interested in estimating the population total:

$$t_y = \sum_{i=1}^{N} y_i$$

We draw a sample s = {1, 2, . . . , n} $\in$ U using a probability sampling design P, where the first and second order inclusion probabilities are π_i =Pr (i $\in$ s) and π_{ij} =Pr (i,j $\in$ s) respectively. An estimate of t_y is the Horvitz-Thompson (HT) estimator:

$$\hat{t}_{HT} = \sum_{i \in s} d_i\, y_i$$

where $d_i = 1/\pi_i$ is the sampling weight, defined as the inverse of the inclusion probability for unit i. An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design P. Its variance under P is given as

$$V_p(\hat{t}_{HT}) = \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i \pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

Now let us suppose that {x_i, i = 1, . . . ,N} is available and $t_x = \sum_{i=1}^{N} x_i$ the population total for x, is known. Ideally, we would like X:

$\sum_{i \in s} d_i\, x_i = t_x$ but often times this is not true.

The idea behind calibration estimators is to find weights wi, i = 1, . . . n close to di, based on a distance function, such that X:

$$\sum_{i \in s} w_i\, x_i = t_x$$

We wish to find weights wi similar to d_i so as to preserve the unbiased property of the HT estimator. Once w_i is found, the calibration estimator for t_y is $\hat{t}_c = \sum_{i \in s} w_i\, y_i$

To formalize this procedure, a distance function to measure proximity of potential calibration factor sets to baseline weights is defined.

By mathematical development, generalized regression expression of estimator is reached (Särndal, Swensson and Wretman,1992).

$$Y_{cal} = Y_{\pi} + B'_{s}(X - X_{\pi}) = \hat{Y}_{GREG} \qquad B_{s} = (\sum_{s} d_{k} x_{k} x'_{k})^{-1} \sum_{s} d_{k} x_{k} y_{k}$$

The CALMAR macro calculates final weights for each of the cases, based on its features, taking into account baseline weights (inverse of sample inclusion probability).

2.1 Objective

The main objective of the survey was estimating the amount of people performing any type of voluntary work in the City of Rosario. And to quantify the time spent working for the market, household work and other activities related to studying, free time and rest, etc.

Voluntary work is defined as the unpaid activity performed based on one's own decision with the intention of rendering benefits to others with no underlying duty nor obligation derived from family relations or friendship.

2.2 Target population

Data about the use of time by men and women living in households in the City of Rosario was collected. Household members who were 15 years old or older were probed, and they were asked to fill in an activity diary and an individual questionnaire.

2.3 Surveying mechanisms

Two surveying mechanisms were used: an individual questionnaire --where household members were recorded, along with the personal traits of each of them-- and the Activity Diary.

The **Activity Diary** covers a24-hour time span, starting at 4.00 a.m. on the day before the interview and ending at 4.00 a. m. of the interview day. It contains 48 30-minute time sets, in which up to three activities can be included, 6 overlapping codes for the second and third activity, and localization codes for all activities. The Activity Diary also included three verification questions, designed to render activities that are usually forgotten, or to help coding; and a control question about the type of day (typical day/non-typical day).

The Activity Diary was filled in by surveyors with activities done by the respondent during the interview performed based on a guide of questions specifically designed. Questions "*What were you doing yesterday between ... and...?* and then, *Were you doing anything else?* repeated for each time set, except for the sleeping time (considering the Diary started at the time the respondent woke up and ended at the time he/she went to sleep), and the time spent on working for the market, for which a special set of questions was designed.

2.4 Sampling

Multi-phased study with primary unit stratification. Primary unit stratification was performed based on a previous paper, where radiuses were clustered according to a cluster statistical analysis, analyzing a set of variables from the 2001 National Population and Household Census ("Social Map").

Primary units: radiuses (areas). Selection was done with a probability which was proportional to the stratum size, measured by the number of households.

Secondary units: blocks, single randomization, with probability based on the number of blocks in each radius.

Tertiary units: households, systematic sampling. Selection of 10 households based on a systematic sampling with random start.

2.4.1 Sample size Allocation

To determine sample size, certain knowledge of the population in terms of basic features of the subject of the study and the disaggregation level intended to analyze data was necessary.
Based on previous studies, data collected about the rate of people performing voluntary work at the level of 2-digit disaggregation were analyzed. Considering the need to provide data at the level of the City of Rosario, a sample size of 1040 households was defined; a sample of 10 households was defined in the selected areas. With such a size, a variation coefficient below 15% was expected for the main activities.

2.4.2 Sample selection

As a framework for sample selection, a geographical area frame was used; such area consisted of census radiuses used in the 2001 Population and Household Census, grouped in 5 strata.
Census radiuses were selected for each stratum with a probability which was proportional to their size; blocks by single randomization and household with probability by systematic sampling with random start.

3. Outcome

The following table depicts final sample of primary units and of second phase, distributed by strata.

Table # 1: Final sample of primary units and second phase, distributed by strata

Stratum	Total	Household Rate in the sample	Amount of selected radiuses
1	7,322	100	10
2	43,512	210	21
3	68,444	270	27
4	144,408	390	39
5	4,385	70	7
Total	268,071	1,040	104

3.1 Baseline weightings and estimators

To calculate weightings, the probability of selection of the areas, blocks and households was taken into account.
When the household is selected and all household members are probed, weighting derived from selection probability is the same for the household and all the people living therein, since no screening process was done in the household.
When calculating totals for baseline weightings, it was noted these did not match with the population total. Therefore, the re-weighting technique was used in order to adjust sample distribution to the population distribution reported by external sources.
Since there was no census recent information available, population projection was used, by age and gender, in the year the survey was performed.
Applying newly calibrated weights, estimations were calculated, and a more consistent pattern was found.

Tables #2 and #3: Estimated population with baseline weightings

Gender	Total	%
Total	789,584	100.0%
Male	374,217	47.4%
Female	415,366	52.6%

Age group	Total	%
Total	789,584	100%
0-14	170,990	22%
15-24	118,837	15%
25-39	170,562	22%
40-49	95,471	12%
50-64	122,053	15%
65-74	56,704	7%
+75	54,967	7%

Table #4: Estimated population with calibrated weightings

Gender	Total	%	Age group	Total	%
Total	909,814	100%	Total	909,814	100%
Male	430,813	47.4%	0-14	212,953	23.4%
Female	479,001	52.6%	15-24	164,821	18.1%
			25-39	177,640	19.5%
			40-49	106,907	11.8%
			50-64	128,023	14.1%
			65-74	66,915	7.4%
			+75	52,555	5.8%

Table #5: Population Projections

	Total			Men		Women	
		Column %	Row %a				
Total	909,814	100	100	430,813	47.4	479,001	52.6
0-14	212,953	23.4	100	107,816	50.6	105,137	49.4
15-24	164,821	18.1	100	82,373	50	82,448	50
25-39	177,640	19.5	100	86,720	48.8	90,920	51.2
40-49	106,907	11.8	100	50,830	47.5	56,077	52.5
50-64	128,023	14.1	100	58,540	45.7	69,483	54.3
65-74	66,915	7.4	100	27,396	40.9	39,519	59.1
+75	52,.555	5.8	100	17,138	32.6	35,417	67.4

4. Conclusion

In spite of all the logical limitations of this study, we can assume that using estimators with auxiliary information generally offers the benefit of accuracy of some of the estimations, and it hardly ever worsens the baseline situation. This is a relevant fact, since surveys usually have multiple objectives and, therefore, it is not easy to find a set of auxiliary information appropriate for all of them. By means of using the Calmar software, total population adjustment was attained by age group and total population based on gender. However, if we aim at estimations at even lower disaggregation levels, such as "Population total by age group and gender", alternative methods will have to be considered.

5. Bibliography

J.C.Deville and C.E.Särndal - (1992) - "Calibration Estimators in Survey Sampling"
Document # F 9310 - Instituto Nacional de Estadística y Estudios Económicos de Francia - (1993) - "Calmar macro"
Arriero C.P. - "Calibrating a household Survey by using the Calmar Program"
Wu C., Stitter R.R. - (2001) - "A model-calibration approach to using complete auxiliary information from survery data"
Esquivel, Valeria - (2009) - "Uso del Tiempo en la Ciudad de Buenos Aires" - 1° Edic.Los Polvorines: Universidad Nacional Gral. Sarmiento.
Budlender D., Chobokoane N. and Mpetsheni Y. - (2001) - "A survey of time use: How South African women and men spend their time" Statistics South Africa. Pretoria
Leslie Kish - (1972) - "Muestreo de Encuestas"
Cochran, W.C. - (1982) - "Técnicas de Muestreo"