This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Balanced Sampling for Multi-Way Stratification

Contents

General section

1. Summary

Balanced sampling is a class of techniques using auxiliary information at the sampling design stage. Many types of sampling designs can be interpreted as balanced sampling, such as simple random sampling with fixed size, stratified simple random sampling and unequal probability sampling.

Furthermore, the balanced sampling can be applied to define a multi-way stratification design also known as incomplete stratification or marginal stratification. Multi-way stratification allows to plan the sample sizes of the domains of interest belonging to two or more non-nested partitions of the population in question without using the standard solution based on a stratified sample in which strata are identified by cross-classifying the variables defining the different partitions (one-way stratified design). The standard solution in many Structural Business Surveys (SBSs) may have drawbacks from the view-point of cost-effectiveness. In fact, SBSs produce typically estimates for a great number of very detailed domains forming several non-nested partitions of the population and creating really small cross-classified strata.

2. General description of the method

Balanced sampling can be applied according to two different inferential approaches: the model based approach (Royall and Herson, 1973, Valliant *et al.* 2000) and the design based or randomisation assisted approach (Deville and Tillé, 2004). The first approach bases the inference on a statistical superpopulation model and it may be performed by probability or non-probability sample. In this framework a sample is balanced when the sample means of a set of auxiliary variables (balancing variables) are equal to the known population means (Valliant et all, 2000). Balanced samples are used to follow a robust sampling strategy. The design based approach needs a sampling frame and uses a probability sample to make inferences. In this second context a sample is balanced when the Horvitz-Thompson (H-T) sample estimates for the auxiliary variables are equal to their known population totals. The selection of a balanced sample generally improves the efficiency of the sampling estimates (Cochran, 1977). This section focuses on this second inferential approach.

A widely used application of the method is stratified simple random sampling. As known, it has been introduced in the sampling methodology to enhance the efficiency of the estimates. Nevertheless, stratified sampling can be used as an operative tool in the surveys as well. An instrumental use of stratified sampling is when the objective of the survey is to produce estimates for some subpopulations (or domains) forming two or more non-nested partitions of the population and a fixed or planned sample size for each domain is required. A standard sampling design solution defines strata by cross-classifying the variables defining the different partitions. In this case stratification is not strictly used to improve estimation quality. It is used to implement a random selection method guaranteeing the selected sample sizes corresponding to the planned ones. This standard solution, hereinafter denoted as one-way stratified design, may have some drawbacks, especially in the SBSs.

When the number of cross-classified strata is too large, there are some immediate consequences, described as follows:

(i) the overall sample size could easily be too large for the survey economic constrains;

(ii) when the population size in many strata is small, the stratification scheme becomes inefficient; in other words the sample allocation may be far from the theoretically desired allocation;

(iii) when there are strata containing only few units in the population, a not equally distributed response burden may arise in surveys repeated over time.

Many methods have been proposed in the literature to keep that the sample size under control in all the domains without using one-way stratified designs. This means that sample size of each cross-classified stratum is a random variable. These approaches may be roughly divided into two main categories. The first category contains methods commonly known as controlled selection. Seminal papers have been proposed by Bryant et al. (1960) and Jessen (1970). Other methods based on controlled rounding problems via linear programming have been proposed by Causey et al. (1985), Rao and Nigam (1990; 1992), Sitter and Skinner (1994) and Winkler (2001).

In the second category there are methods based on sample coordination. A separate sample is selected for each partition in order to guarantee the maximum overlap among the different samples (Ohlsson, 1995; Ernst and Paben, 2002). We define all these methods as multi-way stratified designs.

Literature shows that these methods pose theoretical and operative problems especially for large scale surveys as in the SBSs. A recently proposed method, the Cube algorithm (Deville and Tillé, 2004) overcomes these drawbacks. The method, included in the first category, has been originally defined for drawing balanced samples with a large number of balancing variables for large population size. Multi-way stratification is a special case of balanced sampling. Given the population *U* of size *N*, let π_k be the inclusion probability of *k*-th population unit (*k*=1, …, *N*) and let δ_{dk} be the value of the indicator variable of the domain U_d, being $\delta_{dk}=1$ if the unit *k* belongs to domain U_d and equal to zero otherwise. Then, by definition the sample size of U_d is $\sum_{k=1}^{N} \delta_{dk}\pi_k = n_d$. The Cube method assumes that the inclusion probabilities are known, and it selects a random sample achieving the consistency among the known totals and the H-T estimates. We define the following auxiliary variables $z_{dk} = \delta_{dk}\pi_k$ for each domain. When the sample is balanced on the *z* variables the H-T estimate $\hat{Z}_d = \sum_{k=1}^{N} s_k z_{dk} / \pi_k$ has to be equal to the known population total $Z_d = n_d$ with s_k being a random variable equal to 1 if the *k*-th unit belongs to the sample and equal to 0 otherwise. For satisfying the balancing equations, $\hat{Z}_d = Z_d$, the Cube algorithm has to select n_d units from U_d. When the expected sample sizes are integer numbers the Cube algorithm applied to obtain a multi-way stratification always finds the solution. Some illustrative examples of multi-way sampling designs are given in Falorsi and Righi (2008).

3. Preparatory phase

4. Examples – not tool specific

4.1 *Example: the multi-way stratification design for controlling the sample size*

In order to explain the problem, we consider the population of 165 schools reported in Table 1 (Cochran, 1977, p. 124). We assume that the parameters of interest are the totals of a variable, related to school, separately for the *Size of city* (5 categories: *I,II,III,IV,V*) and for the *Expenditure per pupil* (4 categories: *A,B,C,D*). Two distinct partitions of the population are defined: the size of city (first partition) defining 5 non-overlapping domains, and the expenditure per pupil defining 4 domains. We have 9 domains of interest.

Table 1.

		Expenditure per pupil				
		A	*B*	*C*	*D*	Totals
Size of city	*I*	15	21	17	9	62
	II	10	8	13	7	38
	III	6	9	5	8	28
	IV	4	3	6	6	19
	V	3	2	5	8	18
	Totals	38	43	46	38	165

The standard one-way stratified design (or cross-classification design) defines 20=5×4 strata by crossing the categories of the domains of the two partitions. Due to budgetary constraints, we suppose that the sample size could be up to 10 units. Nevertheless, in each stratum at least one school should be selected (or two schools for estimating the sampling variance without any bias) and, consequently, according to this design the sample size should amount to 20 (or 40) schools at least. Hence, the cross-classification design becomes unfeasible.

4.2 *Example: the multi-way stratification design to retain the sample allocation*

We consider the above population of schools. We plan a sample of 20 schools and we want to allocate the sample proportionally to the domain size. Table 2 shows the planned size and the integer rounded sample allocation. We note that the sample sizes in the cross-classified strata are not constrained to be integers.

Table 2.

		Expenditure per pupil					
		A	*B*	*C*	*D*	Domain weight	Rounded planned sample size
Size of city	*I*	0.0909	0.1273	0.1030	0.0545	0.3757	8
	II	0.0606	0.0485	0.0788	0.0424	0.2303	5
	III	0.0364	0.0545	0.0303	0.0485	0.1697	3
	IV	0.0242	0.0182	0.0364	0.0364	0.1152	2
	V	0.0182	0.0121	0.0303	0.0485	0.1091	2
	Domain weight	0.2303	0.2606	0.2788	0.2303	1.0000	
	Rounded planned sample size	5	5	5	5		20

According to the one-way stratified design of size 20 (Table 3), we obtain a sample allocation far from the planned one.

Table 3.

		Expenditure per pupil					
		A	*B*	*C*	*D*	Domain weight	Rounded planned sample size
Size of city	*I*	1	1	1	1	0.2000	8
	II	1	1	1	1	0.2000	5
	III	1	1	1	1	0.2000	3
	IV	1	1	1	1	0.2000	2
	V	1	1	1	1	0.2000	2
	Domain weight	0.2500	0.2500	0.2500	0.2500	1.0000	
	Rounded planned sample size	5	5	5	5		20

4.3 *Example: the multi-way stratification design for reducing the response burden*

We consider the population of schools described in section 4.1 and we suppose that the population distribution to be fixed over time. We have to select a sample of size 40 on several survey occasions. Moreover, we want to compute unbiased variance estimates. According to the one-way stratified design we have to select 2 schools per stratum (Table 4).

Table 4.

		Expenditure per pupil					
		A	*B*	*C*	*D*	Domain weight	Rounded planned sample size
Size of city	*I*	2	2	2	2	0.2000	8
	II	2	2	2	2	0.2000	8
	III	2	2	2	2	0.2000	8
	IV	2	2	2	2	0.2000	8
	V	2	2	2	2	0.2000	8
	Domain weight	0.2500	0.2500	0.2500	0.2500	1.0000	
	Rounded planned sample size	10	10	10	10		40

Then, we can see that the schools in stratum *V-B* are drawn with certainty on each survey occasion and the schools in strata *IV-B* and *V-A* have a high probability to be included in the samples. That happens because in stratum *V-B* there are only two schools in the population, while in the strata *IV-B* and *V-A* there are three schools in the population and the inclusion probability is 0.67. Hence, the response burden is not equally distributed in the population of schools (is high for the schools belonging to small population size strata) and this burden does not depend on efficiency issues.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Bryant, E. C., Hartley, H. O., and Jessen, R. J. (1960), Design and Estimation in Two-Way Stratification. *Journal of the American Statistical Association* **55**, 105–124.

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985), Applications Transportation Theory to Statistical Problem. *Journal of the American Statistical Association* **80**, 903–909.

Cochran, W. G. (1977), *Sampling Techniques*. Wiley, New York.

Deville J.-C. and Tillé, Y. (2004), Efficient Balanced Sampling: the Cube Method. *Biometrika* **91**, 893–912.

Ernst, L. R. and Paben, S. P. (2002), Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications. *Journal of Official Statistics* **18**, 185–202.

Falorsi, P. D. and Righi, P. (2008), A Balanced Sampling Approach for Multi-Way Stratification Designs for Small Area Estimation. *Survey Methodology* **34**, 223–234.

Jessen, R. J. (1970), Probability Sampling with Marginal Constraints. *Journal of the American Statistical Association* **65**, 776–795.

Lu, W. and Sitter, R. R. (2002), Multi-Way Stratification by Linear Programming Made Practical. *Survey Methodology* **28**, 199–207.

Ohlsson, E. (1995), Coordination of Samples using Permanent Random Numbers. In: *Business Survey Methods* (eds. Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S.), Wiley, New York, Chapter 9.

Rao, J. N. K. and Nigam, A. K. (1990), Optimal Controlled Sampling Design. *Biometrika* **77**, 807–814.

Rao, J. N. K. and Nigam, A. K. (1992), Optimal Controlled Sampling: a Unifying Approach. *International Statistical Review* **60**, 89–98.

Royall, R. and Herson, J. (1973), Robust Estimation in Finite Population. *Journal of the American Statistical Association* **68**, 880–889.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Winkler, W. E. (2001), Multi-Way Survey Stratification and Sampling. RESEARCH REPORT SERIES, Statistics #2001-01, Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.

Specific section

8. Purpose of the method

Balanced sampling is used for selecting a multi-way stratified design, which is a sampling design planning the sample sizes for domains of interest belonging to different partitions of the population without using a one-way stratified or cross-classified stratification design.

9. Recommended use of the method

1. The method can be applied when the one-way stratified designs (those where strata are obtained by combining the domains of different partition of the population) can be inefficient or can produce statistical burden for surveys repeated over time.
2. The method may be applied in large scale surveys, with large population and a lot of domains.
3. The method may be useful in the small area estimation problem when the membership indicator variables for small areas are known at population level. Planning the sample size for each domain allows to estimate specific small area effects improving the efficiency of indirect model based small area estimators.

10. Possible disadvantages of the method

1. The method needs to know the inclusion probabilities. The definition of the optimal inclusion probabilities is less intuitive than in case of one-way stratification design.
2. Analytic expression of the variance of the estimates is unknown. Approximations (shown in the literature) are needed.
3. Some difficulties when a complex estimator is used for the computation of sampling errors.

11. Variants of the method

1. Balanced Sampling for multi-way stratification is defined to select a planned sample size for each domain. In addition, it may be worthwhile including other balancing variables to enhance the estimation efficiency according to the calibration estimation theory.

12. Input data

1. Data including the domain membership indicator variable and the inclusion probability for each population units are needed.

13. Logical preconditions

1. Missing values
    1. Not allowed.
2. Erroneous values
    1. Not allowed.
3. Other quality related preconditions

   1.

4. Other types of preconditions
    1. The sum of the inclusion probabilities over each domain must be an integer.
    2. The sum over population domains of the inclusion probabilities must be equal for each partitions.

14. Tuning parameters

1. Depending on the origin of the inclusion probabilities a calibration step could be needed. The calibration step modifies the probabilities for satisfying sample size consistency among the partitions and for achieving an integer expected sample size in each domain (see 13.4.1).

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. Sample membership indicator variable is added in the input data set.

17. Properties of the output data

1. The sum of the sample membership indicator variable over each domain is equal to the expected sample size.

18. Unit of input data suitable for the method

Processing full data set.

19. User interaction - not tool specific

1. Definition of the set of inclusion probabilities.
2. Before execution of the method, verify that the planned sample sizes for each domain are integer numbers and consistent.
3. When performing a multi-way stratification design considering also other balancing variables in the sample selection process, the indicators of the quality of balancing have to be analysed.

20. Logging indicators

1. No specific indicators.

21. Quality indicators of the output data

1. When used only for multi-way stratification, the theory shows that the method selects exactly a sample satisfying the planned sample size. When other balanced variables are added, the ratio among the H-T estimates and the known totals are used as quality indicators.
2. No other quality indicators are used to strictly evaluate the performances of the methods.

22. Actual use of the method

1. Balanced sampling is widely used in the Insee not specifically for implementing multi-way stratification.
2. Istat has used balanced sampling for a population survey.
3. An Istat research project is studying the optimal allocation for multi-way stratified design.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Sample Selection – Main Module
2. Sample Selection – Sample Co-ordination

24. Related methods described in other modules

1.

25. Mathematical techniques used by the method described in this module

1. The method mainly implements a balancing martingale theory, with the aim to round off each inclusion probabilities randomly to 0 or 1. From the mathematical point of view that corresponds to the maximisation of the entropy measure (maximisation of the randomness) under linear constraints (balancing equations).

26. GSBPM phases where the method described in this module is used

1. 2.4 Design Frame & Sample methodology
2. Partially 4.1 Select sample

27. Tools that implement the method described in this module

1. Sampling R package
2. SAS Macro downloadable Insee site

28. Process step performed by the method

Sample planning and selection

Administrative section

29. Module code

Sample Selection-M-Balanced Sampling

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-12-2011	first version	Paolo Righi	ISTAT
0.2	29-02-2012	second version	Paolo Righi	ISTAT
0.3	22-02-2013	third version	Paolo Righi	ISTAT
0.3.1	06-09-2013	preliminary release		
0.3.2	09-09-2013	page numbering adjusted		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:42